Total variational optical flow for robust and accurate bladder image mosaicing

Sharib Ali

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Total variational optical flow for robust and accurate bladder image mosaicing

THÈSE

présentée et soutenue publiquement le 4 Janvier 2016

pour l'obtention du

Doctorat de l'Université de Lorraine

(Mention: Automatique, Traitement du Signal et des Images, Génie Informatique)

par

Sharib ALI

Composition du jury:

Rapporteurs :	Cédric DEMONCEAUX	PU, Université de Bourgogne Franche-Comté Laboratoire Le2i, UMR CNRS 6306
	João BARETTO	PU, Université de Coimbra, Portugal Electrical and Computer Engineering Department
Examinateurs :	Adrien BARTOLI	PU, Université d'Auvergne ISIT, UMR CNRS 6284 CENTI, Faculté de Médecine
	François GUILLEMIN	PU-PH, Institut de Cancérologie Jean Godinot, Reims
Invités :	Pascal ESCHWEGE	PU-PH, CHU Nancy-Hôpitaux de Brabois
	Ismaël DIDELON	SD-Innovation, Frouard
Directeur de thèse: :	Christian DAUL	PU, Université de Lorraine CRAN, UMR CNRS 7039
Co-Directeur de thèse: :	Walter BLONDEL	PU, Université de Lorraine CRAN, UMR CNRS 7039

Centre de Recherche en Automatique de Nancy

UMR 7039 Université de Lorraine - CNRS

avenue de la forêt de Haye 54516 Vandoeuvre-lès-Nancy
Tel : +33 (0)3 83 59 59 59 Fax : +33 (0)3 83 59 56 44

Acknowledgment

First of all, I would like to express my deepest gratitude towards my advisors Prof. Christian Daul and Prof. Walter Blondel. Without their joint motivation and support, this thesis work would not have been possible. Both of them were very patient and helpful at the same time. I had also an opportunity to learn and enjoy teaching with Prof. Daul. His hard work and passion in both research and teaching have inspired me a lot. He has been constantly encouraging me till the final submission of my thesis. I am overwhelmed by his consistent support throughout my thesis duration. I definitely look forward to working with you in future.

I would like to give my sincere thanks to Prof. François Guillemin for helping me understand the cystoscopic data during my first year of Ph.D. and also for accepting to be in my examination board. I am very thankful to all the jury members and the invitees for agreeing to participate in my thesis defense. I sincerely would like to thank Prof. Cédric Demonceaux and Prof. João Baretto for reviewing this manuscript and Prof. Adrien Bartoli for accepting to be in the examination board. I must not forget to thank Prof. Fabrice Meriaudeau for being there whenever i needed some professional and personal advises.

I would like to mention and thank Ernest for his time in helping me figure out some issues with software programming. I have learnt some very useful tips from him. I would also like to thank Marine for sharing some patient data with me. She has also been quite positive and a helping hand in data acquisition at CRAN. I would like to thank Christine, Carole and all my collegues at CRAN for being wonderful, nice and helpful. I must admit that I will miss you all and I wish you all a very marvelous time in your lives.

I would like to acknowledge and thank the constant support of my family and friends. I would like to thank Binod for giving me shelter for few days during my thesis writing. Yes, i have been jobless and homeless during first weeks of October. I would like to convey my special thanks to one of my most wonderful friend Mariia, who not only patiently listened to my thesis writing frustrations at times but also encouraged me to do my best. She also helped me with my English corrections. Thank you Mariia for your patience and going through my this manuscript. I hope you enjoyed it at times too. I must not miss my brother Shakir Ali and my dearest sister Mona, they both have been the greatest blessings to me. Their constant support and encouragement have always helped me a lot to become who I am today. Thank you both of you. This thesis is also for you all.

I would like to dedicate this thesis
to my late mom Samsun Nesha.

List of Tables

2.1 Variants of gradient-based assumptions according to [Papenberg et al., 2006]. . . 37

2.2 Average end-point error, in pixels and average angular error, in degrees ($\frac{AEPE}{AAE}$) are given for some well known TV-L^1 based methods on Middlebury test dataset. The methods are presented in the order of their rank on this benchmarking for test dataset (refer for details: `http://vision.middlebury.edu/flow/eval/results/results-e1.php`). An average value (avg.) is also provided for validating the algorithm tolerance to varying textures in this dataset. 62

3.1 AEPE/AAE (in pixels/in degrees) degrees) given for different TV methods applied on the Middlebury data-base. AAE and AEPE are mathematically defined in Chapter 2. 80

3.2 Standard deviation in registration parameters of phantom sequence I. First two column presents the standard deviation from constant translation $\sqrt{t_x^2 + t_y^2}$ for approximately 20 pixels and 50 pixels. Last four columns present the deviation from constant pure in-plane rotations for values 1°, 3°, 5° and 7°. 81

3.3 Homography parameter values for simulated sequences I and II. The RGB color channels are blurred for some images of simulated sequence-I. The value of the standard deviation σ_{blurr} of the gaussian function used for blurring [Chadebecq et al., 2012] is 2.5. 83

3.4 Parameter settings for the optical flow determination. The dense point correspondence is used to determine the homography parameters. *NA stands for not applicable parameter to a method.* 84

3.5 Mean registration ($\hat{\epsilon}_{i,i+1}$) and mosaicing ($\hat{\epsilon}_{0,50}$) errors in pixels for 50 image pairs of the simulated video sequence-I. Bladder simulated video sequences without ($\sigma_{blurr} = 0$) and with ($\sigma_{blurr} = 2.5$) additional Gaussian blur (depicting defocus/refocus in cystoscopy [Chadebecq et al., 2012]) are used for quantifying the robustness of the methods. Mean registration time for image pairs with size of 512×512 pixels are also presented (CPU implementation time for images under no blur). 85

3.6 Errors obtained for the methods compared on simulated sequence-II. Mean registration time of image pairs with size 400×400 pixels for CPU implementation are also given unless GPU mentioned. 87

3.7 Error quantification (EPE/AAE in pixels and degrees respectively) for the Grove 3 image pair with and without illumination change. 90

4.1 Comparison of state-of-the-art methods with the proposed **ROF-NND** method with the overall AEPE (in pixels) and AAE (in degrees) on the Middlebury optical flow benchmark [Baker et al., 2011]. Runtimes are provided for the Urban image pair under CPU implementation unless mentioned as GPU. Average ranking is provided online at `http://vision.middlebury.edu/flow/eval/results-e1.php`. . 108
4.2 KITTI flow benchmark comparition for the proposed and existing reference state-of-the-art methods (without incorporating stereo-matching or epipolar geometry). Average runtimes (**t**) are given for CPU implementation unless mentioned. For details see also `http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=flow`. Noc represents errors evaluated for non-occluded regions and occ represents errors evaluated for all image pixels including occlusion. The % of bad pixels and the AEPE for the pixels at AEPE threshold of 3 pixels are given. 109
4.3 Performance on MPI Sintel benchmark (`(http://sintel.is.tue.mpg.de/results)`). Only methods with variational approach implementation are shown here for final pass dataset. The column "s0-10" and "s10-40" represents the AEPE over regions with flow vector magnitudes ranging in $[0, 10]$ pixels and $[10, 40]$ pixels. Average runtime (t) is given for CPU implementation unless mentioned. NA stands for not applicable. 110
4.4 Percentage of bad pixels and AEPE value (in brackets), at AEPE threshold of 3 pixels, for the state-of-the-art methods and the proposed **ROF-NND** method. The results are given for the non-occluded ground truth of the four KITTI training image sequences (**#11, #15, #44 and #74**) which include illumination changes. 114
4.5 Large displacements tests on four training image sequences of the KITTI dataset (pair number 117, 144, 147 and 181) with non-occluded ground truth results. Two criteria at error threshold of 3 pixels (percentage of bad pixels and the AEPE value given in brackets), are used to evaluate state-of-the-art methods and the proposed **ROF-NND** method. See also `http://www.dagm.de/symposien/special-sessions/` for such large displacements tests). 115
4.6 $H_{i,i+1}^{true}$ homography parameter intervals used for computing the displacements between consecutive images. θ, $\{s_x, s_y\}$, $\{f_x, f_y\}$, $\{t_x, t_y\}$ and $\{h_1, h_2\}$ are the in-plane rotation, shear, scale, 2D translation and perspective parameters respectively. 117
4.7 Registration and mosaicing results obtained for datset "data I" (human skin epithelium). 118
4.8 Registration and mosaicing results for dataset "data-II" (pig bladder phantom). . 119

List of Figures

1 Exemple de mosaïque de l'épithelium d'une vessie. Ce champ étendu a été calculé avec 900 images d'une vidéo-séquence d'une durée de 43 secondes. Le cercle blanc en pointillés rep ésente la première image et correspond au champ de vue de l'endoscope. Cette vidéo-séquence (données patient) a été acquise avec un cystoscope rigide durant une procédure clinique standard en lumière blanche. . . xviii

2 Estimation précise du flot optique pour la séquence "marble" (voir le lien `http://i21www.ira.uka.de/image_sequences/`). (a) Image 10 de la séquence, (b) image 20 de la séquence, (c) vérité terrain correspondant au flot optique entre (a) et (b). La teinte représente l'orientation des vecteurs et la saturation de la couleur code la longueur des vecteurs, (d) résultats obtenus avec la méthode variationnelle totale classique basée sur la norme l^1, (e) résultats avec la méthode RFLOW proposée et (f) flot optique obtenu avec la méthode ROF-NDD proposée. xxi

3 Estimation de points homologues dans les images de la vessie. (a) Image source I_{i+1} (image avec le numéro $i+1$ dans la séquence. (b) Image cible ($i^{\text{ième}}$ image de la séquence). La cible a été floutée et la valeur moyenne de ses niveaux de gris abaissée (image assombrie par rapport à celle dans (a)). (c) Champ de déplacements obtenu avec la méthode variationnelle totale classique basée sur la norme l^1. Les vecteurs du champ (flèches) sont visualisés tous les cinq pixels dans les directions x et y des axes des images. (d) Champ de déplacements obtenu avec le modèle proposé. xxii

4 Paire d'images utilisées pour les tests de robustesse vis-à-vis des changements d'illumination. (a) Image originale textures "Grove3" (source). (b) Image cible texturée ("Grove3") dans laquelle des changements importants d'illuminations ont été simulés. (c) Vérité terrain donnant le flow optique exacte entre (a) et (b). (d) Flot optique déterminé avec la méthode ROF-NDD pour la paire d'images dans (a) et (b). xxii

5 Mosaïques d'images de vessie acquises sous deux différentes modalités. (a) Mosaïque construite avec 200 images acquises en lumière blanche. Certaines images sont floues ou affectées par de d'importantes réflexions spéculiares. (b) Mosaïque (modalité de fluorescence) qui visualise une région d'intérêt après une une résection transurétrale. xxiii

6 Bladder epithelium mosaic (extended FOV) using 500 frames of a 20 seconds video-sequence. The black dashed lines represent the camera trajectory and a red rectangle region represents the starting frame (also low FOV seen through cystocopy) xxv

1.1 Bladder wall layers and different stagings of carcinoma. Courtesy: The Urology Group (http://www.urologygroupvirginia.com/). 2
1.2 Sketch illustrating a cystoscopic examination procedure. A cystoscope is inserted into the bladder via the urethra. The images of the acquired video-sequence are displayed on a screen and appear in circular and small field of view (FOV). The image shown in this diagram was obtained for a white light source. Courtesy: The Urology Care Foundation (http://www.urologyhealth.org/). 4
1.3 Endoscopes used in urology. a) Rigid cystoscope, Karl Storz company. b) Flexible cystoscope (EndoEYE model from the Olympus company). 4
1.4 Image mosaicing framework. 6
1.5 Examples of images obtained from different endoscopic applications. (a-b) urinary bladder (WL and FL cystoscopy), (c) near urethral opening (WL cystoscopy), (d) esophagus (gastroscopy), (e) stomach (gastroscopy), (f) larynx (laryngoscopy), (g) pituitary gland (endo-nasal neuro surgery), (h) colon polyp (colonoscopy) and (i) microscopic image of cardiac type epithelium in vivo (confocal laser endomicroscopy, CLE). 7
1.6 Image mosaics obtained for different endoscopic applications given in Fig. 1.5. (a) 2D large extended FOV mosaic for images acquired with a WL cystoscope [Hernandez-Mier et al., 2010], (b) 2D planar panoramic image built generated in real-time using FL cystoscopy video frames [Behrens et al., 2009], (c) 2D image mosaic representing a quasi-planar surface near the urethral opening [Ali et al., 2013b], (d-e) mosaic of gastroscopic quasi-planar image sequences showing extended FOV around angularis [Liu et al., 2015] and pylorus regions, (f) image mosaic of larynx generated with a general-purpose stitching software [Schuster et al., 2012], (g) real time view expansion of an endo-nasal region [Berger et al., 2013], (h) 3D reconstruction of polyp region using a shape-from-motion approach for a colonoscopic image aquisition set-up [Koppel et al., 2007] and (i) extended FOV mosaic of CLE for round cardiac type epithelium in vivo [Vercauteren, 2008]. 9
1.7 Acquisition of cystoscopic data and image texture variability illustration. (a) Schematic sketch of the bladder scene (b) Example of an image with contrasted texture (c) Image with vignetting effect (d) Example with weak contrasted texture image (e) Image with motion blur. 10
1.8 Background removal using median filtering. a) Original image, b) background image obtained with median filtering technique, c) estimated mask from the background image in (b), d) pixel-wise difference of the original gray level image of (a) and the low-pass filtered image (median filter) in (b), e) image normalized to zero mean with pixels from FOV mask only shown in (c), (f) intensity profile V along the red line shown in (a) for the images (b), (d) and (e). 13
1.9 Contrast enhancement in cystoscopic image sequences. (a) poor contrast image (I_{test}) due to view-point, b) target image to which image in (a) need to be registered, c) reference image (I_{ref}), d) enhanced image ($I_{enhanced}$), (e) Singular values profile $V(\sigma)$ in good and bad contrast images. 16

1.10 Mosaicing results based on feature point extraction with and without image pre-processing. a) Impact of large brightness variability affecting few images (the two first images on the right are underexposed). b)The alignment errors are indicated by the arrows which point regions where the textures (vessels) of two images should be perfectly superimposed (these structures are infact shifted). c) Mosaic after SVD enhancement of images. d) After illumination correction with the SVD technique, the structures of the two first images are now perfectly superimposed. 17
1.11 A 3D point $\mathbf{X}$ lying on a plane π has projection $\mathbf{x}_i$ on image I_i an $\mathbf{x}_j$ on image j, ($j = i + 1$ for consecutive image pairs). These points are projectively equivalent and can be mapped by a 2D homography $H^{\pi}_{i,j}$ which can be used to express the points of I_j in the coordinate system of I_i. 18
1.12 An image I showing set of selected points such that points chosen for homography estimation using the 4-point DLT algorithm are well-distributed in the image. . . 18
1.13 Feature extraction under different image quality/texture conditions [Ali et al., 2013b]. a) Contrasted texture: dense matching with SURF feature extraction technique. This image shows its own extracted feature points (in green) and the successfully extracted and matched feature points (in red) of a second contrasted bladder images. b)Blurred textures: the sparse and undistributed feature matching with SURF is illustrated by the too few green-red mark pairs. Registering the image with this poor information lead to inaccurate results. c) Dense correspondence with variational optical flow method on image (b). Red points correspond again to the key feature points in the target image and green points represents the key feature points in the source image. Yellow line connects the matched key feature points in the target and the source images overlaid. 21
1.14 Two composited maps before and after blending. a, c) Without blending, b, d) with Laplacian-Gaussian blending technique described in [Burt and Adelson, 1983, Szeliski, 2006]. Intensity discrepancies along the image transitions during stitching are diminished in the blended mosaics. To limit the contrast expansion with the Laplacian blending algorithm, the background of the blended mosaic has been subtracted form the Laplacian blended mosaic with the weight of 0.1. Structures present in the mosaic are preserved and enhanced while keeping the original texture. In (d) the small structures in the red circles are preserved by the blending technique. 26

2.1 Representation of motion in video-data. a) Video-frame as a function of space (x, y) and time t, b) 2D displacement (u, v) in between consecutive video frames. 30
2.2 Illustration of temporal alliasing effects on optical flow. (a) Small motion giving correct nearest match, *i.e.* no aliasing. (b) Large motion, nearest match is not correct due to aliasing. 31
2.3 Illustration of the ambiguity of the OFC equation. Left: Aperture problem, only flow field (u, v) normal to the edge (denoted by solid arrow) can be computed. However, the two other (dashed) arrows can also represent the actual solution. Right: in images regions without intensity variations all solutions are possible for vector (u, v). 32
2.4 Illustration of intensity order transforms (b-e) in a 3×3 neighborhood patch $\mathcal{P}_{3\times3}$ around the pixel of interest marked in grey. a) Intensity of the original pixels in $\mathcal{P}_{3\times3}$, (b) Rank, (c) Census, (d) Complete rank and (e) Complete census. 40

2.5 Illustration of bahaviour of penalty functions for the spatial terms. In red: Horn and Schunk quadratic term (over-smoothing effect). In blue: Charbonnier convex forulation with ($p = 0.5$). In blue dashed: Charbonnier non-convex penalty with $p < 0.5$. In magenta: Lorenzian function with $\sigma = 0.03$. 42
2.6 Classification of regularizers used in various optical flow models. 44
2.7 Illustration of convexity concept. (a) A convex set. (b) A non-convex (concave) set; (c) A convex function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ is represented by a curve and the line in green represents the convex set of points between the points v_1 and v_2. Linear combination of points which is present on this line gives the convex set in the function domain $F(v_1)$and $F(v_2)$ representing the curve. 47
2.8 a) $F(v)$ is convex and differentiable so $F(v) \geq F(v_1) + \nabla F(v_1)^T(v - v_1)$. b) A function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, and a value $v^* \in \mathbb{R}$ such that it represents the slope of the function F and the conjugate function $F^*(v^*)$ is the maximum gap. 48
2.9 Limitations of gradient descent techniques for non-convex and convex non-differentiable functions. (a) Gradient descent giving local minimal solution for non-convex approach. b) Non-differentiability in convex functions (usually all norms) being modeled as differentiable function by adding a small constant ϵ, $TV(v)$ being a $1D$ representation of the function. 51
2.10 Blob and vessel measures determined by the magnitudes of Eigen values. (a) Blob like structure is obtained when $\lambda_+ \sim \lambda_-$). (b) Elongated vessel structure (with $\lambda_+ >> \lambda_-$) 54
2.11 Structure estimate and its gradient. (a, d) Original image of classical scene and bladder scene. (b, e) Structure estimates of (a) and (d) respectively and (c, f) respective gradient images. 54
2.12 Flow color code. 59
2.13 Visual validation of the improvement of classical TV-L^1 algorithm. (a, b) Frame 16 and 17 respectively of Marble sequence, (c) ground truth flow between (a) and (b), (d) flow field obtained with the classical Horn-Schunck approach, (e) flow field obtained with the classical TV-L^1 algorithm and (f)flow field obtained with the structure estimate (**RFLOW**). 61
2.14 Results on the Middlebury test image sequences. The images (a, d, g) of the first column represents the Ground Truth, the second column (b, e, h) corresponds to the optical flow estimation by the TV-L1-improved and the third column (c, f, i) represents the optical flow estimation using the proposed model. Flow errors are shown adjacent to the color representation of OF field. 62
2.15 Optical flow estimation using the proposed method for dynamic scenes in Middlebury data-set. (a-c) Backyard sequence, (d-e) Basketball sequence. The images pairs are in the first and last column and the flow field is given in the central column. 63
2.16 Homologous point estimation in WL modality. a) Source image I_2, b) target image I_1, c) displacement field obtained using classical TV-l^1 method [Pock et al., 2007], d) displacement field with the proposed model. Target image is blurred and darkened relative to the source image. Flow vectors (arrows) at every 5^{th} pixel in x and y directions are shown. 63

3.1 Effect of various orientation filters on a synthetic test image. (a) Original image (level 0). (b) Image (level 4 of a Gaussian pyramid) obtained after applying Steerable filters having an angular spacing of 15° in the basis filter. (c, d) Riesz filters of 1^{st}-order basis on Gaussian pyramid and Riesz filter with 2^{nd}-order basis filters on DoG pyramid respectively (for level = 4). (e-g) Edge detection on (b), (c) and (d) respectively. (h) 2^{nd} order Riesz wavelet pyramid. 68
3.2 Representation of Riesz wavelet basis filters of order 2 ($N = 2$). 68
3.3 Use of structure information for improved regularization. (a) Representation of the orthogonal major and minor eigenvalues (λ_+, λ_-) and eigenvectors (e_+, e_-) of a structure tensor computed with the grey-levels around $I^j_{i+1}(\mathbf{x} + \mathbf{v}^j)$. (b) Original image with rich texture. (c) Structure tensor image of (b) obtained as $\sqrt{{\lambda_+}^2 + {\lambda_-}^2}$ for each pixel $I^j_{i+1}(\mathbf{x} + \mathbf{v}^j)$. In this example, the background pixels are shown in black (small λ_+ and λ_-, i.e. large D_j in Eq. (3.9)) and the foreground pixels (structures) in white (large λ_+ and/or λ_-, i.e. small D_j in Eq. (3.9)). (d) Anisotropic diffusion tensor image based on the standard diffusion ellipsoids for visualization. It shows that the diffusion tensor acts differently around the torus region: the hollow of the torus (background) and the solid regions of the torus surface (foreground) have different diffusivity in terms of both orientation and their magnitude. 70
3.4 Visual representation of the effect of edge preserving anisotropic regularizer on the Schefflera image. (a) Original image. (b) Ground truth optical flow. (c) and (d) are results for improved TV-L^1 [Wedel et al., 2009b] without and with diffusion tensor respectively. (e) and (f) are the flow results for our TV-approach on wavelet space without and with diffusion tensor regularizer respectively. (g) Flow color code. Flow results are shown for red rectangular part in (a). 71
3.5 Computation of the curl-weight ϕ_w. (a) Angle difference $d\phi$ between the minor Eigenvectors e_- and e'_- of superimposed pixels $\mathbf{x}$ and $\mathbf{x}+\mathbf{v}$ located in the target image I_i and the warped source image I_{i+1} respectively. (b) ϕ_w as a continuous function of $\mid d\phi \mid$. The weight starts at 0.5 for angular difference of 1° which gradually increases in a non-linear fashion reaching the highest curl-weight of 1 for $\mid d\phi^j_{\lambda_-} \mid > 24°$. 72
3.6 Significance of the curl operator in flow regularization energy $E_s(\mathbf{v}^j)$. (a) Original cystoscopic image. (b) Rotation around an axis being perpendicular to the image plane and passing through point O. (c) Image (a) after a 5° pure in-plane rotation. (d) Flow field obtained with the classical regularizer of Eq. (3.7). This flow field corresponds to the pixels of rectangle given in (a). The vectors are quasi-parallel with an over-estimated magnitude and false orientations. (e) Flow field obtained after applying div-curl decomposition (refer to Eq. (3.12)). The vectors correspond to a circular flow and have a magnitude which increased when moving apart from the point O. 73
3.7 Illustration of the impact of the weighted median filtering on the flow field accuracy. (a) Highly textured RubberWhale image [Baker et al., 2011]. (b) Flow ground truth. (c) and (d) Flow estimation using the proposed model without and with weighted median filtering respectively. The rectangular boxes in black surround the regions of interest. In (d), it is visible that the effect of the shadow on the flow accuracy was attenuated (dashed line rectangle) and the torus hole is visible (solid line rectangle). 75

3.8 Parameter settings using the RubberWhale image pair of the Middlebury training dataset. a) AEPE and AAE/10 for different values of λ_s giving the relative importance of the data-term and the regularizer. The tenth of AAE is plotted to represent both the AEPE and the AAE results on a unique decade of values. (b) AEPE and computational time plotted against the scale-factor α. The tenth of the computation time is plotted to represent both performance criteria on a unique decade of values. 78
3.9 Acquisition of bladder phantom video-sequences with controlled displacements. (a) Experimental set-up for acquiring the images of a flattened-out pig bladder with an endoscope. (b) Image of the flattened out pig bladder. 80
3.10 Simulation of video-sequences with known displacements between consecutive images. Two scenes with very different textures are used for the simulation. The black rectangles sketch the extracted sub-images from the high resolution image. These extracted images (with known homographies linking them pairwise) simulate a video-sequence. b) Mars rover curiosity drill of the rock target for sample collection. (*Courtesy: NASA*). 82
3.11 Mean mosaicing error evolution when placing the pixels of image I_i in the coordinate system of image I_0. (a) Pig bladder mosaic built with the proposed AOFW algorithm using simulated sequence I. The trajectory of the simulated image center path is shown in black. (b-e) Detailed view at the start and end of the loop (path closing) for the classical TV-L^1 (in green) [Pock et al., 2007], the graph-cut based method [Weibel et al., 2012b] (in red), the improved TV-L^1 [Wedel et al., 2009b] (in darkblue) and the proposed method (in black) respectively. Visual misalignment is also perceptible along the C-shaped vessel structure indicated by a white line. This error is visually imperceptible for the proposed method in (e). . 86
3.12 Experiments on simulated video-sequence II. a) Mosaic computed with the ground truth homographies. b) Path trajectory of the mosaics computed with the homographies obtained for the reference methods and the ground truth homographies. (c, d) Mosaics with the classical TV-L^1 approach and the proposed AOFW method respectively. The visual misalignments are indicated by arrows and a solid white line in mosaics (c) and (d). Comparing the shape and size of the white "holes" in the mosaic centres shows also that the shape of mosaic (d) is closer to the ground truth shape than that of the map in (c). Misalignment for each method can also be observed at each trajectory point in (b). 88
3.13 Visualization of the effect of strong illumination changes on on the Grove3 sequence of the Middlebury training dataset. (a) Original frame10 of Grove3. (b) Frame11 of Grove3 with strong simulated illumination changes [Drulea and Nedevschi, 2013]. (c) Ground truth flow field. (d) Optical flow obtained with the AOFW method on images without modified illumination. (e) Optical flow obtained with the AOFW method on images with the modified illumination (frame 11 as in (b)). (f) Optical flow obtained with the RFLOW method on images with modified illumination (frame 11 as in (b)). 89

4.1 Visualization of the robustness of the proposed ROF-NND method against illumination changes. (a) Original frame10 of Grove3. (b) Frame11 of Grove3 with illumination. (c) Ground truth flow. (d) Result with ROF-NND before illumination changes ($AEPE/AAE = 0.58/6.05$). (e) Result of ROF-NND after illumination changes ($AEPE/AAE = 0.58/6.05$). 93

4.2 Definition of neighnorhood window $\mathcal{W}$, patches $\mathcal{P}_i$ and of their relative positions in image $I(\mathbf{x})$. (a) The dark blue dot gives the centre of window $\mathcal{W}$ delineated by the orange square. (b) $\mathbf{x}_i$ (light blue dots, $i \in [1, 8]$ in this example) are the neighbors of $\mathbf{x}$ in $\mathcal{W}$. A patch $\mathcal{P}_i$ (having the same size as $\mathcal{W}$) is centered on $\mathbf{x}_i$ and is represented by the red rectangle. The black dots (pixels $p^j_{\mathbf{x}_i}$) correspond to the neighbors of $\mathbf{x}_i$. (c) Illustration of the constant translation between corresponding pixels $\mathbf{x}_j$ and $\mathbf{p}^j_{\mathbf{x}_3}$ ($j \in [0, 8]$) of window $\mathcal{W}$ and patch $\mathcal{P}_3$ respectively. (d) Same "shift" representation for window $\mathcal{W}$ and patch $\mathcal{P}_5$. 96
4.3 Pixel shifts in the 4-connected neighborhood in window $\mathcal{W}_{3\times3}$. On the left: 4-possible shifts in light blue dots $\{\mathbf{x}_1, .., \mathbf{x}_4\}$ around pixel $\mathbf{x}$ in $\mathcal{W}_{3\times3}$. On right: image $I(\mathbf{p}_{\mathbf{x}_3})$ formed with pixel shift on $\mathbf{x}_3 \in \mathcal{W}_{3\times3}$ forming a patch $\mathcal{P}_3$ (in red). The 8-pixels in $\in \mathcal{P}_3$ (black dots) corresponds to the pixel positions of neighborhood pixels in shifted image $I(\mathbf{x}_3)$ around $\mathbf{x}_3$ in original image $I(\mathbf{x})$. 98
4.4 Illustration of the effect of **NND** on edge preservation and under illumination changes due to a shadow. (a) Orginal image with areas affected with shadows in red rectangles. (b) Image of $NND(I, \mathbf{x}, 7)$ (c) Image of $NND(I, \mathbf{x}, 8)$. (d) Ground truth flow field (classical flow color code used). (e) Result obtained with original image in Classical TV-L^1 framework [Pock et al., 2007]. (f) Result obtained with **NND** under similar implementation as the method used for (e). 99
4.5 Tests for setting the optimal size of patch $\mathcal{P}$. (a) Average errors on Middlebury training dataset for varying k. The AAE values in degrees are divided by 10 for representing both errors (AEPE and AAE) on a common range of values (for visualization purpose). (b) % of bad-pixels at AEPE threshold of 3 pixels for the KITTI training dataset with illumination changes (in red) and large displacement (in blue). (c) Average optical flow computation time on the KITTI for increasing k. 105
4.6 Percentage of average bad-pixels at AEPE threshold of 3 pixels (BP3) and average computation time as a function of the scale-factor (α_{scale}). The values are given for a combined MATLAB/C- implementation of the proposed algorithm (**ROF-NND** with $k = 1$) tested on large displacement image pairs (#117, #144, #147 and #181) of the KITTI training dataset. 106
4.7 Optical flow results (in flow color code) obtained with the proposed method (**ROF-NND**) for the training image pairs of the Middlebury dataset along with their corresponding AEPE/AAE (pixels/°). Average overall AEPE/AAE on this training sequence are 0.28 pixels/3.32°. 107
4.8 Optical flow results (in flow color code) obtained with the proposed method (**ROF-NND**) for the Middlebury test dataset (hidden ground truth). 108
4.9 Simulated illumination changes on the second image of the Grove2 image pair. a) Original image I_{in} b) I_{out} with an additive term $a = 30$, c) I_{out} with multiplicative $m = 1.8$ and d) I_{out} with $\gamma = 3.5$. 112
4.10 Illustration of the effect of illumination changes (by varying a, m and γ) on AEPE and AAE for the MLDP method [Mohamed et al., 2014], the census transform [Hafner et al., 2013], the correlation flow [Drulea and Nedevschi, 2013] and the proposed method (ROF-**NND**). 113

4.11 Results obtained for sequence 144 of the KITTI training datasets. First row from the top: two images for which the optical flow has to be determined with known ground truth flow on the right. Second row: results obtained for the MDP-flow2 method [Xu et al., 2012] (on the right: flow field with its usual color code, in the middle: end point error image with small and large values in blue and red respectively, on the right: bar chart of the end point errors in pixels). Third row : same results for the MLDP method [Mohamed et al., 2014]. Fourth row: results for the proposed **ROF-NND** method with $k = 1$. Fifth row : results for the **ROF-NND** method with $k = 2$. 116
4.12 Human skin data-I mosaics. 900×1400 pixels mosaic was obtained with I_0 being the first image. (a) Mosaic with RFLOW method. A visual misalignment is shown between the first frame and the last frame with a red line. (b) Mosaic obtained with the ROF-NND method.. 118
4.13 Results for dataset "data II": pig bladder mosaics built with the homographies $H_{i,i+1}^{est}$ estimated with the proposed ROF-NDD method. This mosaic has a size of 900×1500 pixels. 119

5.1 Strong scene variability inside and between image modalities. (a-b) Intra-patient texture and illumination variability in cystoscopy for images acquired with a rigid (a) and flexible (b) cystoscope repectively. (c-d) Inter-patient illumination variability in dermoscopy. (e-f) Strong specular reflections due to moistness of regions around the organs (stomach and liver respectively). 123
5.2 Illustration of repetitive patterns and of scenes with illumination variability. (a) Non-uniform illumination and repeated texture in underwater scene. The illumination changes occur notably due to reflection of light from moving water during image acquisition. (b) Non-uniform illumination due to the organ depth and viewpoint. (c-d) Illumination changes in bimodality cystoscopic imaging: white light modality (WL) in (b) and fluorescence modality (FL) in (c). 124
5.3 Patient cystoscopic mosaics. (a) First image mosaic of 500 images constructed with the RFLOW algorithm. Scars can be seen on the bladder wall, some of them are shown in black rectangles. (b) Second bladder map of an urethral opening region (200 image pairs). This mosaic was built with the AOFW method. The low textured areas with dashed rectangles represent healed scar regions. A black arrow at loop closing in (b) shows a small misalignment between the vessels. . . . 126
5.4 Mosaic using 500 frames of patient data. Texture and illumination variability is persistent. Illumination changes are due to view-point changes. A polyp is shown in circular black region. Additionally, scale- and perspective- changes are observed when moving from right to left. 127
5.5 Bladder mosaic with strong in-plane rotations and perspective changes. The acquisition of the cystoscopic video-sequence was done after transurethral resection of a bladder tumor. The red circle represents the resected region with blood stains. The black line represents the reconstructed trajectory of the cystoscope projected onto the mosaicing plane. The arrows indicate vessel continuity points (for qualitative/visual estimate of global registration errors) at 30th, 175th and 225th image pairs respectively (left to right). 128

5.6 Mosaic of patient data obtained with flexible cystoscope. (a) Strong specular reflections in all the image pairs additionally with large displacement and strong perspective changes. (b) Strong specular reflection along with non-planar organ surface showing an air bubble in the cavity. 129
5.7 Mosaic of every 10^{th} frame of patient data. Image mosaic showing the transurethral resection (surgical) procedure. White looped wire can be observed at the end (on the right) of the mosaic which is being used for removing the bladder polyp in extended FOV bladder mosaic. Green lines in mosaics represent the reconstructed camera trajectory 129
5.8 Mosaic built with large sequences. (a) Inner bladder wall mosaic using the ROF-NND method. It uses 900 frames corresponding to 35 s of cystoscopic video data. The white circle represents the FOV of a video frame. (b) Video image sequence corresponding to 25 seconds of the same video sequence used in (a) and consisting of 618 frames. A maximal displacement of 30 pixels between the image pairs occurs in this video-sequence extract. 131
5.9 Mosaicing tests under FL modality. (a) Patient data mosaic built with 100 image pairs. (b, c) Bladder region in original small FOV image before and after transurethral resection of bladder tumor (TURBT). The corresponding bladder region in mosaic a) is indicated by an arrow. 132
5.10 Second mosaic with FL data. White circle represents the FOV of the cystoscope and corresponding image under WL and FL are shown respectively in (a) and (b). 132
5.11 Gastroscopy image mosaics of the pyloric antrum region. (a) Image mosaic without strong specular reflections. (b) Mosaic with 70 images with strong specular reflections. Regions in the mosaic having large specular reflections are in the black rectangles and the green arrows mark the structure continuity which demonstrates the quality of image alignment. 133
5.12 Stitching of 100 frames (every 10^{th} frame of the sequence) extracted from a laparoscopic video sequence of the region around the liver. Large specular reflections and brightness changes can be observed in image pairs. The first mosaic image is located at the top right of the map corner. 134
5.13 Human data mosaic of the face and the neck region. The green line represents the camera trajectory. 135
5.14 Underwater mosaics computed for video-sequences described in [Michel J. et al., 2011]. a) Seabed mosaic visualizing sessile fauna, hermit crabs, and horse mussels embedded at the surface of the sediment. b) Mosaic with repeated patterns of brittle stars strewing the seabed (in circle). Blur is perceptible in the left and central mosaic parts in (a). This blur is due to strong tidal currents. Large illumination variability can be observed. 136
5.15 Panorama of the landing site of the NASA's curiosity rover. The black line represents the trajectory of the camera motion. This mosaic was obtained with the AOFW method. 136

Table of contents

Résumé étendu **xviii**

1 Contexte médical . . . xviii
2 Travaux en lien et objectifs de la thèse . . . xix
3 Principales contributions . . . xx
3.1 Terme d'attache aux données incluant des informations de structure . . . xx
3.2 Approche multi-résolution pour la préservation des bords . . . xx
3.3 Descripteurs de voisinages invariant par rapport à illumiation . . . xxi
3.4 Régularisation non locale à l'aide d'un filtrage bilatéral modifié . . . xxi
4 Discussion . . . xxii

General Introduction **xxiv**

Chapter 1 Medical context and scientific objectives **1**

1.1 Medical and scientific motivations . . . 1
1.1.1 Bladder cancer and clinical diagnosis . . . 2
1.1.2 Limitations of cystoscopic examination . . . 3
1.1.3 The need for bladder image mosaicing . . . 5
1.2 Image mosaicing . . . 5
1.2.1 Application to endoscopy . . . 6
1.2.2 Mosaicing trends in endoscopy . . . 8
1.3 2D Bladder image mosaicing . . . 10
1.3.1 Pre-processing . . . 11
1.3.2 Geometrical transformation . . . 15
1.3.3 Bladder image registration and mosaicing . . . 20
1.3.4 Global map correction . . . 22
1.3.5 Post-processing . . . 24
1.4 Thesis objectives and contributions . . . 27
1.4.1 Main contributions . . . 27
List of publication . . . 28

Chapter 2 Optical flow **29**
2.1 Motivation: Fast and Robust establishment of dense correspondences . . . 29
2.2 Optical flow . . . 30
2.2.1 The optical flow constraint . . . 30
2.2.2 Local and global approaches . . . 32
2.3 Modelling of variational optical flow . . . 35
2.3.1 Data-term modeling . . . 35
2.3.2 Regularizer . . . 41
2.4 Mathematical optimization . . . 46
2.4.1 Prerequisites . . . 46
2.4.2 Convex optimization . . . 50
2.5 TV-L^1 optical flow: Background and first contribution . . . 53
2.5.1 Robust energy model (RFLOW) . . . 53
2.5.2 Primal-dual energy minimization . . . 57
2.5.3 Optical flow assessment and benchmarking . . . 59
2.5.4 Results and discussion . . . 60
2.6 Main contributions . . . 64
List of publication . . . 64

Chapter 3 Anisotropic optical flow on edge preserving Riesz wavelet basis **65**
3.1 Motivation: Improved robustness for weak textured images . . . 65
3.2 Optical flow on edge preserving wavelet basis . . . 66
3.2.1 Classical brightness constancy in multi-resolution framework . . . 66
3.2.2 Multi-resolution with Riesz basis . . . 67
3.3 Anisotropic regularization . . . 69
3.3.1 Tensor based anisotropic regularization . . . 70
3.3.2 Div-curl decomposition of $\mathbf{v}$ . . . 72
3.3.3 Weighted non-local median filtering . . . 74
3.4 Optimization . . . 75
3.5 Optical flow algorithm overview and parameter settings . . . 76
3.5.1 Algorithm overview . . . 76
3.5.2 Parameter setting . . . 77
3.6 Results and discussion . . . 79
3.6.1 Evaluation of motion estimation on the Middlebury database . . . 79
3.6.2 Evaluation of different motion types using an endoscopic set-up . . . 79
3.6.3 Evaluation with simulated homographies sparse textured scenes . . . 81
3.6.4 Robustness of algorithm against strong illumination changes . . . 89

3.7 Main contributions and conclusion . . . 90
List of publication . . . 91

Chapter 4 Illumination invariant optical flow using neighborhood descriptors 92
4.1 Motivation: Robustness to illumination changes . . . 93
4.1.1 Related work: Illumination robust methods . . . 94
4.1.2 Aim of the chapter and its organization . . . 95
4.2 Proposed optical flow approach (**ROF-NND**) . . . 95
4.2.1 Neighborhood descriptors . . . 95
4.2.2 Normalized neighborhood descriptor vectors . . . 97
4.2.3 Illustration of effect of neighborhood descriptors . . . 98
4.2.4 **NND** as data term in variational flow . . . 99
4.2.5 Non-local filtering in flow regularization . . . 100
4.2.6 Energy minimization . . . 101
4.3 Coarse-to-fine optical flow approach of the **ROF-NND** algorithm . . . 103
4.4 Experiments on public datasets and algorithm benchmarking . . . 104
4.4.1 Choice of algorithm parameters . . . 105
4.4.2 Experiments on Middlebury benchmark . . . 107
4.4.3 Experiments on KITTI benchmark . . . 109
4.4.4 Experiments on MPI Sintel benchmark . . . 110
4.4.5 Discussion on benchmarking . . . 111
4.5 Experiments for illumination invariance . . . 112
4.6 Experiments for large displacements (chosen $\alpha_{scale} = 0.7$) . . . 114
4.7 Image mosaicing of low textured medical scenes . . . 117
4.7.1 Datasets and evaluation criteria . . . 117
4.7.2 Validation results . . . 118
4.8 Main contributions and conclusion . . . 120
List of publication . . . 121

Chapter 5 Image mosaicing in endoscopy and of other complicated scenes 122
5.1 Motivation: Fast, accurate and robust image mosaicing of various complicated real data sequences . . . 122
5.2 Point correspondence estimation for for complicated scenes . . . 123
5.3 Endoscopic dataset . . . 124
5.4 Qualitative mosaicing results on patient data . . . 125
5.4.1 White light cystoscopy . . . 125
5.4.2 Fluorescence cystoscopy . . . 131

5.4.3 Gastroscopy . . . 133
5.4.4 Laparoscopy . . . 134
5.5 Quality mosaics for other scenes . . . 134
5.5.1 Dermoscopy . . . 134
5.5.2 Underwater scenes . . . 135
5.5.3 Video mosaic of the Mars surface . . . 135
5.6 Main contributions and conclusion . . . 137
List of publication . . . 137

Conclusion and perspectives **138**

Bibliography **141**

Résumé étendu

Mots clés : approches variationnelles totales, flot optique, constance de structure, descripteurs de voisinages, régularisation anisotropique, mosaïquage d'images endoscopiques.

1 Contexte médical

L'objectif de cette thèse est de faciliter le diagnostic du cancer de la vessie. Les lésions cancéreuses sont situées sur la paroi interne de la vessie. L'inspection visuelle de l'épithelium de la paroi de la vessie à l'aide d'un cystoscope (endoscope utilisé en urologie) est la procédure clinique standard pour localiser des lésions, par exemple des polypes ou des tumeurs cancéreuses multi-focales. Les lésions identifiées visuellement sont retirées pr voie chirurgicales à l'aide d'outils insérés à travers le canal opérateur opérateur des cystoscopes rigides. Ainsi montré dans le cercle en pointillés de la Fig.1, une limitation majeure de l'outil cystoscopique (et d'autres endoscopes comme les gastroscopes ou es laparoscopes) est lié au fait que le clinicien a uniquement un accés visuel à un petit champ de vue à un instant donné. La partie visible de la paroi épithéliale correspond à une surface allant de 4 à 7 centimètres carrés. Une visualisation dans leur globalité de lésions

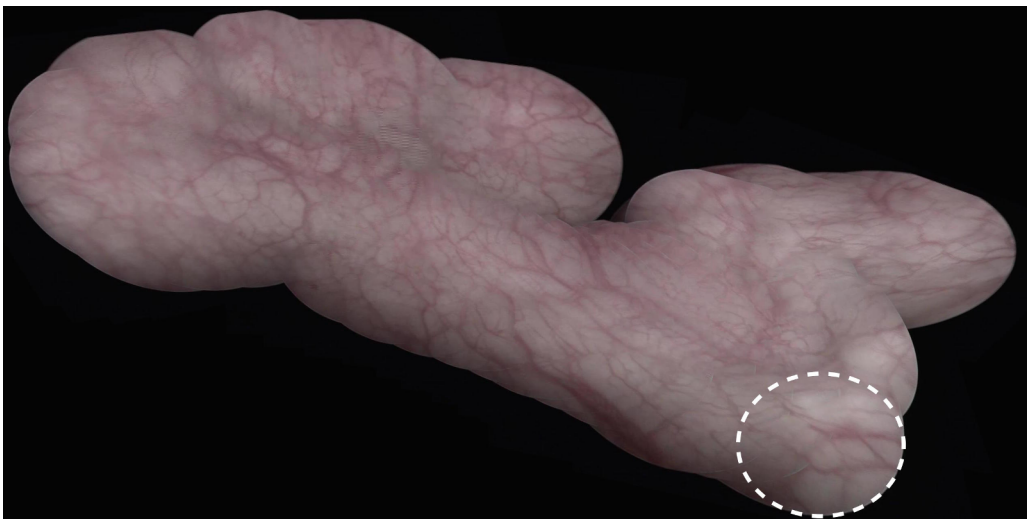


Figure 1: Exemple de mosaïque de l'épithelium d'une vessie. Ce champ étendu a été calculé avec 900 images d'une vidéo-séquence d'une durée de 43 secondes. Le cercle blanc en pointillés rep ésente la première image et correspond au champ de vue de l'endoscope. Cette vidéo-séquence (données patient) a été acquise avec un cystoscope rigide durant une procédure clinique standard en lumière blanche.

.

cancéreuses ou de cicatrices, qui s'étalent en générale sur des grandes surfaces de la paroi de la vessie, est impossible en une image unique de la vidéo-séquence. Le fait que la reconstruction "mentale" de la scène est fastidieuse et très prenante en termes de temps (notamment à cause du manque de visibilité de repères anatomiques comme les uretères ou l'urètre) contribue également à la difficulté du diagnostic.

La construction d'images à large champ de vue (panoramas ou mosaïques) à l'aide des séquences d'images acquises un durant des cystoscopies peut être une réponse aux problèmes médicaux mentionnés précédemment. Le diagnostic peut être amélioré à l'aide des mosaïques qui visualisent des lésions cancéreuses en une image unique. Par ailleurs, comparer des images à champ de vue étendu calculées pour deux (ou plus de) séquences cystoscopiques facilite le suivi des patients et diminue le temps requis pour ce suivi par les urologues ou les chirurgiens (ce suivi est impossible en comparant directement des vidéo-séquences acquises pour un patient à plusieurs semaines ou mois d'intervalle). La visualisation de champs de champs de vue étendus de la paroi interne de la vessie est rendu possible par des algorithmes de mosaïquage d'images. Le mosaïquage d'images est un processus qui place les images correspondant chacune à une vue partielle de la scène dans un repère global unique pour obtenir un champ de vue étendu de l'ensemble de la scène. Une étape clé du mosaïquage d'images réside dans la détermination de la correspondance entre les pixels homologues de paires d'images. Ces correspondances permettent la détermination des transformations géométriques entre paires d'images qui sont requises pour le placement des données dans un système global de coordonnées. La méthode qui permet d'obtenir cette correspondance est désignée par le terme "recalage d'images".

2 Travaux en lien et objectifs de la thèse

Les données cystoscopiques sont caractérisées par des changements d'illumination entre images (par exemple dues à des changements de points-de-vue) et par des variations des caractéristiques de la scène (par exemple variabilité de texture intra et inter-patient). Des méthodes de recalage d'images basées sur l'extraction de caractéristiques images (par exemple basées sur des approches de type SURF ou SIFT) peuvent être appropriées dans le cas de textures contras ées et lorsque ces textures sont bien réparties sur les images. Néanmoins, ce type d'approches sont incapables d'établir une correspondance précise et sans ambiguités dans les scènes compliquées de la vessie dans lesquelles les textures sont faiblement marquées et/ou pré sentent un contraste très variable. Les approches de recalage basées directement sur les donées iconiques (ou sur les valeurs des "pixels") permettent de mieux apréhender les fortes variabilités en termes de textures ansi que les conditions d'illumination non uniformes dans les pairs d'images. Dans la litérature, ces méthodes ont été mises œuvre avec succès dans le contexte du mos ïquage d'images de la vessie. Des travaux menés par le passé au Centre de Recherche en Automatique de Nancy reposaient sur l'utilisation d'une mesure de similarité basée sur l'information mutuelle [Miranda-Luna et al., 2008], sur des méthodes locales de flot optique [Hernandez-Mier et al., 2010] et sur des méthodes de coupes de graphes [Weibel et al., 2012b]. Ces trois approches ont conduit à un mosaïquage précis. Parmi ces méthodes, celles reposant sur l'information mutuelle et sur l'approche des coupes de graphes étaient les plus robustes. Cependant, cette robustesse a été obtenue au détriment des coûts de calculs qui devinrent élevés (environ une minute est requise pour recaler les pairs d'images consécutives de la vidéo-séquence). Dans cette thèse nous avons développé des méthodes précise et robustes de flot optique dense basées sur des méthodes variationnelles totales minimisant des nouveaux modèles d'énergies. Les algorithmes proposés fournissent la correspondance dense (entre pixel homologues) qui est requise pour superposer deux images à l'aide d'une homographie.

Une norme l^1 a été utilisée pour modéliser à la fois le terme d'attache aux données et le terme de régularisation. Une des contributions importantes de ce travail réside dans l'incorporation de textures de vessies (celles des vaisseaux sanguins) dans le terme d'attache au données pour obtenir un modèle robuste et précis. Une régularisation anisotropique a également une régularisation anisotropique a également été proposé une régularisation anisotropique a également été proposé eté proposé afin de préserver les discontinuités des champs de vecteurs au niveau des points de contours. Le modè le d'énergie proposé est convexe et peut ainsi être potentiellement utilisé dans le cadre d'une implémentation parallélisée sur une architecture GPU autorisant des grandes vitesses de calcul.

3 Principales contributions

3.1 Terme d'attache aux données incluant des informations de structure

Les images de vessie contenant usuellement des vaisseaux sanguins avec une forme caractéristique (allongées), nous avons exploité cette information de structure en tant que donnée qui complémente l'hypothèse classique de constance des intensités utilisée dans les approches variationnelles pour le calcul du flot optique (voir le modèle de référence de Horn et Schunck). La modification du terme d'attache aux données est basée sur l'hypothèse qu'une même structure vue dans des images consécutives conserve un aspect constant entre deux instants (acquisitions) proches. Cette hyphothèse de constance de structure ne rend pas seulement l'algorithme robuste et précis, mais diminue également les temps requis pour converger vers la solution (superposition des pixels homologues). Ce nouvel algorithme (baptisé RFLOW, de l'anglais Robust FLOW) a amélioré la qualité des champs de vecteurs, non seulement dans les images cystoscopiques, mais aussi pour d'autres types de scènes. Un exemple de mosaïque de vessie obtenue avec cette méthode est montré dans la Fig. 1. Dans cette approche globale (RFLOW), les champs de vecteurs du movement apparent sont propagés à partir des pixels avec des informations de structure vers des points images situés dans des zones homogènes des images (avec des couleurs avec très peu de variation). Ceci rend possible le recalage de paires d'images extraites vidéo-séquences relativement longues et visualisant des scènes affectées pas une grande variabilité en termes de textures.

3.2 Approche multi-résolution pour la préservation des bords

Une approche pyramidale multi-résolution est utilisée pour traiter les grands déplacements dans les scènes tout en diminuant les temps de calculs requis pour l'estimation du movement apparent. Classiquement, les images brutes sont sous-échantillonnées ou des filtres gaussiens sont employés pour améliorer le sous-échantillonnage dans le but de calculer un flot optique précis. Un problème majeur de ces techniques est qu'aux niveaux à faible résolution de la pyramide les structures visibles dans les images (par exemples des textures ou des contours d'objets) sont sur-lissées. un tel sur-lissage conduit souvent à une mauvaise initialisation du champ de vecteur au niveaux à haute résolution des pyramides. Dans le domaine du flot optique, ce problème est souvent désigné par le terme "d'aplanissement" du mouvement apparent. Pour traiter ce problème, nous avons mis en œuvre des filtres directionnels auto-adaptatifs appelés les ondelettes de Riesz qui sont utilisés pour préserver les structures ou les contours au niveau des images à faible résolution. Ces filtres augmente de manière significative le contraste des structures. Ainsi, mise à part l'amélioration de la décomposition pyramidale, la projection des pixels sur une telle base de

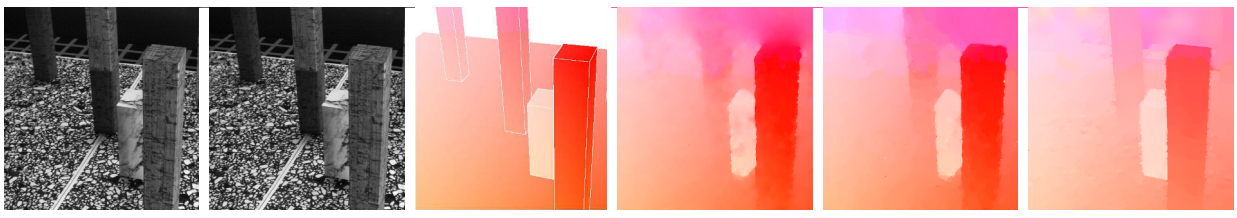
filtres peut être directement utilisé en tant qu'information complémentaire et additionnelle dans les termes d'attache aux données.

3.3 Descripteurs de voisinages invariant par rapport à illumiation

Les contributions proposées jusqu'ici au CRAN dans le domaine du recalage d'images de la vessie ont été conçues pour le mosaïquage d'images acquises dans la modalité de la lumière blanche. Celles-ci sont difficilement adaptables à la modalité complémentaire de fluorescence. La raison principale de ce manque de flexibilité est lié aux termes d'attache aux données utilsés dans ces approches. Ceux-ci rendent les méthodes proposées par le passé sensibles aux changements d'illumination (pas de robustesse vis-à-vis de forts changements d'illumination). Néanmoins, dans certaines séquences d'images, la variabilité en termes d'illumination est forte à cause de grands changements des points-de-vue ou à cause d'un changement de modalité (qui survient par exemple lorsque le clinicien commute entre la lumière blanche et la fluorescence). Nous avons modélisé un nouveau terme d'attache aux données basé sur les propriétés "d'auto-similarité" des descripteurs de voisinage. Pour ce faire, nous avons défini un term d'attache aux données utilisant des vecteurs de descripteurs de voisinage de dimension n calculés pour chaque pixel d'intérêt. Les pixels dans un voisinage de taille n décrivent les caractéristiques de ce voisinage à l'aide d'un modèle gaussien. Par exemple, un pixel entouré de pixels avec des valeurs comparables à celui du point image d'intérêt aura un impact plus important sur le terme d'attache aux données que des pixels non similaires. Cette approche (appelée ROF-NDD, de l'anglais robust optical flow using neighborhood descriptors) est invariant vis-à-vis de forts changements d'illumination et permet le mosaïquage acquises en lumière blanche et/ou dans la modalité de fluorescence.

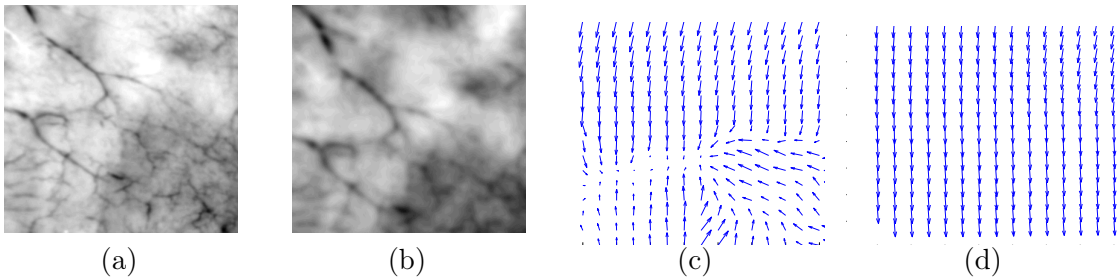
3.4 Régularisation non locale à l'aide d'un filtrage bilatéral modifié

Ainsi mis en lumière par la littérature récente sur le flot optique, les filtres bi-latéraux sont bien connus pour leur préservation précise des discontinuités le long des pixels appartenant à des contours. Ces techniques non-locales de régularisation utilisent trois différentes mesures de corélation, à savoir une mesure de distance spatiale, une distance entre couleurs et une mesure de l'occlusion des pixels d'intérêts dans le voisinage considéré. Une mesure additionnelle de cohérence des structures a été intégrée dans le terme de régularisation afin d'augmenter la précision de ces filtres bi-latéraux classiques. Une intégration directe des filtres bi-latéraux modifiés dans



a) b) c) d) e) f)

Figure 2: Estimation précise du flot optique pour la séquence "marble" (voir le lien http://i21www.ira.uka.de/image_sequences/). (a) Image 10 de la séquence, (b) image 20 de la séquence, (c) vérité terrain correspondant au flot optique entre (a) et (b). La teinte représente l'orientation des vecteurs et la saturation de la couleur code la longueur des vecteurs, (d) résultats obtenus avec la méthode variationnelle totale classique basée sur la norme l^1, (e) résultats avec la méthode RFLOW proposée et (f) flot optique obtenu avec la méthode ROF-NDD proposée.



(a) (b) (c) (d)

Figure 3: Estimation de points homologues dans les images de la vessie. (a) Image source I_{i+1} (image avec le numéro $i+1$ dans la séquence. (b) Image cible ($i^{\text{ième}}$ image de la séquence). La cible a été floutée et la valeur moyenne de ses niveaux de gris abaissée (image assombrie par rapport à celle dans (a)). (c) Champ de déplacements obtenu avec la méthode variationnelle totale classique basée sur la norme l^1. Les vecteurs du champ (flèches) sont visualisés tous les cinq pixels dans les directions x et y des axes des images. (d) Champ de déplacements obtenu avec le modèle proposé.

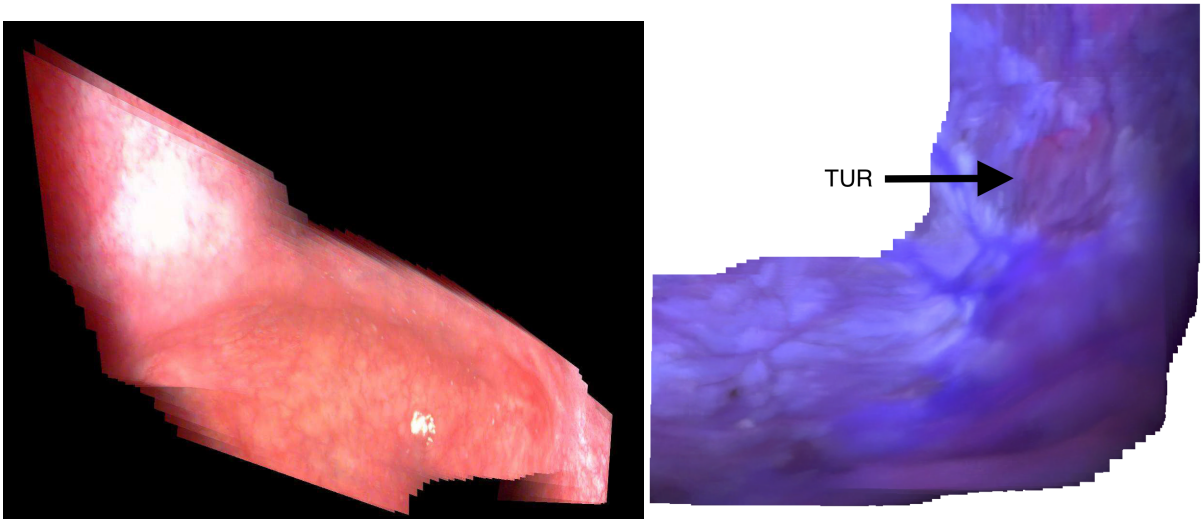
le cadre une minimisation d'énergie par une approche totale variationnelle est également une des contributions de cette thèse. Une telle intégration rend l'algorithme non seulement plus précis, mais augmente également la vitesse de convergence.

4 Discussion

Les résultats donnés dans la Fig. 2 illustrent visuellement, qu'en comparaison avec les approches variationnelles totales classiques, la méthode proposée est plus précise dans la conservation des discontinuités des champs de vecteurs au niveau des pixels des contours. Alors que la méthode dite ROF-NDD a la meilleure précision avec une erreur moyenne d'estimation du déplacement (AEPE) qui vaut seulement 0.21 pixels, les approches variationnelles totales classiques sont les moins précises et conduisent à des AEPE de 1.23 pixels pour les mêmes images (voir la paire d'images dans les Figs. 2(a-b)). La méthode RFLOW proposé dans cette thèse utilise une information de structure présente dans les images de la vessie et est robuste vis-à-vis de petits changements de perspectives ou du floutage des images. Cependant, pour des changements importants d'illumination (comme des réflexions spéculaires ou des changements de modalités), la méthode ROF-NDD peut être utilisée. Les deux méthodes sont illustrées ci-dessous avec des données de référence.

a) b) c) d)

Figure 4: Paire d'images utilisées pour les tests de robustesse vis-à-vis des changements d'illumination. (a) Image originale textures "Grove3" (source). (b) Image cible texturée ("Grove3") dans laquelle des changements importants d'illuminations ont été simulés. (c) Vérité terrain donnant le flow optique exacte entre (a) et (b). (d) Flot optique déterminé avec la méthode ROF-NDD pour la paire d'images dans (a) et (b).



a) b)

Figure 5: Mosaïques d'images de vessie acquises sous deux différentes modalités. (a) Mosaïque construite avec 200 images acquises en lumière blanche. Certaines images sont floues ou affectées par de d'importantes réflexions spéculiares. (b) Mosaïque (modalité de fluorescence) qui visualise une région d'intérêt après une une résection transurétrale.

L'exemple donné dans la Fig. 3 implique une paire d'image ayant entre elles une un déplacement connu pour chaque pixel. la deuxième image de cette paire a de plus été floutée avec un filtre gaussien (l'écart type σ valant 0,25 avec des niveaux de gris dont la dynamique est définie par l'intervalle [0, 1]) . Les changements de contrast dans la deuxième image ont également été modifiés suivant une loi logarithmique. La vérité terrain (champ de vecteur qui doit être retrouvé) est constitué de vecteurs parallèles avec une amplitude quasi-constante. La méthode RFLOW décrite dans cette thèse a utilisé avec succès l'information de structure et conduit de manière robuste à des résultats précis. En effet, dans la Fig. 3(d), les vecteurs sont parallèles. A l'opposé, dans la Fig. 3(c), la méthode variationnelle totale classique conduit à de nombreux faux vecteurs (vecteurs non parallèls, donc mal orientès). Les tests d'invariance vis-à-vis de l'illumination avec la méthode ROF-NDD sont montrés dans la Fig. 4. Ainsi observé dans a Fig. 4(d), le champ de vecteur estimé avec cette méthode est très proche de la vérité terrain visualisée dans la Fig. 4(c). Une mosaïque en lumière blanche, avec de forte réflexions spéculaires et des textures faiblement contrastée est également montrée dans la Fig. 5(a), alors qu'une deuxième séquence d'images acquises avec un agent de contraste de fluorescence est donnée dans la Fig. 5(b). Les résultats obtenus pour ces deux séquences témoignent de la robustesse de algorithmes ROF-NDD pour le mosaïquage d'images acquises sous différentes conditions d'illumination.

General Introduction

This thesis was written at the CRAN laboratory (Centre de Recherche en Automatique de Nancy, UMR 7039 CNRS/Université de Lorraine) in the department of SBS (Santé-Biologie-Signal). The work was co-funded by the Conseil Régional de Lorraine and the Agence nationale de la recherche (ANR) in the framework of the CyPaM2 project (ANR-11-TECS-001). The aim of this project is to facilitate bladder cancer (BCa) diagnosis, notably through the extended field of view (FOV) mosaics. This thesis aims at the development of robust and accurate image registration algorithms suitable for various challenging conditions observed in cystoscopic sequences. However, the application of these algorithms is not kept limited to cystoscopic scenes but have been extended to scenes with similar severe complications like in gastroscopy, laparoscopy and under-water scenes. Medical expertise and cystoscopic data were supplied by the Institut de Cancérologie de Lorraine (ICL), which is a comprehensive cancer center located in Nancy, France.

BCa lesions are found on the internal bladder wall. Visual inspection of the epithelium of the bladder wall done with a cystoscope (endoscope used in urology) is the standard clinical procedure for localizing lesions like polyps or multifocal cancerous tumors. Visually identified lesions are surgically removed using tools inserted through the operative channel of the rigid cystoscopes. As shown by the red rectangle in Fig. 6, a major limitation of a cystoscopic device (and of other endoscopes such as gastroscopes or laparoscopes) is that the clinician can only visualize a small field of view (FOV) at a particular time instant. The visible part usually corresponds to an area varying from one up to several (nearly $4-7$) square centimeters. The complete visualization of lesions or scares, which are usually spread over large areas on the bladder wall, is not possible in one unique image. The fact that it is tedious and time consuming to reconstruct the bladder scene mentally (notably due to the lack of visible landmarks like ureters or urethra) contributes to the difficulty in diagnosis.

Constructing large FOV images (panoramas or mosaics) using image sequences acquired during cystoscopy can be a solution to the previously mentioned medical problems. With mosaics, the diagnosis can be improved since the lesions are completely seen in one unique image. Moreover, comparing extended FOV images computed for two or more cystoscopic scans make the follow-up easier and less time consuming for the clinicians (urologists or surgeons). Such an extended visualization of the internal bladder wall is done by the technique called "image mosaicing". Image mosaicing is the process by which images of partial views of a scene are placed into one global coordinate system to obtain an extended FOV of the whole scene. The key step in image mosaicing is the determination of the correspondence between homologous pixels of image pairs. This correspondence allows for computing the geometrical transformations between image pairs required to place all frames into a global mosaic coordinate system. The method to be used to establish the correspondence between homologous pixels is referred as "image registration".

Exploiting two large FOV mosaics simplifies the work of urologists since, on the one hand, landmarks like ureters or urethra directly locate the regions to be compared and, on the other hand, complete lesions can be visually compared. Moreover, videos contain highly redundant data

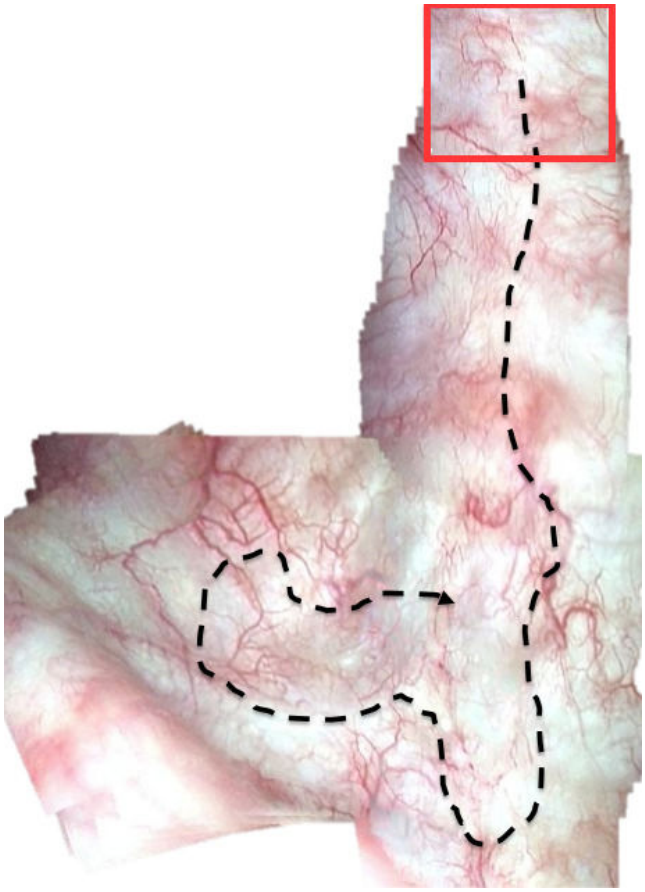


Figure 6: Bladder epithelium mosaic (extended FOV) using 500 frames of a 20 seconds video-sequence. The black dashed lines represent the camera trajectory and a red rectangle region represents the starting frame (also low FOV seen through cystocopy)

which complicate their exploitation after examination and their archiving. A mosaic will contain the most valuable information in a compact way facilitating archiving and use for follow-up. Mosaics represent also a valuable medium facilitating the discussion between different medical specialists (radiophysicians, oncologists, or urologists). Thus, the main objective of this thesis is to facilitate bladder cancer (BCa) diagnosis and treatment through accurate and robust image registration algorithm/s.

Cystoscopic video-sequences are affected by illumination changes between images (i.e. due to viewpoint changes) and varying scene characteristics (e.g. high inter- and intra- patient texture variability). "Feature based" image registration approaches (e.g. based on SIFT or SURF) can be appropriate when the textures are well contrasted and well spread in the image pairs. But, such approaches are unable to establish unambiguous correspondences in complicated bladder scenes with weakly pronounced texture and/or presence of varying contrast. "Pixel based" registration methods are more suitable to deal with large texture variability and non-uniform illumination conditions which are present in bladder image pairs. These methods have been successfully used in the literature of the bladder image mosaicing. Previously, at the CRAN laboratory, a mutual information based similarity measure [Miranda-Luna et al., 2008], a local optical flow technique [Hernandez-Mier et al., 2010] and a graph-cut based optimization method [Weibel et al., 2012b] were exploited for accurate and robust bladder image mosaicing. Among these methods, mutual

information and graph-cut based implementations are the most robust ones. But, this robustness was obtained at the cost of high computational time (nearly a minute was needed for registering one consecutive image pair of the video-sequences). In this thesis, we have implemented accurate, robust and dense global optical flow approaches based on first-order prima-dual energy minimization of total variational (TV) energy models. These proposed algorithms provide dense correspondences (between homologous pixels) required for superimposing two images with an homography. A l^1-norm has been used in both the data-term and the regularizer. One of the important contributions is the modeling of robust and accurate data-terms by incorporating in them some bladder texture information like blood vessels. In order to preserve flow field discontinuities along the edge pixels, anisotropic regularization schemes have also been proposed to optimally constrain the flow fields. The proposed model is convex and thus can be potentially used for parallel processing of the proposed energy minimization on high speed GPU architecture.

This thesis is organized into following chapters:

Chapter 1. This chapter first introduces the need for image registration and mosaicing in cystoscopic image data, and more generally for endoscopic data. Then, a literature review on image registration of endoscopic videos and their trends for extending of endoscopic FOVs for facilitating diagnosis and treatment is presented. It has been shown that the objective of achieving extended FOV can be achieved by performing either 2D or 3D cartography. However, for scenes which are planar or quasi-planar, the most convenient and interesting is 2D image mosaicing. Several steps of image mosaicing are discussed separately and image registration is identified as the key-step of this mosaicing framework. It explains how achieving sub-pixel image registration accuracy between image pairs can lead to coherent mosaics without demanding computationally expensive bundle adjustment methods for global error correction in mosaics. Moreover, pre-processing steps for distortion correction and contrast enhancement are presented for the specific case of endoscopic images. An efficient technique for estimating an accurate homography by choosing correspondences from four different quadrants of an image is also presented. In gist, this chapter focuses on both medical and scientific challenges in endoscopic data and presents classical methods so far described in the literature. It also supports the choice of total variational methods for dense optical flow estimation for cystoscopic image registration.

Chapter 2. This chapter is dedicated to the literature review on global optical flow methods. One major aim of this chapter is to show that the dense optical flow methods are the best choice for establishing dense correspondences between image pairs in complicated medical scenes. To do so, it describes different energy models that have been developed in the past. A thorough overview on designing robust data-terms and on the efforts for accurately preserving the flow field vectors via constrained hypothesis (regularization) is presented. A section of this chapter is dedicated to the state-of-the-art methods for illumination robust and large displacement optical flow methods estimation. A general total variational energy minimization using a first-order primal-dual framework in convex optimization is also presented. This chapter also integrates the first contribution of this thesis work. Since bladder images contain usually elongated blood vessels or blob-like vessel nodes, structure information has been incorporated as the complementary information to the classical brightness constancy assumption of variational approaches in optical flow. The proposed structure constancy is based on the assumption that the textures existing in consecutive images remain of similar shape. This assumption not only makes the algorithm robust and accurate, but also speeds up the algorithm convergence time. An example of

the bladder mosaic obtained by this novel method (RFLOW) is shown in Fig. 6. In the RFLOW approach, the flow fields are propagated from pixels with texture information to pixels in homogeneous image regions. These first results validate the choice of total variationnal approaches for the establishment of dense correspondences between bladder images.

Chapter 3. A coarse-to-fine multiresolution scheme is used to handle large displacements in the scene as well as to decrease the computational time of the flow field estimation. Classically, down-sampling of raw images is done or Gaussian filters are used to improve the down-sampled image quality so that accurate optical flow fields can be estimated. A major problem in these techniques is that the image structures can be over-smoothed at coarser levels. Such over-smoothing often results in false flow-field initialization to the finer levels. This effect is also referred as "flattening-out" problem in optical flow. In order to address this problem, in this chapter, a self-adaptive directional filters called Riesz wavelets have been used to preserve the edges of structures. This algorithm, referred to as "AOFW", also uses a novel anisotropic regularization technique of the flow fields. This regulariser is based on bilateral filters for preserving flow field discontinuities accurately. This non-local regularization technique uses 3 different correlation measures based on spatial information, color values and occlusion of a pixel of interest with respect to the neighboring pixels. An additional structure coherence measure has been integrated in the regularizer to increase the accuracy of such classical bilateral filters. A direct integration of the modified bilateral filter into the framework of the TV approach energy minimization is also one of the contribution of this thesis. Such an integration makes the algorithm even more accurate and speeds up its convergence time.

Chapter 4. Previous contributions on bladder image registration at CRAN were only developed for data acquired under white light (WL) modality. They are difficult to adapt to the complementary fluorescence light (FL) modality. The major reason of this lack of flexibility lies in the data-term used in the previous methods. As these methods are sensitive to illumination changes, they are not robust enough for handling such varying scene conditions. However, in some sequences, illumination variabilities are strong because of large view-point changes and modality changes (e.g. during switch between WL and FL modalities).
This chapter proposes an illumination robust algorithm, referred as "ROF-NND", using neighborhood descriptors in the data-term. A direct integration of a weighted median filtering as the TV-regularizer in the prima-dual first-order energy minimization has also been done in this chapter. Such an integration well preserves the motion discontinuities in the flow fields leading to very accurate flow fields. The optical flow results obtained by the proposed method is benchmarked on three publicly available optical flow databases (Middlebury, MPI Sintel and KITTI datasets). Additionally, dedicated illumination invariance tests are done using classical techniques in the literature. Other robustness and accuracy tests were performed on data with ground truth information (realistic medical phantom data).

Chapter 5. This chapter presents the application of the proposed methods. Each algorithm presented in this thesis has its own advantages with regard to the robustness against given varying scene conditions. For instance, in case of strong illumination changes in image pairs, the ROF-NDD method should be used since the other methods are less robust towards such varying scene conditions. The application of our algorithms to other

endoscopic data and other scenes like underwater and space exploration images are also presented.

Finally, a general conclusion summarizes the main contributions of this thesis and gives perspectives which can improve the potential of the described work.

Chapter 1

Medical context and scientific objectives

Contents

1.1 Medical and scientific motivations 1
1.1.1 Bladder cancer and clinical diagnosis 2
1.1.2 Limitations of cystoscopic examination 3
1.1.3 The need for bladder image mosaicing 5
1.2 Image mosaicing 5
1.2.1 Application to endoscopy 6
1.2.2 Mosaicing trends in endoscopy 8
1.3 2D Bladder image mosaicing 10
1.3.1 Pre-processing 11
1.3.2 Geometrical transformation 15
1.3.3 Bladder image registration and mosaicing 20
1.3.4 Global map correction 22
1.3.5 Post-processing 24
1.4 Thesis objectives and contributions 27
1.4.1 Main contributions 27
List of publication 28

1.1 Medical and scientific motivations

Image mosaicing algorithm consists of different image processing steps. Besides the medical context, this thesis chapter also presents the state-of-the-art methods in the frame of each step of the bladder mosaicing application. However, this chapter is not restricted to the literature analysis. On one hand it identifies the steps in which no strong improvements are required and the solution chosen and implemented for each step in mosaicing chains presented in Chapters 3, 4 and 5 are justified. On the other hand, the aim of the chapter is also to identify the steps in which it is crucial to develop new methods for improving the registration algorithms. To reach this goal, this chapter is already based on preliminary personal works and results validated by publications. For this reason, the presented literature includes (may be in an unusual way for a thesis) own and early publications accepted in the first year of thesis.

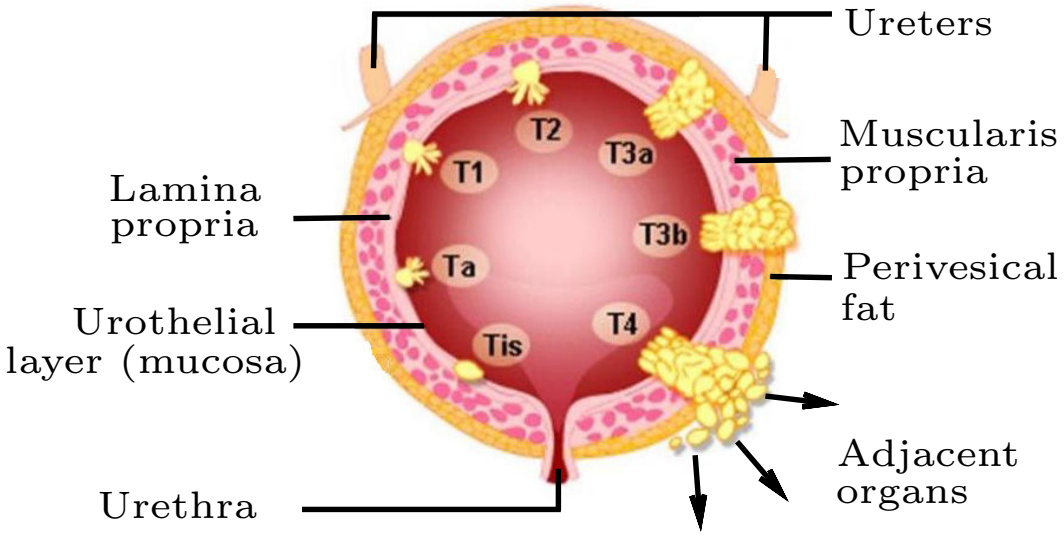


Figure 1.1: Bladder wall layers and different stagings of carcinoma. Courtesy: The Urology Group (http://www.urologygroupvirginia.com/).

1.1.1 Bladder cancer and clinical diagnosis

Bladder Cancers (BCa) are the 4^{th} most prevalent cancers in men and 9^{th} most common cancers among women in developed countries [Ferlay et al., 2013]. BCa are the 2^{nd} most frequent malignant cancers of the urinary tract after prostrate cancer [Pezaro et al., 2012]. In 2012, more than 151,000 new cases of BCa were diagnosed in Europe and accounted to nearly 52,400 deaths. During the same year, 429,000 new cases of BCa were detected worldwide with varying mortality rates across the globe. Approximately 70% of patients with BCa are over 65 years of age [Bellmunt et al., 2014]. Most patients (almost 90%) are diagnosed with Non-Muscle-Invasive BCa (NMIBCa) which has a high recurrence rate of nearly 78% within 5 years. There is a significant risk (nearly 45% chance in 5 years) of progression of NMIBCa to Muscle-Invasive BCa (MIBCa). The lifetime cost of managing Bca is among highest of all cancers.

A Trans-Urethral Resection of Bladder Tumor (TURBT) is performed during cystoscopic examination (visualization of the bladder inner wall through an endoscope) for determining the stage of BCa. Both NMIBCa and MIBCa are graded according to their location of occurrence, their progression site and affect on other organs. The bladder wall is made up of the following 4 layers (refer to Fig. 4.13):

- the innermost lining of epithelium cells called transitional epithelium or urothelial layer (urothelium),
- the thin layer of connective tissue, blood vessels and nerves called lamina propria,
- the thicker muscular layer referred as muscularis propria and
- the outermost layer of fatty connective tissue separating the bladder from nearby organs.

NMIBCa include Carcinoma *In Situ* (CIS), Ta tumors and T1 tumors. CIS is a very early and difficult to detect cancer types, with cancer cells found on the innermost layer of bladder epithelium. Ta tumors are found superficially located at the bladder lining, while in the T1 stage the

cancerous cells start to enter in the muscular layer by crossing the connective tissue. However, if the cancerous cells have invaded the muscle wall of the bladder and/or have spread to nearby organs then the patient has MIBCa type of cancer (labeled from T2 to T4 stages).

Early stages of bladder cancer are found in the first layers of the internal bladder walls. Therefore, visual scanning of the internal bladder walls with a cystoscope is a standard procedure for clinical examination. The diagnosis of BCa is based on the histological evaluation of bladder tissue biopsis taken during TURBT. A cystoscope is a slender tube like with a light source to illuminate inside the orgna, optical lenses and a CCD (charged coupled device) sensor for image acquisition embedded at the distal tip of the tube. During the examination procedure, the cystoscope is inserted through the urethral opening and advancing into the bladder (Fig. 1.2). This procedure is done either through a rigid cystoscope (Fig. 1.3 (a)) or a flexible fiber-optic cystoscope (Fig. 1.3 (b)). During the examination, the bladder is filled with an isotonic saline solution (Fig. 1.2) which temporarily inflates the organ and limits its warping. The aim of it is to helps the clinicians to navigate inside the organ and to have access to all bladder parts to be able to visualize them on a monitor screen.

Depending upon the type of light source used, there are two well-known modalities in cystoscopic imaging: 1) White Light modality (WL) and 2) Fluorescence Light modality (FL). WL cystoscopy is the exploratory examination of reference and TURBT is performed under this modality to remove visible NMIBCa. FL induced cystoscopy is a complementary modality recommended by the European Association of Urology for CIS patients because it improves the detection rates and allows a more complete removal of tumors. During fluorescence cystoscopy, a drug called Hexaminolevulirate (HAL) is instilled into the bladder using a catheter passing through the urethra. After an hour, an HAL-induced level of porphyrins is accumulated in rapidly proliferating cells (cancer). Illuminated by blue light with a wavelength band ranging in $[380, 450]$ nm, a higher fluorescence can be detected from the cancerous sites with a clear contrast over the healthy non fluorescence tissue (healthy tissue appears blue while the porphyrins that have been accumulated in cancer cells appear red).

1.1.2 Limitations of cystoscopic examination

As shown in Fig. 1.2, a major limiting factor of a cystoscopic device (like any other endoscopic method of visualization) is that the clinician is able to visualize only a limited field of view (FOV) at a particular time instant, usually varying from an area of one up to several square centimeters. As a result he cannot visualize complete lesions, or scares, as they may be spread over tens or hundreds of images. The partial view of lesions do not facilitate the diagnosis. Moreover, due to the small FOV, lesions cannot be localized with respect to landmarks (e.g. urethra, ureters or air bubble). This limitation impedes clinicians from memorizing the exact location in the bladder of regions of interest (e.g with lesions) observed at different time instances. Due to the small FOV, the urologist has to displace the endoscope with back and forth and/or zigzag movements. The fact that it is tedious and time consuming to reconstruct the bladder scene in mind (notably due to the lack of landmark visibility) contributes to the difficulty of diagnosis. Another important limitation of this small FOV is that clinicians can never be sure that the complete bladder wall has been scanned ,without overlooking any part.

To document cystoscopic examinations, a procedure broadly used by surgeons or urologists is to do a manual sketch of the bladder including the anatomical landmarks (usually ureters and urethral openings). During the cystoscopic examination, clinicians mentally get the insight of the bladder scene and keep track of the instrument's positions. Whenever a region of interest is observed, its position is noted in the sketch and also screen-shots are sometimes printed and

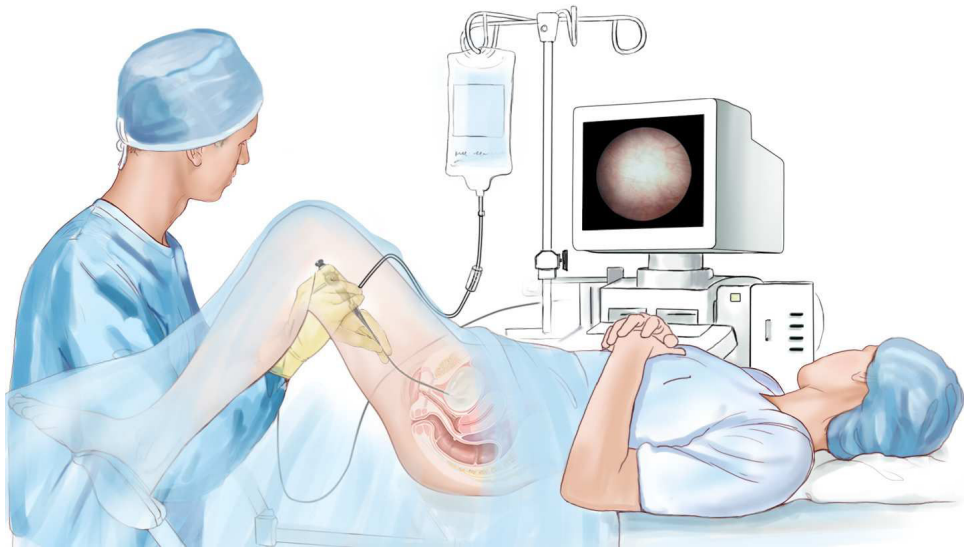


Figure 1.2: Sketch illustrating a cystoscopic examination procedure. A cystoscope is inserted into the bladder via the urethra. The images of the acquired video-sequence are displayed on a screen and appear in circular and small field of view (FOV). The image shown in this diagram was obtained for a white light source. Courtesy: The Urology Care Foundation (http://www.urologyhealth.org/).

pasted on the manual map (the printed images have either a small FOV or a low resolution when the FOV is large). Comments and diagnosis of the clinicians are also classically sound recorded. However, the videos themselves are not recorded since they are hardly exploitable after the examination. Indeed, it is difficult for clinicians who did not acquired the data to use these videos to perform a diagnosis, notably because they have no exact information about the cystoscope trajectory and positions (this information can be mentally and more or less accurately evaluated by clinicians during the examinations by displacing the instrument). Even for those who have facilities for registering the video-sequence, the exploitation of the latter is difficult after the examination.

BCa patients usually have to undergo regular control examinations, the time interval between these controls being often of some months. The archived data do not allow for an optimal patient follow-up since small FOV photographs are not appropriate for an accurate lesion evolution assessment. Moreover, even by archiving the videos, the follow up remains difficult since

(a) (b)

Figure 1.3: Endoscopes used in urology. a) Rigid cystoscope, Karl Storz company. b) Flexible cystoscope (EndoEYE model from the Olympus company).

comparing two image sequences is very tedious (the classical duration of a video is about 6 to 10 minutes) and the evaluation of lesion evolution is very difficult. Moreover, for the same reasons, the archived data are also not appropriate for ensuring examination and treatment traceability.

1.1.3 The need for bladder image mosaicing

Constructing large FOV images (panoramas or mosaics) using image sequences, while preserving at best the initial data resolution can be a solution for the previously mentioned medical problems. With mosaics, the diagnosis can be improved since the lesions are completely seen in one unique image. Moreover, the reoccurence rate of BCa is very high (almost 90%) in organ regions close to tissues which were treated for cancers. Thus, it becomes important to observe the complete nearby areas of lesion occurrence in wide FOV images. Moreover, comparing extended FOV images between them makes lesion follow-up and data archiving easier, efficient and very interesting to the clinicians for both diagnosis and follow-up.

Indeed, exploiting two large FOV mosaics simplifies the work of urologists since, on the one hand, landmarks like ureters or urethra directly locate the regions to be compared and, on the other hand complete lesions can be visually compared. Moreover, videos contain highly redundant data which complicate their exploitation after examination and their archiving. A mosaic will contain the most valuable information in a compact way facilitating archiving and use for follow-up. Mosaics represent also a valuable medium of medical data facilitating the discussion between different medical specialists (radio-physicians, oncologists, or urologists). Such extended FOV images can be obtained using image mosaicing techniques.

1.2 Image mosaicing

Image mosaicing also referred to as "image stitching" is the technique of aligning partially overlapping images of different parts of the same scene into one global coordinate system to obtain a wider field of view. This technique also improves the image resolution of the scene. Several steps are required to build visually coherent mosaics.A complete image mosaicing chain often consists of four main steps (refer to block diagram in Fig. 1.4):

1) a pre-processing step (e.g. blurred frame rejection or image sequence selection, image distortion correction, vignetting effect handling and compensation of scene illumination differences between images),

2) paired image registration of the selected sequence,

3) a stitching step consisting in placing all images in a common and global mosaic coordinate system and

4) a final map correction step for obtaining visually coherence maps.

One crucial step in image mosaicing is the determination of the correspondence between homologous pixels of image pairs. This correspondence allows for computing the geometrical transformations between image pairs required for placing all frames in one global mosaic coordinate system (image stitching). The method to be used to establish the correspondence between homologous pixels depend on the image content and quality and is referred as "image registration".

Image stitching has been used widely in photography and photogrammetry. This technique

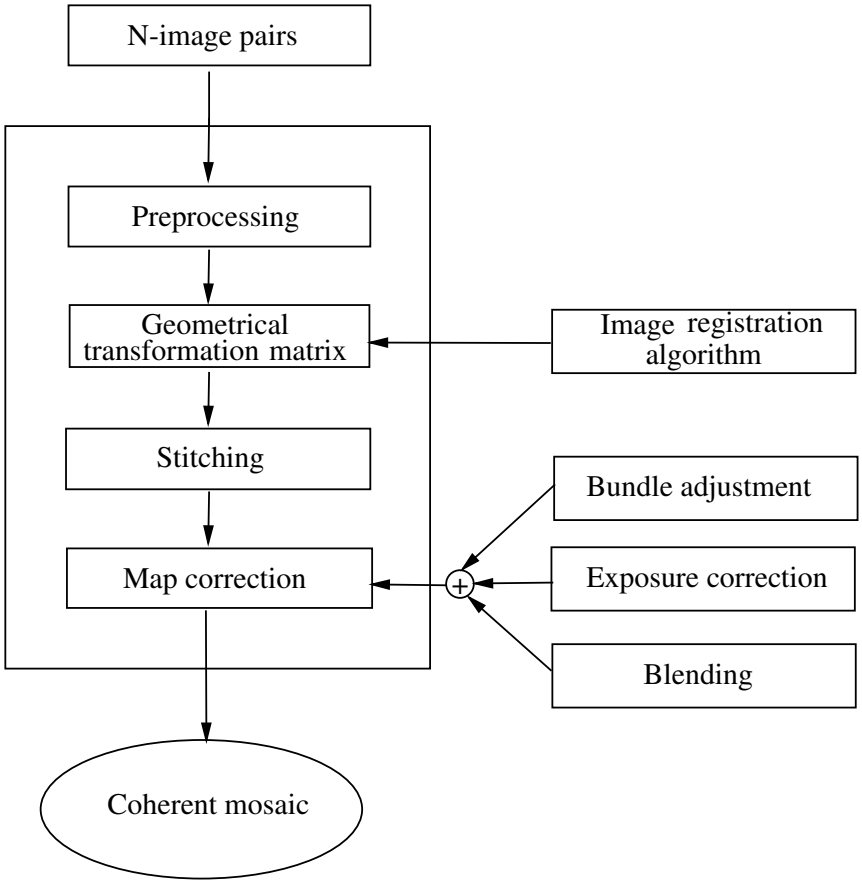


Figure 1.4: Image mosaicing framework.

has served medical community in the recent few decades. One of the most prominent and growing example is in the case of mosaicing of endoscopic images for extending their low field-of-view acquisition features.

1.2.1 Application to endoscopy

Endoscopes have been widely used to inspect various hollow organs like the bladder, urethra and ureter (in cystoscopy, Fig. 1.5 (a-b)); the liver (in laparoscopy, Fig. 1.5 (c)); the stomach and esophagus (in gastroscopy, Fig. 1.5(c-d)); the throat (in laryngoscopy, Fig. 1.5 (f)); the pituitary gland (in endocrinology, Fig. 1.5 (g); the colon (in colonoscopy, Fig. 1.5(h))). Endoscopic procedures form a part of routine clinical practices for minimally invasive examination and interventions. These procedures help clinicians to visualize inside the organ without surgery and allow them to perform Minimally Invasive Surgery (MIS). This technique reduces the surgical trauma in patients and additionally speeds the recovery time preventing serious infections. However, since clinicians are able to see only small region of the organ (small FOVs) in most of the endoscopic examination, they have to control precisely the endoscope tip displacement which requires a high degree of orientation, coordination and fine motor skills to be able to scan all required area of tissues. Moreover, endoscopic images have low resolution varying from 0.25 megapixels (PAL) to 2 megapixels (in HD). Extended and enhanced visualization of the tissue (organ) is important for clinicians. This task can be achieved by using image mosaicing tech-

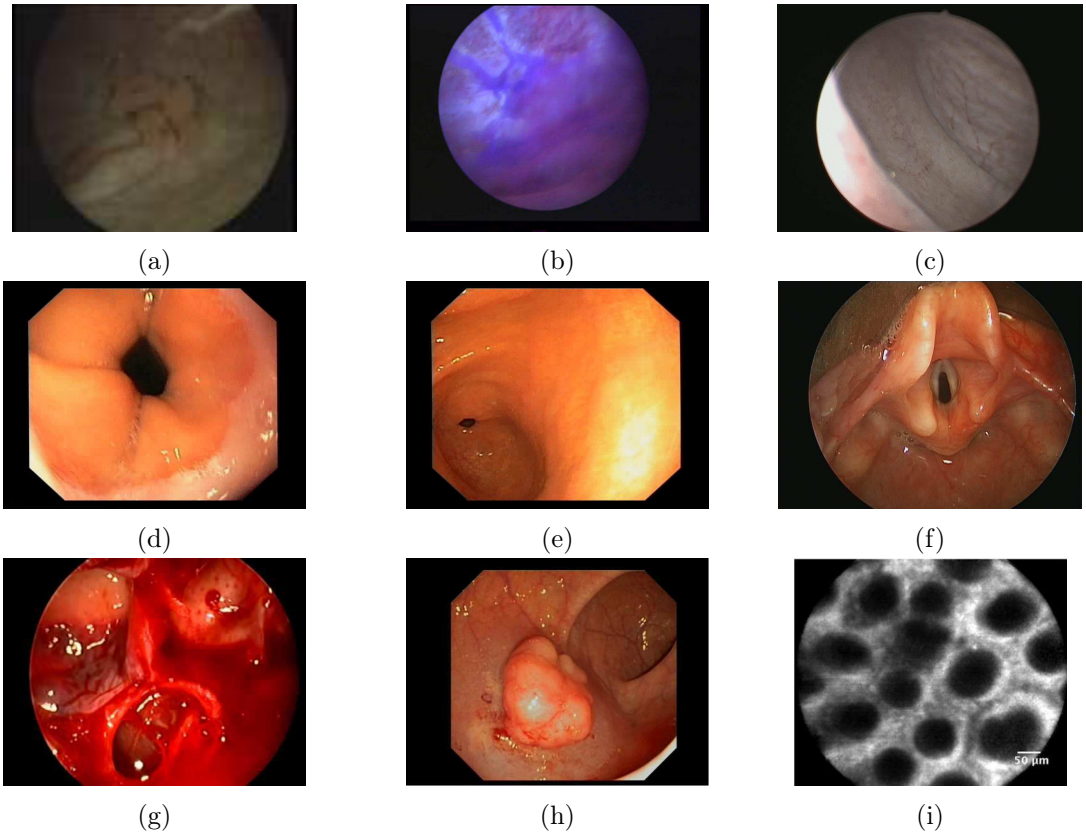


Figure 1.5: Examples of images obtained from different endoscopic applications. (a-b) urinary bladder (WL and FL cystoscopy), (c) near urethral opening (WL cystoscopy), (d) esophagus (gastroscopy), (e) stomach (gastroscopy), (f) larynx (laryngoscopy), (g) pituitary gland (endonasal neuro surgery), (h) colon polyp (colonoscopy) and (i) microscopic image of cardiac type epithelium in vivo (confocal laser endomicroscopy, CLE).

nique.

Robust image mosaicing for endoscopic images is still an open problem for the image processing and computer vision community. This is due to numerous factors to deal with which include: 1) the geometrical shape of the organ, 2) the local and global organ deformations, 3) the texture characteristics of organ's wall, 4) the endoscopic navigation procedures and 5) the image quality. The shape of organs vary from concave hollow structures (as for the internal bladder wall) to hollow cylindrical urethra, funnel-like pylorus opening in the stomach and vase-like regions near colon. The shapes of the surface parts observed through an endoscope appears often to be planar or quasi-planar as the distal tip is in general hold close to the organ walls in many of these organs [Miranda-Luna et al., 2008, Vercauteren, 2008, Behrens et al., 2009]. This is one reason why endoscopic image sequences can be used to built 2D wide field-of-view mosaics. Depending upon the examination, various techniques are sometimes used to limit the organ deformations. In cystoscopy for instance, the bladder is filled with an isotonic saline solution which contributes to the organ stiffening so that the surface can be considered as rigid between two image acquisitions (the surface deforms itself slowly and the acquisition rate is as high as 25 images/second). Large FOV maps for WL cystoscopy and FL cystoscopy are shown in Fig. 1.6 (a) and Fig. 1.6 (b) respectively. However, robust registration of these image pairs is often difficult because of various

reasons: 1) the textures of the tissue under examination are little or poorly contrasted, 2) moist tissue surfaces implies specular reflections (refer Fig. 1.5 (d-h)) and 3) organ parts are often hidden in some image regions so that homologous data determination can fail. Mosaics shown in Figs. 1.6 (d-g) respectively indicate strong specular reflections, imaging artefacts, texture variability and non-planar maps.

Local deformations in some organs (e.g. colon or liver) can sometimes be large enough to impede 2D image registration. Large scale changes or large perspective changes also cannot be handled by existing state-of-the-art registration schemes. Even though for several hollow organs the surface seen in images can be considered as planar, this "planar assumption" does not hold for all organ types (for instance in case of the liver and the colon endoscopy except confocal microscopic images of them).

1.2.2 Mosaicing trends in endoscopy

According to the geometry and complexity of organs and depending upon the available information in the images, several endoscopic image mosaicing algorithms are found in the literature [Maurin et al., 2009, Bergen et al., 2013a, Miranda-Luna et al., 2008, Maier-Hein et al., 2013]. A major part is proposed for endoscopic examinations (e.g. laparoscopy or urology) while some are almost non-existent for other examinations (e.g. intestine). When the FOV in the images is quasi-planar and homologous points can be established between images, 2D mosaicing algorithms can potentially lead to mosaics with significantly increased FOV. But, when the "local planar assumption" is not fulfilled and the endoscope is close to the surface under observation, building 3D mosaics is more appropriate (as for 3D mosaicing of the internal abdominal cavity [Maurin et al., 2009, Mahmoud et al., 2012, Maier-Hein et al., 2014]). However, such 3D approaches are often only feasible when homologous points can be robustly extracted and matched. When only based on 2D image information, such approaches are difficult to implement in different endoscopic modalities (like in urology, see the work detailed in [Soper et al., 2012] which highlights these difficulties in the specific case of the bladder).

Different approaches have been proposed in the field of internal bladder mosaicing in urology. Bladder is a hollow organ which can be assumed as quasi-planar surface in the FOV, when the cystoscope is close to the bladder wall. Local deformations are limited by filling the hollow cavity with saline solution (also mentioned in Section 1.1.1). Previous works related to both FL cystoscopy [Behrens et al., 2009, Behrens et al., 2010, Ali et al., 2015a] and WL cystoscopy [Bergen et al., 2013a, Weibel et al., 2012b, Miranda-Luna et al., 2008, Hernandez-Mier et al., 2010] have shown advances and significance of endoscopic 2D image mosaicing in urology. The images of video-sequences are aligned to one global mosaicing coordinate system by concatenation of homographies estimated by dedicated pairwise image registration techniques. A detailed overview on such registration methods dealing with bladder image mosaicing will be discussed in Section 1.3. Considerable efforts have also been made to generate mosaics of tubular-shape like organs such as ureters and the urethra, trachea, intestine and esophagus. Igarashi et al. [Igarashi et al., 2009] presented opened panorama of the ureter and the urethral cavity. A fusion of 2D-panorama of laparoscopic image sequence with 3D computed tomography (CT) image was also shown in this contribution. A real-time panorama based on homography estimation in the gastrointestine (GI) image pairs acquired from capsule endoscopy was presented recently by Yi et al. [Yi et al., 2013].

Simultaneous Localization And Mapping (SLAM), widely used in robotics, has been applied for surface reconstruction of deformable organs or non-rigid surfaces under examination. SLAM approach based on Extended Kalman Filtering (EKF) using a single camera was proposed by

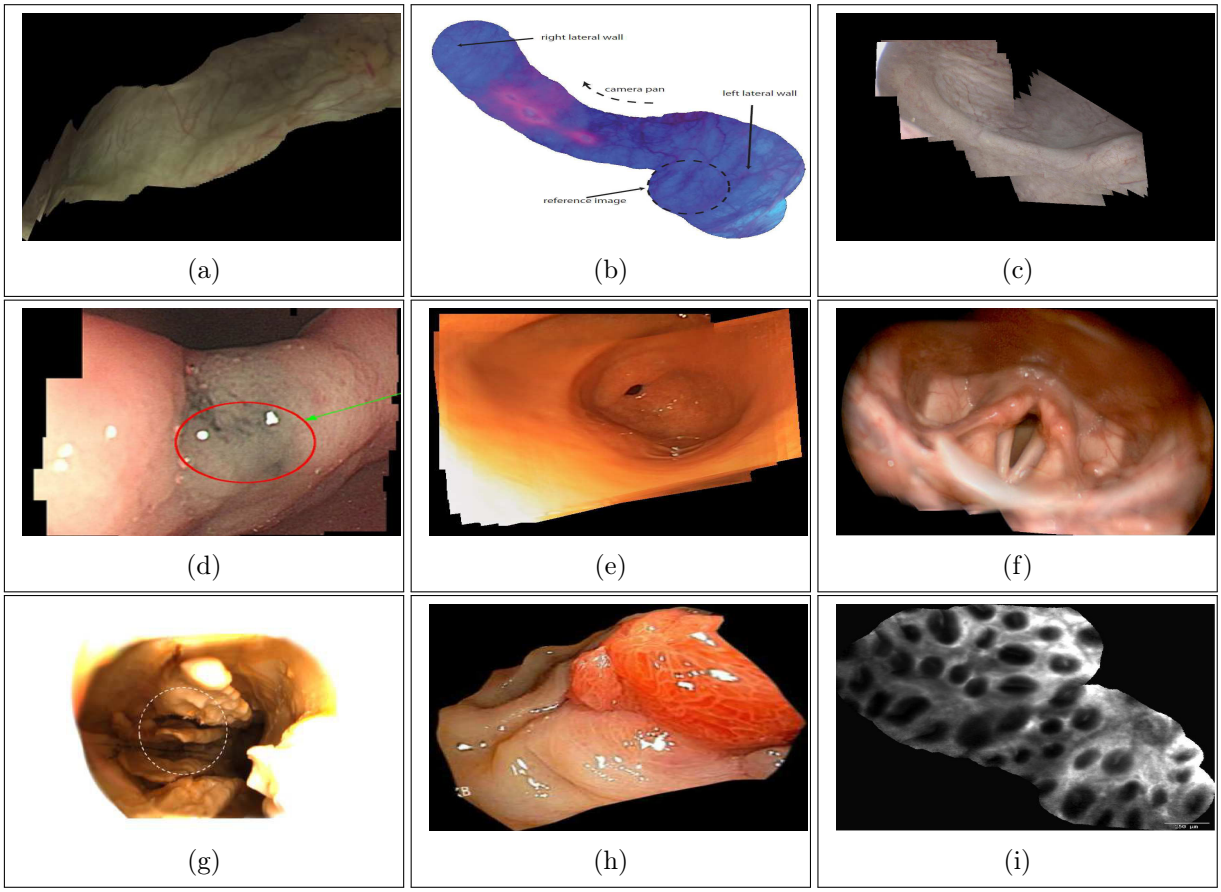
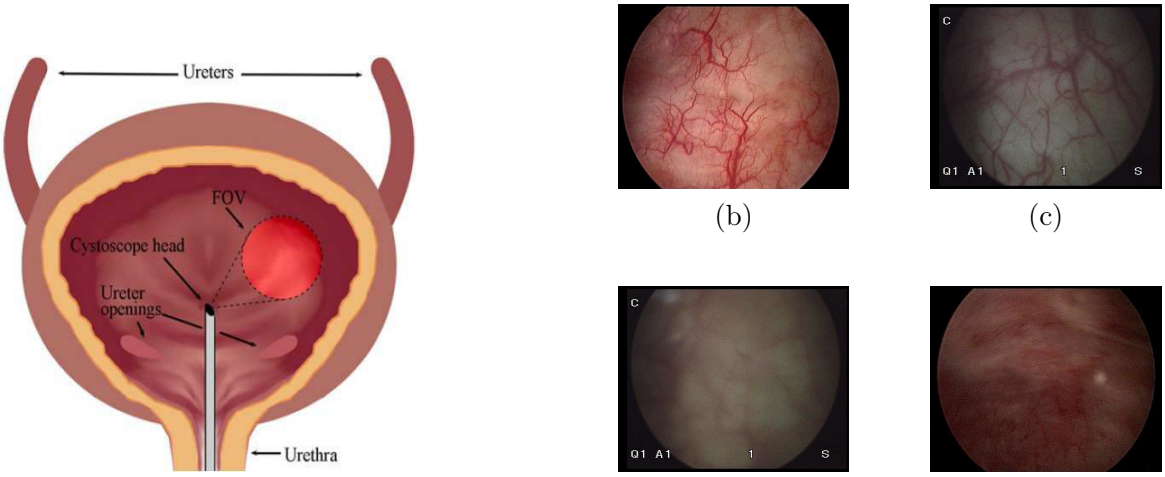


Figure 1.6: Image mosaics obtained for different endoscopic applications given in Fig. 1.5. (a) 2D large extended FOV mosaic for images acquired with a WL cystoscope [Hernandez-Mier et al., 2010], (b) 2D planar panoramic image built generated in real-time using FL cystoscopy video frames [Behrens et al., 2009], (c) 2D image mosaic representing a quasi-planar surface near the urethral opening [Ali et al., 2013b], (d-e) mosaic of gastroscopic quasi-planar image sequences showing extended FOV around angularis [Liu et al., 2015] and pylorus regions, (f) image mosaic of larynx generated with a general-purpose stitching software [Schuster et al., 2012], (g) real time view expansion of an endo-nasal region [Berger et al., 2013], (h) 3D reconstruction of polyp region using a shape-from-motion approach for a colonoscopic image aquisition set-up [Koppel et al., 2007] and (i) extended FOV mosaic of CLE for round cardiac type epithelium in vivo [Vercauteren, 2008].

Davusion et al. [Davison et al., 2007]. Methods mentioned in [Totz et al., 2012, Maier-Hein et al., 2013, Grasa et al., 2009] uses SLAM based approach with EKF and stereo-endoscopic image data for extending FOV during MIS. An overview of state-of-the-art methods for 3D surface reconstruction in computer-assisted laparascopy for MIS is presented in [Maier-Hein et al., 2013]. Malti et al. [Malti et al., 2012] published his investigations for shape reconstruction of deformable surfaces using shape-from-motion-and-shading approach for laparoscopy.

Due to the complex geometry and lack of rigidity most of the work of mosaicing is focused as 3D surface reconstruction in colonoscopy. In this medical application, shape-from-motion (SfM) was used by Thormaehlen et al. [Thorsten et al., 2002] to reconstruct the polyp in colon. SfM was also used by Fan et al. [Fan et al., 2010] for capsule endoscopic images of the colon. Mi-



(a) (b) (c) (d) (e)

Figure 1.7: Acquisition of cystoscopic data and image texture variability illustration. (a) Schematic sketch of the bladder scene (b) Example of an image with contrasted texture (c) Image with vignetting effect (d) Example with weak contrasted texture image (e) Image with motion blur.

croscopic images of different organs in the body like the urinary tract, the esophagus, stomach, colon, liver, lungs are obtained using confocal laser endo-microscopes. These images are planar irrespective of the organ under examination as they are at microscopic scale and hence closest to the surface. This is why only 2D mosaicing algorithms suffice the FOV extension of such endoscopic procedures. A major contribution in mosaicing of these image sequences is detailed in Ph.D. thesis of Vercauteren [Vercauteren, 2008].

1.3 2D Bladder image mosaicing

Building large FOV mosaics using the images of a video-sequence has several advantages. Diagnosis can be facilitated with the visualisation of the whole lesion area in an increased FOV image. Lesions can be localized with respect to anatomical landmarks leading to faster examination trial. Lesion follow-up over weeks or months become possible by comparing bladder wall mosaics of the same patient. Moreover, archiving large FOV mosaics improve examination traceability (video sequences are not archived because they are often difficult to be interpreted by urologists after examination).

The quality of each mosaic is strongly influenced by the registration accuracy of images I_i and I_j of a video-sequence. In the regions of interest of the bladder wall, the organ most often presents smooth surface (i.e. without surface discontinuities) which are observed by maintaining the endoscope's distal tip close to the epithelium (in most of the endoscopic procedure). With such scene geometry, the small FOV images encompasses quasi-planar surfaces. In the frame of standard cystoscopy, the bladder wall must be distended for proper visualization. To do so the bladder is filled with a saline isotonic solution. Moreover, considering the high acquisition rate of 25 images/s, bladder wall deformation is imperceptible between two consecutive acquisitions (image pair (I_i, I_j) with $j = i + 1$). Global map corrections (bundle adjustment) are inevitable in some maps specially in mosaics with the loop closing trajectory where texture misalignment occur between the first and last loop images. For computing visually coherent maps, blending and contrast adjustment techniques also play significant role in better visualization of the 2D

mosaics. All these inevitable steps are presented in the mosaicing block diagram given in Fig. 1.4. Each step required for building visually coherent and interpretable bladder maps is explained thoroughly in this section along with its related literature.

1.3.1 Pre-processing

Endoscopic images are affected by several factors which include image distorsion due to the optics, vignetting effect, large illumination gradients due to view-point changes and occurrence of spatially periodic patterns induced by optical fibers in flexible cystoscopy (this last effect is less and less visible for recent fiberscopes). As a consequence, the images are usually degraded in quality both in appearance and in signal-to-noise ratio. For robust image registration and coherent map visualization these problems need to be addressed. This step can be taken as a crucial step for endoscopic image mosaicing.

Distortion correction

Endosccopic images have strong radial distortion with straight lines being projected as curves in the image (barrel or fish-eye effects [Miranda-Luna et al., 2004]). Bouget's implementation of Zhang's method is used for calibrating a camera from minimum of three grid images [Zhang et al., 2003]. Unfortunately, the toolbox has usability issues for regular clinical practice. For instance it is very impractical for clinicians to calibrate their endoscopes before use by taking multiple grid images. The same drawback can be noticed for the barrel distortion method specifically proposed in [Miranda-Luna et al., 2004] for endoscopes. Therefore, a more practical way proposed by Yahir et al. [Hernandez-Mier et al., 2010] was to crop and to exploit only the central image regions (400× 400 pixels). This was done to avoid the computational complexity of the distortion correction algorithms and achieve practical usability in clinics (i.e., no need to run calibration toolboxes). Moreover, according to the author, the radial distortions are negligible in the central part of the images. Though this method shows a strong support in terms of usability, it has encounters two major drawbacks: 1) loss of small FOV parts which might contain useful information and help in diagnosis and 2) cropping of image region parts is empirical and therefore can not guarantee the diminished effect of radial distortion. Barreto et al. [Barreto et al., 2009] proposed an automatic camera calibration for endoscopes.
This approach has been used in this thesis for minimizing the effect of camera lens distortions in addition with epipolar geometry constraints (explained in later Section 1.3.2) for obtaining putative matched correspondences in image pairs.

Let $\mathbf{Q}$ be the non-homogeneous coordinates of a point in the 3D space. Then, the projection point of $\mathbf{Q}$ with lens distortion of camera into the image plane Ω is given by:

$$\mathbf{x}_q \sim K\Gamma_\xi P[Q \quad 1]^T, \tag{1.1}$$

where, P is a 3×4 projection matrix depending on the camera rotation R and translation t. K is the intrinsic camera matrix with camera center (c_x, c_y), focal length f, the skew s and aspect ratio a:

$$K \sim \begin{pmatrix} af & sf & c_x \\ 0 & a^{-1}f & c_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.2}$$

Γ_ξ is the unknown projection function to be estimated for obtaining the undistorted points (radial distortion correction). Radial distortion correction is inevitable to obtain dense putative points

matches mostly when the whole image region is considered. If $\mathbf{q}_d$ is the distorted point due to lens alone and ξ is the distortion coefficient then, the undistorted point $\mathbf{q}_u$ can be expressed as:

$$\mathbf{q_u} = \Gamma_\xi \ \mathbf{q_d} = (1 + \xi \ \mathbf{q}_d^T \ \mathbf{q}_d)^{-1} \ .\mathbf{q}_d. \tag{1.3}$$

Let $\mathbf{x}$ be the point on the image plane then for a skewless camera (i.e. $s = 0$) with aspect ration unity ($a = 1$), then the distorted point in the image is given by:

$$\mathbf{x}_d = K\mathbf{q}_d = K{\Gamma_\xi}^{-1}\mathbf{q}_u, \tag{1.4a}$$

$$\mathbf{x}_d = f{\Gamma_\xi}^{-1}\mathbf{q}_u + \mathbf{c}. \tag{1.4b}$$

Let $\hat{\mathbf{x}_d} = \mathbf{x}_d - \mathbf{c}$ be the image points in the coordinate frame centered at the principal point, so Eq. 1.3 becomes,

$$\mathbf{q}_u = \Gamma_\xi \ \underbrace{(f^{-1}\hat{\mathbf{x}_d})}_{\mathbf{q}_d} \tag{1.5}$$

Solving Equations 1.3 and 1.5, we obtain an undistorted image point $\mathbf{x}_q$ in pixel given by,

$$\mathbf{x}_q = f \ \mathbf{q}_u = (1 \ + \ \frac{\xi}{f^2} \ \hat{\mathbf{x}_d}^t \ \hat{\mathbf{x}_d})^{-1} \ .\hat{\mathbf{x}_d}. \tag{1.6}$$

With the principal point $\mathbf{c}$ usually being coincident with the image center and the constant parameter ξ being estimated using offline calibration method mentioned in [Barreto et al., 2009], one can estimate the varying focal length f (as a variable for camera zoom) using uncalibrated feature tracking based algorithm (like uRD-KLT algorithm mentioned in [Lourenço et al., 2014]). As in practice, the zoom of the camera is not constant in many endoscopes like in the case of laparascopes, so it is important to autocalibrate it for retrieving focal length of the lens at a specific instant i.e. when the device is being used (online).

Vignetting effect and image normalization

Scene brightness can be defined as the power per unit foreshortened area emitted into a unit solid angle by a surface reffered as "radiance". After passing through the camera lens, the power of this radiant energy falling on the image plane is called irradiance. Irradiance is then transformed to image brightness. The amount of light (radiance) hitting the image plane varies spatially due to multiple factors causing decreased illumination and contrast at the image periphery [Kim and Pollefeys, 2008]. The loss of illumination at image periphery is referred to as "vignetting". While vignetting appears inside an image, another important non-uniformity of brightness in endoscopic images occur between images due to viewpoint changes. Vignetting is one of the effects contributing to brightness changes between homologous pixels of two images (homologous pixels should have the same brightness value). These factors affect both the registration robustness and the visual quality of mosaics.

Endoscopic images are largely affected by "vignetting effect" of the optical sensors. It is visible in cystoscopic images as in Figs. 1.8(a) and 1.9(a). These are very low spatial frequency distribution of image intensities. It can be observed that there is an illumination gradient while going from the center of the image towards the periphery region. A line profile ($V(orig.)$) plot in Fig. 1.8(f) for pixels along red line of cystoscopic image in Fig. 1.8(a) shows large gradients in intensities from periphery towards the center region (i.e. higher values near the image center).

Feature based registration approaches are less sensitive to illumination changes as they are

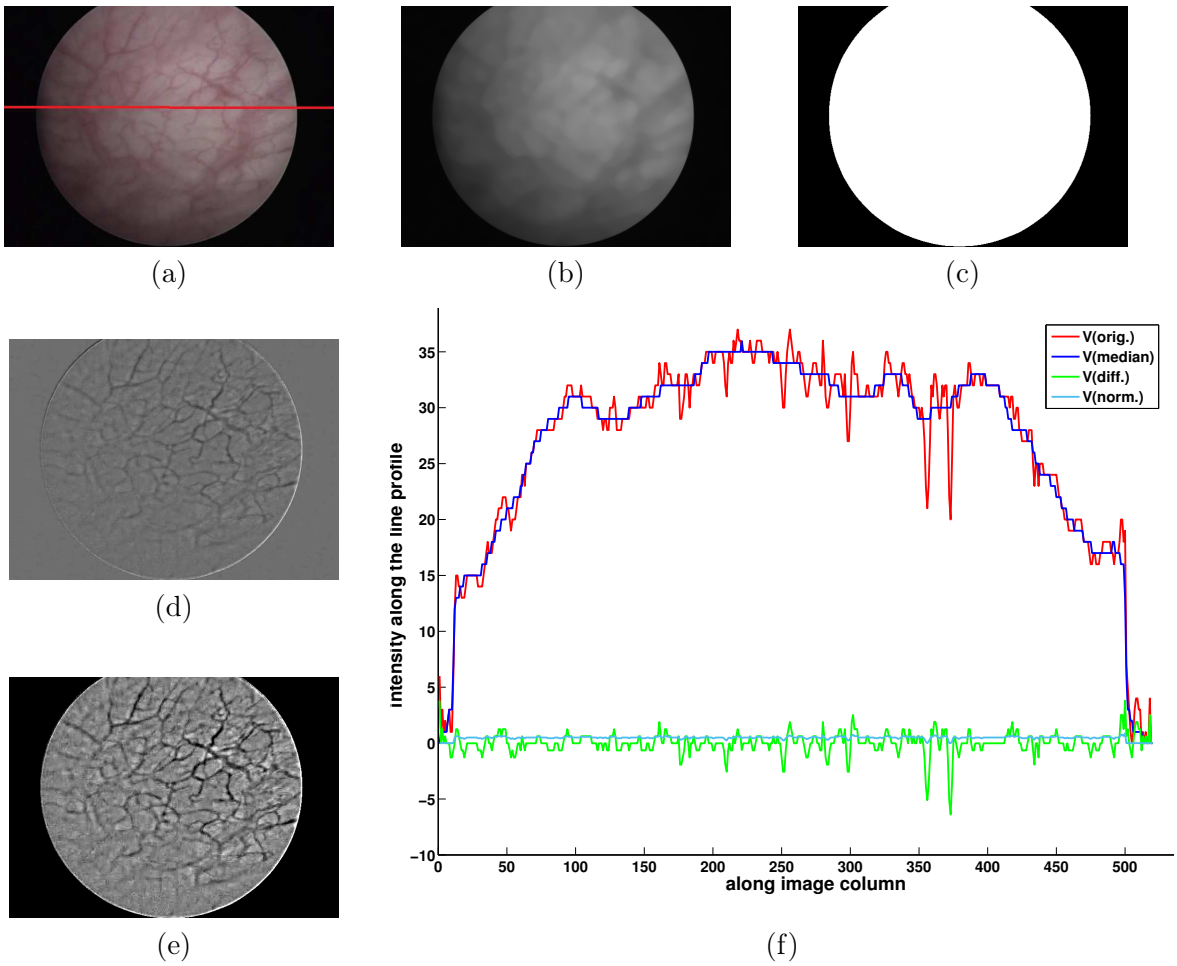


Figure 1.8: Background removal using median filtering. a) Original image, b) background image obtained with median filtering technique, c) estimated mask from the background image in (b), d) pixel-wise difference of the original gray level image of (a) and the low-pass filtered image (median filter) in (b), e) image normalized to zero mean with pixels from FOV mask only shown in (c), (f) intensity profile V along the red line shown in (a) for the images (b), (d) and (e).

usually applied for detecting corners or edges in each image. However, pixel based registration approaches which are used to minimize or maximize a cost function based on iconic data information are prone to illumination or brightness changes. It is therefore important to preprocess images before running any pixel based image registration scheme as inhomogeneous illumination strongly influences registration robustness and accuracy. Therefore, inhomogeneous illumination strongly influences registration robustness and accuracy. In [Hernandez-Mier et al., 2010], images are homogenized by subtracting a band-pass filtered image from the original image. This band-pass filter is designed to eliminate both low frequencies due to vignetting (frequency of illumination gradients are supposed to be much lower than the interesting bladder textures), as well as high frequencies due to the fiberpattern of flexible cystoscopes (the regular optical fiber pattern consists of frequencies being higher than the bladder textures). In the corrected image, all vascular structures remain visible, while exposure gradients and fiber-patterns are strongly attenuated.

In Fig. 1.8 an image normalization approach is sketched for attenuating the vignetting effect (the removal of high frequencies due to fibers of flexible cystoscopes is not detailled). The color image in Fig. 1.8(a) is first converted in grey-levels. The "background" image is obtained by processing this grey-level image with a median filter acting as a low pass filter (low frequencies corresponding to vignetting is preserved while higher frequency signals correponding to texture are attenuated). Fig. 1.8(c) shows the mask which can be computed to delineate the circular FOV. Fig. 1.8(d) gives the pixel-wise difference between the original grey-level image and the background image. Signal parts relating to vessels are enhanced. In Fig. 1.8(e) the image is normalized to have a mean grey-level value equal to 0 and a unit variance. A line profile $V(.)$ is shown in Fig. 1.8(f) for each images along the similar locations marked by red line in Fig. 1.8(a). It is noticeable that profile V_{median} is a low-pass filtered version (i.e. with less spatial variations) of profile $V_{orig.}$. Profile $V_{diff.} = (V_{orig.} - V_{median.})$ and $V_{norm.}$ correspond to strong texture signals.

Though this filtering technique improves the registration robustness of some algorithms [Hernandez-Mier et al., 2010, Weibel et al., 2012b], it is not a general and efficient pre-processing method for many other registration approaches, notably for those based on global energy minimization techniques like optical flow. Indeed, changing both locally and globally the pixel intensities in the images impedes the use of many approaches optimizing global energies able to handle the weak texture information often occurring in bladder images. Due to this drawback, another pre-processing algorithm is discussed in the next section.

Contrast enhancement using SVD (Background exposure correction)

Image textures in bladder video-sequences are often weakly contrasted due to vignetting effect, viewpoint changes and image defocussing due to instrument depth changes (variation of the distance between the distal tip and the epithelial surface) cause the illumination patterns in images to shift. While the effect of vignetting is usually only seen on the periphery, illumination variations induced by defocus or viewpoint changes affect the whole image. It is crucial to correct these images because even less sensitive feature based approaches can fail to detect features and matching ambiguities between homologous points could increase. An example of an image pair with strong illumination difference is shown in Fig.1.10 (a-b). Due to large change in brightness between this image pair, registration methods based on brightness consistency assumptions will fail to retrieve correct transformation parameters needed for registering the data. Also too weak and/or varying contrast between images affects the visual quality and coherence of the mosaics gathering data from hundreds or thousands of images. So, the contrast of the image has to enhanced before registration and stitching (as seen later, the contrast between neighboring images in the mosaics should also be homogenized). In addition, the contrast should resemble the texture of the same bladder scene. Well known histogram equalization can be used for this purpose. However, such an approach stretches also the contrast in regions where there is no contrast and are also prone to increase noise in the image.

Singular values possess the background illumination information of the scene. So, a less illuminated image can be brightened with by exploiting the singular values of a well contrasted and well-illuminated image in the dataset without stretching the contrast over all image regions (as histogram equalization do) and affecting iconic data information. A background exposure correction using singular value decomposition (SVD) starts with the selection of a visually well contrasted image. This image (I_{ref}, Fig.1.9(c)) is taken as reference for the whole data set (in a video sequence). Singular values of both reference image I_{ref} and poorly illuminated test image I_{test} are shown in Fig. 1.9(e). In this figure, σ_i is the sum of the singular values computed in each

color band (R,G and B) and are plotted along Y-axis while X-axis represents the position of non-zero singular values. A threshold $\sigma_{th.}$ is set for a particular dataset which is usually empirically set (experimentally found value is 5.5×10^4, for most of cystoscopic image sequences). Constrast of each image under consideration is estimated as the sum of their singular values ($\sum \sigma_i$) and compared with $\sigma_{th.}$. If the sum of the singular values of image I_{test} is over threshold obtained for the reference image, then the contrast of I_{test} is improved using the SVD technique. To do so, the singular value matrix $\sum$ of I_{test} (Fig. 1.9 (a)) is replaced with the $\sum_{ref}$ of I_{ref}. After reconstruction, an enhanced image $I_{enhanced}$ (Fig. 1.9 (d)) with consistent illumination distribution similar to well contrasted images in the dataset is obtained. For details see the method detailed in [Adal et al., 2013].

The example in Fig. 1.10 shows that both the registration accuracy and the mosaicing quality are affected by changing illumination conditions over a video-sequence. Fig. 1.10(a) shows a small bladder video-sequence (the first images are on the right and the last images are on the left). The two first images are underexposed and with a lack of contrast (in patient data). The Speeded Up Robust Features (SURF, [Bay et al., 2008]) method was used to find the homologous points which are used to register the images. The misaligned textures in Fig. 1.10(a) show that the registration of the images failed without the SVD decomposition based preprocessing (the contrast of the two images on the right was improved only for visualization purpose). In the mosaic of Fig. 1.10(a) the error between two first images are by far too large. The mosaic given in Fig. 1.10(c) was obtained by performing an SVD based pre-processing on the images which are badly illuminated. It is visible in this figure that the mosaic is now without abrupt change in illumination over the video-sequence. Moreover the first two images have an appropriate size in the mosaic. An almost perfect texture alignment is seen between the image pairs in Fig. 1.10(d). This example shows that the global mosaicing quality is strongly dependent on the accuracy of each individual registration of image pairs and that the exposure correction is essential in some image pairs before performing registration.

Motivation of illumination change compensation in this thesis

The development of registration algorithms being independent of brightness or contrast variations is one of the major goal of this thesis. In this context, algorithms compensating such illumination variations plays a significant role. However, the later SVD based method has only been adopted for the stitching step, i.e. after the images have been already registered (the registration step is based on other proposed solutions to deal with strong illumination changes in the cystoscopy).

A distortion correction method has not been used in the proposed mosaicing algorithms as the patient data used in this thesis were provided by the "Institut de cancérologie de Lorraine" for uncalibrated rigid and flexible cystoscopes. The aim of the work was also to propose a robust mosaicing algorithm without distortion correction algorithms to preserve to remain the clinical procedure unchanged. However, if a flexible calibration procedure (acceptable in a standard cytoscopy) allows for distortion correction, all presented mosaicing methods remain unchanged and their results can only improve themselves.

1.3.2 Geometrical transformation

Under the assumption that the small FOV images visualize almost planar surfaces (Section sec:Mosaicing-trends)), the common regions between images I_i and I_{i+1} (consecutive images in the sequence) are geometrically linked by a homography $H^{\pi}_{i,i+1}$. This geometrical transformation superimposes homologous pixels with homogeneous coordinates $(x_{i+1}, y_{i+1}, 1)$ and

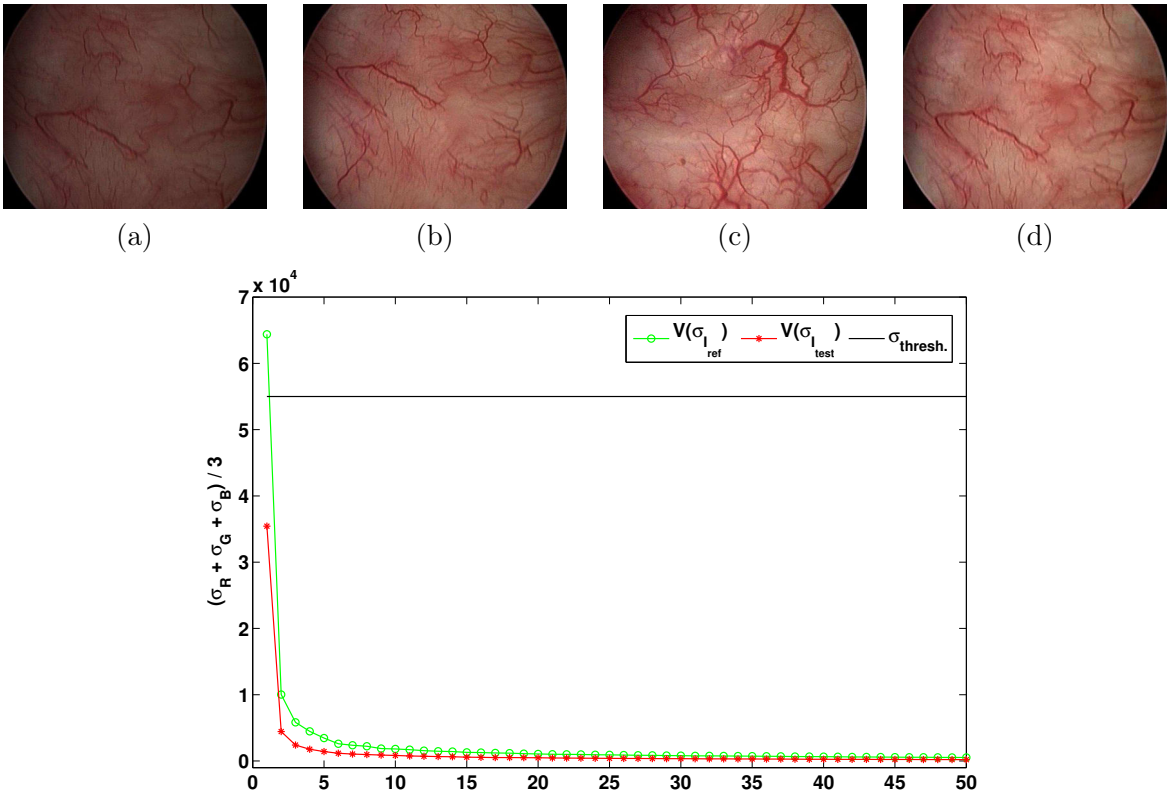


Figure 1.9: Contrast enhancement in cystoscopic image sequences. (a) poor contrast image (I_{test}) due to view-point, b) target image to which image in (a) need to be registered, c) reference image (I_{ref}), d) enhanced image ($I_{enhanced}$), (e) Singular values profile $V(\sigma)$ in good and bad contrast images.

$(w_i x_i, w_i y_i, w_i)$ respectively in the source (I_{i+1}) and target (I_i) images. Writing Eq. (3.16) with homogeneous coordinates, the homography is defined by the parameters f, $(s_{\mathbf{x}}, s_{\mathbf{y}})$, ϕ, $(t_{\mathbf{x}}, t_{\mathbf{y}})$ and (h_1, h_2) denoting the focal length, shearing factors, in-plane rotation, $2D$ translations and perspective changes respectively. Subscripts x and y stand for the image axes. w_i is completely defined by the perspective parameters h_1 and h_2.

$$\begin{pmatrix} w_i x_i \\ w_i y_i \\ w_i \end{pmatrix} = H^{\pi}_{i,i+1} \begin{pmatrix} x_{i+1} \\ y_{i+1} \\ 1 \end{pmatrix} = \begin{pmatrix} f cos\phi & -s_{\mathbf{x}} sin\phi & t_{\mathbf{x}} \\ s_{\mathbf{y}} sin\phi & f cos\phi & t_{\mathbf{y}} \\ h_1 & h_2 & 1 \end{pmatrix} \begin{pmatrix} x_{i+1} \\ y_{i+1} \\ 1 \end{pmatrix} \tag{1.7}$$

In the rest of this contribution, the transformations $H^{\pi}_{i,i+1}$ between consecutive image pairs are called "local" homographies, whereas "global" homographies $H^{g}_{0,i}$ place the pixels of images I_i in the coordinate system of the first image I_0 taken as mosaic start. Global matrices $H^{g}_{0,i}$ are the product of local homographies defined as,

$$H^{g}_{0 \leftarrow i} = \prod_{k=i-1}^{k=0} H^{\pi}_{i-k-1,i-k}. \tag{1.8}$$

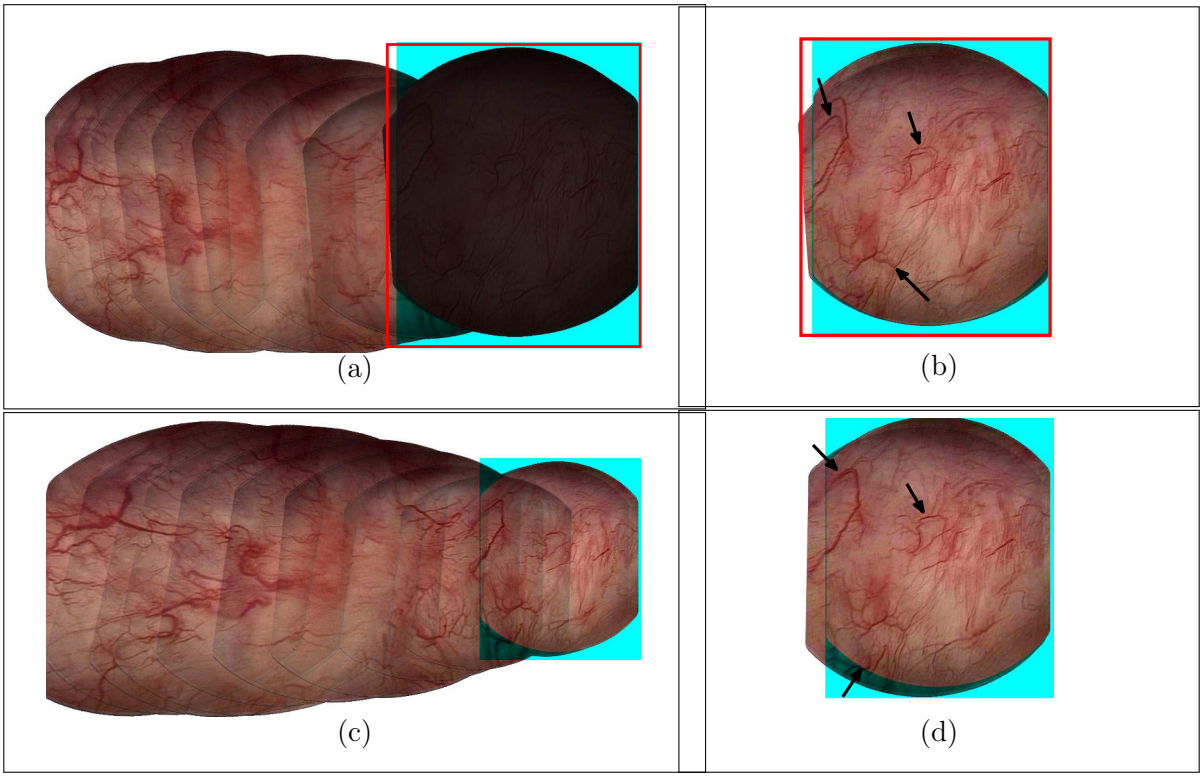


Figure 1.10: Mosaicing results based on feature point extraction with and without image pre-processing. a) Impact of large brightness variability affecting few images (the two first images on the right are underexposed). b)The alignment errors are indicated by the arrows which point regions where the textures (vessels) of two images should be perfectly superimposed (these structures are infact shifted). c) Mosaic after SVD enhancement of images. d) After illumination correction with the SVD technique, the structures of the two first images are now perfectly superimposed.

Eq. (3.17) is used to place the pixels of each image I_i (with $i \in [1, N]$) of a video-sequence of N frames into the coordinate system of I_0.

Estimation of Homography using homologous points

A homography can be represented as a non-singular 3×3 matrix H. Let us consider two points $\mathbf{x}_i$ and $\mathbf{x}_j$ of two images I_i and I_j respectively (with $j = i + 1$ for consecutive image pairs) corresponding to the same 3D point in the bladder scene. Then, these homologous points are related with a homography $H_{i,j}$ as (also refer to Fig. 1.11):

$$\mathbf{x}_i = H_{i,j}\mathbf{x}_j. \tag{1.9}$$

The planar homography constraint is:

$$\mathbf{x}_i \times H_{i,j}\mathbf{x}_j = 0. \tag{1.10}$$

As sketched in Fig. 1.11, $\mathbf{x}_j$ must lie on the epipolar line l_j corresponding to the point $\mathbf{x}_i$ (i.e. $\mathbf{x}_i$ and $\mathbf{x}_j$ are coplanar). The point correspondences found can be refined (*i.e.* outliers in dense

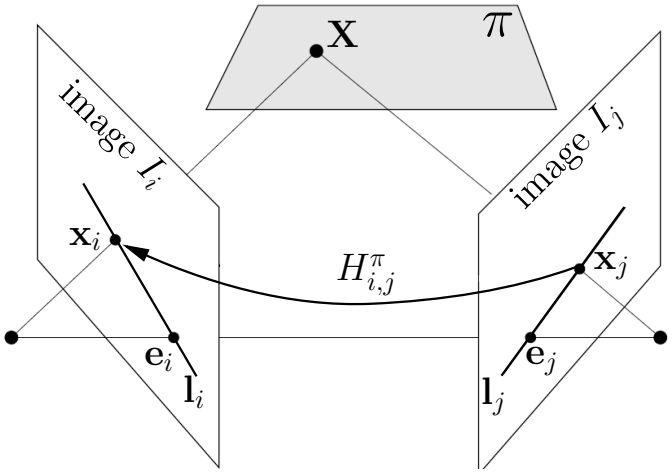


Figure 1.11: A 3D point $\mathbf{X}$ lying on a plane π has projection $\mathbf{x}_i$ on image I_i an $\mathbf{x}_j$ on image j, ($j = i + 1$ for consecutive image pairs). These points are projectively equivalent and can be mapped by a 2D homography $H^{\pi}_{i,j}$ which can be used to express the points of I_j in the coordinate system of I_i.

correspondences) by using a 3×3 fundamental matrix F of rank 2 which is defined upto scale with $det(F) = 0$ and related to:

$$\mathbf{x}_i \cdot F\mathbf{x_j} = 0 \tag{1.11}$$

Epipolar lines are essential in dealing with outliers when view points are changing. In two viewpoint case, the image coordinate $\mathbf{x}_i$ in image I_i can be related to homologous coordinate $\mathbf{x}_j$ in image I_j if and only if it lies on the epipolar line l_j for that point and is related to each other by:

$$l_j = F\mathbf{x_i}. \tag{1.12}$$

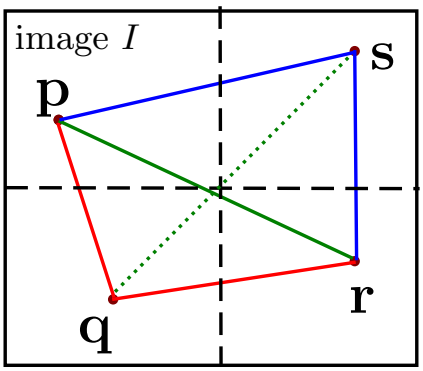


Figure 1.12: An image I showing set of selected points such that points chosen for homography estimation using the 4-point DLT algorithm are well-distributed in the image.

Epipolar lines are also essential in rejecting the coplanar points (*i.e.*, points lieing on the same epipolar line). We have chosen a RANSAC variant based on the least median of squares (LMedS) [Rousseeuw and Leroy, 1987]. This technique relates the quality of the fundamental matrix F estimated for each chosen subset with the median of its residuals for the remaining samples. The epipolar inliers are then found from the dense correspondence. Using these correct matches, a 4−point normalized (Direct Linear Transformation) DLT algorithm [Hartley and Zisserman, 2003] is used to solve the linear least square overdetermined system of $2n$ equations that can be formulated as:

$$\| \mathbf{A}h_{1\times 9} \| = 0 \qquad \text{subject to } \| h_{1\times 9} \| = 1 \tag{1.13}$$

where, $h_{1\times 9} \in \mathbb{R}^9$ is null vector with the components of homography $H_{i,j}$ and $\mathbf{A}$ is a $2n \times 9$ matrix with $n-$correspondences for solving Eq. (1.10).

Motivation for false point matches rejection during homography computation

Homologous points provided by dense correspondence establishment algorithms provided in this thesis mostly provide true matches. However, independently of the used method, a given number of matches are wrong. The homography computation method employed in this thesis rejects the points that do not lie on the epipolar lines and then uses RANSAC for finding local homographies. The methods developed in the framework of thesis is capable of estimating dense correspondences robustly and with sub-pixel accuracy. But, due to the rigid and planar (distal tip of cystoscope close to bladder epithelium) assumption of the surface in view, we have chosen to use homography with 8-DOFs (Degrees of Freedom) for stitching purpose. However, the methods investigated in this thesis for dense correspondence estimation are not limited to rigid surface consideration and hence deformable registration of the surfaces can also be achieved with the investigated method in this thesis. It is to be noted that the complicated texture and imaging conditions discussed above demands a robust algorithm for pairwise registration between consecutive frames in order to allow large and continuous mosaics. This is a major drawback for many previously proposed algorithms as they are unable to compute correct matches between the frames. One idea to essentially improve the stitching accuracy would be to warp the image with large overlaps using direct interpolation of each pixel to obtain sub-maps which can result in an improved texture distribution. Then the accurate sub-maps can be stitched together using homography for obtaining bigger mosaics. However, in the case of cystoscopic sequences: 1) the surface is mostly rigid and 2) planar. So, using a homography already solves the problem if and only if we solve the major problem which is dense correspondence establishment for complicated scene conditions in cystoscopy. For this purpose, we have isolated homography computation from direct integration to the correspondence finding. The reason behind this is to make our solution more general, robust and applicable to even deformable surfaces whch is often the case in many endoscopic sequences.

Two additional constraints are applied for faster and reliable estimation of local homographies: 1) the images are divided into quadrants and 4-matched points are selected such that the area of triangles formed by them is greater than some threshold area $\triangle_{th.}$ which is empirically set and depends upon the image size (e.g., in Fig. 1.12, $\{\triangle \boldsymbol{pqs}, \triangle \boldsymbol{qrs}, \triangle \boldsymbol{prs}, \triangle \boldsymbol{pqr}\} \geq \triangle_{th.}$). 2) For the established homography to be valid, we set another constraint as: $0.75 < det(H_{i,j}) < 1.2$. This is done for discarding image pairs with degenerate cases.

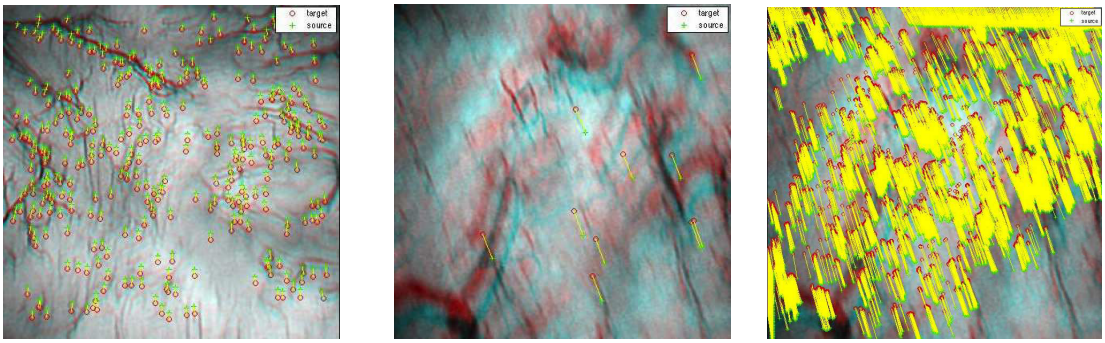
Finally, a planar mapping is used to locate pixels of the images into a single global coordinate frame using their respective global homographies. Linear interpolation has been used to minimize the gap between the pixels during the mapping procedure.

1.3.3 Bladder image registration and mosaicing

Most of the works in the literature on image registration and mosaicing of internal bladder is concentrated in five different places: 1) RWTH Aachen University, Aachen, Germany, 2) Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, 3) University of Washington, Seattle, WA (Human Photonics Laboratory), 4) Technical University of Munich, Garching, Germany and 5) Centre de Recherche en Automatique de Nancy (CRAN), Nancy, France. Bladder cartography literature began only in 2004 with the work of Miranda et al. [Miranda-Luna et al., 2004] in the field of white light cystoscopy (reference diagnosis) and in 2008 by Behrens et al. [Behrens et al., 2009] in the field of fluorescence cystoscopy (photodynamic diagnostics, PDD). Since then, different solutions have been proposed for 2D cystoscopic image registration and mosaicing for facilitating clinical examinations. Several techniques have been proposed for obtaining bladder mosaics of cystoscopy in WL and FL modalities separately. However, recently we have proposed few algorithms that are capable of registering image pairs independent of their modality changes. A feasibility study of high resolution bladder wall map generation is also provided by Shevchenko et al. [Shevchenko et al., 2012]. In this Section, we will first look into feature based methods that were shown feasible for either WL cystoscopy or FL cystoscopy. Then, we will observe various pixel based (direct) methods that have been used in the field of WL cystoscopic bladder image registration.

Feature based registration approaches for 2D bladder cartography

Behrens et al. [Behrens et al., 2011] successfully implemented a mulit-threaded image registration algorithm using SURF feature extraction method for building real-time mosaics with images acquired under FL endoscopy. A pre-processing step for correcting lens distortion in the images were used and the map obtained after global alignment were blended using linear blending technique for coherent visualization. In [Behrens et al., 2009], authors of the same university (RTWM Aachen) provide an alternative solution for distortion free visualization of the bladder maps by projecting local panorama onto the planar faces of a hemi-cube. The complete hemicube represented the global panorama while each face of it denoted local panorama of bladder wall. Well contrasted bladder image pairs under FL modality were used for obtaining these mosaics. Usually, under this modality, the bladder structures (vascular structures or other textures of the epithelium) are less contrasted than in WL modality and largely depend upon observation distance. Images are often darker and less observant to the clinicians in FL. FL modality is thus used only as a complimentary modality to WL (*i.e.* WL is the reference modality). The image characteristics under the standard WL modality possess major challenges: i) occurrence of images without well-pronounced textures as in Fig. 1.7(d), ii) illumination variability due to change in camera view-points and vignetting effects, iii) low contrast, iv) blurry images due to cystoscope movement and/or de-focus or re-focus of the camera lens (see Figs. 1.7(e)) and v) large scale or perspective changes. For these reasons, very few methods in the literature are based on feature based methods [Bergen et al., 2013b, Soper et al., 2012, Ali et al., 2013b]. Bergen et al. [Bergen et al., 2013b] used feature based approach for images acquired under WL cystoscopy. The authors used SIFT feature extraction and RANSAC method for estimating homographies which were used to stitch images together for obtaining local panoramic maps. A graph-based and hierarchical approach were also introduced to obtain the geometrical link between many other sub-maps generated in the procedure of internal bladder wall mosaicing. This technique was used to deal with challenging condition of in-vivo data of WL cystoscopy like large perspective changes, violation of planar assumption etc (i.e. such frames are rejected and new sub-maps are initialized automatically). A similar approach was used by Soper et al. [Soper et al., 2012]



(a) (b) (c)

Figure 1.13: Feature extraction under different image quality/texture conditions [Ali et al., 2013b]. a) Contrasted texture: dense matching with SURF feature extraction technique. This image shows its own extracted feature points (in green) and the successfully extracted and matched feature points (in red) of a second contrasted bladder images. b)Blurred textures: the sparse and undistributed feature matching with SURF is illustrated by the too few green-red mark pairs. Registering the image with this poor information lead to inaccurate results. c) Dense correspondence with variational optical flow method on image (b). Red points correspond again to the key feature points in the target image and green points represents the key feature points in the source image. Yellow line connects the matched key feature points in the target and the source images overlaid.

for obtaining matched feature points between image pairs which were then used to extract 3D points using a structure-from-motion (SfM) technique. This method was tested on an excised pig bladder with images acquired with an ultra-thin scanning fiber endoscope (SFE). This offline method was proposed as an alternative data acquisition procedure by a nurse or ancillary care provider allowing the urologist to review endoscopic image data postoperatively. However this technique was not developed for standard cystoscopes and is not usable in a standard procedure carried out by a an urologist or a surgeon.

Feature based methods are however not reliable in case of poorly textured image sequences and can not robustly handle large intra- and inter- patient bladder texture variability. Moreover, such feature extraction techniques often fail due to blur (e.g. originating in fast camera motion) or local deformation of the organ. The lack of robustness of feature based approaches (illustrated in Fig. 1.13) was thoroughly studied in [Ali et al., 2013b]. However, for dealing with challenging cystoscopic images acquired in WL modality for robustly registering image pairs, other authors applied more robust direct methods for bladder mosaicing [Miranda-Luna et al., 2008, Hernandez-Mier et al., 2010, Weibel et al., 2012b, Ali et al., 2014, Ali et al., 2015a].

Pixel based approaches for 2D bladder cartography

Miranda et al. [Miranda-Luna et al., 2008] maximized the grey-level mutual information with a stochastic steepest gradient descent optimization method for registering consecutive images of a cystoscopic video sequence. This algorithm is robust and give sub-pixel accuracy but it is very slow. Indeed, one minute is required to register image pairs with an average number of 250 iterations of the stochastic gradient. Higher-order discrete energy functions were minimized in [Weibel et al., 2012b] using graph-cuts. This method is robust and image alignment with sub-pixel accuracy can be achieved. However, the major drawback of this method is that the

computational time remains high. About 20 seconds are required to register an image pair with a C++ program running on a quad core processor computer. Yahir et al. [Hernandez-Mier et al., 2010] used a faster local optical flow (OF) approach based on Baker and Matthews algorithm [Baker et al., 2003]. Although the results in [Hernandez-Mier et al., 2010] are promising, local OF methods lead to rank deficient matrices in homogeneous regions and provide only a very sparse flow field in images with few textures. Computing accurately the parameters of the homography of Eq. (3.16) requires a dense and precise homologous point assessment. Robustness and accuracy cannot be guaranteed with local OF approaches. However, local methods are fast and give sub-pixel accuracy in the bladder image pairs with high contrast and well-distributed texture. Ali et al. [Ali et al., 2013a] proposed to use a global optical flow method based on total variational approach for finding the homologous pixels between the image pairs. The author used these dense pixel correspondences to establish a homography between images. A feasibility test was made in [Ali et al., 2013b] using TV-L^1 based dense OF estimation for low textured cystoscopic images. A switch between a feature-based approach and the pixel-based OF method depending on the presence of image contrast and texture was used to check the interest of combining different approaches for fast, robust and accurate mosaicing. In presence of distributed texture, the SURF feature extraction method was used for establishing correspondence in image pairs as in Fig. 1.13 (a), otherwise the OF approach was automatically selected for determining dense point correspondence in image pairs (refer Fig. 1.13 (b-c)). The TV- L^1 based OF technique was used in coarse-to-fine image resolution approach for handling large cystoscope displacements.

None of the above described methods in literature were tested for both WL and FL cystoscopic images (i.e. robustness of algorithms independent of their modality). It was shown that global methods like graph-cut based method presented in [Weibel et al., 2012b] and optical flow techniques used in [Hernandez-Mier et al., 2010, Ali et al., 2013a, Ali et al., 2013b] had larger errors in presence of larger illumination variabilities and change in cystoscopic modality.

Motivation of dense correspondance establishment for bladder image registration

The literature of bladder image registration discussed above indicates that pixel based registration approaches are the most appropriate for registering bladder images as they consist of images with poor contrast and high intra- and inter- patient texture variability. Preliminary works and results in this thesis [Ali et al., 2013a, Ali et al., 2013b] have shown that a total variational approach in optical flow has a high potential for robust and accurate image registration. Total variational approaches are flexible for data-cost modeling and can also be interesting in terms of computational time (they can notably be parallelized). One major contribution of this thesis is to validate energy minimization techniques in a total variational framework for computing accurate motion fields between image pairs. The robustness of the formulated and implemented methods are validated on several realistic data-sets with varying texture and severe lighting conditions. The dense estimated motion vectors are then used for estimating homographies for accurately registering of image pairs. It is to be noted that homographies are essential for compositing all frames to one global coordinate for obtaining mosaic. Another challenge of this thesis is to shown that algorithms based on total vatiotional optical flow can be a solution for robust and accurate registration of image sequences of both WL modality and FL modality (i.e., modality independent).

1.3.4 Global map correction

The bundle adjustment (BA) technique [Triggs et al., 2000] was initially developed to find 3D point positions and camera motion parameters that minimize the reprojection error from a given set of putative point correspondences. Thus, a classical application of BA is to construct a 3D

model of a scene using sequences of 2D images. However, BA can also be usefull as correction step in the frame of a 2D mosaicing algorithm. Indeed, BA can be used as the last step of feature based multiview structure and motion estimation algorithms applicable in 2D or 3D image stitching [Brown and Lowe, 2007, Marzotto et al., 2004], structure-from-motion (SfM) [Snavely et al., 2006]. The principle of BA is explained below using its original application (3D scene/structure modeling/reconstruction using a sequence of images).

This optimization problem is usually formulated as a non-linear least squares problem, where the error is the squared L2 norm of the difference between the observed 2D feature location and the projection of the corresponding 3D point on the image plane of the camera. The idea is to most accurately predict the locations of N observed 2D points in a set of M available images.

Let $\mathbf{x}_{qi}$ be the projection of the q^{th} 3D point on image i and $\hat{\mathbf{x}}_{qi}$ be the predicted projection of point q on image i then, the minimization of this reprojection error over all points and camera parameters (i.e. over all camera positions or acquisitions) is represented by a least square problem as:

$$\min_{K_i,\Gamma_\xi R_i,t_i,\mathbf{X}_q} \sum_{q=1}^{N}\sum_{i=1}^{M} \parallel \mathbf{x}_{qi} - \hat{\mathbf{x}}_{qi} \parallel^2, \tag{1.14}$$

where K_i, Γ_ξ, $\{R_i, t_i\}$ stand respectively for the intrinsic camera parameters (in uncalibrated case), lens distortion function, the 3D rotation and translation between viewpoints and $\mathbf{X}_q$ being the 3D point. In order to estimate the reprojection point, camera intrinsic parameters K_i and distortion function Γ_ξ are need to be known (see Eq. (1.1) and Eq. (1.14)) which is not often practical in case of endoscopes. However, a generalization of BA with application to 2D video mosaicing was formulated in [Marzotto et al., 2004]. The objective was to find global homography $H_{0\leftarrow i}$ placing the pixels of I_i in reference image I_0 and minimizing the misalignment of the pre-defined set of M grid points.

Let $\mathbf{x_g}$ be a grid-point and L_g be the set of edges $(i,j) \in L$ such that $\mathbf{x_g}$ belongs to the overlapped region between frames i and j then the error at the grid-point $\mathbf{x_g}$ is:

$$\epsilon = \frac{1}{\mid L_g \mid} \sum_{(i,j)\in L_g} \left(\mathbf{x_g} - H_{0\leftarrow i} H_{i\leftarrow j} H_{0\leftarrow j}^{-1} \mathbf{x_g} \right), \tag{1.15}$$

where $H_{i\leftarrow j}$ is the local homography relating frame j to frame i and is associated with their global homographies as $H_{i\leftarrow j} = H_{0\leftarrow i}^{-1} H_{0\leftarrow j}$. Since, the BA principle is the simultaneous minimization of error $\epsilon\epsilon^t$, the Levenberg and Marquard algorithm [Marquardt, 1963] is used to minimize Eq. (1.16).

$$\min_{H_{0\leftarrow i}} \sum_{g=1}^{M} \parallel \epsilon \parallel^2 . \tag{1.16}$$

Motivation for minimizing the use of BA in bladder mosaicing algorithms

Registration errors either small or large exists between image pairs. These errors accumulates themselves and corresponding mosaicing error increases with the number of images used to build the mosaic. Mosaicing errors are not visible in case of straight line trajectories of the camera. However, a visible bladder texture misalignment is usually noticeable in closed loops or zig-zag trajectories while depending largely upon the registration accuracy. Additionally, the amount of time required for finding accurate global homographies $H_{0\leftarrow i}$ using BA methods largely depends upon number of images, the value of initial misalignment errors, overlap and number of grid-points taken into account for BA [Marzotto et al., 2004].

In cystoscopic image sequences, an overlap greater than 90% is usually found between consecutive image frames and to obtain larger FOV it is important to take large number of image frames which will consequently lead to larger global misalignment. The magnitude of this global misalignment will largely depend upon both the robustness and accuracy of the paired image registration method. For bladder image pairs, due to numerous challenges explained in previous sections, it is difficult to find an optimal and robust algorithm. Our objective is to propose algorithms which are robust and very accurate for paired bladder image registration, thus minimizing the global error accumulation on go itself without needing for such complex and time consuming BA algorithm. BA is done in this thesis only if there is large misalignment (i.e. greater than misalignment thresholds τ_e in trajectory paths).

1.3.5 Post-processing

After several images are composited to one global coordinate frame with their computed global homographies (corrected homographies if BA used) $H_{0\leftarrow i}, i \in \{0, ..., N-1\}$, there are visible color (mostly brightness) discontinuities in the mosaic which are due to the light gradients between the stitched images. A general practice in image mosaicing is to correct the colors of the overlapped image regions using blending techniques. The border between the stitched images become imperceptible with such methods. An efficient way of doing this is to first find the optimal scene path (seam detection) between the image transitions and then finally using suitable blending techniques [Szeliski, 2006, Agarwala et al., 2004].

Optimal seam finding

Optimal seam selection is performed sequentially as new images are added. 1) If a scene region is seen in several images, the problem is to select the pixels among these images so that their colors minimize the color differences in the mosaic and 2) this task is a labeling problem involving a map which has the same size in pixels as the mosaic and each value (or label) in the map corresponds to the number i of the image I_i which provides the color for that particular point in the mosaic. To do so two objective functions are optimized together for optimal seam finding. The first is per-pixel image $\mathcal{C}_D$ objective determining likelihood of pixels to produce good composites:

$$\mathcal{C}_D = \sum_{\mathbf{p}} D_{I(p)}(\mathbf{p}), \tag{1.17}$$

where $D_{I(p)}(\mathbf{p})$ is the data penalty associated with choosing image I at pixel $\mathbf{p}$. The second objective is to penalize differences in labeling between adjacent images defined as seam objective $\mathcal{C}_S$ and represented as:

$$\mathcal{C}_S = \sum_{(\mathbf{p_i},\mathbf{p_j})\in\mathcal{N}} S_{I_i(p),I_j(p)}(\mathbf{p_i},\mathbf{p_j}), \tag{1.18}$$

where $S_{I_i(p),I_j(p)}$ is the seam cost and $\mathcal{N}$ is set of 4-connected neighborhood pixels. Agarwala et al. [Agarwala et al., 2004] used graph-cut based optimization method to find the seam and later used a gradient based blending approach. A similar approach was used by Thomas et al. [Weibel et al., 2012b] for seamless composition of WL cystoscopic images. Graph-cut based seam finding approach is slow because the optimization is usually done over a large amount of pixels. A relatively faster approach is that proposed by Uyttendaele et al. [Uyttendaele et al., 2001] and produces similar result as that in [Agarwala et al., 2004]. The algorithm compares all overlapping image pairs to determine regions where the images disagree (RODs, i.e. regions

where overlapping pixels have a large color differences usually due to object motion). A graph is then constructed with the RODs as vertices and edges representing ROD pairs in the final composite image. A feathering approach was later used to blend the images after removal of undesired ghost artefacts in images. However, this method is sensitive to region of difference (ROD) threshold.

Feathering or alpha blending

Feathering (also called as alpha blending) is the most simple blending approach in which the pixel values in the blended mosaic regions are the weighted average from overlapping images. This is straightforward and computationally the most efficient with respect to other blending approaches.

Let p be the pixels in overlapping region $I_{I_m \cap I_j}$ of the current mosaic I_m and the new image I_j which has to be placed in the mosaic. With the weight w, the mosaic is updated by compositing the overlapping region as follows:

$$I_{I_i \cap I_j}(p) = (1 - w)I_i(P) + (w)I_j(p) \tag{1.19}$$

Weighted coefficient w is computed as the function of the Euclidean distance between the pixel position in image I_j and the mosaic center [Danielsson, 1980, Saito and Toriwaki, 1994]. For pixels of I_j which are "inside" the mosaic, the mosaic colors strongly impact the color of the newly added pixel while for pixels of I_j at the border of the current mosaic the color of I_j can contribute to a smooth color change in the mosaic. However, due to the low quality of some cystoscopic images and large light gradients between the image pairs, this weighting function do not suffice for obtaining coherent mosaics. Therefore, Yahir et al. [Hernandez-Mier et al., 2010] used a Gaussian function for computing weight $w(p) = 0.9e^{\frac{-r}{2\sigma}} + 0.1$, with σ equal to the one fourth of the image width and r being the distance of $w(p)$ to the center of image I_j. Blurring and ghost artefacts can still be problems in this modified algorithm. Exposure difference between overlapped images and ghosting artefacts are tackled in [Uyttendaele et al., 2001, Agarwala et al., 2004] before feathering or blending. Additionally, due to vignetting effects in cystoscopic images, it is difficult to achieve sharp transitions suppressing low frequencies. Blurring is also evident when Gaussian weight is taken as in [Hernandez-Mier et al., 2010].

Laplacian pyramid blending

Burt and Adelson [Burt and Adelson, 1983] proposed a blending method based on a band-pass Laplacian pyramid. It uses a Gaussian to blend the image while keeping the significant image features. Each level of the pyramid is smoothed with the $\frac{1}{16}[1\ 4\ 6\ 4\ 1]$ binomial kernel and subsampled by a factor of 2. Laplacian image pyramids generated for images (I_i, I_j) are combined at different Laplacian levels by weighting both I_i and I_j with a blurred mask $K_{(ij)}$. This blurred mask is the Gaussian pyramid of the region associated with only I_j. Finally, the composite image is reconstructed by interpolating and summing all of the band-pass images in the pyramid.

Since light gradient differences are both inevitable and inconsistent phenomena in cystoscpic images, selecting a proper weight can be very challenging for feathering. As a result, mosaics obtained due to feathering did not give satisfying (visual) results. So, we have not used such an approch in the global mosaicing scheme. Laplacian pyramid based blending technique has been used to obtain visually coherent mosaics shown in Fig. 1.14. Sharp intensity variation along the

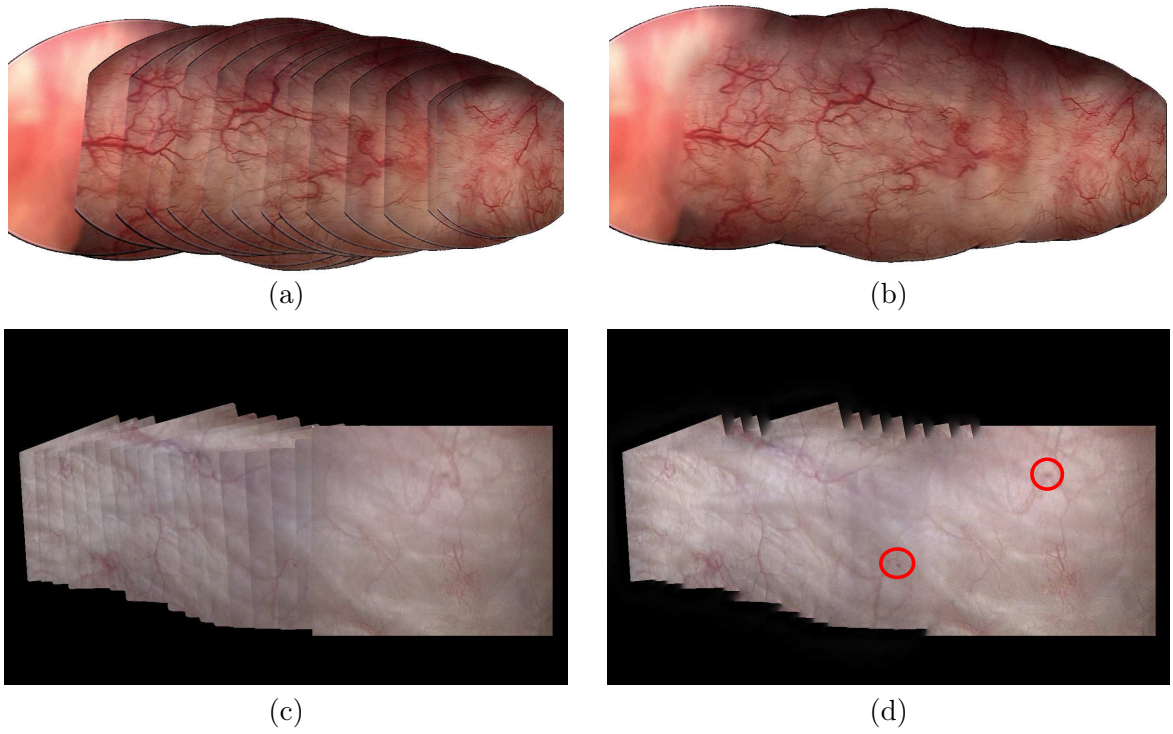


Figure 1.14: Two composited maps before and after blending. a, c) Without blending, b, d) with Laplacian-Gaussian blending technique described in [Burt and Adelson, 1983, Szeliski, 2006]. Intensity discrepancies along the image transitions during stitching are diminished in the blended mosaics. To limit the contrast expansion with the Laplacian blending algorithm, the background of the blended mosaic has been subtracted form the Laplacian blended mosaic with the weight of 0.1. Structures present in the mosaic are preserved and enhanced while keeping the original texture. In (d) the small structures in the red circles are preserved by the blending technique.

image transitions in the mosaics are seen in Fig. 1.14(a) and Fig. 1.14 (c). Moreover, due to vignetting effect in the images, the mosaics are darker along each transition. After Laplacian pyramid based blending technique, relatively smooth transitions are seen. Less visible edge discontinuities are observzable in Fig. 1.14(b) and Fig. 1.14 (d)). Additionally, these mosaics look more brighter than the unblended ones due to strong attenuation of vignetting effects in the stitched images after blending.

Motivation for using blending techniques

Since, seam finding techniques adds to the computational time of the global mosaicing algorithm, we have not used such a technique in this thesis. However, seam finding approaches can be interesting for offline image composition where computational time is not of major concern. Blending techniques minimize the visibility of the intensity discrepency between images stitched together, especially along the transition between them. But, their adverse effect is that they homogenize also all the discontinuities to be preserved along structures (in our case mostly vessels). Moreover, image misalignment (visible due to texture/vessel discontinuities) at image borders are visually attenuated by blending techniques. Thus to evaluate visually the registration accuracy no blending is applied to many of the presented registration results. Image blending in this thesis has therefore been made limited and have been discussed with final results only at the end part of

this manuscript.

1.4 Thesis objectives and contributions

Among all mosaicing steps described in this chapter, the pairwise registration of bladder images is the most crucial step for obtaining in a robust way accurate mosaics facilitating lesion diagnosis and follow-up. A analysis of the literature presented in this chapter shows that image registration is also the step in which most algorithm improvements are required.
Thus the major objective of this thesis is to propose a fast, robust and accurate image registration approach for obtaining large FOV 2D mosaics in cystoscopy. Image registration algorithms as independent as possible to intra- and inter- patient texture variability, illumination conditions, modality changes and cystoscope displacement variability are investigated. The application of algorithms proposed in this thesis are not only limited to cystoscopic images and thus can be applied to other scenes.

Existing pixel based approaches in bladder image registration focus mostly on WL modality (reference modality in cystoscopy). This thesis will deal with both WL and FL image sequences where FL being a complimentary diagnostic modality. The motivation behind developing such an algorithm comes from the fact that FL modality has increased cancer lesion detection rate of over 90% which is much more than WL cystoscopy (only upto 60% sensitivity). Two mosaics in similar scans under different modalities when superimposed on each other can reveal significant information important for reliable diagnosis. Such motivation was not mentioned before and hence remained an open problem in multi-modal cystoscopy. Ideally, a modality independent and illumination invariant image registration approach should be proposed and adapted to cystoscopic image mosaicing. To reach this objective, one major contribution of this thesis is to retain dense homologous pixel correspondences similar to that of a most recent bladder image mosaicing algorithm [Weibel et al., 2012b] while minimizing the computation cost with much more improved accuracy and robustness. Obtaining a such a dense correspondence in a robuste and very accurate manner reduces also the need of expensive bundle adjustment algorithms. Even in loop closing, zig-zac or crossing paths, the estimated geometrical transformation parameters from computed dense pixel correspondences in between image pairs should lead to coherent maps without a strong need for re-estimation of homographies.

Optical flow approaches have been actively used in the framework of paired image registration of bladder images [Weibel et al., 2012b, Hernandez-Mier et al., 2010] and registration of other scenes [Lucas and Kanade, 1981, Brox et al., 2004, Pock et al., 2007]. This thesis proposes dense optical flow estimation methods based on variational energy minimization scheme.

Dense optical flow was chosen for correspondence establishment since among the bladder literature optical flow methods gave the most promise results in terms of compromise between robustness, accuracy and computation time. Moreover, the analysis given in Chapter 2 about the state of the art in optical flow shows that previously proposed bladder registration algorithms do not exploit all possibilities offered by optical flow. Estimated dense flow vectors have been essentially used for computing homographies in image pairs.

1.4.1 Main contributions

Following summarizes the work/contribution of this chapter:

- General overview on all the steps required for a robust and accurate 2D mosaicing of bladder scenes.

- Identification of the mosaicing step (image registration) which will potentially lead to the best improvements of the whole mosaicing procedure results.
- Preliminary tests done with optical flow methods in the framework of this thesis [Ali et al., 2013a, Ali et al., 2013b] provided us with the mind-set of making choice of appropriate methods for dense homologous point correspondence establishment required for robust image registration.
- Automatic switching between feature based approaches and pixels based approaches for fast and robust 2D cystoscopic cartography [Ali et al., 2013b].

List of publications

[ADWB13] Sharib Ali, Christian Daul, Thomas Weibel and Walter Blondel "Fast mosaicing of cystoscopic images from dense correspondence: combined SURF and TV-L1 optical flow method," 20^{th} *IEEE Int. Conf. on Image Processing, (ICIP)*, pp. 1291–1295, Melbourne, Australia, September 2013.

[AWD13] Sharib Ali, Walter Blondel and Christian Daul, "TV-L1 based fast and robust mosaicing of cystoscopic images," $XXIV^{em}$ *Colloque GRETSI Traitment du Signal and des Images (GRETSI)*, CDROM, Brest, France, August 2013.

Chapter 2

Optical flow

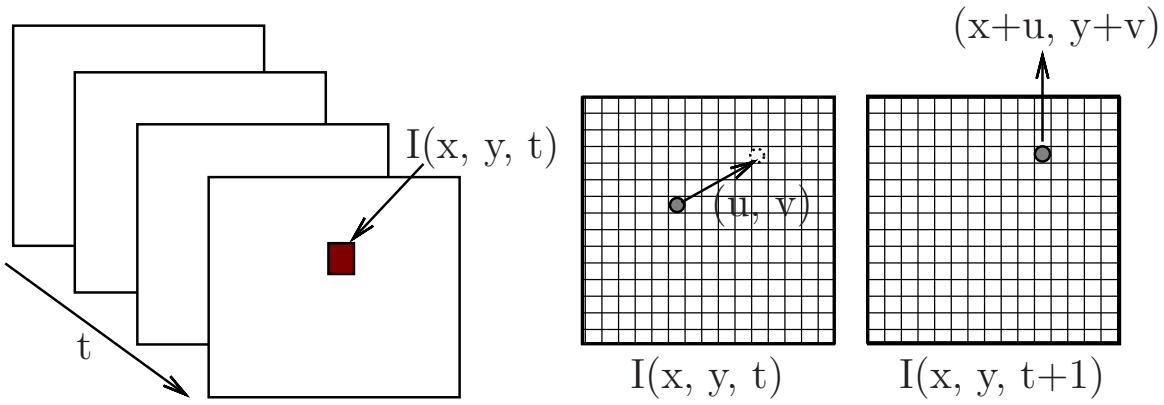
Contents

2.1 Motivation: Fast and Robust establishment of dense correspondences 29
2.2 Optical flow 30
2.2.1 The optical flow constraint 30
2.2.2 Local and global approaches 32
2.3 Modelling of variational optical flow 35
2.3.1 Data-term modeling 35
2.3.2 Regularizer 41
2.4 Mathematical optimization 46
2.4.1 Prerequisites 46
2.4.2 Convex optimization 50
2.5 TV-L^1 optical flow: Background and first contribution 53
2.5.1 Robust energy model (RFLOW) 53
2.5.2 Primal-dual energy minimization 57
2.5.3 Optical flow assessment and benchmarking 59
2.5.4 Results and discussion 60
2.6 Main contributions 64
List of publication 64

2.1 Motivation: Fast and Robust establishment of dense correspondences

Estimating the motion of scene points is an important task which can be seen as a dense pixel correspondence problem in consecutive frames of a video sequence or in left and right image pairs in passive stereo-vision. Due to the large flexibility in defining robust data-costs, optical flow methods can provide significant improvements in establishing putative dense point correspondences between image pairs. In this thesis, we are interested in solving the correspondence problem using robust formulation of total variational approach for images of a static scene taken with a single moving camera.

This chapter consists of an introduction to some well established formulations in optical flow domain for motion estimation: trends in data-terms utility and smoothness terms have been revisited. Additionally, a new total variational based energy modeling has been formulated and



(a) (b)

Figure 2.1: Representation of motion in video-data. a) Video-frame as a function of space (x, y) and time t, b) 2D displacement (u, v) in between consecutive video frames.

optimized to give more robust and accurate flow fields comparative to some classical TV-L^1 models. We have also introduced well established flow error quantification techniques and some well known public datasets used for flow field evaluation and benchmarking of the algorithms.

2.2 Optical flow

In computer vision, motion estimation is one of the major open problem. This is because motion estimation is useful in a wide range of image processing techniques like image registration, video-compression, super resolution, object tracking, structure from motion, motion segmentation and many more. The 2D displacement vector field describing the apparent motion of a scene in two images is popularly known as optical flow. It is the 2D projection of some 3D motion onto the image plane. It was developed as an idea to notice the magnitude and direction of the change in intensity/brightness patterns between consecutive frames. This is why, even without actual motion of the scene displacement, a non null optical flow field can occur, for instance due to moving shadows, reflections or view-point changes (so an optical flow does not always correspond exactly to a motion field). Although the assumption of constant intensities over time is incorrect, it has been widely used in the literature of optical flow leading to often satisfying results.

2.2.1 The optical flow constraint

Let us consider $I(x, y, t) : \Omega \to \mathbb{R}$ be an image sequence captured over time t with I being the function of space $\mathbf{x} = (x, y)$ and time $t \in [0, T]$. Now, determining the optical flow means calculating the motion field vectors between the images of the sequence given by $\vec{\mathbf{v}} = (u, v) : \Omega \to \mathbb{R}^2$. According to the classical brightness constancy assumption, the intensity function over time is considered to remain constant [Lucas and Kanade, 1981, Horn and Schunck, 1981], that is,

$$\frac{dI(x, y, t)}{dt} = 0. \tag{2.1}$$

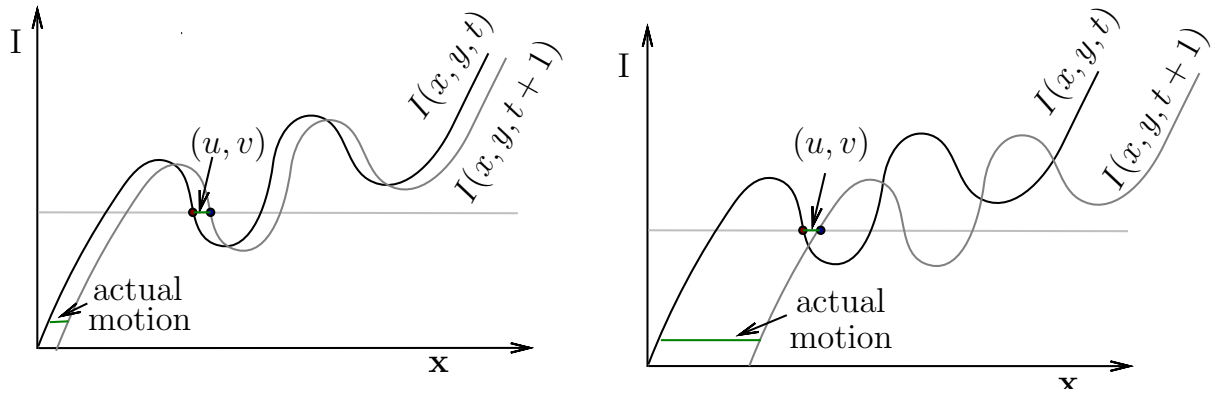


Figure 2.2: Illustration of temporal alliasing effects on optical flow. (a) Small motion giving correct nearest match, *i.e.* no aliasing. (b) Large motion, nearest match is not correct due to aliasing.

Using Taylor's expansion and neglecting 2^{nd} and higher order terms, we get:

$$\frac{\partial I(x,y,t)}{\partial x}\cdot\frac{dx}{dt} + \frac{\partial I(x,y,t)}{\partial y}\cdot\frac{dy}{dt} + \frac{\partial I(x,y,t)}{\partial t} = 0. \tag{2.2}$$

Let u_0 and v_0 be the given initial flow fields, then the incremental motion vector is given by,

$$d\vec{\mathbf{v}} = (u - u_0, v - v_0) = \left(\frac{dx}{dt}, \frac{dy}{dt}\right). \tag{2.3}$$

The spatial image gradient is given by:

$$\nabla I(x,y,t) = \left(\frac{\partial I(x,y,t)}{\partial x}, \frac{\partial I(x,y,t)}{\partial y}\right) = (I_x, I_y), \tag{2.4}$$

and the temporal image derivative is given by:

$$I_t(x,y,t) = \frac{\partial I(x,y,t)}{\partial t} = I(x,y,t) - I(x,y,t+1). \tag{2.5}$$

From equations (2.2), (2.3), (2.4) and (2.5), we obtain, the classical *Optical Flow Constraint* (OFC) equation as:

$$\nabla I(x,y,t)^T (u - u_0, v - v_0) + I_t(x,y,t) = 0. \tag{2.6}$$

Two problems are most dominant in the OFC given in Eq. (2.6):

1 Temporal aliasing can lead to false optical flow vectors. It is more persistent if the motion vector is large (refer to Fig. 2.2),

2 For each pixel there are two unknowns (u, v) and one equation. So, only the flow component normal to image edges $\mathbf{v}_\perp = (-\frac{I_t}{|\nabla I|})(-\frac{\nabla I^T}{|\nabla I|})$ can be determined as represented on the left of Fig. 2.3 by a solid arrow. This ambiguity in direction of motion perception is called aperture problem. Moreover, in homogeneous regions (with no information) motion vectors cannot be estimated (refer to right of Fig. 2.3).

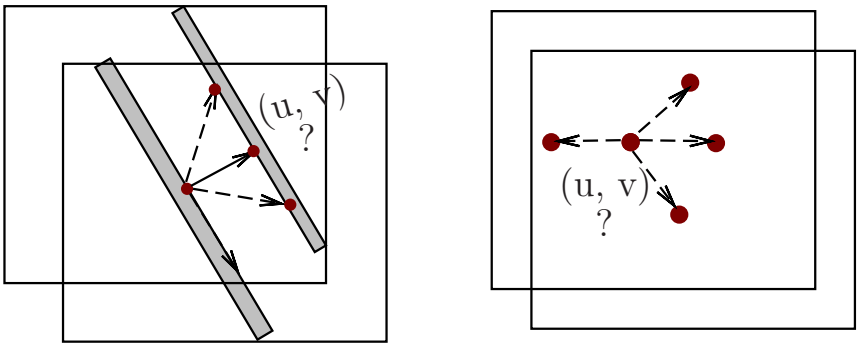


Figure 2.3: Illustration of the ambiguity of the OFC equation. Left: Aperture problem, only flow field (u, v) normal to the edge (denoted by solid arrow) can be computed. However, the two other (dashed) arrows can also represent the actual solution. Right: in images regions without intensity variations all solutions are possible for vector (u, v).

Differential methods that make this ill-posed Eq. (2.6) to a well posed problem can be divided into two classes namely local and global approaches. A brief overview of both is given in next subsections.

2.2.2 Local and global approaches

a) Local approach

Lucas and Kanade [Lucas and Kanade, 1981] proposed a *local* method that is based on the assumption that the optical flow field is constant over the neighborhood of the current pixel location $p(x, y)$. Let $\mathbf{x} = (x, y)$ be the center pixel of the neighborhood and $\hat{\mathbf{x}} = (\hat{x}, \hat{y})$ be the neighboring pixels having a weighting function $g(\hat{\mathbf{x}}, \mathbf{x})$ then the optical flow at position $\mathbf{x}$ in the image can be estimated by minimizing the squared errors in the neighborhood $\mathcal{N}(\mathbf{x})$ given as:

$$\mathbf{v}E(\mathbf{v}) = \sum_{\hat{\mathbf{x}} \in \mathcal{N}(\mathbf{x})} g(\hat{\mathbf{x}}, \mathbf{x}) \left(\nabla I(\hat{\mathbf{x}}, t)^T \mathbf{v} + I_t(\hat{\mathbf{x}}, t) \right)^2 . \tag{2.7}$$

Since Eq. (2.7) is convex, the global minimum is reached, if $\dfrac{\partial E(\mathbf{v})}{\partial \mathbf{v}} = 0$, *i.e.*,:

$$\begin{cases} \dfrac{\partial E(\mathbf{v})}{\partial u} = \sum g(\hat{\mathbf{x}}, \mathbf{x}) \left(I_x u + I_y v + I_t \right) I_x = 0 & \text{(2.8a)} \\ \dfrac{\partial E(\mathbf{v})}{\partial v} = \sum g(\hat{\mathbf{x}}, \mathbf{x}) \left(I_x u + I_y v + I_t \right) I_y = 0, & \text{(2.8b)} \end{cases}$$

which can be formulated in a matricial form as:

$$A \mathbf{v}^T = b, \tag{2.9}$$

where, $A = \begin{bmatrix} \sum g I_x I_x & \sum g I_x I_y \\ \sum g I_x I_y & \sum g I_y I_y \end{bmatrix}$, $\mathbf{v}^T = \begin{pmatrix} u \\ v \end{pmatrix}$ and $b = - \begin{bmatrix} \sum g I_x I_t \\ \sum g I_y I_t \end{bmatrix}$.
Thus, the flow for the image pixel $p(x, y)$ can be solved using least-squares estimate. When the rank of matrix A is equal to 2, the image structure in the local neighborhood $\mathcal{N}(x)$ has enough information to solve the aperture problem and the solution becomes $\mathbf{v} = A^{-1} b$. However, to

ensure the non-singularity and obtain a unique solution, in practice large neighborhood window in pixels $\mathbf{x}$ has to be considered. The reliability of such methods largely depends upon the eigenvalues of A (assuming $\lambda_+ \leq \lambda_-$), representing locally the image structures. If both eigenvalues are large then the flow can be uniquely determined. However, when $\lambda_+ >> \lambda_-$ then only normal flow can be determined, while for $\lambda_- = 0$ or too large eigenvalues ratio ($\frac{\lambda_+}{\lambda_-}$) no flow fields can be determined.

In order to deal with the temporal aliasing of the optical flow, a coarse-to-fine approach has to be used. The basic idea is to initially compute the optical flow at a coarse image resolution. The resulting optical flow can then be upsampled and refined at successively higher resolutions, up to the resolution of the original input images. This approach can be used to recover larger displacements.

Drawbacks of local methods: To compute the unique solution in pixel $\mathbf{x}$, the size of the neighborhood is increased assuming a constant displacement in $\mathcal{N}(x)$. However, in image frames with large homogeneous regions, rank deficiency of matrix A will still persist. A more trivial way to solve this problem is to compute the optical flow only where a unique solution can be found. This yields to a sparse optical flow field, in-turn resulting in sparse point correspondences between consecutive video frames. Moreover, the assumption of constant velocity within a region is generally not valid in case of realistic data.

Since, one the hand, the velocity field has two components and , on the other hand, the change in brightness gives only one constraint, the Optical flow cannot be computed at a point in the image independently of its neighboring points without introducing additional constraints. Global approaches introduces this constraint without considering only neighborhood pixels with assumption of constant velocity in them.

b) Global Approach

A global method uses a smoothness term that propagates optical flow fields from well–posed regions to regions with poorly conditioned data fidelity. Such a method utilize the complete available image data to estimate flow fields and yields a dense result.

Horn and Schunck [Horn and Schunck, 1981] formulated the optical flow determination as an optimization problem of the form:

$$\min_{(u,v)} \{ \int_\Omega \underbrace{(I_x.u + I_y.v + I_t)^2}_{data\ term\ (\mathcal{D})} d\mathbf{x} \quad + \quad \lambda_s^2 \int_\Omega \underbrace{(\| \nabla u \|^2 + \| \nabla v \|^2)}_{smoothness\ term\ (\mathcal{R})} d\mathbf{x} \}, \tag{2.10}$$

where λ_s defines the trade–off between the smoothness ($\mathcal{R}$, regularization) and the data term $\mathcal{D}$. The vector field smoothness over the whole image space Ω is calculated by the sum of the squares of the of the gradients of the flow field vector components u and v as defined by the second term in Eq. (2.10). It can also be expressed as:

$$\lambda_s^2 \int_\Omega (\| \nabla u \|^2 + \| \nabla v \|^2) d\mathbf{x} = \lambda_s^2 \int_\Omega (u_x^2 + u_y^2 + v_x^2 + v_y^2) d\mathbf{x}, \tag{2.11}$$

with $\nabla u = (u_x, u_y)$ and $\nabla v = (v_x, v_y)$. This smoothness term can also be represented as the sum of the squares of the Laplacians of the x- and y-components of the flow. The Laplacians of u and v are defined as:

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \nabla^2 v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}. \tag{2.12}$$

Thus, Eq. (2.10) can be modified as the following global energy minimization problem:

$$\min_{(u,v)} E = \min_{(u,v)} \{ \int_{\Omega} (I_x.u + I_y.v + I_t)^2 d\mathbf{x} + \lambda_s^2 \int_{\Omega} (\nabla^2 u + \nabla^2 v) d\mathbf{x} \}. \tag{2.13}$$

The energy E can be minimized by solving the associated multi-dimensional Euler-Lagrange equations:

$$\begin{cases} \dfrac{\partial L}{\partial u} - \dfrac{\partial}{\partial x}\dfrac{\partial L}{\partial u_x} - \dfrac{\partial}{\partial y}\dfrac{\partial L}{\partial u_y} = 0 & \text{(2.14a)} \\ \dfrac{\partial L}{\partial v} - \dfrac{\partial}{\partial x}\dfrac{\partial L}{\partial v_x} - \dfrac{\partial}{\partial y}\dfrac{\partial L}{\partial v_y} = 0, & \text{(2.14b)} \end{cases}$$

where L is the integrand of the energy E giving:

$$I_x^2 u + I_x I_y v = \lambda_s^2 \nabla^2 u - I_x I_t \tag{2.15a}$$

$$I_x I_y u + I_y^2 v = \lambda_s^2 \nabla^2 v - I_y I_t \ . \tag{2.15b}$$

A convenient way to solve for the Laplacians is the numerical approximation is by using finite differences as:

$$\nabla^2 u = \bar{u} - u \quad \text{and} \quad \nabla^2 v = \bar{v} - v, \tag{2.16}$$

where the mean values $\bar{u}$ and $\bar{v}$ are computed by convoluting u and v with the mask M, *i.e.*, $\bar{u} = M * u$ and $\bar{v} = M * v$, with,

$$M = \begin{bmatrix} \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \end{bmatrix}. \tag{2.17}$$

Thus, equations (2.15–2.16) can be rewritten as:

$$(\lambda_s^2 + I_x^2) u + I_x I_y v = (\lambda_s^2 \bar{u} - I_x I_t) \tag{2.18a}$$

$$I_y I_x u + (\lambda_s^2 + I_y^2) v = (\lambda_s^2 \bar{v} - I_y I_t) \ . \tag{2.18b}$$

The determinant of the coefficient matrix of Eqns. (2.18) is equal to $\lambda_s^2(\lambda_s^2 + I_x^2 + I_y^2)$. Solving for u and v, we find:

$$(\lambda_s^2 + I_x^2 + I_y^2) u = (\lambda_s^2 + I_y^2)\bar{u} - I_x I_y \bar{v} - I_x I_t \tag{2.19a}$$

$$(\lambda_s^2 + I_x^2 + I_y^2) v = -I_x I_y \bar{u} + (\lambda_s^2 + I_x^2)\bar{v} - I_y I_t \ . \tag{2.19b}$$

Rewriting Eqns. (2.19a) and (2.19b), we get the equations of the form:

$$(\lambda_s^2 + I_x^2 + I_y^2)(u - \bar{u}) = -I_x \left[I_x \bar{u} + I_y \bar{v} + I_t \right] \tag{2.20a}$$

$$(\lambda_s^2 + I_x^2 + I_y^2)(v - \bar{v}) = -I_y \left[I_x \bar{u} + I_y \bar{v} + I_t \right] \ . \tag{2.20b}$$

We can see from the Eqns. (2.20a) and (2.20b) that λ_s^2 plays a significant role only for the areas where the brightness gradient is small. An iterative solution is proposed by Horn and Schunck to

find a new set of velocity estimates (u^{n+1}, v^{n+1}) from the estimated derivatives and the average of the previous velocity estimates $(\bar{u}^n, \bar{v}^n)$ as:

$$\begin{cases} u^{n+1} = \bar{u}^n - \dfrac{I_x \left[I_x \bar{u}^n + I_y \bar{v}^n + I_t\right]}{(\lambda_s^2 + I_x^2 + I_y^2)} & \text{(2.21a)} \\ v^{n+1} = \bar{v}^n - \dfrac{I_y \left[I_x \bar{u}^n + I_y \bar{v}^n + I_t\right]}{(\lambda_s^2 + I_x^2 + I_y^2)} \; . & \text{(2.21b)} \end{cases}$$

The major advantage of this method is to handle the image regions with few textures where local method gives ill-conditioned system of equations. The regularization term propagates the optical flow into such regions by filling them with the neighboring estimates. The Horn and Schunck model [Horn and Schunck, 1981] combines a quadratic penalization of the classical optical flow constraint (OFC) for modeling the data fidelity and a quadratic penalization of the flow gradients enforcing a smooth flow field.

Drawbacks of the Horn ad Schunck model: The major drawback of this method is that the regularization term penalizes high gradients of u and v that results in over–smoothing effect. In other words, this method disallows the discontinuities in the flow field. Secondly, prior assumption of smooth flow field does not hold for images with strong gradients (edge pixels), and only allows for small displacements. Additionally, this method cannot handle robustly the outliers in the data fidelity term. Moreover, the computation effort to optimize energy equation (2.10) has limited the use of this approach in many applications. However, as mentioned in [Baker et al., 2011], the accuracy of such global variational methods are among the best performing algorithms available for optical flow estimation.

2.3 Modelling of variational optical flow

Optical flow algorithms mimic human visual perception by estimating the motion of objects and/or of complete scene parts. Robust and dense optical flow computation is challenging in scenes with strong texture variability, changing illumination conditions, topology differences, local deformations, large displacements and occlusions. In addition, imaging artifacts like blur due to camera motion and defocus or refocus of images foster these challenges leading to an inaccurate estimation of optical flow fields. Therefore, obtaining a dense and accurate optical flow field under challenging scene and imaging conditions is still an open problem in computer vision. Due to the flexibility of energy formulation and optimization of the variational model [Horn and Schunck, 1981], it has been an active research area since past few decades. The majority of the proposed algorithms vary in the way the data-fidelity and/or the regularization term/s are modeled. This section briefly describes the general trends that are relevant to the work accomplished in this thesis.

2.3.1 Data-term modeling

Different constancy assumptions were formulated in the past for modeling robust data-term in the framework of variational approach. The data-term has often been modeled as the choice of one or more scene dependent constancy assumptions. Both quadratic and non-quadratic penalizations of these constancy assumptions have been used in the literature. Also, some approaches have used modified version of OFC equation (refer to Eq. (2.6)). In this section, we discuss each of these changes with the related literature that leads to our motivation for designing robust cost functions for solving the correspondence problem between image pairs.

a) Constancy assumptions

Let $\nabla_2 := (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})^T : \mathbb{R}^2 \supset \Omega \rightarrow \mathbb{R}$ be with two spatial derivatives along x and y (with, $\mathbf{x} = (x, y)$) and $\nabla_3 := (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial t})^T : \mathbb{R}^3 \supset \Omega \rightarrow \mathbb{R}$ be two previous spatial derivatives and one temporal derivative at ($\mathbf{x}$, t). We wish to determine the displacement vector between two frames t and $t+1$, *i.e.* to find $\mathbf{u} = (u, v, 1)^T$, with $u : \Omega \rightarrow \mathbb{R}$ and $v : \Omega \rightarrow \mathbb{R}$ being displacements along the $x-$direction and $y-$direction respectively.

Brightness Constancy Assumption (BCA): This assumption is based on the idea that the degree of similarity between the two corresponding pixels or regions is high in terms of intensity. Ideally, the brightness in the neighborhoods of two consecutive images is assumed to remain constant. Thus, according to this brightness constancy assumption (BCA):

$$\rho_{bca}(\mathbf{x}, \mathbf{v}, t) = I(\mathbf{x} + \mathbf{v}, t + 1) - I(\mathbf{x}, t) = 0. \tag{2.22}$$

The linearization of Eq. (2.22) using the Taylor's expansion lead to the well-known OFC equation in Eq. (2.6), *i.e.*,

$$\rho_{bca}(\mathbf{x}, \mathbf{v}, t) \approx \nabla I(\mathbf{x}+\mathbf{v}, t + 1)(\mathbf{v} - \mathbf{v}^0) + I(\mathbf{x}+\mathbf{v}, t + 1) - I(\mathbf{x}, t). \tag{2.23}$$

Then the L^2 data-term can be written as:

$$\mathcal{D}_1(\mathbf{x}, \mathbf{v}, t) = \mid \rho_{bca}(\mathbf{x}, \mathbf{v}, t) \mid^2 = \mid \nabla I(\mathbf{x}+\mathbf{v}, t + 1)(\mathbf{v} - \mathbf{v}^0) + I(\mathbf{x}+\mathbf{v}, t + 1) - I(\mathbf{x}, t) \mid^2 = 0, \tag{2.24a}$$

which can also be written as:

$$\mathcal{D}_1(\mathbf{x}, \mathbf{u}, t) = (\mathbf{u}^T \ \nabla_3 I(\mathbf{x}+\mathbf{v}, t + 1))^2. \tag{2.24b}$$

The classical data-term of Eq. (2.24a) has been successfully used when the brightness constancy assumption is fulfilled [Lucas and Kanade, 1981, Horn and Schunck, 1981, Aubert et al., 1999, Zach et al., 2007]. However, in case of variable illumination in scenes, this assumption does not hold. This may occur in both classical photography (where local or global illumination changes between images are for instance due to large camera viewpoint changes, moving objects, shadows or day light variations in outdoor scenes) and in medical images (*e.g.* imaging artifacts due to organ specular reflections, images obtained under different imaging protocols, vignetting effects due to diffuse lighting conditions in endoscopes etc).

Gradient Constancy Assumption (GCA): In order to tackle global illumination changes (all grey-levels are changed by a nearly constant factor between two acquisitions), the spatial brightness gradient ∇I is assumed to be constant over time [Uras et al., 1988, Papenberg et al., 2006, Tistarelli, 1996] instead of brightness constancy. That is,

$$\nabla I(\mathbf{x} + \mathbf{v}, t + 1) - \nabla I(\mathbf{x}, t) = 0. \tag{2.25}$$

Thus, after linearization of Eq. (2.25) and similar calculus steps as for BCA, we get, the data-term as:

$$\mathcal{D}_2(\mathbf{x}, \mathbf{u}, t) = \left(\mathbf{u}^T \nabla_2 (\nabla_3 I(\mathbf{x}+\mathbf{v}, t + 1))\right)^2. \tag{2.26}$$

Some recent algorithms [Brox et al., 2004, Brox and Malik, 2011, Rashwan et al., 2013] have used GCA as a complementary data-term along with BCA to deal with illumination changes between image pairs.

Hessian Constancy Assumption (HCA): Higher order derivatives can be considered for constancy assumption formulation [Papenberg et al., 2006], one choice is to use Hessian matrix $\mathcal{H}_2$ of images $I(\mathbf{x},t)$ and $I(\mathbf{x}+\mathbf{v},t+1)$. So, the data-term with Hessian constancy assumption can be written as:

$$\mathcal{D}_3(\mathbf{x},\mathbf{v},t) = \mid \mathcal{H}_2 I(\mathbf{x}+\mathbf{v},t+1) - \mathcal{H}_2 I(\mathbf{x},t) \mid^2 = 0, \tag{2.27a}$$

after linearizing and similar calculus steps as above, it becomes:

$$\mathcal{D}_3(\mathbf{x},\mathbf{u},t) = \left(\mathbf{u}^T \nabla_2(\nabla_2(\nabla_3 I(\mathbf{x}+\mathbf{v},t)))\right)^2. \tag{2.27b}$$

Data-terms with higher-order derivatives than two are sensitive to noise and hold sparse information (usually null). Both gradient and Hessian provide directional information which can be exploited for improving robustness of the algorithm. However, such directional information is not valid in case of image in-plane rotation as the gradients in this case changes and are not comparable between two frames. Thus, translational and divergent motions can be well estimated but rotational motion fields impose problems in its efficiency. Brox et al. [Brox et al., 2004] used GCA as a complementary term to BCA (High Accuracy Optical Flow, HAOF). A weighting factor was used to compensate the adverse effect of one on the other in extreme cases (either illumination change or rotation). Other variants of gradient based constancy assumptions according to Papenberg et al. [Papenberg et al., 2006] are listed in Table 2.1.

Table 2.1: Variants of gradient-based assumptions according to [Papenberg et al., 2006].

Constancy assumption	**Data-term** $(\mathcal{D})$
Laplacian	$\mathcal{D}_4 : \mid \triangle I(\mathbf{x}+\mathbf{v},t+1) - \triangle I(\mathbf{x},t) \mid^2$
Norm of gradient	$\mathcal{D}_5 : (\mid \nabla I(\mathbf{x}+\mathbf{v},t+1) \mid - \mid \nabla I(\mathbf{x},t) \mid)^2$
Norm of Hessian	$\mathcal{D}_6 : (\mid \mathcal{H}_2 I(\mathbf{x}+\mathbf{v},t+1) \mid - \mid \mathcal{H}_2 I(\mathbf{x},t) \mid)^2$
Determinant of the Hessian	$\mathcal{D}_7 : (\det \mathcal{H}_2 I(\mathbf{x}+\mathbf{v},t+1) - \det \mathcal{H}_2 I(\mathbf{x},t))^2$

b) Data-term penalisation

In the above cases, the data-terms penalize deviations from constancy assumptions in a quadratic way, *i.e.*$\Psi(s^2) = s^2$ with $\Psi(.)$ as the function of a term defined by s. This was also used by Horn-Schunck model [Horn and Schunck, 1981]. Such quadratic models cannot handle outliers in the data-term. It is thus desirable to penalise the outliers less severely so as to preserve the discontinuities in the data-term. Thus, most recent approaches [Aubert et al., 1999, Brox et al., 2004, Zach et al., 2007, Brox and Malik, 2011, Rashwan et al., 2013] compute the L^1 norm of the data term. In order to guarantee well-posedness $\Psi_\rho(s^2)$, the penalizer is used such that it is convex in s:

$$\mathcal{D} = \Psi_\rho(s^2) = \sqrt{s^2+\epsilon^2}, \tag{2.28}$$

where, ϵ is a small positive constant. This leads to a non-smooth optimization problem allowing for motion discontinuities along the edges. Such a minimization technique is also robust to outliers. In non-convex functions, finding the global minimum is difficult because of presence of multiple minima. Thus, transforming the non-convex function to convex one makes the computation more efficient and easier to obtain global minimal solution of the energy formulated. This method in variational optical flow has been well established as TV-L^1 models.

c) Data-term for large displacements

Classically, coarse-to-fine strategies are used to deal with large displacements in image pairs [Brox et al., 2004, Aubert et al., 1999, Zach et al., 2007, Wedel et al., 2009b, Brox and Malik, 2011, Rashwan et al., 2013]. However, the proposed techniques often suffer from the intrinsic limitation of multi-scale approaches which is due to disappearance of small objects or weakly contrasted textures at coarser-levels. To tackle such limitation of the pyramidal approaches, large displacement optical flow methods (SIFT flow, LDOF, Deep Flow) described in [Liu et al., 2011, Brox and Malik, 2011, Weinzaepfel et al., 2013] integrate sparse key-point correspondences obtained by SIFT descriptors at each pyramid level of the variational framework. The confidence provided locally by the matched key-points for each region guides the variational scheme (initialization of optical flow vectors at each pixel) to retrieve flow fields for large displacements.

Liu et al. [Liu et al., 2011] defined an objective function similar to total variational optical flow framework but using Scale-Invariant Feature Transform (SIFT) descriptors [Lowe, 2004] instead of raw pixels. The idea was to obtain dense correspondences between image pairs using discrete and discontinuity preserving flow algorithm in a coarse-to-fine approach. Two SIFT descriptors were created for each pair of homologous pixels of two images. The vectors associated to each pixel consists of $4 \times 4 \times 8$ features (vector with 128 components). The data-term was defined as truncated L^1 norm between these descriptors, and also for the regularization terms to encourage sharp discontinuity. If $s_1(\mathbf{x})$ and $s_2(\mathbf{x})$ are the two descriptor vectors then the energy associated with the data-term can be written as:

$$E_{SIFTflow}(\mathbf{v}) = \sum_{\mathbf{x}} \min \left(\| \, s_2(\mathbf{x}+\mathbf{v}) - s_1(\mathbf{x}) \, \|_1 \right). \tag{2.29}$$

Such discrete methods have few major drawbacks: 1) they do not provide sub-pixel accuracy and 2) they are inaccurate at motion discontinuities and for non-translational motions [Brox and Malik, 2011]. To overcome these limitations, Brox and Malik [Brox and Malik, 2011] proposed to combine descriptor matching with classical data-terms based on brightness constancy ρ_{bca} (defined in Eq. (2.22)) and gradient constancy ρ_{gca} (defined in Eq. (2.25)) of variational model and solved using Euler-Lagrangian approach by making the function convex in s. The energy modeled using this combined approach can be written as:

$$E_{LDOF}(\mathbf{v}) = E_{bca}(\mathbf{v}) + \gamma E_{gca}(\mathbf{v}) + \beta E_{match}(\mathbf{v}, \hat{\mathbf{v}}) + E_{desc.}(\hat{\mathbf{v}}) + \alpha E_{smooth}(\mathbf{v}), \tag{2.30}$$

where, E_{bca} and E_{gca} are the energy associated with the data-terms $\mathcal{D}_\infty$ and $\mathcal{D}_\in$ respectively in Eqns. (2.24a) and (2.26), E_{smooth} is the regularizer and $\{\gamma, \beta, \alpha\}$ are the tunning parameters. E_{match} and $E_{desc.}$ are related to the descriptor matching cost in the energy minimizing the data-term and are defined as:

$$E_{match} = \int \delta_i(\mathbf{x}) d_i(\mathbf{x}) \Psi \left(| \, \mathbf{v} - \hat{\mathbf{v}} \, |^2 \right) d\mathbf{x} \quad \text{and} \tag{2.31a}$$

$$E_{desc.} = \int \delta_i(\mathbf{x}) d_i(\mathbf{x}) \left(| \, f_2(\mathbf{x}+\hat{\mathbf{v}}) - f_1(\mathbf{x}) \, |^2 \right) d\mathbf{x}. \tag{2.31b}$$

In Eq. (2.31a) and Eq. (2.31b), $\delta_i(\mathbf{v}) = 1$ if descriptor is available in the first image at $\mathbf{x}$, otherwise, it is 0, $d_i(\mathbf{x}) = \dfrac{d_2 - d_1}{d_1}$ is the matching score having d_1 and d_2 as the first and second best matches respectively, and $\hat{\mathbf{v}}$ is the correspondence vector obtained by descriptor matching at $\mathbf{x}$. In Eq. (2.31b), f_1 and f_2 represents sparse fields of feature vectors in image 1 and image 2

respectively. The main reason behind modeling these continuous functions is that the descriptor matching has important drawbacks: 1) since it is a discrete method so it does not provide sub-pixel accuracy and 2) the fixed spatial extent of rich descriptors is responsible for inaccuracies at motion discontinuities and in case of all non-translational motions. Hence, the auxiliary variable $\hat{\mathbf{v}}$ is used which allows to integrate discrete descriptor matching into a continuous approach in the form of soft constraints. In the functions, E_{match} and $E_{desc.}$, the descriptors are only available on a fixed spatial grid defined by the $\delta_i(\mathbf{x})$ function. At other points these terms do not contribute (*i.e.* their value is null).

Another similar approach is presented by Weinzaepfel et al. [Weinzaepfel et al., 2013] as the "DeepFlow" algorithm. In this approach, a quasi-dense correspondence is obtained by a new descriptor based matching algorithm. SIFT descriptors are used in this algorithm. Each obtained SIFT patch is split into 4 different quadrants which can independently move in the image pair (both target and source image) during matching. Good matches correspond to the local maxima in the response maps of corresponding image patches. A max-pooling algorithm is used to recover the path of response values that generated the maximum response retrieving dense correspondences from every matched patch. Thus, this descriptor based matching algorithm is designed upon a multi-stage architecture (depending upon the image size), interleaving convolutions to obtain response maps and max-pooling from these response maps. Finally, the obtained quasi-correspondences are embedded into the variational energy minimization framework for optical flow field estimation. However, it is to be noted that these large displacement optical flow algorithms are only robust for image pairs with well distributed textures. Such approaches are not optimized for complicate (e.g. medical) scenes with few and/or weakly contrasted textures. Other scenes with repetitive patterns in them, blur due to motion or image defocus will also largely affect the performance of these algorithms.

d) Data-term based on patch matching

Self similarity patterns have been successfully used for handling large motion, blur and illumination variations [Werlberger et al., 2010, Drulea and Nedevschi, 2013, Chen et al., 2013]. Nearest Neighbor Field (NNF) using PatchMatch [Barnes et al., 2009] for optical flow estimation followed by a motion segmentation step is used in [Chen et al., 2013]. Since NNF algorithms are not limited by the magnitude of displacement fields, they provide an efficient technique for handling large displacements. However, due to the non-convex formulation and integration of several steps like outlier rejection, multi-label graph-cut and fusion, this NNF based model [Chen et al., 2013] is computationally expensive.

A fast and parallelizable minimization problem using *zero-mean normalized cross-correlation (ZNCC)* as a matching cost (data-term) was formulated in [Drulea and Nedevschi, 2013]. A similar approach using truncated ZNCC was also formulated by Werlberger et al. [Werlberger et al., 2010]. It has been shown in [Drulea and Nedevschi, 2013] that the ZNCC based matching cost is invariant to illumination changes in contrast to brightness constancy assumption (BCA). This is because it penalizes severe deviations from the correct match more than SSD based approaches does (like BCA or GCA based data-terms). The ZNCC distance between blocks of two images can be formulated as *(the details for this generalization is presented in [Drulea and Nedevschi,*

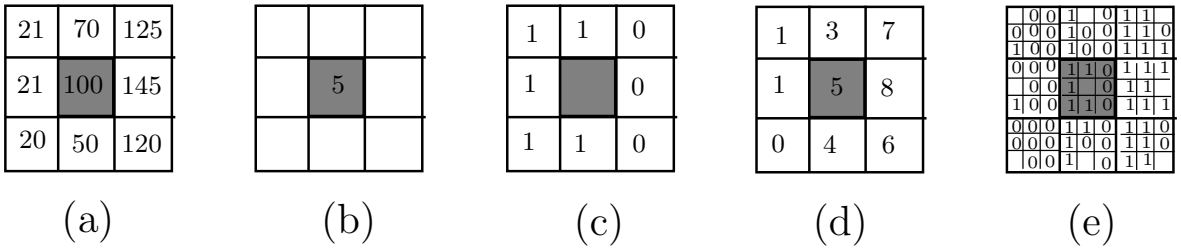


Figure 2.4: Illustration of intensity order transforms (b-e) in a 3×3 neighborhood patch $\mathcal{P}_{3\times3}$ around the pixel of interest marked in grey. a) Intensity of the original pixels in $\mathcal{P}_{3\times3}$, (b) Rank, (c) Census, (d) Complete rank and (e) Complete census.

2013]):

$$E_{ZNCC}(\mathcal{P}, \mathbf{v}) = \sum_{\mathcal{P}\in\Omega}\{1 - \frac{1}{\mid \mathcal{N}_{\mathbf{p}} \mid}\sum_{\mathbf{x}\in\mathcal{N}_{\mathbf{p}}} \frac{(I(\mathbf{x}+\mathbf{v}, t+1) - \mu(\mathbf{p}+\mathbf{v}_{\mathbf{p}}, t+1))\,(I(\mathbf{x}-\mu(\mathbf{p}, t)))}{\sigma(\mathbf{p}+\mathbf{v}_{\mathbf{p}}, t+1)\sigma(p, t)}\}, \tag{2.32}$$

where μ and σ are mean and variance respectively computed over the same neighborhood region $\mathcal{N}_{\mathbf{p}}$ at location $\mathbf{p} \in \mathcal{P}$ patch. The energy associated with truncated ZNCC [Werlberger et al., 2010] can be therefore written as:

$$E_{TZNCC}(\mathcal{P}, \mathbf{v}) := \min(1, E_{ZNCC}(\mathcal{P}, \mathbf{v})). \tag{2.33}$$

However, for large displacements in the scene, these methods largely depend on their window size for constituting ZNCC patches in images which are computationally expensive. A more faster but sparse approach (SimpleFlow) was presented by Tao et al. [Tao et al., 2012] which uses local cues for building a probabilistic representation of the motion flow. However, this model is unable to estimate accurate motion fields in images with repetitive patterns and in image pairs with very large motions.

Rank transform and census transform [Zabih and Woodfill, 1994] are the class of patch-based intensity order descriptors and are invariant to monotonically increasing grey-value rescalings. The relation of the pixel of interest $\mathbf{x}$ to the neighborhood pixels is encoded in a signature descriptor s. The *rank transform* (rank-T) holds the position of its grey value in the ranking of the pixels which is considered as the number of counts of the pixels in a given patch which is smaller than the pixel of interest. For example, in Fig. 2.4 (a), five neighborhood pixels (50, 20, 21? 21, 70) in a 3×3 patch are smaller than the intensity of the pixel of interest 100, so the signature image will have 5 at this location ($s_{RT} = 5$)as shown in Fig. 2.4(b). The rank-T will map each pixel in the image with its rank signature (scalar). In case of *census transform* (census-T), one bit information is stored for each pixel in the neighborhood. The pixel in the neighborhood which is smaller than the pixel of interest is assigned 1, otherwise it's bit is 0 forming a binary signature. For example, in Fig. 2.4 (c), the binary signature s_{CT} is $\{1, 1, 1, 1, 1, 0, 0, 0\}^T$.

In stead of assigning rank to a single pixel of interest, Demetz et al. [Demetz et al., 2013] proposed a *complete rank transform* (CRT) which is based on the signature obtained by assigning the rank to each element in the neighboring pixels of the patch $\mathcal{P}$ (refer to Fig. 2.4(d)). The

signature obtained by CRT for Fig. 2.4(a) is: $s_{CRT} = \{4, 0, 1, 1, 3, 7, 8, 6\}^T$. The *complete census transform* (CCT) can be seen as similar to CRT but with replacement of census signatures assigned for each element in the neighborhood patch as represented in Fig. 2.4(e). However, this will increase computational time because of very large bit length compared to both census-T and CRT descriptors.

Census transform has been successively used for computing optical flow in outdoor scenes with varying lighting conditions and large displacements. The Census cost is a piece-wise constant with its gradient either 0 or ∞ everywhere. The Census data erergy term at location $\mathbf{p}$ is defined as [Stein, 2004, Hafner et al., 2013]:

$$E_{Census}(\mathbf{p}, \mathbf{v}) = \sum_{\mathbf{p} \in \Omega} \mathbb{1}_{Ce(I(\mathbf{x}+\mathbf{v}),\mathbf{p}+\mathbf{v},t+1) \neq Ce(I(\mathbf{x}),\mathbf{p},t)}, \tag{2.34}$$

where $Ce(I(\mathbf{x}), \mathbf{p}, t) = sgn(I(\mathbf{p}) - I(\mathbf{x})).\mathbb{1}_{|I(\mathbf{p})-I(\mathbf{x})|>\epsilon}$ with $\mathbb{1}$ as the indicator function. Since the linearization of this Census transform is not easy, integrating it into the variational approach is a non-trivial task. However, using "soft" L^1 norm Vogel et al. [Vogel et al., 2013] reformulated a convex approximation as sum of centralized absolute differences (CSAD) and integrated into variational approach. CRT signatures were used as the data-cost modeling in [Demetz et al., 2013] with the assumption that such descriptors coincide in the corresponding pixels in consecutive image pairs. The energy minimization equation was modeled as TV-L^1 model and solved using Euler-Lagrange method in variational calculus. These methods have gain popularity in recent years because of their invariance to illumination changes.

Ranftl et al. [Ranftl et al., 2014] proposed a scale-invariant census descriptor by sampling the radial stencils with different radii. A second-order total generalized regularization (TGV), orginally proposed by Bredies et al. [Bredies et al., 2010], was used along with non-local weights similar to [Werlberger et al., 2010, Drulea and Nedevschi, 2013]. It was shown that the TGV regularizer gives more accurate results than a TV-regularization. TGV regularization was also used by Demetz et al. [Demetz et al., 2015] confirming the increased accuracy of their previous CRT method [Demetz et al., 2013].

2.3.2 Regularizer

In order to tackle the aperture problem in optical flow, Horn and Schunck [Horn and Schunck, 1981] incorporated a smoothness constraint in the energy minimization scheme to penalize deviations of data-term from piecewise smoothness. This prior information not only helps to compute unique solution locally but also influences the quality of flow field estimation. Over past few decades, many works have been done in modeling an optimal regularizer for obtaining robust and accurate optical flow fields. These works are detailed in [Nagel and Enkelmann, 1986, Weickert and Schnörr, 2001, Weickert et al., 2006]. Weickert et al. [Weickert et al., 2006] divided the regularization into flow driven and image driven regularization models. Depending upon the type of their impact, they were further classified into isotropic and anisotropic (or non-isotropic) regularizers. However, spatial regularizer was first introduced in optical flow domain by Horn and Schunck [Horn and Schunck, 1981]. This regularizer is modeled as a quadratic penalization of the gradients of the flow field as proposed by Tikhonov [Tikhonov, 1963]:

$$\mathcal{R}(\mathbf{v}) = \int_{\Omega} | \nabla \mathbf{v} |^2 \, d\mathbf{x}. \tag{2.35}$$

The homogeneous regularization defined by Eq. (2.35) oversmooths the flow field discontinuities resulting in blurry optical flow field. It is obvious that flow fields should be obtained such that

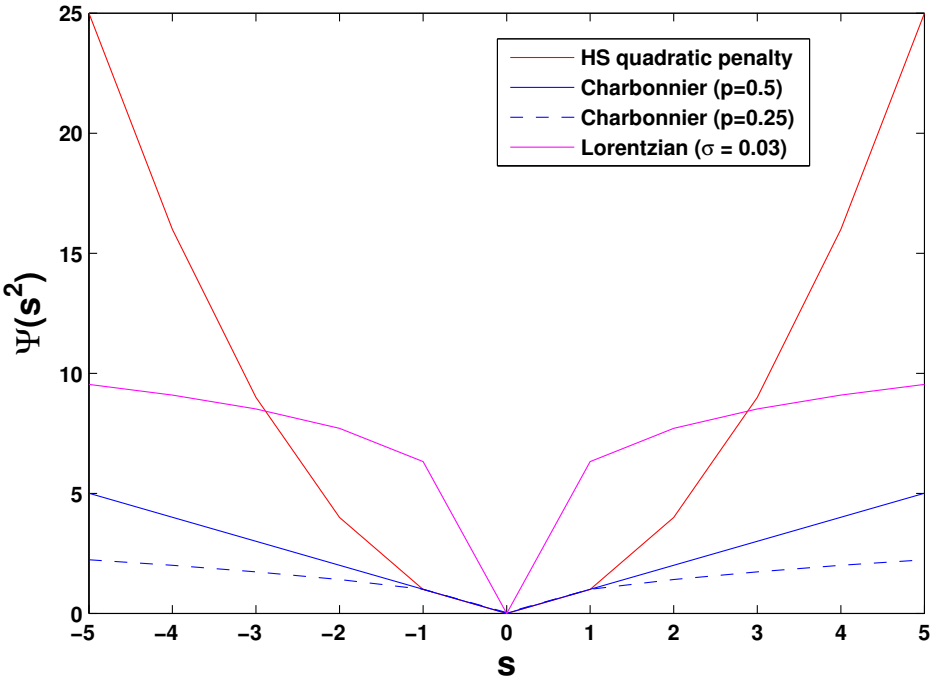


Figure 2.5: Illustration of bahaviour of penalty functions for the spatial terms. In red: Horn and Schunk quadratic term (over-smoothing effect). In blue: Charbonnier convex forulation with ($p = 0.5$). In blue dashed: Charbonnier non-convex penalty with $p < 0.5$. In magenta: Lorenzian function with $\sigma = 0.03$.

motion discontinuities across the edge boundaries are preserved. Various works have been done to model such discontinuity-preserving regularizers.

One way of modeling robust regularizers is by replacing the quadratic factor in the regularizer with non-quadratic approaches [Black and Anandan, 1996, Brox et al., 2004, Bruhn et al., 2005, Zach et al., 2007, Sun et al., 2010]. The non-quadratic models includes: 1) the Charbonnier penalty function, $\Psi(\mathbf{s}^2) = (\mathbf{s}^2 + \epsilon^2)^p$ [Bruhn et al., 2005, Sun et al., 2010] and 2) the non-convex penalty using the Lorentzian function, $\Psi(\mathbf{s}^2) = \log(1 + 0.5 * \frac{\mathbf{s}^2}{\sigma^2})$ [Black and Anandan, 1996], with σ as the scalar constant to be fixed. Charbonnier penalty is a convex and a differentiable variant of the L^1 norm when $p = 0.5$. This convex modeling is the most popular model for robust formulations. Eventually, when $p < 0.5$ then the penalty function becomes non-convex. The behavior of these penalty functions is shown in Fig. 2.5. Sun et al. [Sun et al., 2010, Sun et al., 2014] showed that the convex formulation of Charbonnier penalty performed better than its non-convex form and more robustly than Lorenzian penalty. One particular reason for the popularity of convex penalty functions is that they are easier to optimize and a global optimal minimum solution is obtained without getting trapped in local minima as in the case of non-convex functions. Thus, we can write a robust regularizer in total variational framework using L^1 norm as:

$$\mathcal{R}(\mathbf{v}) = \int_{\Omega} \sqrt{|\ \nabla \mathbf{v}\ |^2 + \epsilon^2}\ \ d\mathbf{x} \quad \text{with} \quad \epsilon \approx 0 + . \tag{2.36}$$

In this section, we will present a brief taxonomy of regularizers which are mostly based on the work of Weickert et al. [Weickert et al., 2006] and Sun et al. [Sun et al., 2014]. The smoothness term has to be penalized for obtaining robust optical flow estimation with preserved

motion discontinuities. However, this process is highly experimental and depends upon the type of image data. For instance, in case of image pairs with low texture it is required that the regularization propagates the flow field from the regions with information to more homogeneous regions (*i.e.* without information). In such case, image-driven regularizers are needed and the propagation of the displacement vectors are usually linear as they do not depend on the flow field but rather information content of images. However, in images with many textured regions, image-driven regularizers over-segment the flow field. In this case, flow-driven regularizers are effective for penalizing the flow field gradients directly leading to non-linear diffusibility of flow vectors. The regularization taxonomy is summarized in the block diagram in Fig. 2.6 and discussed thoroughly hereafter.

a) Image-driven regularizer

In order to reduce the over-smoothing effect due to the Horn-Schunck [Horn and Schunck, 1981] regularizer, image gradient and its variant was used to penalize the regularizer [Alvarez et al., 1999, Alvarez et al., 2000, Nagel and Enkelmann, 1986]. The idea behind exploiting image gradient based information is because flow discontinuities are a sub-set of image edges (*i.e.* often appear at edges of objects or textures). However, depending upon the smoothing behaviour due to incorporated image gradients, these regularizers can be further classified into:

1. **Isotropic image-driven regularization:** Such an approach [Alvarez et al., 1999] uses a direct penalization of the classical regularizer along image edges. However, since the directional information of image edges are not used, these type of regularizer have an isotropic behavior. It can be formulated as :

$$\mathcal{R}_1(\mathbf{v}) = \int_\Omega g(\mid \nabla I \mid^2) \mid \nabla \mathbf{v} \mid^2 \quad d\mathbf{x}, \tag{2.37}$$

   where g is a strictly positive weighting funtion depending upon image gradient ∇I. In Eq. 2.37, $g(.)$ is such that its value is high for non-edge pixels giving more weight to $\mid \nabla \mathbf{v} \mid^2$ which leads to smoother flow field while the weight function $g(.)$ is close to 0 at the edge pixels of image I.

2. **Anisotropic image-driven regularization:** Nagel et al. [Nagel and Enkelmann, 1986] used an anisotropic image-driven penalizer which takes into account the orientation of image gradients so that the classical smoothness-term is penalized only along the image edges while keeping the smoothness in homogeneous regions. This penalization was modeled as a diffusion tensor $D(\nabla I)$, as in technique for edge enhancement and denoising. Such approaches can be formulated as [Nagel and Enkelmann, 1986, Alvarez et al., 2000]:

$$\mathcal{R}_2(\mathbf{v}) = \int_\Omega \left((\nabla \mathbf{v})^T D(\nabla I)(\nabla \mathbf{v})\right) \quad d\mathbf{x}, \tag{2.38}$$

   with D as the projection matrix perpendicular to ∇I (refer to [Weickert et al., 2006] for details). This matrix consists of directional information in the form of Eigenvalues λ_+ and λ_- (of corresponding eigenvectors, e_+ and e_-). Depending upon the magnitude of these Eigenvalues, the effect of the regularizer in image region/s are either isotropic or anisotropic. Anisotropic smoothing is obtained along the edges (i.e. when λ_+ and λ_- tend towards 0 and 1 respectively), while an isotropic smoothing behavior is performed inside homogeneous region/s.

Figure 2.6: Classification of regularizers used in various optical flow models.

b) Flow-driven regularizer

Image-driven regularizers can lead to a region partition, even if these regions have the same (constant) flow fields (over-segmentation) *i.e.* flow discontinuities are not preserved at appropriate pixels. The idea of flow-driven methods is to penalize the smoothness term using the information of the flow field which is basically the flow gradients itself. Similar to image-driven methods, they can be classified into:

1. **Isotropic flow-driven regularization:** The smoothing is reduced only at the motion boundaries. A non-quadratic formulation was adapted to attenuate over-smoothing along the motion boundaries. A well-established approach is presented by Schnörr et al. [Schnörr and Sprengel, 1994].

$$\mathcal{R}_3(\mathbf{v}) = \int_\Omega \Psi\left(\mid \nabla \mathbf{v} \mid^2\right) \quad d\mathbf{x}, \tag{2.39}$$

   where $\Psi(s^2)$ represents a non-quadratic approach which is differentiable and an increasing function that is convex in s (also refer Fig. 2.5 and explanations above).

2. **Anisotropic flow-driven regularization:** It is modeled as a structure tensor $D(\nabla \mathbf{v})$ depending upon the eigenvalues and eigen vectors of the flow field. $D(\nabla \mathbf{v})$ is a symmetric, positive, semi-definite and strictly convex 2×2 matrix with non-negative singular values $(\sigma_i, ..., \sigma_n)$. Since, the singular values hold the contrast information of the flow field, it can be used to obtain relevant flow discontinuity information. Thus, the regularizer with anisotropic behaviour with flow-driven penalization can be written as:

$$\mathcal{R}_4(\mathbf{v}) = \int_\Omega \mathrm{tr}\left(\Psi\left((\nabla \mathbf{v})(\nabla \mathbf{v})^T\right)\right) \quad d\mathbf{x}, \tag{2.40}$$

   where $D(\nabla \mathbf{v}) = (\nabla \mathbf{v})(\nabla \mathbf{v})^T$ and $\mathrm{tr}(D(\nabla \mathbf{v})) := \sum_i \sigma_i$. This penalization behaves non-linearly performing isotropic regularization along homogeneous regions and anisotropically along the regions with flow field discontinuities [Weickert et al., 2006]. It takes into account the flow field regularization along flow discontinuities without over-segmenting the flow regions unlike the image-driven methods.

c) Weighted non-local regularization

Several authors [Werlberger et al., 2010, Drulea and Nedevschi, 2013, Ranftl et al., 2014] proposed to use a non-local regularizer in a total variational L^1 framework for robustly preserving the motion discontinuities. The idea behind such regularization is the assumption that the pixels belonging to the same object have almost the same flow; *i.e.* the flow field is constant in same object region. In order to establish this penalization, bilateral filters were used to provide the weights to the flow field. These filters measure the belonging of pixels $\mathbf{x}$ and their neighborhood pixels $\hat{\mathbf{x}}$ exposing how likely $\mathbf{x}$ belongs to the same object. Such regularization can be modeled as:

$$\mathcal{R}_5(\mathbf{v}) = \sum_{\mathbf{x} \in \Omega} \sum_{\hat{\mathbf{x}} \in \mathcal{N}_\mathbf{x}} w_\mathbf{x}^{\hat{\mathbf{x}}} \parallel \mathbf{v}_\mathbf{x} - \mathbf{v}_{\hat{\mathbf{x}}} \parallel_1, \tag{2.41}$$

where $w_\mathbf{x}^{\hat{\mathbf{x}}}$ is the correlation entity based on (i) the spatial distance $\mid \mathbf{x} - \hat{\mathbf{x}} \mid^2$ and (ii) the color distance in CIE-Lab color space $\mid I(\mathbf{x}) - I(\hat{\mathbf{x}}) \mid^2$:

$$w_\mathbf{x}^{\hat{\mathbf{x}}} = e^{-|\mathbf{x}-\hat{\mathbf{x}}|^2/2\sigma_1^2} . e^{-|I(\mathbf{x})-I(\hat{\mathbf{x}})|^2/2\sigma_2^2}. \tag{2.42}$$

In Eq. 4.8, σ_1, σ_2 are normalization factors and $I(\mathbf{x})$ is the color vector in CIE Lab colorspace. Sun et al. [Sun et al., 2010, Sun et al., 2014] used a similar approach. The classic-NL algorithm described in [Sun et al., 2014] uses such a non-local median filtering, along the edges found with a Sobel edge detector, for obtaining flow boundary regions more accurately. The optical flow field obtained by using this regularization technique is stated to be almost the most accurate among the techniques in the literature of variational approaches [Werlberger et al., 2010, Ranftl et al., 2014, Xu et al., 2012].

2.4 Mathematical optimization

A general mathematical optimization problem can be written as:

$$\min F_0(v) \qquad \text{subject to } F_i(v) \leq b_i,\ i = 1, ..., k, \tag{2.43}$$

where $v = (v_1, ...v_n)$ is the optimization variable of the problem, F_0 and F_i are the objective function and the inequality constraint function respectively and $b_1, ..., b_m$ are the constants. If both the objective function $F_0 : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $F_i : \mathbb{R}^n \to \mathbb{R}$ in Eq. (2.43) have positive curvature then the function F_0 is convex. The local solution of this objective function is the global optimal solution. Least square problems are convex and they can be solved analytically. But, in general, no analytical solution can find a global minimal solution of minimization problems so only local solution is possible. Unlike, convex optimization algorithms are reliable and efficient techniques which guarantee the global optimal solution. However, they are difficult to identify and are computationally more expensive than the analytical methods.

In this thesis, we are interested in finding the global optimal solution. The flexibility of transforming a non-convex problem to a convex problem is a strong motivation of choosing a convex optimization scheme for determining the optical flow field. We will therefore restrict ourselves in this thesis in solving convex problems. In this section essential definitions are introduced to understand the convex problem formulation and convex optimization approach. At the end of this section, we present a total variational approach for optical flow estimation using a convex optimization approach modeling robust data-term and regularizer.

2.4.1 Prerequisites

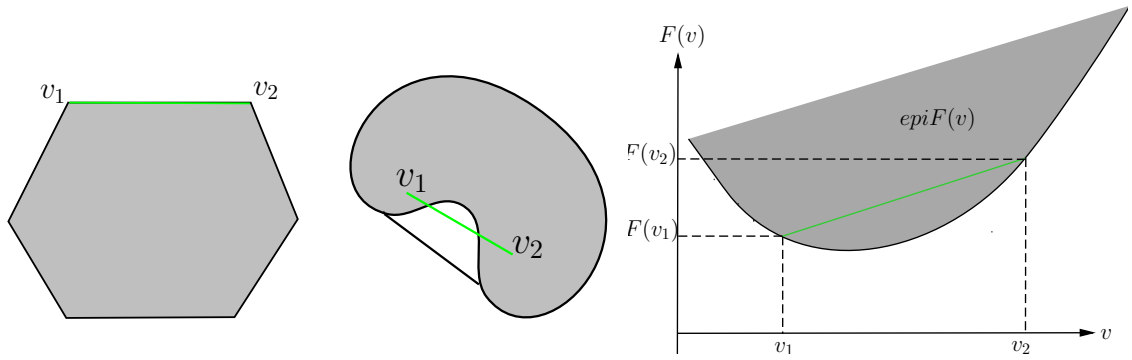
Convex set

A set C is convex if it contain line segments between any two points in that set. If v_1, $v_2 \in C$ are any two points, then all the other points on this line segment is a convex set if they entirely lie in C. Mathematically, we can write a convex set as the linear combination of two weighted points such that:

$$\theta v_1 + (1 - \theta)v_2 \in C, \quad \forall \theta \in (0, 1). \tag{2.44}$$

In Fig. 2.7a, all the points on the green line connecting v_1, $v_2 \in C$ is a convex set while in Fig. 2.7b some points on this green line are not contained in C so this set of points do not give the convex set. The set of all convex combinations of points $v_1, ..., v_k \in C$ span the convex hull of a set C, *i.e.*,

$$\left\{ \sum_{i=1}^{k} \theta_i v_i \right\}, \quad \text{with} \quad \theta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{k} \theta_i = 1. \tag{2.45}$$



(a) Convex set (b) Non-convex set (c) Convex function

Figure 2.7: Illustration of convexity concept. (a) A convex set. (b) A non-convex (concave) set; (c) A convex function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ is represented by a curve and the line in green represents the convex set of points between the points v_1 and v_2. Linear combination of points which is present on this line gives the convex set in the function domain $F(v_1)$and $F(v_2)$ representing the curve.

Euclidean balls and ellipsoids

An Euclidean ball with center v_c and radius r is a convex set if $\| v_1 - v_c \|_2 \leq r$, $\| v_2 - v_c \|_2 \leq r$ and $0 \leq \theta \leq 1$, then,

$$\| \theta v_1 + (1-\theta)v_2 - v_c \|_2 \leq r. \tag{2.46}$$

For a convex set $C = \{v \mid \| v \|_2 \leq 1\}$, the projection on Euclidean ball is given as:

$$P_C(v) = \frac{v}{\| v \|_2} \quad if \| v \|_2 > 0. \tag{2.47}$$

Ellipsoids also belong to the family of convex sets and is represented as:

$$\mathbb{F} = \{v \mid (v - v_c)^T P^{-1} (v - v_c) \leq 1\}, \tag{2.48}$$

where $P = P^T > 0$, i.e., P is a symmetric and positive definite matrix. The vector $v_c \in \mathbb{R}^n$ is the *center* of the ellipsoid. The matrix P determines how far the ellipsoid extends in every direction from v_c while the lengths of semi-axes are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P.

Convex functions

$F(v) : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function if the domain of F contains a convex set such that

$$F(\theta v_1 + (1-\theta)v_2) \leq \theta F(v_1) + (1-\theta)F(v_2) \quad \forall v_1, v_2 \in C, \quad \forall 0 \leq \theta \leq 1. \tag{2.49}$$

Fig. 2.7c represents a convex function $F(v)$ with the points v_1 and v_2 on this function satisfying Eq. (2.49). The left hand side of this equation represents the points on the curve and the right hand side indicates the points on the green line contained in the epigraph of $F(v)$ (i.e. the area above the function graph). $F(v)$ is strictly convex if it lies inside the epigraph but not on the boundaries. Thus, the convex set consists of points formed with $0 < \theta < 1$ and $v_1 \neq v_2$. Some examples of convex functions are norms and affine functions on $\mathbb{R}^n$.

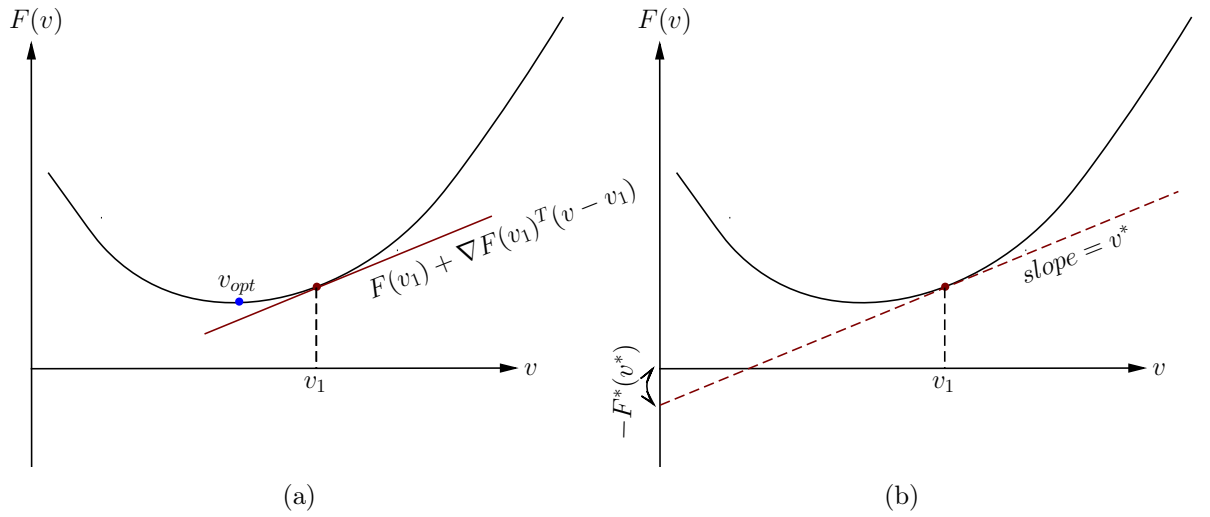


Figure 2.8: a) $F(v)$ is convex and differentiable so $F(v) \geq F(v_1) + \nabla F(v_1)^T(v - v_1)$. b) A function $F : \mathbb{R}^n \to \mathbb{R}$, and a value $v^* \in \mathbb{R}$ such that it represents the slope of the function F and the conjugate function $F^*(v^*)$ is the maximum gap.

Norms and affine functions

All norms are convex. The p-norm of a n-dimensional vector is represented as:

$$\| v \|_p = \left(\sum_{i=1}^{n} | v_i |^p \right)^{\frac{1}{p}} \text{for} \quad p \geq 1. \tag{2.50}$$

An affine function $F :\to \mathbb{R}^n$ defined as $F(v) = Ax + b$ with A being some matrix in $\mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ is some translation. An affine function is both convex and concave and its set C can thus be expressed as:

$$C = V + b = v + b \mid v \in V, \tag{2.51}$$

where V is a subspace associated with the affine set C (linear function, Ax) and b is an offset contained in C.

Convexity conditions

First-order convexity condition: A function F is differentiable if domain of F (**dom** F) is open and the gradient exists for each point $v \in$ **dom** F given as:

$$\nabla F(v) = \left(\frac{\partial F(v)}{\partial v_1}, \frac{\partial F(v)}{\partial v_2}, ..., \frac{\partial F(v)}{\partial v_n} \right). \tag{2.52}$$

That is, F is convex if and only if **dom** F is convex and $\forall\ v, v_1 \in$ **dom** F and given by:

$$F(v_1) + \nabla F(v_1)^T(v - v_1) \leq F(v). \tag{2.53}$$

Eq. (2.53) is the first-order Taylor approximation of F near v_1 which gives a global underestimator of the function as illustrated in Fig. 2.8a. Conversely, if first-order Taylor approximation of a function is always a global underestimator then the function is convex.

The inequality in Eq. (2.53) shows that from local information about a convex function (i.e. its value and derivative at a point) we can derive global information (i.e. global underestimator of it). For example, the inequality in Eq. (2.53) shows that if $\nabla F(v_1) = 0$, then for all $v \in \mathbf{dom}\ F$, $F(v) \geq F(v_1)$, i.e., v_1 is the global minimizer of the function $F(v)$. In Fig. 2.8a, a global optimal solution is obtained at $v_1 = v_{opt}$ represented by a blue dot. At this point $\nabla F(v) = 0$.

Second-order convexity condition: F is convex if and only if **dom** F is convex and its Hessian is positive semi-definite, *i.e.*, for all $v_1 \in \mathbf{dom}\ F$:

$$\nabla^2 F(v_1) \geq 0. \tag{2.54}$$

Geometrically, it means that the requirement of the graph of the function is a positive (upward) curvature of v_1 since the derivative is non-decreasing as in Eq. (2.54). Strong convexity holds for $\nabla^2 F(v_1) > 0$ but the converse may not be true.

Operations that preserves convexity

Following operations preserve the convexity of a function F:

- *Non-negative weighted sum.* If $F_1(v), ..., F_k(v)$ are convex functions and $\lambda_1, ..., \lambda_k$ be the non-negative weights such that $F = \sum_{i=1}^{k} \lambda_i F_i(v)$, then the function F is also convex.
- *Composition with the affine function.* An affine mapping of a convex function $F : \mathbb{R}^n \to \mathbb{R}$ of the form $G(v) = F(Av + b)$, with a matrix $A \in \mathbb{R}^{n \times m}$ and a vector $b \in \mathbb{R}^n$, then $G(v)$ is also convex.
- *Pointwise maximum and supremum.* Computing the pointwise maximum of a set of convex functions $F_1(v), ..., F_k(v)$ with $F(v) = \max\{F_1(v), ..., F_k(v)\}$, results in a convex function. The pointwise supremum over an infinite set of convex functions is defined as function $g(v) = \sup_{v_1 \in \mathcal{A}} F(v, v_1)$. g is a convex function if $F(v, v_1)$ is convex in v.
- *Minimization.* If a minimum exists, the convexity ensures that the minimum is global. The set of global minima again form a convex set.
- *Perspective.* If $F : \mathbb{R}^n \to \mathbb{R}$, then the perspective of F is the function $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by $g(x, t) = tF(v/t)$, with domain **dom** $g = \{(x, t) \mid x/t \in \mathbf{dom}\ F, t > 0)\}$. If F is a convex function then its perspective function g is also convex.

The conjugate function

Let us consider functions $F : \mathbb{R}^n \to \mathbb{R}$ and $F^* : \mathbb{R}^n \to \mathbb{R}$ defined as:

$$F^*(v^*) = \sup_{v \in domF} \{\langle v^*, v \rangle - F(v)\}, \tag{2.55}$$

where F^* is the conjugate of function F. The auxiliary variable v^* also called *dual variable* represents the slope of the original function F as illustrated in Fig. 2.8b. Since F^* depends upon the slope of the original function F and it is also the pointwise supremum of a family of convex function of v^* so F^* is always convex whatever is the shape of F.

Properties: We mention here, few important properties of conjugate functions that will be required in other sections of this thesis:

- **Fenchel's inequality.** According to the Fenchel-Young inequality:

$$F(v) + F^*(v^*) \geq \langle v, v^* \rangle, \quad \forall v, v^*. \tag{2.56}$$

- **Conjugate of the conjugate.** The conjugate of a conjugate function is the original function, *i.e.* $F^{**}(v^{**}) = F(v)$, also known as bi-conjugate. It forms a convex envelop of its original function F and is both convex by definition and independent of the shape of F (*i.e.* lower semi-continuous). However, if F is a closed curve then the biconjugate is equal to the original function F.

2.4.2 Convex optimization

Convex optimization has been very popular in recent years in computer vision. It has been widely used in solving variational models that are convex or their approximates. It is found that this optimization tool is reliable and efficient in estimating the global minimum point of a wide range of complex continuous energy functions. The beauty of convex optimization technique is that it guarantees to find an optimally global solution to convex models.

Let us assume that we want to minimize a convex funtion $E(v)$ ($E : \mathbb{R}^n \rightarrow R$) with E being $\mathcal{C}^1$ differentiable and its gradient ∇E is Lipschitz continous. Let $v = (v_1, v_2, ..., v_n)$ be the search space and v_{opt} be the optimum point to be found then according to the first-order convexity condition $\nabla E(v_{opt}) = 0$ and $E(v_{opt}) \leq E(v)$, $\forall\ v$ (refer Section 2.4.1). However, to retrieve the optimal minimal solution, an iterative procedure is done such that the global minimizer (vanishing gradients) suffice the condition for obtaining the optimal point. Thus, an iterative search in v is done. One most widely used technique is to use gradient descent methods. The update direction for the gradient descent method can be defined as:

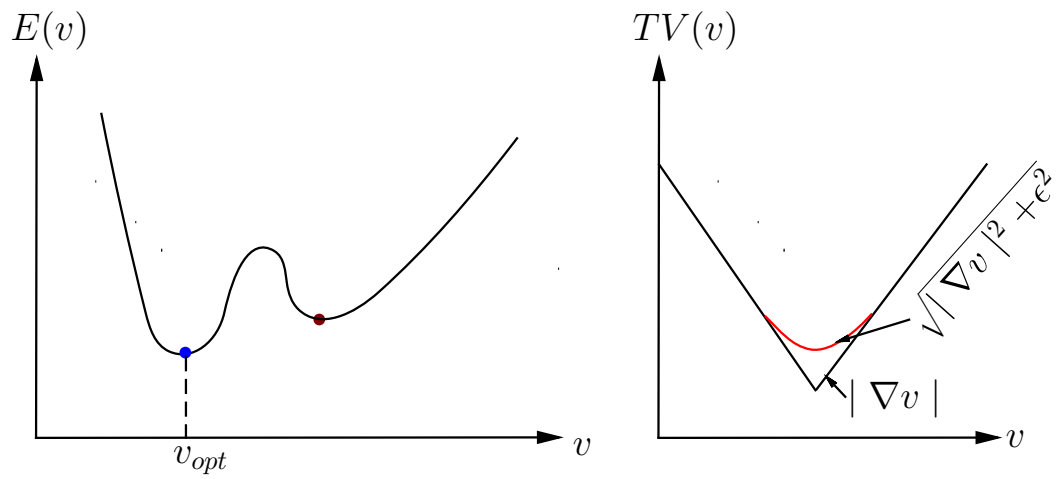
$$v^{n+1} = v^n - \tau \nabla E(v^n), \tag{2.57}$$

where τ is the step size that must be chosen to maintain the balance between convergence speed and algorithm robustness. An initial guess v^0 in Eq. (2.57) is used to begin the iteration. The algorithm follows N_{max} iterations if convergence criteria ensuring the optimal solution is not reached.

Non-differentiable TV-regularizer

If the function E is not convex then the gradient descent may lead only to a local minimum (refer to Fig. 2.9a where dark brown dot represents the local solution while dark blue dot on the same function is the global optimal minimum solution). Variational methods were most popular in 1980's but their energies are usually non convex functions. A pioneering work was done in the field of variational optical flow by Horn and Schunck [Horn and Schunck, 1981] by modeling a non-convex function in the data-term and a convex in the smoothness term. Another well-known model that has been widely used in recent years is the *total variational* (TV) model introduced by Rudin et al. [Rudin et al., 1992], also known as ROF-model. The total variational regularizer classically used for image denoising is defined as:

$$TV(\mathbf{v}) = \int_\Omega \mid \nabla \mathbf{v} \mid d\mathbf{x}. \tag{2.58}$$



(a) (b)

Figure 2.9: Limitations of gradient descent techniques for non-convex and convex non-differentiable functions. (a) Gradient descent giving local minimal solution for non-convex approach. b) Non-differentiability in convex functions (usually all norms) being modeled as differentiable function by adding a small constant ϵ, $TV(v)$ being a $1D$ representation of the function.

There are two major short-comings for this equation to be solved:

1) gradient methods only work if $\mathbf{v}$ is differentiable which is not the case in Eq. (2.58) and

2) the TV energy itself in Eq. (2.58) is not differentiable at $v \to 0$.

In spite of these limitations, the non-quadratic formulation of the TV-regularizers reduces the effect of over-smoothing and the flow field edges (flow discontinuities) are preserved. The ROF-model has thus become very popular in many domains like image segmentation, image denoising, image deblurring and optical flow estimation. The above mentioned limitations have been solved by authors in various ways. Brox et al. [Brox et al., 2004] replaced the $\nabla\mathbf{v}$ of the TV-regularizer as:

$$TV(\epsilon\mathbf{v}) = \sqrt{|\ \nabla\mathbf{v}\ |^2 + \epsilon^2}, \tag{2.59}$$

where $\epsilon = 0.001$. The TV-regularizer in Eq. (2.59) is now differentiable and hence a gradient descent approach can be used to obtain the global minimal solution. Additionally, $TV(\epsilon\mathbf{v})$ can also avoid certain jumps in the functions. The new function $TV(v)$ in 1D is shown in red in Fig. 2.9b. Even though this approach gives an optimally acceptable global solution, a more efficient way is to introduce a dual variable.

Let $\mathbf{q} : \Omega \to \mathbb{R}^2$ be a dual variable, then,

$$|\ \nabla\mathbf{v}\ | = \sup_{|\mathbf{q}|\leq 1} \mathbf{q}\nabla\mathbf{v}. \tag{2.60}$$

In Eq. (2.60), the vector that maximizes the scalar product $\mathbf{q}\nabla\mathbf{v}$ is attained at $\mathbf{q}$ if $\nabla\mathbf{v} \neq 0$ and the vector $\mathbf{q}$ is bounded by a unit ball such that the unit vector in gradient direction maximizes the function (i.e. $\mathbf{q} = \frac{\nabla\mathbf{v}}{|\nabla\mathbf{v}|}$). This allows for the generalization of both discontinuous functions

and non-differentiable functions. Thus, the TV-regularizer with a dual variable can be written as:

$$TV(\mathbf{v}) := \sup_{|\mathbf{q}|\leq 1} \mathbf{q}\ \nabla\mathbf{v}\ d\mathbf{x}, \tag{2.61a}$$

$$TV(\mathbf{v}) := \sup_{|\mathbf{q}|\leq 1} \mathbf{v}\ \text{div}\ \mathbf{q}\ d\mathbf{x}. \tag{2.61b}$$

Eq. (2.61a) is differentiable in $\mathbf{v}$ and Eq. (2.61b) is differentiable in dual variable $\mathbf{q}$ with $\mathbf{q}$ constrained to the unit disc at every point $\mathbf{x} \in \Omega$.

Quadratic relaxation

Recent optical flow methods have shown increased robustness and accuracy with a L^1-norm formulation of the classical Horn-Schunck model. The general total variational optical flow formulation using L^1-norm [Zach et al., 2007, Pock et al., 2007, Wedel et al., 2009b, Wedel et al., 2009a] can be written as the following energy minimization problem:

$$\min_{\mathbf{v}} \int_{\Omega} \{| \underbrace{I_2(\mathbf{x}+\mathbf{v}) - I_1(\mathbf{x})}_{non-convex} | + \underbrace{| \nabla\mathbf{v} |}_{convex}\} d\mathbf{x}. \tag{2.62}$$

Eq. (2.62) represents a classical TV-L^1 approach which has a non-convex data-term because of the non-linearity in $I_2(\mathbf{x}+\mathbf{v})$ and a regularizer which is convex. Wedel et al. [Wedel et al., 2009a] used a quadratic relaxation term to decouple the data-term and the regularizer. Following such decoupling, we can write the minimization problem as:

$$\min_{\mathbf{v},\mathbf{U}} \int_{\Omega} \{| I_2(\mathbf{x}+\mathbf{v}) - I_1(\mathbf{x}) | + | \nabla\mathbf{U} | + \frac{1}{2\theta}(\mathbf{U}-\mathbf{v})^2\} d\mathbf{x}. \tag{2.63}$$

Now, according to Eq. (2.63), the smoothness term is no more in a function of $\mathbf{v}$ and the minimization of $\mathbf{U}$ is similar to the convex ROF-model (see the last two terms in the sum of Eq. (2.63)). Even though the data-term is non-convex it can be solved pointwise at fixed $\mathbf{U}$. It thus becomes a joint minimization problem that can be solved globally for both $\mathbf{U}$ and $\mathbf{v}$.

Primal dual approach

Referring to the work of Chambolle and Pock [Chambolle and Pock, 2011], this section summarizes important aspects of the primal-dual optimization approaches and their relevant solutions. The general first-order nonlinear primal problem can be formulated as:

$$\min_{v} F(Kv) + G(v), \tag{2.64}$$

with $F : \mathrm{Y} \to [0, +\infty)$ being a convex function, $G : \mathrm{Y} \to [0, +\infty)$ being another convex lower semi-continuous function and K is the linear operator. The primal-dual formulation of Eq. (2.4.2) can be written in the form of a generic saddle point problem:

$$\min_{v\in \mathrm{X}} \max_{q\in \mathrm{Y}} \langle Kv, q\rangle + G(v) - F^*(q), \tag{2.65}$$

where q is the dual variable, F^* is the convex conjugate of F. The solution of this objective function is thus achieved by the subsequent minimization in the primal variable v (gradient

descent) and the maximization in the dual variable q (gradient accent). The idea of primal-dual optimization is to minimize the primal-dual gap such that the a saddle-point (v_{opt} for the primal solution and q_{opt} for the dual variable) is found. The saddle-point is contained in search points (or search-space in 2D grid) v and q. The updates for $n \geq 0$ is given by the set of equations:

$$\begin{aligned} \text{dual-update:} \quad & q^{n+1} = (I + \sigma \partial F^*)^{-1}(q^n + \sigma K \bar{v^n}) \\ \text{primal-update:} \quad & v^{n+1} = (I + \tau \partial G^*)^{-1}(v^n - \tau K^* q^{n+1}) \\ \text{extrapolation:} \quad & \bar{v}^{n+1} = v^{n+1} + \theta(v^{n+1} - v^n) \end{aligned} \tag{2.66}$$

In Eq. (2.66), $\tau > 0$ and $\sigma > 0$ are the step sizes for the primal and dual updates respectively. The initialization is done with $\bar{v}^0 = v^0$ and $\theta \in (0, 1)$, typically chosen to be 1.

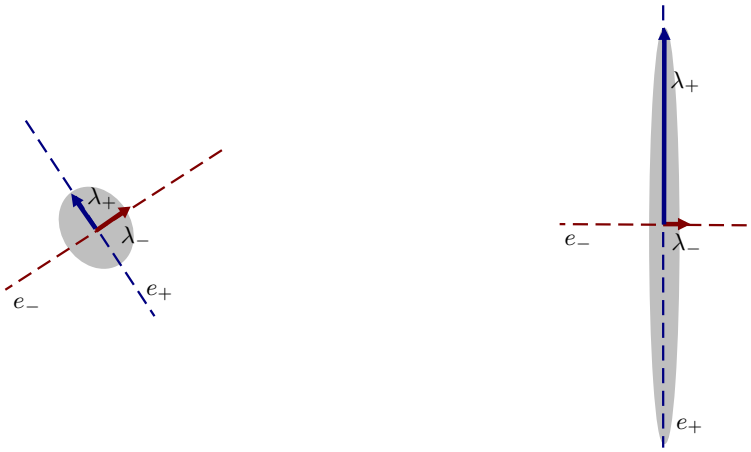
2.5 TV-L^1 optical flow: Background and first contribution

The majority of dense optical flow algorithms are originally motivated from Horn and Schunck model [Horn and Schunck, 1981]. They include a data-fidelity term with some constancy assumption(s) and a regularizer governing the flow variation across the image. However, the Horn-Schunck model does not allow for discontinuities in the displacement field due to quadratic regularizer and cannot robustly handle outliers in the data-term. Thus, most of the recent methods use L^1-norm in the total variation energy framework [Aubert et al., 1999, Zach et al., 2007, Wedel et al., 2009b, Brox et al., 2004]. In this section, a brief background on some of these methods is first given, then the first contribution of this thesis is presented as robust energy modeling in TV-L^1 framework and its minimization.

Although the method presented here suits for various scene types, it was particularly conceived for endoscopic bladder scenes [Ali et al., 2014] with textures corresponding mostly to elongated and thin structures (such as blood vessels). One of the aim of this section is to definitively validate the fact that the dense optical flow technique is an appropriate basis for dense correspondence establishment between bladder images characterized by texture variabilities and illumination changes.

2.5.1 Robust energy model (RFLOW)

Using classical data-term $\mathcal{D}_1$ based on BCA is often not robust to illumination changes or in case of image pairs with shadows. A structure-texture decomposition of the input images was incorporated by Wedel et al. [Wedel et al., 2009b] to improve the accuracy of TV-L^1 approach for small illumination changes occurred due to shadows. Such decomposition was done using the ROF model proposed by Rudin et al. [Rudin et al., 1992] classically developed for image denoising. Werlberger et al. [Werlberger et al., 2009] suggested to replace isotropic TV-regularization with an image driven anisotropic Huber regularization. A robust data term was formulated using both the brightness constancy and a gradient constancy assumption in [Brox et al., 2004]. It was shown by Papenberg et al. [Papenberg et al., 2006] that using complementary constancy assumptions makes the data-term robust. This section shows the importance of structure estimates when formulating the data-term. To do so, it is illustrated how classical data-terms can be completed and used in the efficient primal-dual approach in convex optimization. The reason behind it is to show the flexibility of variational model for improving both the accuracy and robustness in the flow field estimation.

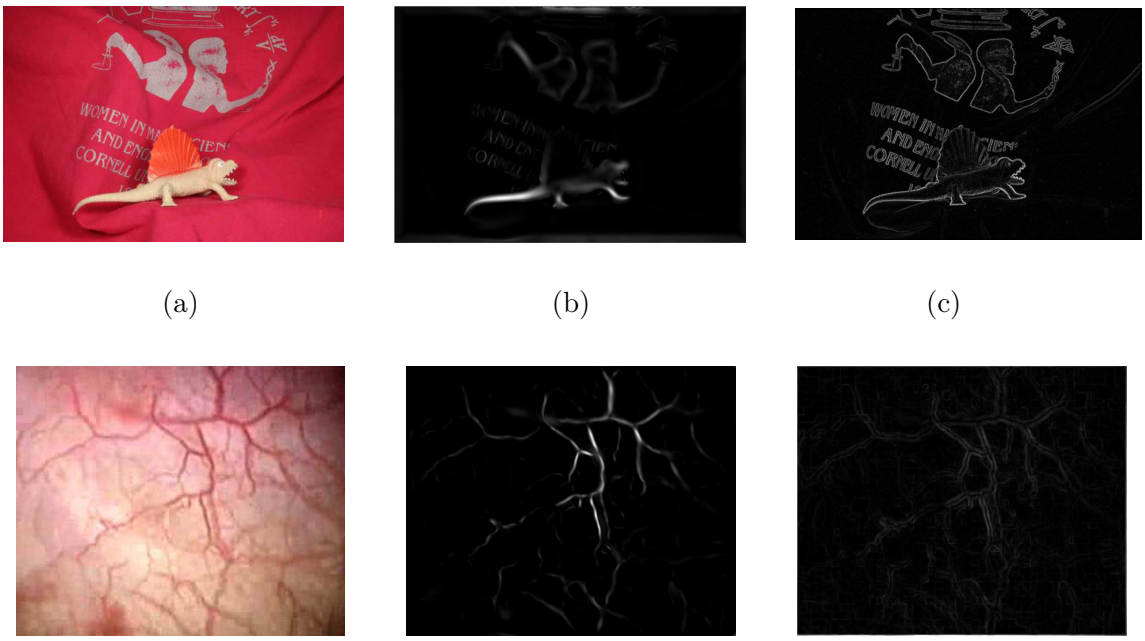


(a) (b)

Figure 2.10: Blob and vessel measures determined by the magnitudes of Eigen values. (a) Blob like structure is obtained when $\lambda_+ \sim \lambda_-$). (b) Elongated vessel structure (with $\lambda_+ >> \lambda_-$)

Robust structure estimate based data-term

The global idea is to improve the data-cost by increasing the impact of image structures in the optical flow estimation procedure. Let I_1 and I_2 be an image pair in an image sequence



(a) (b) (c)

(d) (e) (f)

Figure 2.11: Structure estimate and its gradient. (a, d) Original image of classical scene and bladder scene. (b, e) Structure estimates of (a) and (d) respectively and (c, f) respective gradient images.

corresponding to time instances t and $t+1$ respectively.A scale-space approach is used to exploit the image structure information. This approach uses a symmetric 2×2 Hessian matrix $H_2 I_1$ given hereafter for image I_1:

$$H_2 I_i(\mathbf{x}, s) = s^\gamma I * \begin{pmatrix} \partial^2 G(x,y,s)/\partial x^2 & \partial^2 G(x,y,s)/\partial x \partial y \\ \partial^2 G(x,y,s)/\partial x \partial y & \partial^2 G(x,y,s)/\partial y^2 \end{pmatrix}, \tag{2.67}$$

where s is the scale, γ is the parameter for normalized derivatives in scale-space model [Lindeberg, 1994] and G is the $2D$ Gaussian kernel with standard deviation ($\sigma = \sqrt{s}$). Lorenz et al. [Lorenz et al., 1997] proposed to use second-order derivative of Gaussian function for a better approximation of the optimal filter to deal with noisy images. Eigenvalue analysis of the Hessian matrix in Eq. (2.67) is used to select the strongest response at different scales and orientations for reconstruction of the substantial structure. A similar approach was proposed by Frangi et al. [Frangi et al., 1998]. Let λ_+ and λ_- be the two Eigenvalues ($\lambda_+ \geq \lambda_-$) of the Hessian matrix $H_2 I_i$ then the structure estimate is given by $SI_i(\mathbf{x}, s)$:

$$SI_i(\mathbf{x}, s) = exp(-\frac{{R_b}^2}{2\beta_1^2})(1 - \exp(\frac{-{R_s}^2}{2\beta_2^2})) \tag{2.68}$$

where, $R_b = \lambda_-/\lambda_+$ is the blobness measure, $R_s = \sqrt{\lambda_+^2 + \lambda_-^2}$ represents the structureness and $\{\beta_1, \beta_2\}$ are sensitivity control parameters empirically set to 0.5. Here, R_b relates to the eccentricity of the ellipse and becomes maximum when an elongated elliptical structure is found (refer to Fig. 2.10 (b)) and a circular blob like structure is obtained when R_b is close to one (refer to Fig. 2.10 (a)). R_s is low for the background, *i.e.* when no structure is present. This means both Eigenvalues are low and noisy background pixels are suppressed. The sign of Eigenvalues indicates the brightness or darkness of the pattern determined. The structure estimate of an image is also shown in Fig. 2.11 (b, e). It is visible in the structure estimate of the classical scene in Fig. 2.11 (b) and in the bladder structure estimate Fig. 2.11 (e) that the elongated structures have the strongest response with Eq. (2.68). This is due to the fact that the product of two terms have simultaneously the highest values for elongated structures: a low R_b value of stretched shapes lead to large exponential and high R_s values (presence of textures) make the exponential value included with the structure estimate close to zero.

Using the assumption that the structures estimated between the images (refer Eq. 2.68) do not change in between two consecutive frames, then the Structure Constancy Assumption (SCA) can be written as the following constrained equation:

$$\rho_{sca}(\mathbf{x}, \mathbf{v}, s) = \mid SI_2(\mathbf{x}+\mathbf{v}, s) - SI_1(\mathbf{x}, s) \mid = 0. \tag{2.69}$$

The first-order Taylor expansion of Eq. (2.69) leads to the following quadratic data-term:

$$\mathcal{D}_8(\mathbf{x}, \mathbf{v}, s) = \mid \rho_{sca} \mid^2 \quad = \quad \mid \nabla SI_2(\mathbf{x} + \mathbf{v}^0).(\mathbf{v} - \mathbf{v}^0) + \underbrace{SI_2(\mathbf{x} + \mathbf{v}^0) - SI_1(\mathbf{x})}_{SI_t} \mid^2, \tag{2.70}$$

with $\mathbf{v}^0$ being the close approximation of $\mathbf{v}$.

Towards an image-driven robust regularizer

The major drawback of these shape estimation alone is that, in overlapped regions, the edges can suffer from oversmoothing. Thus, at each scale s, we introduce an edge discriminator which

will weight the importance of edge pixels and non edge pixels in the regularization (also refer [*Lorenzet al.*, 1997]):

$$\eta_{edge}(\mathbf{x}, s) = \frac{\mid \nabla I_1 \mid}{\sqrt{\sigma}\lambda_-(\mathbf{x}, s)}. \tag{2.71}$$

$\eta_{edge}(\mathbf{x}, s)$ is close to 1 for boundaries and tends towards 0 for non-edge pixels. Adding an edge discriminator can preserve the flow discontinuities across edges (see edges in Fig. 2.11 (c)). However, to constraint the strong influence of the edge discriminator over the regularizer, we assign a threshold value for each detected edge pixel:

$$\eta_{edge}(\mathbf{x}, s) := \begin{cases} 0.65, & \forall \quad \text{edge pixels} \\ 0 & \text{otherwise.} \end{cases} \tag{2.72}$$

The weights assigned to the regularizer penalize the edges of image structures less severely than that in the classical TV-regularization.

The Proposed Model-I: RFLOW

A first proposed model in this thesis is the total variational approach using L^1-norm with a data-term including a weighted combination of a brightness constancy assumption ρ_{bca} and the structure constancy assumption ρ_{sca} between image pairs. The brightness constancy assumption is robust in many scenarios but fails when there is change in light gradients between image pairs. Thus, in such cases we have used structure constancy assumption. A scale-space based approach for robustly estimating the structures in images are used. These structures are not affected by the change in light gradients and/or the presence of noise. However, in case of weak textured image pairs, SCA cannot be effective and the change in illumination can affect the data-term globally. So, in order to reduce such drift, an additional illumination compensating function $\mathbf{L} : \Omega \rightarrow \mathbb{R}$ was used to complete the data-term in L^1 framework at pixel $\mathbf{x}$, written as:

$$\mathcal{D}(\mathbf{x}, \mathbf{v}, \mathbf{L}) = \phi \parallel \rho_{bca}(\mathbf{x}, \mathbf{v}) \parallel_1 + (1 - \phi) \parallel \rho_{sca}(\mathbf{x}, \mathbf{v}) \parallel_1 + \parallel \gamma \mathbf{L} \parallel_1, \tag{2.73}$$

with ϕ as the weight governing the effect of constancy assumptions which is adaptive and computed using the method detailed in Xu et al. [Xu et al., 2012]. γ controls the influence of the illumination compensating function $\mathbf{L}$ in the data-term.

In order to prevent the smoothness across the boundaries and to distinguish between edge and structure, an edge discriminator (refer Eq. (2.71)) is incorporated in the regularizer. This edge discriminator is used to penalize the large influence of the smoothness term on the image edges, because usually, flow discontinuties are along the image edges and have to be preserved. Boundary pixels will almost not contribute to the regularization since $(1 - \eta_{edge})$ will tend towards smaller values at edge-pixel locations. Integrating them all, our proposed model can be formulated as the following constrained energy minimization equation for all pixels in image domain Ω:

$$\min_{\mathbf{v},\mathbf{L}} \int_\Omega \mid \mathcal{D}(\mathbf{v}, \mathbf{L}) \mid d\Omega + \lambda_s \int_\Omega \{(1 - \eta_{edge}) \mid \nabla \mathbf{v} \mid + \mid \nabla \mathbf{L} \mid\} d\Omega, \tag{2.74}$$

where λ_s is the trade-off between the data-term and the smoothness term respectively. The penalty function $(1 - \eta_{edge})$ is added in discrete setting with values given by Eq. (2.72), λ_s is set experimentally to be 50.

2.5.2 Primal-dual energy minimization

In Eq. (2.74), the first integral represents the image data fidelity term, the second integral corresponds to the regularization term with an additional smoothness for illumination ($\nabla \mathbf{L}$). The discrete version of this primal motion model can be written as:

$$\min_{(\mathbf{v}\in Y, \mathbf{L}\in X)} \sum_{x,y} \{\mathcal{D}(\mathbf{v}_{x,y}, \mathbf{L}_{x,y}) + \lambda_s(1-\eta_{edge}) \parallel \nabla \mathbf{v}_{x,y} \parallel_1 + \lambda_s \parallel \nabla \mathbf{L}_{x,y} \parallel_1\} \tag{2.75}$$

The discretization of the robust data-term in Eq. (2.73) can be written as below using Eqns. (2.23) and (2.70):

$$\begin{aligned}\mathcal{D}(\mathbf{v}_{x,y}, \mathbf{L}_{x,y}) = \phi \parallel (\nabla I_2)_{x,y}(\mathbf{v}_{x,y} - \mathbf{v}^0_{x,y}) + (I_t)_{x,y} \parallel_1 \\ +(1-\phi) \parallel (\nabla S_2)_{x,y}(\mathbf{v}_{x,y} - \mathbf{v}^0_{x,y}) + (SI_t)_{x,y} \parallel_1 + \gamma \parallel \mathbf{L}_{x,y} \parallel_1,\end{aligned} \tag{2.76}$$

where I_t and SI_t are the temporal difference of the gray-level images (I_1, I_2) and the structure images (S_1, S_2) respectively.

The vectorial gradient ($\nabla \mathbf{v}$) $\nabla u \times \nabla v$ is in $Z-$space and the L^1-norm of this gradient is given by:

$$\begin{aligned}\parallel \nabla \mathbf{v} \parallel_1 = \sum_{x,y} \mid \nabla \mathbf{v}_{x,y} \mid_1, \quad \text{with} \\ \nabla \mathbf{v}_{x,y} = \sqrt{(\nabla u^1_{x,y})^2 + (\nabla u^2_{x,y})^2 + (\nabla v^1_{x,y})^2 + (\nabla v^2_{x,y})^2},\end{aligned} \tag{2.77}$$

with ∇u^1 and ∇u^2 being derivatives of u and ∇v^1 and ∇v^2 derivatives of v along x and y respectively, and u and v being components of $\mathbf{v}$. Similarly, the L^1-norm of $\nabla \mathbf{L} \in Y$ is given by:

$$\begin{aligned}\parallel \nabla \mathbf{L} \parallel_1 = \sum_{x,y} \mid \nabla \mathbf{L}_{x,y} \mid_1 \quad , \quad \text{with} \\ \nabla \mathbf{L}_{x,y} = \sqrt{(\nabla \mathbf{L}^1_{x,y})^2 + (\nabla \mathbf{L}^2_{x,y})^2}\end{aligned} \tag{2.78}$$

Let p and q be the dual variables of convex sets $P \in Y$ and $Q \in Z$ respectively which are defined as: $P = \{p \in Y : \parallel p \parallel_\infty \leq 1\}$ and $Q = \{q \in Z : \parallel q \parallel_\infty \leq 1\}$. $\parallel p \parallel_\infty$ and $\parallel q \parallel_\infty$ are the discrete maximum norms defined as:

$$\begin{aligned}\parallel p \parallel_\infty = \max_{x,y}(\mid p_{x,y} \mid_1), \\ p_{x,y} = \sqrt{(p^1_{x,y})^2 + (p^2_{x,y})^2}\end{aligned} \tag{2.79}$$

$$\begin{aligned}\parallel q \parallel_\infty = \max_{x,y}(\mid q_{x,y} \mid_1), \\ q_{x,y} = \sqrt{(q^1_{x,y})^2 + (q^2_{x,y})^2 + (q^3_{x,y})^2 + (q^4_{x,y})^2}\end{aligned} \tag{2.80}$$

The saddle-point formulation of the primal motion estimation model in Eq. (2.75) is given by a primal-dual theorem as [Chambolle and Pock, 2011]:

$$\min_{\mathbf{v}\in Y, \mathbf{L}\in X} \max_{p\in Y, q\in Z} \langle \nabla \mathbf{v}, q\rangle_Z + \langle \nabla \mathbf{L}, p\rangle_Y + \lambda_s \parallel \rho(\mathbf{v}, \mathbf{L}) \parallel_1 - \delta_P(p) - \delta_Q(q) \quad \text{where,} \tag{2.81}$$

$$\delta_P(p) = \begin{cases} 0 & \text{when} \quad , p \in P \\ \infty & \text{otherwise} \end{cases} \tag{2.82}$$

and

$$\delta_Q(q) = \begin{cases} 0 & \text{when} \quad , q \in Q \\ \infty & \text{otherwise} \end{cases} \tag{2.83}$$

We can rewrite Eq. (2.81) as:

$$\min_{\mathbf{v}\in Y, \mathbf{L}\in X} \max_{p\in Y, q\in Z} \langle \nabla \mathbf{v}, q\rangle_Z + \langle \nabla \mathbf{L}, p\rangle_Y + G(\mathbf{v}, \mathbf{L}) - F^*(p, q), \tag{2.84}$$

with $G(\mathbf{v}, \mathbf{L}) = \mathcal{D}(\mathbf{v}, \mathbf{L})$ and $F^*(p, q) = \delta_P(p) + \delta_Q(q)$.

To solve the primal-dual Eq. (2.84), we compute the resolvent (proximation) operators using the Moreau-Yosida Theorem [Moreau, 1965]. The resolvent operator w.r.t $F^*(p, q)$ is given by simple pointwise projection onto L^2 ball:

$$(p, q) = (I + \sigma\delta F^*)^{-1}(\tilde{p}, \tilde{q}), \tag{2.85}$$

where $\sigma \in [0, 1]$, I is identity matrix and $(\tilde{p}, \tilde{q})$ are the initial vectors. The solutions for p and q for the $(n + 1)^{th}$ iteration is given by:

$$p_{x,y}^{n+1} = \frac{\tilde{p}_{x,y}}{\max(1, \mid \tilde{p}_{x,y} \mid)} \quad \text{and} \quad q_{x,y}^{n+1} = \frac{\tilde{q}_{x,y}}{\max(1, \mid \tilde{q}_{x,y} \mid)}, \tag{2.86}$$

where $\tilde{p}_{x,y} = \dfrac{(p_{x,y})^n + \sigma\nabla L^n}{1 + \sigma * \epsilon_L}$ and $\tilde{q}_{x,y} = \dfrac{(q_{x,y})^n + \sigma\nabla \mathbf{v}^n}{1 + \sigma * \epsilon_{\mathbf{v}}}$ with constants ϵ_L and $\epsilon_{\mathbf{v}}$. $\mid \tilde{p}_{x,y} \mid$ and $\mid \tilde{q}_{x,y} \mid$ are defined respectively in Eqs. (2.79) and (2.80). The next resolvent operator with respect to $G(\mathbf{v}, \mathbf{L})$ is given by:

$$(\mathbf{v}, \mathbf{L}) = (I + \tau\delta G)^{-1}(\tilde{\mathbf{v}}, \tilde{\mathbf{L}}). \tag{2.87}$$

The solutions for $\mathbf{v}$ and L are:

$$(\mathbf{v}^n, \mathbf{L}^n) = (\tilde{\mathbf{v}}, \tilde{\mathbf{L}}) + \begin{cases} (\tau\lambda_s\nabla I_2, \tau\lambda_s\gamma) & if \quad \rho(\tilde{\mathbf{v}}, \tilde{\mathbf{L}}) < -\tau\lambda_s \mid a \mid^2_{x,y} \\ -(\tau\lambda_s\nabla I_2, \tau\lambda_s\gamma) & if \quad \rho(\tilde{\mathbf{v}}, \tilde{\mathbf{L}}) > \tau\lambda_s \mid a \mid^2_{x,y} \\ -(\frac{\rho(\tilde{\mathbf{v}}, \tilde{L})\nabla I_2}{\mid a \mid^2_{x,y}}, \frac{\rho(\tilde{\mathbf{v}}, \tilde{L})\gamma}{\mid a \mid^2_{x,y}}) & if \quad \rho(\mid \tilde{\mathbf{v}}, \tilde{\mathbf{L}} \mid) \leq \tau\lambda_s \mid a \mid^2_{x,y}, \end{cases} \tag{2.88}$$

where $\mid a \mid^2_{x,y} = \gamma^2 + \mid \nabla I_2 \mid^2_{x,y} + \mid \nabla \mathcal{S}_2 I_2 \mid^2_{x,y}$.

Primal-dual updates

$$\left.\begin{aligned} p^{n+1} &\approx \frac{p^n + \sigma\nabla\tilde{\mathbf{L}}^n}{max(1, |p^n + \sigma\nabla\tilde{\mathbf{L}}^n|)} \\ q^{n+1} &\approx \frac{q^n + \sigma\nabla\tilde{\mathbf{v}}^n}{max(1, |q^n + \sigma\nabla\tilde{\mathbf{v}}^n|)} \end{aligned}\right\} \text{dual variable update} \tag{2.89}$$

$$\left.\begin{aligned} \mathbf{v}^{n+1} &= \mathbf{v}^n + (1 - \eta_{edge})\tau div(q^{n+1}) \\ \mathbf{L}^{n+1} &= \mathbf{L}^n + \tau div(p^{n+1}) \end{aligned}\right\} \text{primal variable update} \tag{2.90}$$

$$\left.\begin{aligned} \mathbf{v}^{n+1} &= 2\ \mathbf{v}^{n+1} - \mathbf{v}^n \\ \mathbf{L}^{n+1} &= 2\ \mathbf{L}^{n+1} - \mathbf{L}^n \end{aligned}\right\} \text{extrapolation} \tag{2.91}$$

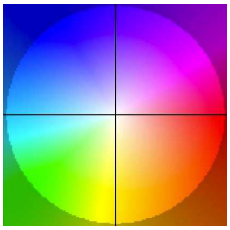


Figure 2.12: Flow color code.

In the above Eqns. (2.88)-(2.91), $\tilde{(.)}$, $(.)^n$ and $(.)^{n+1}$ represents initial values, current estimates and update values respectively. At the beginning of our algorithm in the coarsest pyramid level, both the $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{L}}$ are set to zero. The initial values for the dual variables $\mathbf{p}$ and $\mathbf{q}$ are set to zero for first iteration at all the pyramid levels during the energy minimization procedure. The scalar positive terms τ and σ, responsible of deciding the step-size of the algorithm, are set to 0.35. The weighting terms λ_s and γ are set to 50 and 0.02 respectively.

2.5.3 Optical flow assessment and benchmarking

We recall most important notations here. For our correspondence problem we are interested in computing optical flow field $\mathbf{u} = (u, v, 1)^T$, $(u, v) : \Omega \rightarrow \mathbb{R}$ between two consecutive images $I_1 = I(\mathbf{x}, t)$ and $I_2 = I(\mathbf{x}, t + 1)$ taken at time instances t and $t + 1$ respectively.

a) Visualization

Optical flow field can be visualized using arrow plots (*quiver in MATLAB*). However, to visualize the dense optical flow motion, a better choice is to visualize with the color code as represented in Fig. 2.12. In this figure, the hue and the saturation of colors encode respectively the direction and the magnitude of the flow vectors.

b) Error metrics

Ground truth flow fields are available for simulated sequences or synthetic image pairs. In order to evaluate the algorithms it is important to measure the quality of the estimated flow field. Two most popular measures are the Average Angular Error (AAE) [Barron et al., 1994] and the Average End-Point Error (AEPE) [Baker et al., 2011]. If $\mathbf{u}_c$ is the ground truth flow, then the AAE of the field of vectors $\mathbf{u}$ is defined as:

$$AAE := \frac{1}{\mid \Omega \mid} \int_{\Omega} \arccos\left(\frac{\mathbf{u}^T \mathbf{u}_c}{\mid \mathbf{u} \mid\mid \mathbf{u}_c \mid}\right) d\mathbf{x}, \tag{2.92}$$

with $\mid \Omega \mid$ representing the total number of pixels and $\mid \mathbf{u} \mid = \sqrt{u^2 + v^2 + 1}$ is the Euclidean norm of the flow vector. The EEP is defined as the Euclidean distance of the difference between the estimated and the ground truth flow field. The average EPE (AEPE) can be written as:

$$AEPE := \frac{1}{\mid \Omega \mid} \int_{\Omega} \mid \mathbf{u} - \mathbf{u}_c \mid d\mathbf{x}. \tag{2.93}$$

Another, widely used error measure in stereo-vision is the *Bad Pixel Error* measure (BPE) [Scharstein and Szeliski, 1998, Geiger et al., 2013]. It gives the percentage of points that deviate

more than the threshold δ from the ground truth flow and can be formulated as:

$$BPE := \frac{100}{|\ \Omega\ |} \int_{\Omega} \mathcal{K}_{\delta}(|\ \mathbf{u} - \mathbf{u}_c\ |) d\mathbf{x}, \tag{2.94}$$

with $\mathcal{K}_{\delta}(.) = 1$ if $(.) > \delta$ else 0.

c) Public datasets

Variational approaches have become very popular for recovering the scene motion. However, the formulation of these methods largely affects their ability to estimate accurately either small motion or large motion displacements. This thesis work focuses on both of these aspects with propositions dealing with both kinds of motion. We briefly present the available datasets that are popularly used for benchmarking flow algorithms for both small and large displacements cases.

Small motion: The Middlebury flow dataset [Baker et al., 2011] consists of 8 image pairs having known ground truth (GT) each in training and in test datasets. These include images with hidden texture, synthetic and stereo pair. Since this dataset is used for estimation of small motion fields for varying texture conditions, it is interesting for robustness evaluation of our algorithms for small displacement case.

Large motion: The KITTI optical flow dataset [Geiger et al., 2013] consists of 194 color image pairs in each training and test datasets. The images were acquired with a wide-view camera fixed on a moving vehicle. These image pairs have challenging characteristics of an outdoor scene like illumination variability, large perspective changes, repeated texture patterns and large displacements.

Another large displacement dataset is the MPI Sintel dataset [Butler et al., 2012]. It consists of 12 different synthetic image sequences each of them further classified into clean pass and final pass. The clean pass images are data simulating optimal acquisition conditions: contrasted images without blur and noise. The final pass sequences consist of more challenging dataset which are obtained by degrading the clean pass images. The 12 final pass sequences simulate large motion between image pairs, specular reflections, blur due to camera defocus/refocus, motion blur and atmospheric effects like fog and smoke. This dataset includes also having some image pairs with local deformations.

2.5.4 Results and discussion

The set of parameters used for the experiments for the first proposed method (RFLOW) are $\epsilon_{\mathbf{u}} = 0.0001$, $\epsilon_{\mathbf{L}} = 0.0001$, $\lambda_s = 50$ and $\gamma = 0.02$. τ and σ are set such that $\tau \times \sigma = \frac{1}{8}$.

Illustration of improved optical flow accuracy

First, an image pair from the Marble sequence [1] (see Fig. 2.14(a-b)) is taken to visually illustrate the improvements of the estimated flow field with respect to some classical dense optical flow methods. The known ground-truth flow field between these images is shown in Fig. 2.14 (c). Fig. 2.14 (d) illustrates the oversmoothing effect of the Horn and Schunck model along the boundaries. The non-quadratic energy minimization of the TV-L^1 model [Pock et al., 2007]

[1] The Marble sequence has been created by Otte and Nagel http://i21www.ira.uka.de/image_sequences/.

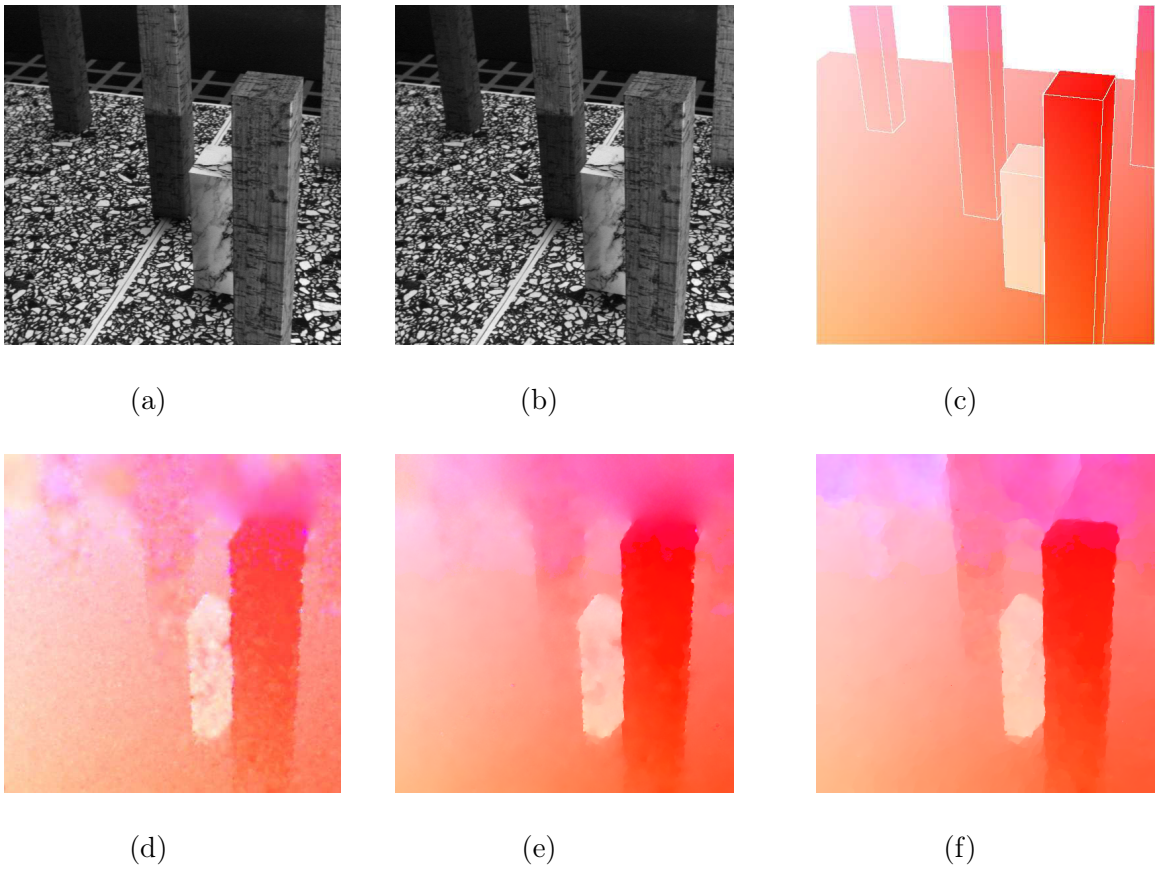


Figure 2.13: Visual validation of the improvement of classical TV-L^1 algorithm. (a, b) Frame 16 and 17 respectively of Marble sequence, (c) ground truth flow between (a) and (b), (d) flow field obtained with the classical Horn-Schunck approach, (e) flow field obtained with the classical TV-L^1 algorithm and (f)flow field obtained with the structure estimate (**RFLOW**).

minimizes this over-smoothness effect as can be seen in Fig. 2.14 (e) with relatively less outliers in the flow field computation (see the small rectangular block in light orange). However, in the classical TV-L^1 model the oversmoothing effect is still a problem which can be observed around the large rectangle (red in flow color code in Fig. 2.14 (e)). The proposed robust TV approach (**RFLOW**) preserves the edges more sharply and allowing fewer outliers in the flow field as visible in Fig. 2.14 (f).

Benchmarking and results of RFLOW on Middlebury dataset

The benchmarking of the proposed RFLOW model is done on the Middlebury dataset [Baker et al., 2011]. The result of this benchmarking against some well-known TV-L^1 improvements is provided in Table 2.2. In this table, the performance of algorithms varies in different image pairs. Algorithms like TV-L^1 improved [Wedel et al., 2009b] and Brox et al. [Brox et al., 2004] have AEPE values greater than 1 pixel for the Urban and Grove data respectively. Although, anisotropic Huber-L^1 [Werlberger et al., 2009] performs better in many image pairs on this dataset. Comparatively, the RFLOW method performs consistently well for AEPE errors in the whole dataset (see last column for average error) and outperformed all the methods giving only

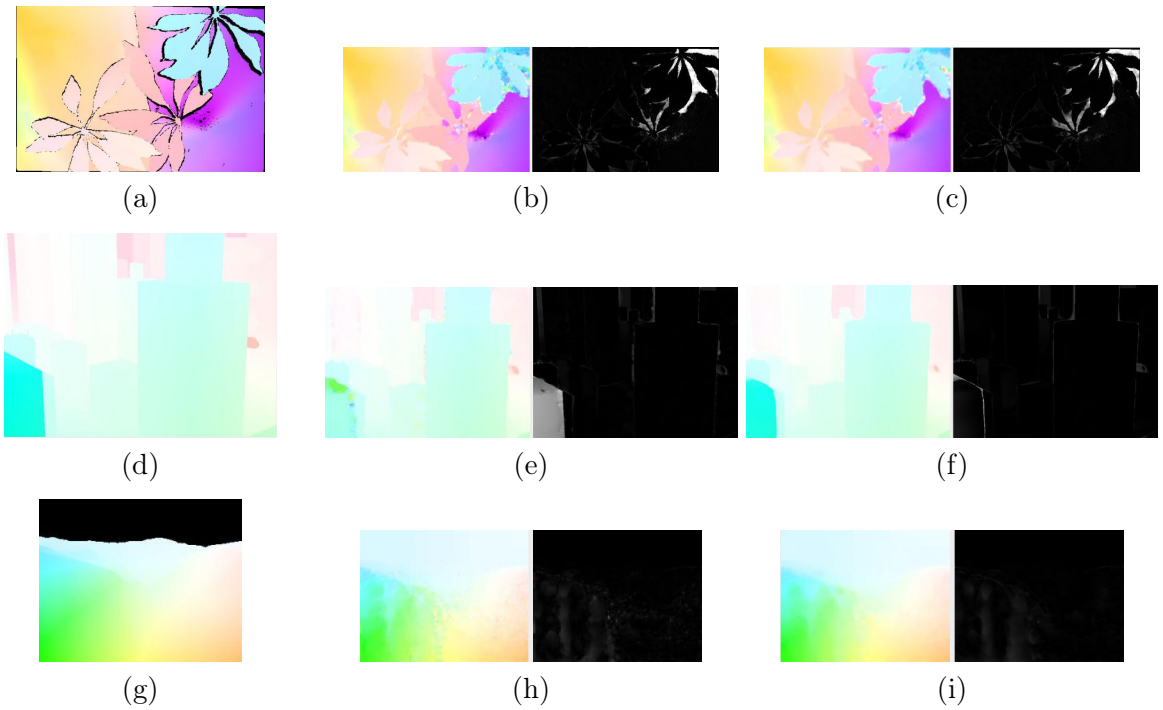


Figure 2.14: Results on the Middlebury test image sequences. The images (a, d, g) of the first column represents the Ground Truth, the second column (b, e, h) corresponds to the optical flow estimation by the TV-L1-improved and the third column (c, f, i) represents the optical flow estimation using the proposed model. Flow errors are shown adjacent to the color representation of OF field.

3.80° errors for AAE criterion and 0.40 pixels for AEPE.

Ground truth and estimated flow vectors represented by improved-TV-L^1 method and the proposed method are shown in Fig. 2.14. The third column representing the proposed method has more accurate flow estimation than the improved TV-L^1 by Wedel et al. [Wedel et al., 2009b]. The flow errors are represented in gray images on the right side of the color-coded figures. The improved TV-L^1 method has highest flow error in Urban dataset while our method works consistently better for all images in the dataset. The proposed model also computes an accurate motions vectors for the dynamic scenes (refer Fig. 2.15).

Methods	Army	Mequon	Scheffl.	Wooden	Grove	Urban	Yos.	Teddy	Avg.
RFLOW	0.10 3.82	**0.19** **2.61**	0.46 5.66	0.22 3.93	0.92 **3.24**	0.42 4.12	0.14 2.61	0.77 4.48	**0.40** **3.80**
Aniso. Huber-L^1	0.10 3.71	0.31 4.36	0.56 6.92	**0.20** **3.54**	**0.84** 3.38	**0.39** **3.88**	0.17 3.37	**0.64** **3.16**	**0.40** 4.04
TV-L^1 improved	**0.09** **3.36**	0.20 2.82	0.53 6.50	0.21 3.80	0.90 3.34	1.51 5.97	0.18 3.57	0.73 4.01	0.54 4.17
Brox et. al.	0.11 4.44	0.27 3.72	**0.39** **4.97**	0.24 4.58	1.10 3.79	0.89 3.91	**0.10** **2.22**	0.91 4.62	0.50 4.03

Table 2.2: Average end-point error, in pixels and average angular error, in degrees ($\frac{AEPE}{AAE}$) are given for some well known TV-L^1 based methods on Middlebury test dataset. The methods are presented in the order of their rank on this benchmarking for test dataset (refer for details: http://vision.middlebury.edu/flow/eval/results/results-e1.php). An average value (avg.) is also provided for validating the algorithm tolerance to varying textures in this dataset.

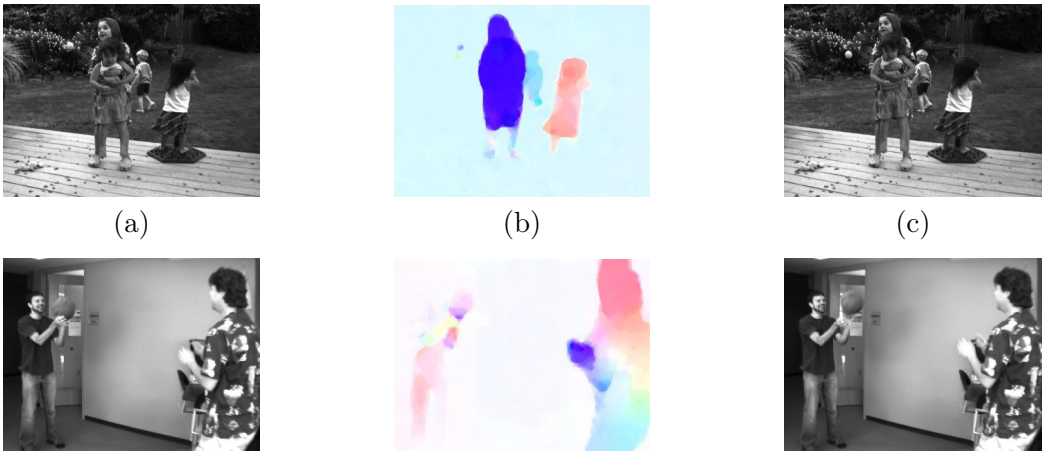


Figure 2.15: Optical flow estimation using the proposed method for dynamic scenes in Middlebury data-set. (a-c) Backyard sequence, (d-e) Basketball sequence. The images pairs are in the first and last column and the flow field is given in the central column.

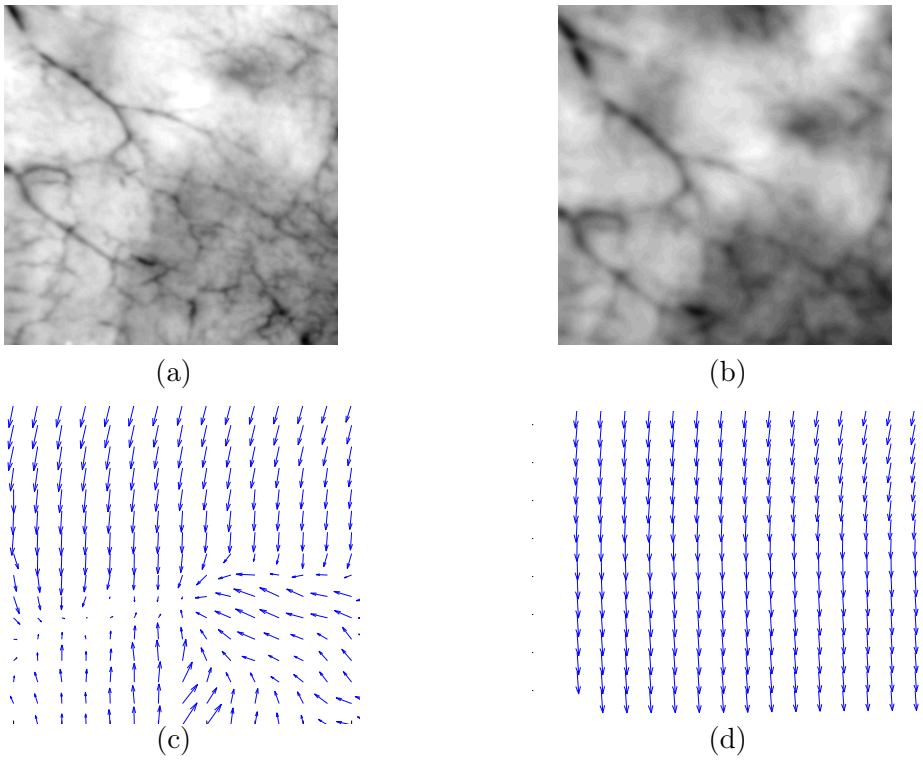


Figure 2.16: Homologous point estimation in WL modality. a) Source image I_2, b) target image I_1, c) displacement field obtained using classical TV-l^1 method [Pock et al., 2007], d) displacement field with the proposed model. Target image is blurred and darkened relative to the source image. Flow vectors (arrows) at every 5^{th} pixel in x and y directions are shown.

First validation of the RFLOW method on bladder images

The example given in Fig. 2.16 involves a pair of images acquired under the WL modality, separated by large pixel displacements and affected by blur. This image pair has known motion vector field between I_1 and I_2. All the motion vectors were established such that they represent the same motion at all the pixels. This was done by simulating a known homography between them. However, when the classical TV-l^1 model [Pock et al., 2007] was applied, it failed to compute accurate dense flow vectors (refer Fig. 2.16(c)) leading to a large AEPE of 6.5 pixels and a local registration error $\epsilon_{1,2}$ of 3.5 pixels. In Fig. 2.16(d) the proposed method represents an accurate and robust motion estimation resulting in precise registration ($\epsilon_{1,2} = 0.2$ pixels and AEPE $= 0.12$ pixels respectively). It is evident from Fig. 2.16(d) that all the vectors are oriented in the same direction with exactly similar magnitude (as that of the ground truth flow vectors).

2.6 Main contributions

Following summarizes the work/contribution of this chapter:

- Overview on local and global optical flow methods for solving correspondence problem.
- Overview on variational approaches: data-term and smoothness term formulations.
- Overview on Mathematical optimization in regard to convex optimization
- Modeling of robust data-term: Structure constancy assumption in addition to the classical brightness constancy assumption. This is the first orginal contribution in this thesis.
- Benchmarking of the proposed robust model-I (RFLOW) in this thesis as a contribution to improvement of TV-L^1 approach.

In this section we presented the main aspects of implementing TV-L^1 methods with appropriate data terms ad regularizers. We also highlighted their advantage in terms of accuracy over other reference approaches. The main contribution lies in the definitive validation of their usefulness for optical flow computation in cystoscopic images. This first results described here and in [Ali et al., 2014] inspired us to develop a more sophisticated TV-L^1 based optical flow methods presented in Chapters 3 and 4.

List of publications

[ADB14] Sharib Ali, Christian Daul and Walter Blondel "Robust and accurate optical flow estimation for weak texture and varying illumination condition: Application to cystoscopy," 4^{th} *Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, pp. 140–145, Paris, France, October 14-17, 2014.

Chapter 3

Anisotropic optical flow on edge preserving Riesz wavelet basis

Contents

3.1 Motivation: Improved robustness for weak textured images 65
3.2 Optical flow on edge preserving wavelet basis 66
3.2.1 Classical brightness constancy in multi-resolution framework 66
3.2.2 Multi-resolution with Riesz basis 67
3.3 Anisotropic regularization 69
3.3.1 Tensor based anisotropic regularization 70
3.3.2 Div-curl decomposition of **v** 72
3.3.3 Weighted non-local median filtering 74
3.4 Optimization 75
3.5 Optical flow algorithm overview and parameter settings 76
3.5.1 Algorithm overview 76
3.5.2 Parameter setting 77
3.6 Results and discussion 79
3.6.1 Evaluation of motion estimation on the Middlebury database 79
3.6.2 Evaluation of different motion types using an endoscopic set-up 79
3.6.3 Evaluation with simulated homographies sparse textured scenes 81
3.6.4 Robustness of algorithm against strong illumination changes 89
3.7 Main contributions and conclusion 90
List of publication 91

3.1 Motivation: Improved robustness for weak textured images

Variational approaches are classically based on coarse-to-fine energy minimization scheme in order to deal with large displacements in image pairs. In such multi-resolution approaches, original images are downsampled with some scaling factor. A major drawback of this downsampling on image space is that it leads to a "flattening-out" problem in images with weakly contrasted textures and/or presence of few image structures. "Flattening-out" here thus means that the

textures are affected during the downsampling process and hence structures like object edges are not preserved at coarse pyramid levels. As a result, flow field discontinuties are oversmoothed and most likely lead to poor or inaccurate initialization of flow vectors when passing from coarse to a finer pyramid level. On contrary, wavelet based methods can deal with such problems and can be directly integrated in multi-resolution energy minimization schemes due to their inbuilt coarse-to-fine property. Such an approach has been used by Liu et al. [Liu et al., 2003] to overcome both the flattening-out problem and for attenuating the effect of point-noise. However, this contribution is used in the frame of local optical flow field computation which is not suitable for dense correspondence establishment in low textured image pairs (refer to Chapter 2). Even in coarse-to-fine approaches, local methods lead to a large sparsity in their estimated flow fields.

In this chapter, we propose a total variational optical flow algorithm using edge preserving Riesz wavelet basis filters in a multi-resolution energy minimization framework. Riesz wavelet filters are used so that edges of the image structures are preserved even at coarse pyramid levels (i.e. "flattening-out" problem is overcomed). These filters also reduces the effect of weak local illumination differences in image pairs. The second aim of this chapter is to model an anisotropic regularizer based on: 1) image structure tensor (Image driven regularizer) and 2) weighted median filtering (non-local flow driven regularization). The weighted median filtering is used as the flow field refinement technique similar to Sun et al. [Sun et al., 2010]. The third goal of this chapter is to validate the improvement in robustness and accuracy of the method on publically available Middlebury dataset [Baker et al., 2011] (consisting of image pairs with large texture variability) and on simulated video-sequences dominated with weakly pronounced textures.

3.2 Optical flow on edge preserving wavelet basis

In this section, a novel approach using wavelet-based pyramids for variational energy minimization is presented. Subsection 3.2.1 first briefly presents the drawback of classical data terms (*i.e.* based on brightness constancy computed directly with the raw images). Subsection 3.2.2 discusses important issues about the choice of the adequate filtering process to be used in the frame of multi-resolution image decomposition techniques.

3.2.1 Classical brightness constancy in multi-resolution framework

Let $\mathbf{x}=\{x, y\}$ be the pixel-coordinates in space $\Omega \in \mathbb{R}^2$ of image pair $\{I_i^0, I_{i+1}^0\}$. The exponent 0 denotes images I_i and I_{i+1} at original resolution (scale 0). Let $\mathbf{v} = \{u, v\}$ be the optical flow field with vector components u and v along the $x-$axis and $y-$axis respectively. The classical brightness constancy equation used in optical flow estimation can then be written as a following non-linear formulation:

$$I_t(\mathbf{x}, \mathbf{v}) = \| I_{i+1}^0(\mathbf{x} + \mathbf{v}) - I_i^0(\mathbf{x}) \|^2 = 0. \tag{3.1}$$

This equation can be linearized and approximated using the first order Taylor expansion:

$$\rho(\mathbf{v}) = I_{i+1}^0(\mathbf{x} + \mathbf{v}^{ap.}) - I_i^0(\mathbf{x}) + \nabla I_{i+1}^0(\mathbf{x} + \mathbf{v}).(\mathbf{v} - \mathbf{v}^{ap.}) = 0, \tag{3.2}$$

with $\mathbf{v}^{ap.}$ as a close approximation to $\mathbf{v}$ calculated at pixel $\mathbf{x}$ and $\rho(\mathbf{v})$ is an approximation of I_t.

Eq. (3.2) assumes that the intensities of homologous pixels are constant over small time interval. This is too strong assumption for real images, mostly in the case of scenes with illumination changes leading to light gradients. For this reason, data-terms may associate an additional gradient [Brox et al., 2004] or Hessian constancy assumptions [Ali et al., 2014] to the classical brightness constancy term for robust optical flow computation. But, gradient or Hessian

constancy assumptions are only realistic for camera motion leading to small in-plane rotations. In industrial or medical images, first and second derivatives of grey-levels or colors are often not constant due to large in-plane rotations. Moreover, whatever the motion type, none of the mentioned constancy assumptions or their combinations lead to a robust optical flow estimation at coarse levels when using the traditional pyramidal approaches (*e.g.* a Gaussian pyramid). Indeed, in image regions with weak texture contrasts, the information is gradually oversmoothed when going from finer to coarser levels in the image pyramids. This will result in a flow field with false vectors (*i.e.* "flattening out problem") at coarser levels leading to poor initialization of the vector field at finer levels. As detailed in this contribution, an edge preserving and illumination invariant wavelet-based approach with an added anisotropic regularization is used to tackle the problem related to non-constant scene illumination and texture flattening out at coarser levels.

3.2.2 Multi-resolution with Riesz basis

Mallat's multi-resolution analysis of $L_2(\mathbb{R}^2)$ [Mallat, 1989] has been used for the decomposition of images into wavelet sub-spaces. However, due to the uneven angular responses of this approach (*i.e.* strongly biased responses towards vertical and horizontal directions), a filter bank with an adaptive orientation property is used. This gives a composite image with the detailed information along the edges. The choice of appropriate wavelets is critical since structure information has to be preserved as much as possible in images with weak texture, especially at coarse levels. Moreover, at each pyramid level, the chosen wavelet transform has to attenuate high frequency noise (e.g. due to camera sensors) and, at the same time, preserve shape and edges of the structures in the images. Weakly pronounced edges and structures are usually not preserved by separable orthogonal wavelet decomposition, mainly because the corresponding filters are not rotation invariant (*i.e.* they are typically not optimal for representing oblique structure orientations). As justified in the next paragraph, the appropriate decomposition approach has to be multi-oriented.

The synthetic image of Fig. 3.1(a) was built with following objectives. Firstly, a white disc over a gray background is used to test the ability of filter banks to preserve edge structures with different local orientations. Secondly, the line edge of the square perimeter is used to test the performance of filter banks on thin line structures in the image at coarser levels. At first multi-scale and multi-oriented image decomposition was performed with the steerable filter banks [Simoncelli et al., 1992] popularly known as "Simoncelli's pyramid". Fig. 3.1(b) shows that the edge of the disc is blurred and the thin structure of the square is almost imperceptible. The result of an edge detection using canny filter on the image in Fig. 3.1(b) is shown in Fig. 3.1(e). The horizontal edge parts of the disc and almost the whole square structure are missing. Fig. 3.1(c) and Fig. 3.1(f) show the results obtained by applying the steerable pyramid proposed by Unser et al. [Unser et al., 2011] on Fig. 3.1(a). This pyramid is built with shift-, translation- and rotation-invariant tight wavelet frames with Riesz transform. In practice, our wavelet-based multi-resolution approach uses a Riesz Gaussian band-pass pyramid approach. For each pyramid level the original image is first convoluted with a "Difference of Gaussian" (DoG) kernel and steered using Riesz basis filters. Fig. 3.1(c) was obtained with a 1^{st}-order Riesz transform. Although the edge of the disc is less smoothed than in Fig. 3.1(b), its over-sized width after edge detection visible as the two concentric circles in Fig. 3.1(f)) is due to the ringing effect. The square detection also remains incomplete. In Fig. 3.1(d), obtained with a 2^{nd}-order Riesz transform shown in Fig. 3.2, the disc edge is sharper and leads to a complete circular edge in Fig. 3.1(g). In the latter, it is also noticeable that the complete square line borders are detected. To sum up, a steerable pyramid built with an orthonormal basis of 2^{nd}-order Riesz wavelets

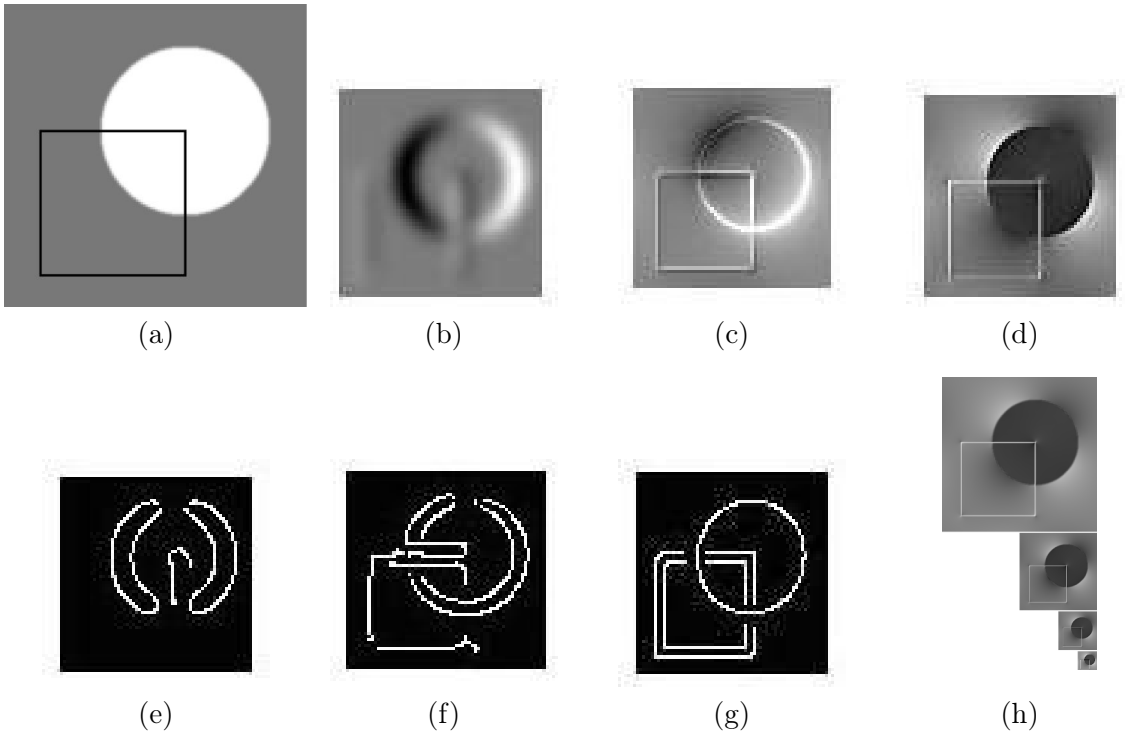


Figure 3.1: Effect of various orientation filters on a synthetic test image. (a) Original image (level 0). (b) Image (level 4 of a Gaussian pyramid) obtained after applying Steerable filters having an angular spacing of 15° in the basis filter. (c, d) Riesz filters of 1^{st}-order basis on Gaussian pyramid and Riesz filter with 2^{nd}-order basis filters on DoG pyramid respectively (for level = 4). (e-g) Edge detection on (b), (c) and (d) respectively. (h) 2^{nd} order Riesz wavelet pyramid.

will optimally preserve edges and ridges in the images. Such wavelet transform has notably the advantage of dealing with arbitrary rotations by forming a suitable linear combinations of its steering matrix unlike Simoncelli's basis which has equiangular orientation.

For a wavelet subspace V_j, the decomposition of image $I \in \mathbb{R}^2$ at scale j can be represented as the projection of I onto V_j with the Riesz basis filters $\mathcal{R}^N$ (see Fig. 3.2):

$$I^j = proj._{V_j} I = \sum_{k \in \mathbb{Z}^2} \langle \mathcal{R}^N \varphi_{j,k}, I * \mathcal{G}_j^{DoG} \rangle \hat{\varphi}_{j,k}, \tag{3.3}$$

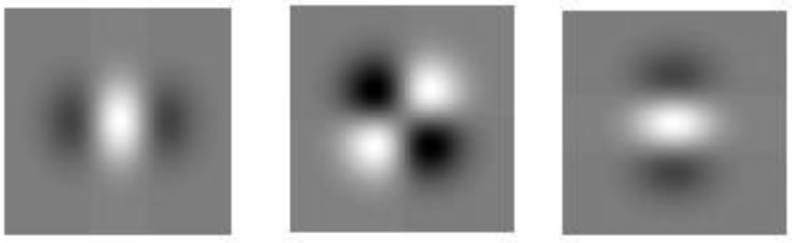


Figure 3.2: Representation of Riesz wavelet basis filters of order 2 ($N = 2$).

where $\mathcal{G}_j^{DoG}$ is the DoG kernel at scale j and shift k. A linear combination of an adaptive directional image is thus obtained using such filter-banks $\mathcal{R}^N$ for a given scale j. Unlike in [Unser et al., 2011], we have used a DoG kernel instead of a Gaussian low-pass kernel. This is done in order to avoid the loss of details in the coarser resolutions while attenuating very high frequency noise. The Riesz basis $\mathcal{R}^N\varphi_{j,k}$ is given by:

$$\mathcal{R}^N\varphi_{j,k} = 2^{-j}\mathcal{R}^N\hat{\varphi}_{0,k}(2^{-j}\mathbf{x} - k), \tag{3.4}$$

with shift $k = (k_x, k_y)$ along $x-$ and $y-$ axes. $\mathcal{R}^N\hat{\varphi}_{j,k}$ is the dual basis of $\mathcal{R}^N\varphi_{j,k}$. The coarse scale image at resolution j is obtained by isotropic band-pass filtering of the original image with the kernel $\mathcal{G}_j^{DoG}$ followed by the projection of such convoluted image on V_j with the steering Riesz basis $(\mathcal{R}^N\varphi_{j,k})$. For the order 2, the steering filter has a combination of $\{\cos^2\theta, \sqrt{2}\sin\theta\cos\theta, \sin^2\theta\}$ with θ computed as $\theta = \frac{1}{2}\arctan(\frac{2\cdot I_{xx}}{I_{yy}-I_{xx}})$. This will result in a steered downsampled images (I^j) at scale j. The fine scale coefficients $(j = 0)$ are first computed for initializing the wavelet coefficients at coarser level $(j = 1)$ and then the expansion coefficients are iteratively computed with successive downsampling by the scale factor of 0.7.

3.3 Anisotropic regularization

This section focuses on two key aspects of the algorithm: 1) the optical flow computation using an image-driven adaptive regularization scheme and 2) the correction of the determined optical flow field using a weighted median filtering. From Sections 3.2.1 and 3.2.2, the brightness constancy in Eq. (3.2) can be placed in the wavelet multi-resolution framework:

$$\rho(\mathbf{v}^j) = \nabla I_{i+1}^j(\mathbf{x}^j + \mathbf{v}^j).(\mathbf{v}^j - \mathbf{v}^{ap.}) + I_{i+1}^j(\mathbf{x}^j + \mathbf{v}^{ap.}) - I_i^j(\mathbf{x}^j) = 0, \tag{3.5}$$

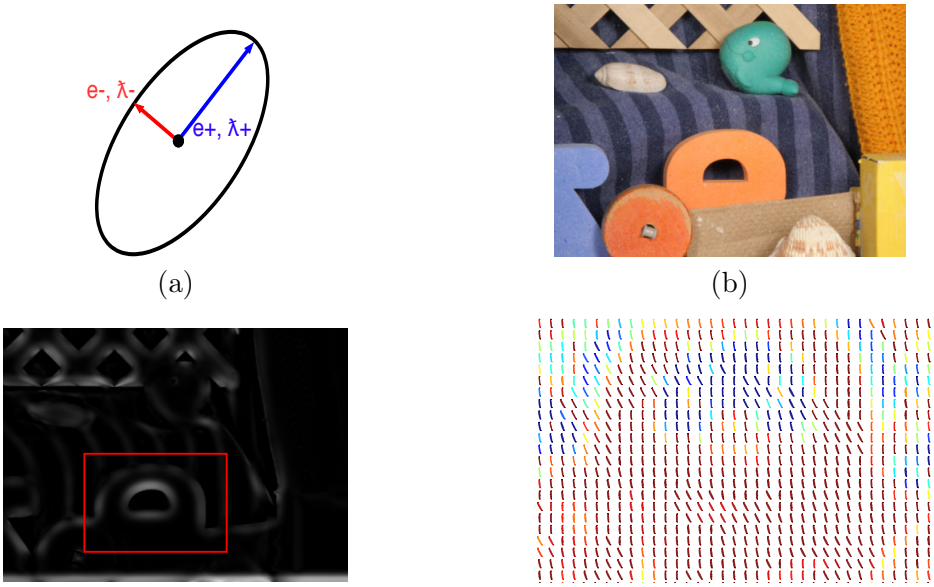
where, (I_i^j, I_{i+1}^j) are the approximation images obtained during decomposition by projection of the pair of original images (I_1^0, I_2^0) on the approximation subspace V_j with a Riesz basis. $\mathbf{v}^j$ is the flow field obtained at resolution j. The data term using L^1-norm $|\,.\,|_1$ can be written as:

$$E_D(\mathbf{v}^j) = \int_\Omega |\ \rho(\mathbf{v}^j)\ |_1\ d\Omega. \tag{3.6}$$

Classical TV- regularizer using L^1-norm is given by

$$E_S(\mathbf{v}^j) = \int_\Omega |\ \nabla\mathbf{v}^j\ |_1\ d\Omega. \tag{3.7}$$

The regularizer of Eq. (3.7) has two major drawbacks. First, during the energy minimization, there is no distinction between the contribution of edge pixels and pixels of homogeneous regions (like background) by the regularizer. Section 3.3.1 and 3.3.3 presents an anisotropic regularization solution in order to deal with such a problem. Second, the gradient in the regularizer of Eq. (3.7) can not distinguish between translations and in-plane rotations. In other words, pure rotations will be enforced as translations. This problem is addressed by Section 3.3.2.



(a) (b)

(c) (d)

Figure 3.3: Use of structure information for improved regularization. (a) Representation of the orthogonal major and minor eigenvalues (λ_+, λ_-) and eigenvectors (e_+, e_-) of a structure tensor computed with the grey-levels around $I^j_{i+1}(\mathbf{x} + \mathbf{v}^j)$. (b) Original image with rich texture. (c) Structure tensor image of (b) obtained as $\sqrt{{\lambda_+}^2 + {\lambda_-}^2}$ for each pixel $I^j_{i+1}(\mathbf{x} + \mathbf{v}^j)$. In this example, the background pixels are shown in black (small λ_+ and λ_-, i.e. large D_j in Eq. (3.9)) and the foreground pixels (structures) in white (large λ_+ and/or λ_-, i.e. small D_j in Eq. (3.9)). (d) Anisotropic diffusion tensor image based on the standard diffusion ellipsoids for visualization. It shows that the diffusion tensor acts differently around the torus region: the hollow of the torus (background) and the solid regions of the torus surface (foreground) have different diffusivity in terms of both orientation and their magnitude.

3.3.1 Tensor based anisotropic regularization

The structure tensor $S(\mathbf{v}^j, \mathbf{x})$ at pixel $\mathbf{x} + \mathbf{v}^j$ of the warped image I^j_{i+1} at scale j is a 2×2 positive semi-definite and symmetric matrix given by:

$$S(\mathbf{v}^j, \mathbf{x}) = \begin{pmatrix} \dfrac{\partial^2 I^j_{i+1}(\mathbf{x}+\mathbf{v}^j)}{\partial x^2} & \dfrac{\partial^2 I^j_{i+1}(\mathbf{x}+\mathbf{v}^j)}{\partial x \partial y} \\ \dfrac{\partial^2 I^j_{i+1}(\mathbf{x}+\mathbf{v}^j)}{\partial x \partial y} & \dfrac{\partial^2 I^j_{i+1}(\mathbf{x}+\mathbf{v}^j)}{\partial y^2} \end{pmatrix}. \tag{3.8}$$

Let $\lambda_+ \geq \lambda_-$ be the eigenvalues of $S(\mathbf{v}^j, \mathbf{x})$ and e_+ and e_- be their corresponding unit eigenvectors as represented in Fig. 3.3(a), then the structure diffusion tensor D^j at scale j [Tschumperlé and Deriche, 2003] can be written as:

$$D^j = \frac{1}{\sqrt{\epsilon + \lambda_+}}.e_+ \otimes e_+ + \frac{1}{\sqrt{\epsilon + \lambda_-}}.e_- \otimes e_-, \tag{3.9}$$

with $\epsilon \leq 0.001$ and where $\otimes$ is the tensor product. For pixels belonging to homogeneous regions both λ_- and λ_+ are small. In Eq. (3.9) this will result in strong values of diffusion tensor D^j.

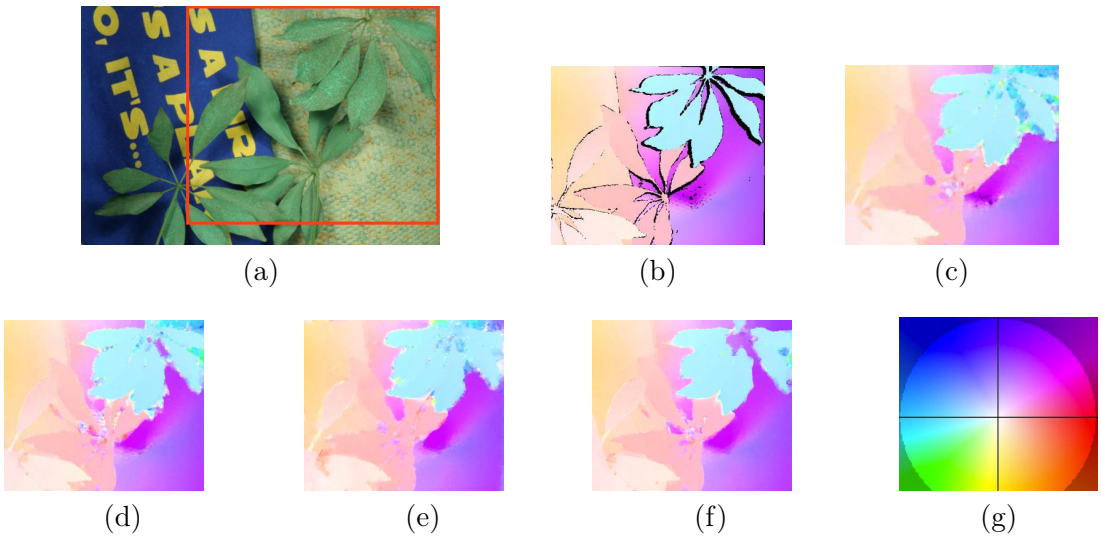


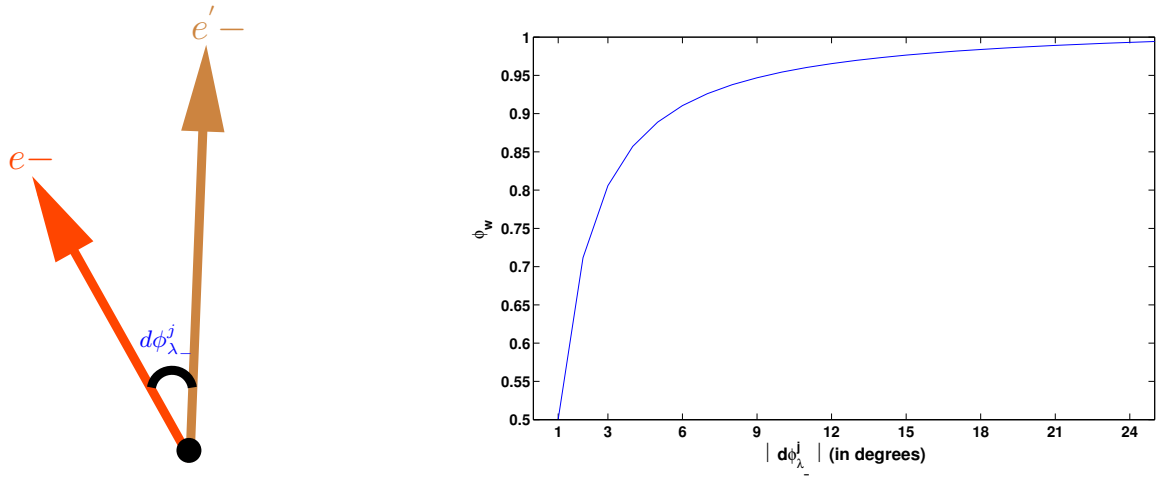
Figure 3.4: Visual representation of the effect of edge preserving anisotropic regularizer on the Schefflera image. (a) Original image. (b) Ground truth optical flow. (c) and (d) are results for improved TV-L^1 [Wedel et al., 2009b] without and with diffusion tensor respectively. (e) and (f) are the flow results for our TV-approach on wavelet space without and with diffusion tensor regularizer respectively. (g) Flow color code. Flow results are shown for red rectangular part in (a).

In contrary, edge pixels will lead to greater λ_+ and/or λ_- values resulting in small D^j values (bright pixels in Fig. 3.3(c)). Consequently, D^j can be used as a weight ensuring that only non-edge pixels are enforced for homogeneous vector field while the pixels belonging to edges do not contribute much to the regularization energy thus allowing for preserving edge discontinuities. D^j is thus used as a weight in Eq. (3.7) forming a modified TV-regularizer:

$$E_S(\mathbf{v}^j) = \int_\Omega \sqrt{D^j}. \mid \nabla \mathbf{v}^j \mid_1 d\Omega. \tag{3.10}$$

In Fig. 3.3(c-d), the effect of the diffusion tensor is shown on the RubberWhale image pair (see Fig. 3.3(b)). The structure tensor image of the RubberWhale in Fig. 3.3(c) represents the structure information variations in bright (edge pixels) and dark pixels (homogeneous or non-edge pixels). Fig. 3.3(d) represents the diffusion ellipsoids representing the background and foreground pixels along with their orientations and magnitude. The color represents the magnitude of the diffusion and the orientation of these ellipsoids represent the direction of the diffusion of the flow field. It is observable in Fig. 3.3(d) that due the motion of the torus region in the image pair of RubberWhale, the torus region has different color than the background pixels (see inside hollow region of torus) and the direction of motion is seen to be towards the left. The ellipsoid color in brown and red (observable for torus) represents larger motion relative to other colors (like blue, green and yellow).

The Schefflera images of the Middlebury database [Baker et al., 2011] were used to illustrate the effect of the new regularizer on the optical flow computation. This image contains thin structures, shadows and texture transitions with little contrast. The right hand side of the image (referring to the region inside the red rectangle in Fig. 3.4(a)) is difficult to handle by many flow algorithms due to small motion and similar texture colors between image background



(a) (b)

Figure 3.5: Computation of the curl-weight ϕ_w. (a) Angle difference $d\phi$ between the minor Eigenvectors e_- and e'_- of superimposed pixels $\mathbf{x}$ and $\mathbf{x}+\mathbf{v}$ located in the target image I_i and the warped source image I_{i+1} respectively. (b) ϕ_w as a continuous function of $\mid d\phi \mid$. The weight starts at 0.5 for angular difference of 1° which gradually increases in a non-linear fashion reaching the highest curl-weight of 1 for $\mid d\phi^j_{\lambda_-} \mid > 24°$.

and foreground [Baker et al., 2011]. It can be seen in Fig. 3.4 (c) and (d) that the reference method [Wedel et al., 2009b] has an over-smoothing effect on the foreground flow. However, little improvement is achieved when adapting the TV-regularizer of [Wedel et al., 2009b] with the proposed diffusion tensor. Figs. 3.4 (e) and (f) show large improvement in TV-L^1 model when a diffusion tensor on the Riesz wavelet space is used (refer to the top right part of the flow images). The use of diffusion tensor image on the Schefflera image pair shows improvement over optical flow result obtained with improved TV-L^1 method. As there is presence of shadows and similar textures on the background, the diffusion tensor plays an important role in discriminating different motion of the structures. It is observed in Fig. Figs. 3.4 (f) that the value of $\sqrt{D^j}$ gives improved weighting coefficients on Riesz wavelet space than the original image space. This is because such steering basis filters helps to make a better distinction between the background and the foreground pixels.

3.3.2 Div-curl decomposition of v

In optical flow estimation, the divergence is related to the rate of approach of the object like in case of scale changes whereas the curl mathematically describes the in-plane rotations of the camera [Suter, 1994]. For constant flow fields both the divergence and the curl are zero. When all the pixels have the same motion between two images then the displacement vectors are both divergence and curl free. However, such divergence and curl free motion are not systematic in scenes with fluid motion or in medical scenes like cystoscopy. In order to precisely model the behavior of the flow field, Eq. (3.7) is used to decomposed the smoothness function into a part related to the divergence and a part related to the curl.

$$E_S(\mathbf{v}^j) = \int_\Omega \{\mid \nabla div\ \mathbf{v}^j \mid_1 + \mid \nabla curl\ \mathbf{v}^j \mid_1\} d\Omega \tag{3.11}$$

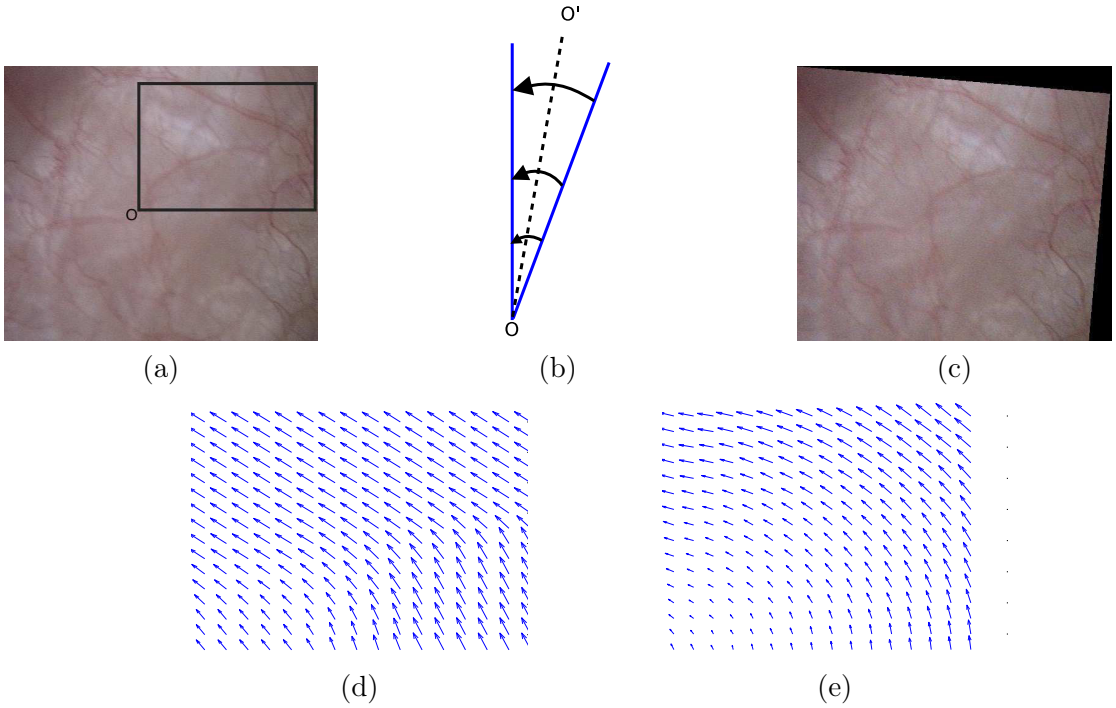


Figure 3.6: Significance of the curl operator in flow regularization energy $E_s(\mathbf{v}^j)$. (a) Original cystoscopic image. (b) Rotation around an axis being perpendicular to the image plane and passing through point O. (c) Image (a) after a 5° pure in-plane rotation. (d) Flow field obtained with the classical regularizer of Eq. (3.7). This flow field corresponds to the pixels of rectangle given in (a). The vectors are quasi-parallel with an over-estimated magnitude and false orientations. (e) Flow field obtained after applying div-curl decomposition (refer to Eq. (3.12)). The vectors correspond to a circular flow and have a magnitude which increased when moving apart from the point O.

However, such a decomposition has to be approximately tuned to ensure a meaningful impact of the divergence and the curl terms on the smoothness energy. To do so, the parameter ϕ_w weights the curl component of the energy $E_s(\mathbf{v}^j)$ in Eq. (3.11). ϕ_w is defined in Eq. (**??**) and computed as the measure of orientation difference of homologous vectors between the target and the warped source images. The homologous vector orientation difference is sketched in Fig. 3.5(a). From Eq. (**??**) and Fig. 3.5 (b), it can be seen that large angular misalignment of homologous Eigenvectors (i.e. large $d\phi_{\lambda\_}$) will lead to higher weight ϕ_w. Ideally, $\mid d\phi^j_{\lambda\_} \mid = 0$ when the two minor Eigenvectors are collinear (*i.e.* without in-plane rotation, there is no need for regularization of curl component as the displacement is curl-free). Moreover, since in most scenes, like in endoscopy, the in-plane rotation between the frames is usually restricted to a extreme value of 10°, ϕ_w will take values in the range $[0, 0.94]$ (see Fig. 3.5 (b)). For in-plane rotations below 1°, the curl weight is zero which means that the curl component is not considered in the regularizer for $\mid d\phi \mid <= 1°$. However, for other values a continuous non-linear increment in the curl weight can be observed in the Fig. 3.5 (b) for given angles $[1°, 24°]$. By combining the tensor diffusion parameter D^j and the weighted divergence and the curl terms, the final

regularizer is described by the energy defined by Eq. (3.12).

$$E_S(\mathbf{v}^j) = \int_\Omega \sqrt{D^j}.\{\mid \nabla div\ \mathbf{v}^j \mid_1 + \phi_w. \mid \nabla curl\ \mathbf{v}^j \mid_1\} d\Omega \tag{3.12}$$

As shown in Fig. 3.6, the classical TV-regularizer in Eq. (3.7) does not lead to the true orientation of the flow vectors (Fig. 3.6(d)) when it is applied to endoscopic image pairs with pure in-plane rotation (which happens often in endoscopic examinations and in other scene types). However, with the regularizer formulated in Eq. (3.12) the actual in-plane rotation field was observed in Fig. 3.6(e).

3.3.3 Weighted non-local median filtering

The flow field is refined after each source image warping in order to remove the outliers (vectors whose magnitude and/or orientation are very different from those of the neighbor vectors). These inaccurate flow vectors are mainly present at the structure edges as seen in the two rectangles of Fig. 3.3 (c). The filtering process is mathematically formulated by the L^1 minimization given in Eq. (4.7). The flow vector in pixel $\mathbf{x}$ is replaced by the vector leading to the smallest difference between the estimated flow field at $\mathbf{x}$ and the neighborhood flow field $\mathbf{v}'_\mathbf{x}$.

$$\min_{\mathbf{v_x}} \sum_{\mathbf{x}} \sum_{\mathbf{x}' \in \mathcal{N}_\mathbf{x}} w_\mathbf{x}^{\mathbf{x}'} \mid \mathbf{v_x} - \mathbf{v}'_\mathbf{x} \mid_1, \qquad \forall \mathbf{x}' \neq \mathbf{x}. \tag{3.13}$$

$w_\mathbf{x}^{\mathbf{x}'}$ is the weight assigned to each pixel $\mathbf{x}'$ in the neighborhood $\mathcal{N}_\mathbf{x}$ of pixels $\mathbf{x}$. The principle of the minimization is to adjust the weights $w_\mathbf{x}^{\mathbf{x}'}$ according to the "similarity" of pixels $\mathbf{x}$ and $\mathbf{x}'$. This similarity is defined in Eq. (4.8). A large weight value ensures that $\mathbf{v}'_\mathbf{x}$ has a high impact on the corrected value of $\mathbf{v_x}$ only when $\mathbf{x}'$ and $\mathbf{x}$ are effectively in same structures. Such an anisotropic smoothing term was first presented in [Li and Osher, 2009] for image denoising and later adapted for optical flow field filtering by Sun et al. [Sun et al., 2010, Sun et al., 2014]. In our approach however, we have defined weight $w_\mathbf{x}^{\mathbf{x}'}$ as the correlation entity based on the (i) spatial distance (the first exponential in Eq. (4.8)), (ii) color distance (the second exponential) and (iii) coherence measure of the structure tensor (the third exponential):

$$w_\mathbf{x}^{\mathbf{x}'} = e^{-|\mathbf{x}-\mathbf{x}'|^2/2\sigma_1^2}.e^{-|I(\mathbf{x})-I(\mathbf{x}')|^2/2\sigma_2^2}.e^{-|R_s(\mathbf{x})-R_s(\mathbf{x}')|^2/2\sigma_3^2}, \tag{3.14}$$

where σ_1, σ_2, and σ_3 are normalization factors empirically set to 5, 7 and 11 respectively for all tests in this thesis, $I(\mathbf{x})$ is the color vector and $R_s(x,y) = \left(\frac{\lambda_+ - \lambda_-}{\lambda_+ + \lambda_-}\right)$ is the geometrical similarity between the objects in $\mathbf{x}$ and $\mathbf{x}'$ and λ_+, λ_- are the major and the minor eigenvalues respectively computed from the structure tensor (refer to Fig. 3.3 (a)).

Weight $w_\mathbf{x}^{\mathbf{x}'}$ is at the highest value (≈ 1) when three conditions are simultaneously fulfilled: the distance between pixels $\mathbf{x}$ and $\mathbf{x}'$ is small, the color difference $|I(\mathbf{x}) - I(\mathbf{x}')|$ is weak and the pixels belong to the same structure (e.g. in an edge $R_s(\mathbf{x}) \approx R_s(\mathbf{x}')$). The optical flow field is refined in each warp using Eq. (4.7) and by calculating the weight $w_\mathbf{x}^{\mathbf{x}'}$ at each pixel $\mathbf{x}$ using the pixels $\mathbf{x}' \in (x', y')$ of the neighborhood of $N_\mathbf{x}$ centered at $\mathbf{x}$.

The advantage of the weighted median filtering is illustrated in Fig. 3.7 with the Rubber-Whale image of the Middlebury training dataset [Baker et al., 2011]. The image scene has large texture distribution with many shapes close to each other but with different displacements in the image pair. The minimization scheme with the proposed regularizer was able to compute accurate motion in almost all image regions without the weighted median filtering. However, this

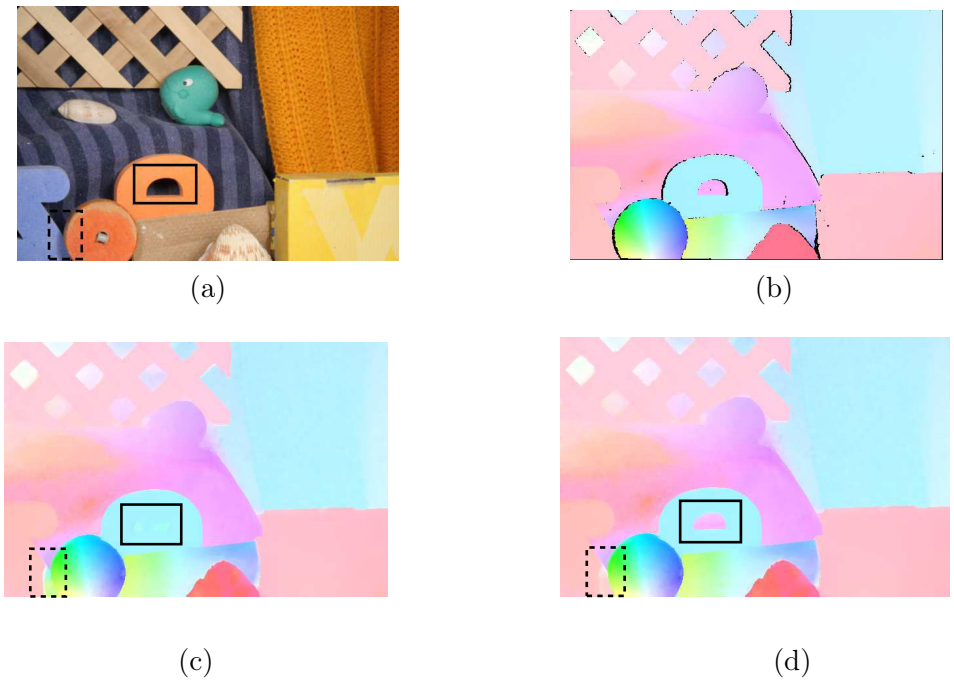


Figure 3.7: Illustration of the impact of the weighted median filtering on the flow field accuracy. (a) Highly textured RubberWhale image [Baker et al., 2011]. (b) Flow ground truth. (c) and (d) Flow estimation using the proposed model without and with weighted median filtering respectively. The rectangular boxes in black surround the regions of interest. In (d), it is visible that the effect of the shadow on the flow accuracy was attenuated (dashed line rectangle) and the torus hole is visible (solid line rectangle).

approach faced a major challenge inside the torus region in Fig. 3.7(c) where flow vectors are not accurate (background pixels surrounded by foreground pixels located in the drawn rectangles). In the shadow region and in the torus hole region, the flow fields were significantly improved by using the weighted median filtering with our additional coherence measure term in the weight $w_{\mathbf{x}}^{\mathbf{x}'}$ (see Fig. 3.7(d)). The motion estimation using the modified regularization steps is close to the ground truth flow when comparing Figs. 3.7(b) and 3.7(d).

3.4 Optimization

From Sections 3.2 and 3.3, the total variational energy minimization using L^1 optimization scheme can now be formulated as follows for each pyramid level:

$$E(\mathbf{v}^j) = \min_{\mathbf{v}^j} \Big[\int_\Omega \{ \mid \rho(\mathbf{v}^j) \mid_1 + \mid \gamma \mathbf{L}^j \mid_1 \} d\Omega + \lambda_s \int_\Omega \sqrt{D^j} . \{ \mid \nabla div \ \mathbf{v}^j \mid_1 + \mid \nabla div \ \mathbf{L}^j \mid_1 + \phi_w . \mid \nabla \ curl \ \mathbf{v}^j \mid_1 \} d\Omega \Big]. \tag{3.15}$$

In the first integral of the right hand term of Eq. (4.9), $\rho(\mathbf{v}^j)$ is the classical BCA defined in Eq. (3.5) with $\gamma \mathbf{L}^j$ as the illumination compensation term [Chambolle and Pock, 2011]. The second integral regularizes the flow field and λ_s is the trade-off between the data-term and the regularization term. In the frame of the proposed Riesz multi-resolution approach, the energy

$E(\mathbf{v}^j)$ is minimized using an iterative primal-dual approach in convex optimization [Chambolle, 2004] for a number of warps N_{warps}. The detailed steps for such primal-dual based energy minimization are presented in the Section 2.5.2 of Chapter 2 of this thesis. The computed flow field $\mathbf{v}^j$ during at each warp is used to displace the source image I_{i+1} to the target I_i.

3.5 Optical flow algorithm overview and parameter settings

This section first presents an overview of the proposed multi-resolution, anisotropic and edge preserving optical flow approach. The values of the most important algorithm parameters for obtaining an optimal compromise between robustness, accuracy and computational speed are then justified through an experimental analysis.

3.5.1 Algorithm overview

Algorithm 1: Anisotropic optical flow estimation on edge preserving wavelet (AOFW). Optimal parameter values are explained latter in this section.

Input : Image pairs, (I_i^0, I_{i+1}^0)
Output : Optical flow field $\mathbf{u}$
Parameters : $\alpha = 0.7$ (scale factor) , $N_{warps} = 2$, $N_{iter.} = 20$, $\lambda_s = 50$ (parameter in Eq. (4.9)), $\Delta = 0.001$
Initialization: $\mathbf{u}^0 = 0$

compute the number of levels N_{levels} with the chosen α
for *scale* $j = 0$ **to** N_{levels} **do**
  $I_i^j = \langle \mathcal{R}^N \varphi_{j,k}, (I_i^0 * \mathcal{G}_j^{bp}) \rangle . \mathcal{R}^N \hat{\varphi}_{j,k}$
  $I_{i+1}^j = \langle \mathcal{R}^N \varphi_{j,k}, (I_{i+1}^0 * \mathcal{G}_j^{bp}) \rangle . \mathcal{R}^N \hat{\varphi}_{j,k}$
for *level* $j = N_{levels}$ **to** 0 **do**
  $\mathbf{u}^j = \mathbf{u}^0$
  for $l = 1$ **to** N_{warps} **do**
    warp image I_{i+1}^j with flow field $\mathbf{u}^j$
    compute diffusion tensor D^j of I_{i+1}^j-warped with Eq. (3.9)
    compute curl weight ϕ_w^j between I_i^j and I_{i+1}^j-warped
    m=1
    do
      compute $\mathbf{u}^j$ using primal-dual energy minimization of $\mathsf{E}(\mathbf{u}^j)$ in Eq. (4.9)
      m=m+1
    while $((m \leq N_{iter})$ && $(\| \boldsymbol{u}^j - \boldsymbol{u}^0 \|^2 > \Delta))$
    Update $\mathbf{u}^0 = \mathbf{u}^j$
    perform weighted median filtering (see Eq. (4.7))
  if $j > 0$ **then**
    $\mathbf{u}^0 = \text{median}(\uparrow \circ \frac{\mathbf{u}^j}{\alpha})$
  else
    return opticalflow $(\mathbf{u} = \mathbf{u}^0)$

An overview of the complete optical flow minimization scheme is given in algorithm 1. The algo-

rithm starts by decomposing the images into N_{levels} using the method described in Section 3.2.2 and with a scale factor value $\alpha = 0.7$. Then for each scale j, from the coarsest to the finest level (*i.e.* $j = 0$ is the original scale) an iterative primal-dual energy minimization is done to solve the formulated convex problem in Eq. (4.9) with the redefined TV-regularizer as in Eq. (3.12). To do so, the optical flow between image pairs at each scale j are computed for N_{warps} warps ($N_{warps} = 2$ in our case). At each scale j, following steps are sequentially performed for each warping of source image I_{i+1}^{j} with the current optical flow $\mathbf{v}^{j}$:

- computation of both the diffusion tensors D^{j} and the curl weights ϕ_w for current warp number l,
- minimization of energy $E(\mathbf{v}^{j})$ in N_{iter} iterations or less iterations if convergence is reached (*i.e.*, $\Delta <= 0.001$) for current warp l and
- optical flow smoothing with the proposed weighted median filtering approach for flow preservation at edges.

The up-sampling of the optical flow is done to obtain the initial flow fields $\mathbf{v}^{j-1}$ at the finer scale $j-1$, followed by a 3×3 classical median filtering.

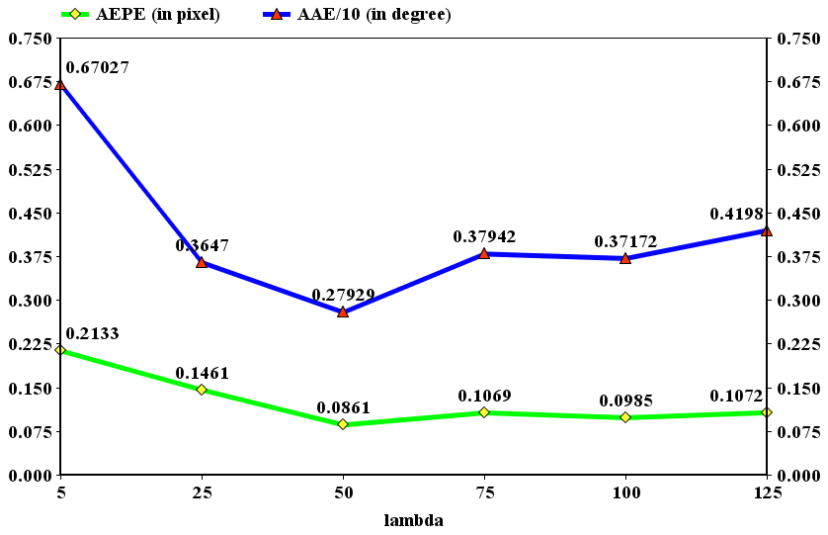
3.5.2 Parameter setting

The RubberWhale image pair of the Middlebury dataset [Baker et al., 2011] was chosen to optimally adjust the parameter values used in the proposed AOFW algorithm. This image pair has known ground truth flow field and possesses large variability in textures and small shadow regions. Average end point error (AEPE, in pixels) and average angular error (AAE, in degrees) are classically used to quantify the accuracy of the estimated flow field against the ground truth field. These accuracy criteria, together with the computation time, enable to find the best compromise between robustness, accuracy and speed of optical flow estimation. Among the important parameters to be adjusted, the weight λ_s (relative importance of data-term and regularizer, see Eq. 4.9) is directly related to algorithm robustness and accuracy, while the scale factor of the pyramid α has a strong impact on accuracy and speed.

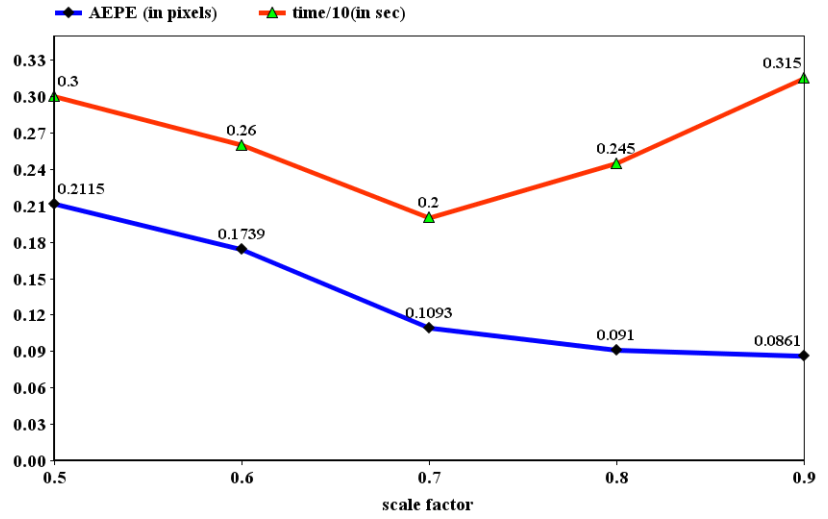
The smoothness weight λ_s in Eq. (4.9) plays an important role in the algorithm convergence. Relaxing the scale-factor α as 0.9 (*i.e.* a large number of pyramid levels minimizes the influence of α on the accuracy and robustness), the flow field $\mathbf{v}$ is estimated for different λ_s values as shown in Fig. 3.8(a). The results given in this Figure were obtained for the RubberWhale data. Both AEPE and AAE are lowest for λ_s equal to 50. There is a 3.7° deviation in AAE for both λ_s equal to 25 and 75. For lower and higher values of λ_s, both AAE and AEPE strongly increase.

For adjusting the optimal number of pyramid levels (*i.e.* the value of α), the optimal smoothness weight $\lambda_s = 50$ was chosen and kept constant for all tests. Referring to Fig. 3.8(b), the algorithm gives the worst result at scale $\alpha = 0.5$. This setting led to the lower computational speed than for $\alpha = 0.9$ but this performance was obtained at the expense of accuracy (high AEPE of 0.2115 pixels). The shortest computational time is obtained for α equal to 0.7. This is due to the fact at this scale the algorithm converges better towards the solution than at lower levels. At α equal to 0.7, AEPE is only 0.1093 pixels which is very close to AEPE values obtained at scale factors of 0.8 and 0.9 but with shorter computational time of 2 s (3.15 s for $\alpha = 0.9$ and 2.45 s for $\alpha = 0.8$).

A robustness test was also performed by running the algorithm with different λ_s on all the real training images of the Middlebury training dataset. It was experimentally found that the



(a)



(b)

Figure 3.8: Parameter settings using the RubberWhale image pair of the Middlebury training dataset. a) AEPE and AAE/10 for different values of λ_s giving the relative importance of the data-term and the regularizer. The tenth of AAE is plotted to represent both the AEPE and the AAE results on a unique decade of values. (b) AEPE and computational time plotted against the scale-factor α. The tenth of the computation time is plotted to represent both performance criteria on a unique decade of values.

overall AEPE/AAE of 0.15/2.95 ($pixels/°$) for λ_s equal to 50 was the most accurate result, while large increase in both AEPE and AAE values were recorded when choosing λ_s different from 50.

Thus, from the experimental results illustrated above, the parameter values for λ_s and α were fixed to 50 and 0.7 respectively. These settings result in an optimal compromise between accuracy and speed while enabling the AOFW algorithm to be robust against different scene variability. The algorithm efficiency using these adjusted (constant) parameter values have been assessed on a large variety of images in Section 3.6.

3.6 Results and discussion

Different datasets were used for the assessment of the algorithm performance: (i) the reference Middlebury database allows for the comparison of the AOFW method with other TV-L^1 optical flow methods, (ii) images acquired under controlled constant translation and constant in-plane rotation allowing for robustness check of the algorithm under various motion types and finally (iii) simulated video-sequences with known homographies are used for quantitative assessment in terms of the accuracy and registration speed of the AOFW method. During all these experiments, the AOFW method is compared with those of reference registration algorithms and robust optical flow techniques suited for image registration.

3.6.1 Evaluation of motion estimation on the Middlebury database

The ground truth information of the Middlebury database is used to test the optical flow accuracy in image regions with particular texture quality (weak textures, strong image blur, regions with shadows, etc.). For this database, the ground truth relates directly to the known optical flow field to be computed. The details on image acquisition and ground truth estimation are given in [Baker et al., 2011] for this database.

The Middlebury optical flow database was used in Sections 3.3 and 3.5 was used to show how the AOFW algorithm locally improves the optical flow in specific regions of the image. This section presents the quantitative results focusing on the Middlebury training dataset [Baker et al., 2011] images with hidden textures.

The AAE (in degrees) and AEPE (in pixels) measures are estimated to quantify the accuracy of the optical flow [Baker et al., 2011]. Table 3.1 shows that the proposed method is the most accurate according to both error criteria. It can be observed that the A-Huber-L^1 [Werlberger et al., 2009] method and the correlation flow [Drulea and Nedevschi, 2013] have performance similar to that of the proposed AOFW method. However, for similar accuracy, the energy minimization speed with our method is approximately 10 times faster than that of Werlberger et al. [Werlberger et al., 2009] and notable deviations in Dimetrodon and Venus image pairs are observed for correlation flow. The fastest of all the algorithms presented in Table 3.1 is the EPPM flow approach [Bao et al., 2014]. This method is almost 10 times faster than the AOFW method. However, this performance in terms of computation time is obtained at the expence of accuracy and robustness as can be observed in Table 3.1. The AAE for RubberWhale and Dimetrodon image pairs for the EPPM approach is comparatively the highest of all the methods with a difference of more than 3 degrees in each. However, since the EPPM is able to handle large displacements, we have included it in our test with phantom sequences-I and -II (displacements between 10 pixels to 300 pixels).

3.6.2 Evaluation of different motion types using an endoscopic set-up

Setup: The experimental set-up of Fig. 3.10(a) was used to acquire image sequences by displacing either an endoscope with a rotational stage or the flattened bladder phantom with two micro-metric stages. The endoscope axis is almost perpendicular to the phantom plane. The displacements of the two micro-metric stages mainly lead to changes of the translation parameters t_x and t_y in the homography matrix described in Chapter 1 and recalled in Eq. (3.16). The values of the other parameters of the homography are very weakly impacted by the displacement between consecutively acquired images. In a similar way, the rotational stage mainly impacts the in-plane rotation parameter (ϕ in Eq. (3.16)) and affects the other parameters less. One way

Algorithm	RubberW.	Hydra.	Dimet.	Venus
EPPM [Bao et al., 2014]	0.30/8.194	0.28/3.37	0.33/7.25	0.28/3.37
TV-L^1 [Pock et al., 2007]	0.13/4.23	0.19/4.20	0.24/4.56	0.43/6.44
I-TV-L^1 [Wedel et al., 2009b]	0.12/3.86	0.17/3.63	0.17/3.22	0.37/4.96
A-Huber-L^1 [Werlberger et al., 2009]	0.09/3.10	0.16/3.30	0.16/2.98	0.34/4.26
corr-flow [Drulea and Nedevschi, 2013]	**0.08**/**2.70**	0.17/**2.01**	0.21/4.54	0.26/4.02
Proposed (AOFW)	**0.08**/2.74	**0.15**/3.15	**0.15**/**2.70**	**0.23**/**3.22**

Table 3.1: AEPE/AAE (in pixels/in degrees) degrees) given for different TV methods applied on the Middlebury data-base. AAE and AEPE are mathematically defined in Chapter 2.

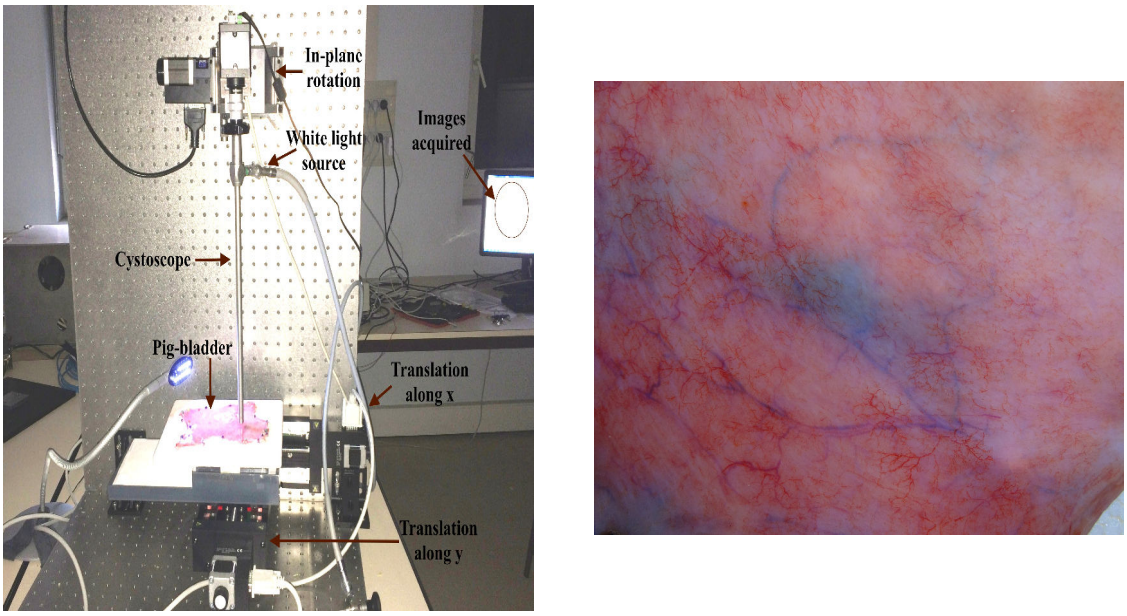


Figure 3.9: Acquisition of bladder phantom video-sequences with controlled displacements. (a) Experimental set-up for acquiring the images of a flattened-out pig bladder with an endoscope. (b) Image of the flattened out pig bladder.

to obtain GT for the reference matrix $H_{i,i+1}^{l,GT}$ linking image i and $i+1$ would be to match homologous points of these images (l refers to the local transformation between consecutive images). The known point correspondence given by the dense optical flow can be used to determine the 2D parameters of the homography corresponding to a given 3D rigid displacement of either the endoscope or the bladder phantom. In practice, this way to proceed does not lead to compute exactly known GT parameters since both small localization errors of the matched points and correction of the strong image distortion inherent to endoscopes affect the accuracy of the GT values of the homographies (distortion parameters are never exactly calibrated for cystoscopes [Miranda-Luna et al., 2004]). For these reasons, the t_x, t_y and ϕ values are not directly used as GT. Indeed, with the experimental set-up configuration of Fig. 3.10, the t_x and t_y parameters have to be rigorously constant when the micro-metric stages generate constant translations of the phantom (i.e. $\sqrt{t_x^2 + t_x^2}$ must be constant). Similarly, when only the rotational stage is used for generating constant rotation of the endoscope, the ϕ parameter keeps exactly the same value. In consequence, the GT information relates to the fact that the translation or rotation value variations must be null respectively for constant displacements or in-plane rotations of the

Method	$\sigma_{\sqrt{t_x^2+t_y^2}}$		σ_{ϕ°			
	$\approx$ **20**	$\approx$ **50**	$\approx$ **1°**	$\approx$**3°**	$\approx$**5°**	$\approx$**7°**
GC based [Weibel et al., 2012b]	0.37	0.25	**0.03**	0.22	0.75	0.77
TV-L^1 [Pock et al., 2007]	1.42	2.56	0.35	1.01	1.25	1.16
HAOF [Brox et al., 2004]	1.36	2.40	0.35	0.92	3.96	4.07
I-TV-L^1 [Wedel et al., 2009b]	0.86	1.27	0.10	0.23	1.03	2.49
Proposed (AOFW)	**0.32**	**0.13**	0.04	**0.20**	**0.27**	**0.33**

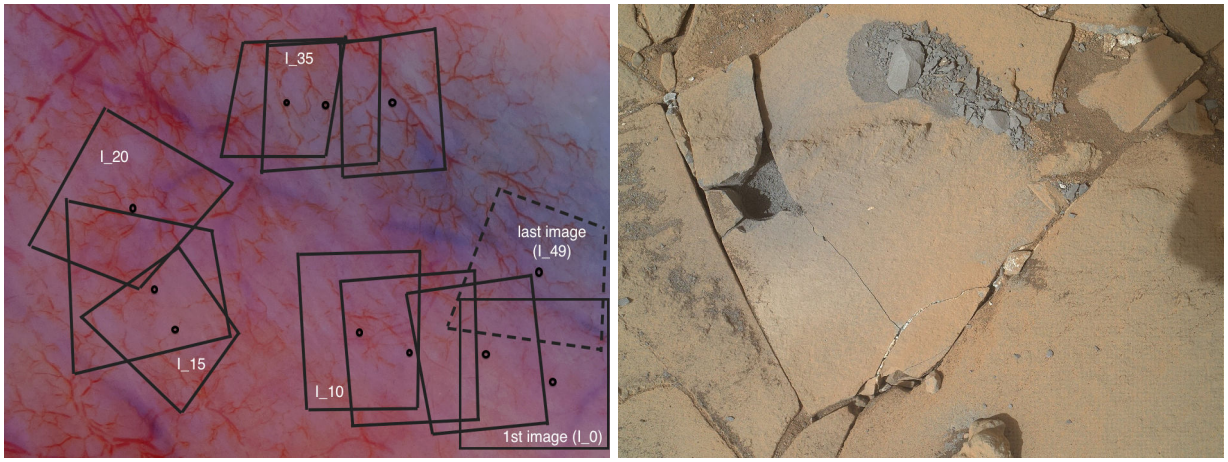
Table 3.2: Standard deviation in registration parameters of phantom sequence I. First two column presents the standard deviation from constant translation $\sqrt{t_x^2+t_y^2}$ for approximately 20 pixels and 50 pixels. Last four columns present the deviation from constant pure in-plane rotations for values 1°, 3°, 5° and 7°.

endoscope. These variations are measured through the standard deviations $\sigma_{\sqrt{t_x^2+t_y^2}}$ and σ_{ϕ° obtained respectively with constant displacements $\sqrt{t_x^2+t_x^2}$ and constant in-plane rotations ϕ.

Quantitative results: As justified, for constant micro-metric stage displacements or constant rotational stage movement, the variations of the translations and the in-plane rotations are ideally null. Table. 3.2 presents the estimate of standard deviation values for two translation magnitudes $\sigma_{\sqrt{t_x^2+t_y^2}}$ and four different in-plane rotations σ_{ϕ°. It has been shown in [Weibel et al., 2012b] that graph-cut based methods are robust to in-plane rotations and can handle large motions when implemented in coarse-to-fine approach. We have therefore used graph-cut based registration scheme for retrieving the transformation parameters. It can be observed that for constant translations of approximately 20 pixels and 50 pixels, the proposed method gives the smallest standard deviation of 0.32 and 0.13 pixels respectively. On the contrary, the classical TV-L^1 [Pock et al., 2007] approach has the largest standard deviation of 1.42 and 2.56 pixels for constant translation of nearly 20 and 50 pixels respectively. Similarly, for in-plane rotations the graph-cut [Weibel et al., 2012b] and the proposed AOFW method give the smallest standard deviations even for large angles of 7° (proposed: 0.33° and graph-cut based method [Weibel et al., 2012b]: 0.77°). Standard deviations of 3.96° and 4.07° for 5° and 7° respectively were noted for the HAOF method [Brox et al., 2004]. While this method lead to acceptable results for in-plane rotations up to 3°, it failed to estimate flow fields enabling the determination of accurate in plane rotations for $\phi \geq 5°$. This experiment thus validates that the proposed algorithm can handle large in-plane rotations and gives accurate alignment in large cystoscopic translations. It can be noticed that the graph-cut method is globally the second best method in terms of accuracy.

3.6.3 Evaluation with simulated homographies sparse textured scenes

The experiments in previous section with real (acquired) data allowed only for having ground truths of simple translations or in plane-rotations (no combination of different motion types). In this section we simulate sequences with more complicate displacements between consecutive image pairs. The ground truth information of the simulated video-sequences corresponds to homography parameters f, $(s_\mathbf{x}; s_\mathbf{y})$, ϕ, $(t_\mathbf{x}; t_\mathbf{y})$ and $(h_1; h_2)$ denoting the focal length, shearing factors, in-plane rotation, $2D$ translations and perspective changes respectively, as defined in



(a) (b)

Figure 3.10: Simulation of video-sequences with known displacements between consecutive images. Two scenes with very different textures are used for the simulation. The black rectangles sketch the extracted sub-images from the high resolution image. These extracted images (with known homographies linking them pairwise) simulate a video-sequence. b) Mars rover curiosity drill of the rock target for sample collection. (*Courtesy: NASA*).

Eq. (3.16). For a given pixel, w_i is defined by the perspective parameters h_1 and h_2.

$$\begin{pmatrix} w_i x_i \\ w_i y_i \\ w_i \end{pmatrix} = H_{i,i+1} \begin{pmatrix} x_{i+1} \\ y_{i+1} \\ 1 \end{pmatrix} = \begin{pmatrix} f cos\phi & -s_{\mathbf{x}} sin\phi & t_{\mathbf{x}} \\ s_{\mathbf{y}} sin\phi & f cos\phi & t_{\mathbf{y}} \\ h_1 & h_2 & 1 \end{pmatrix} \begin{pmatrix} x_{i+1} \\ y_{i+1} \\ 1 \end{pmatrix} \tag{3.16}$$

$H^l_{i,i+1}$ geometrically links consecutive images I_i and I_{i+1} of simulated video-sequences and are called "local" homographies, whereas "global" homographies $H^g_{0,i}$ place the pixels of images I_i in the coordinate system of the first image I_0 taken as mosaic start. Global matrices $H^g_{0,i}$ are the product of local homographies defined as,

$$H^g_{0,i} = \prod_{k=i-1}^{k=0} H_{i-k-1,i-k}. \tag{3.17}$$

Eq. (3.17) is used to place the pixels of each image I_i (with $i \in [1, N]$) of a video-sequence of N frames into the coordinate system of I_0. Below, both the simulation procedure and the evaluation criteria of the simulated sequences are presented.

Simulated datasets

Simulated sequence-I. A high resolution image of an incised and flattened out pig bladder was first acquired (see Fig. 3.10(a)). It was then used for simulating a video-sequence with complicated homographies (*i.e.* homographies with all parameters varying simultaneously). A sub-image I_0 of size 512×512 pixel size was first extracted from the high resolution image. Image I_0 acts as a reference frame in the mosaic to be built (see Fig. 3.10.(a)). Then, for $i + 1 \in [1, 50]$, the position of sub-image I_{i+1} was computed by applying a known local homography $H^{l,true}_{i,i+1}$ to I_i. Table 4.6 gives the homography parameter value interval used to generate randomly ground

Simulated sequence	ϕ°	f_x, f_y	s_x, s_y	h_1, h_2	$\sqrt{t_x^2 + t_y^2}$ **(pixels)**
I	$[-5, 5]$	$[0.95, 1.05]$	$[0.95, 1.05]$	$\pm\ 10^{-5}$	$[20, 250]$
II	$[-7.5, 7.5]$	$[0.90, 1.10]$	$[0.90, 1.10]$	$\pm\ 10^{-4}$	$[10, 300]$

Table 3.3: Homography parameter values for simulated sequences I and II. The RGB color channels are blurred for some images of simulated sequence-I. The value of the standard deviation σ_{blurr} of the gaussian function used for blurring [Chadebecq et al., 2012] is 2.5.

truth displacements. Some of the images were also blurred to simulate de-focusing/re-focusing of endoscopic examination.

Simulated sequence-II. A similar procedure was also used to build another video-sequence of a scene with different (complicated) texture characteristics than the previous medical scene. The high resolution image in Fig. 3.10.(b) is with relatively sparse texture and large illumination variability (presence of shadows) but with well contrasted regions. The 51 sub-images of size 400×400 pixels each are extracted from the high resolution image of a drilled rock surface taken by Mars curiosity rover. All homography parameters were simultaneously set to high values to simulate large in-plane rotations, scale changes, perspective changes and camera displacements between image acquisitions. The parameters value intervals of sequence II are also given in Table 4.6.

For both sequences I and II, the optical flow between consecutive images are computed with the proposed method. This dense homologous point correspondence is then used to estimate the local homographies $H_{i,i+1}^{l,est}$ which are ideally equal to the local ground truth homographies $H_{i,i+1}^{l,true}$.

Image registration evaluation on simulated video-sequences

The simulated sequences have been used for comparing quantitatively the accuracy of the AOFW optical flow method with the reference TV-L^1 methods and a graph-cut based method. In the literature of image mosaicing, the registration method based on graph-cuts has been experimentally justified to be the most accurate and robust method [Weibel et al., 2012b] for low texture bladder video sequences (represented by sequence-I) while TV-L^1 based image registration techniques have also been successfully used for various complicated scenes [Ali et al., 2013b, Pock et al., 2007, Brox et al., 2004]. The proposed AOFW method is also compared with a recent illumination independent and robust TV-L^1 based approach [Drulea and Nedevschi, 2013]. Additionally, a recent self-similarity based approach with patch matching concept has also been included in our evaluation tests [Bao et al., 2014]. Even though this method gives local solution, an analysis of its robustness for rapid registration of the scene is interesting. The main advantage of this approach is its computational efficiency over global methods used in this thesis.

Parameter settings and evaluation criteria: The parameter settings of the reference methods are presented in Table 3.4 while those of the proposed AOFW algorithm are given in Algorithm 1. The parameters of the EPPM algorithm [Bao et al., 2014] were set to the defauts values provided in the executable file downloadable at `https://sites.google.com/site/linchaobao/home/eppm`. All parameters are kept constant for all simulated and real video sequences presented in this chapter.

Both the local $H_{i,i+1}^{l,true}$ and the global $H_{i,i+1}^{g,true}$ homographies are known for simulated

Parameters	λ_s	γ	τ	θ	λ_p	α	ϵ
Description	regularizer weight	constancy weight	time step	tightness	planarity weight	scale factor	stopping criteria
TV-L^1 [Pock et al., 2007, Wedel et al., 2009b]	50	NA	0.2	0.3	NA	0.9	0.001
corr-flow [Drulea and Nedevschi, 2013]	20	NA	0.2	0.3	NA	0.7	0.001
HAOF [Brox et al., 2004]	0.2	50	0.1	NA	NA	0.9	NA
GC [Weibel et al., 2012b]	1.5	NA	NA	NA	10	0.7	NA

Table 3.4: Parameter settings for the optical flow determination. The dense point correspondence is used to determine the homography parameters. *NA stands for not applicable parameter to a method.*

video-sequences I and II. The optical flow between images I_i and I_{i+1} was estimated with the HAOF method [Brox et al., 2004], the classical TV-L^1 method [Pock et al., 2007], the improved TV-L^1 method [Wedel et al., 2009b], the graph-cut based approach [Weibel et al., 2012b] and the proposed AOFW algorithm. The pixel correspondence given by these flow vectors are used to estimate the local homographies $H_{i,i+1}^{l,est}$ as described in [Ali et al., 2013b]. Global transformations were computed by replacing in Eq. (3.17) the $H_{i,i+1}^{l,true}$ matrices by the estimated $H_{i,i+1}^{l,est}$ homographies. The registration errors ($\epsilon_{i,i+1}$ in pixels) between images I_i and I_{i+1} were computed with Eq. (3.18):

$$\hat{\epsilon}_{i,i+1} = \frac{1}{50}\sum_{i=0}^{i=50}\epsilon_{i,i+1}, \text{ with } \epsilon_{i,i+1} = \frac{1}{\mid I_i \cap I_{i+1} \mid}\sum_{p\in I_i\cap I_{i+1}} \parallel H_{i,i+1}^{l,true}p - H_{i,i+1}^{l,est}p \parallel^2, \tag{3.18}$$

where p stands for the pixels in image I_{i+1} placed onto the image I_i, $H_{i,i+1}^{l,true}p$ and $H_{i,i+1}^{l,est}p$ are respectively the exact and estimated placements in I_i, and $\hat{\epsilon}_{i,i+1}$ is the mean registration error computed for the 51 image pairs of the simulated sequences. The mosaicing error given in Eq. (3.19) relates to the placement error $\epsilon_{0,50}$ of image I_{50} into the coordinate system of image I_0.

$$\epsilon_{0,50} = \frac{1}{\mid I_0 \cap I_{50} \mid}\sum_{p\in I_0\cap I_{50}} \parallel H_{0,50}^{g,true}p - H_{0,50}^{g,est}p \parallel^2 . \tag{3.19}$$

The registration error $\hat{\epsilon}_{i,i+1}$ quantifies the mean error when placing a pixel from the coordinate system of I_{i+1} in that of I_i. Error $\epsilon_{0,50}$ corresponds to the Euclidean distance between the ground truth pixel positions $H_{0,50}^{g,true}p$ and the estimated pixel positions $H_{0,50}^{g,est}p$.

Evaluation on simulated sequence-I (bladder epithelium): The mean error values $\hat{\epsilon}_{i,i+1}$ presented in Table 3.5 show that the proposed method outperforms all the reference methods under "no blur condition" (*i.e.* with no additional blur superimposed on the images). The local mean registration error in 50 image pairs ($\hat{\epsilon}_{i,i+1}$) is 0.12 pixel for the proposed AOFW algorithm and the related standard deviation ($\hat{\sigma} = 0.1$ pixels) remains very small. Both the mean registration errors and the standard deviations are larger for all other methods, which means that they are less accurate in the tested simulated image sequence-I.

Fig. 3.11.(b-e) shows the mosaicing error along the loop closing of the simulated sequence I for different methods (Classical TV-L^1 [Pock et al., 2007], graph-cut based [Weibel et al., 2012b], improved TV-L^1 [Wedel et al., 2009b] and the proposed method). As indicated in Table 3.5, the final global mosaicing error $\hat{\epsilon}_{0,50}$ is 26.5 pixels for the graph-cut method (visually illustrated in Fig 3.11(c)). Except for the classical TV-L^1 method [Pock et al., 2007], the other tested

Method	$\hat{\epsilon}_{i,i+1} \pm \hat{\sigma}$	$\hat{\epsilon}_{i,i+1} \pm \hat{\sigma}$	$\epsilon_{0,50}$		$\bar{t}$ (s.)
	$\sigma_{blurr} = 0$	$\sigma_{blurr} = 2.5$	$\sigma_{blurr} = 0$	$\sigma_{blurr} = 2.5$	
GC based [Weibel et al., 2012b]	0.54 ± 0.35	$1.02 \pm \mathbf{0.8}$	26.5	56.7	25
TV-L^1 [Pock et al., 2007]	1.23 ± 1.3	4.23 ± 3.21	61.57	210.9	7
HAOF [Brox et al., 2004]	0.45 ± 0.36	2.62 ± 1.3	22.6	130.70	12
I-TV-L^1 [Wedel et al., 2009b]	0.40 ± 0.3	2.15 ± 1.21	19.24	107.9	8
corr-flow [Drulea and Nedevschi, 2013]	0.33 ± 0.21	1.27 ± 1.03	16.81	68.56	5
EPPM [Bao et al., 2014]	0.24 ± 0.19	0.87 ± 0.90	10.55	**35.28**	**0.23** (GPU)
Proposed (AOFW)	$\mathbf{0.12} \pm \mathbf{0.10}$	$\mathbf{0.86} \pm 0.92$	**5.9**	42.84	2.5

Table 3.5: Mean registration ($\hat{\epsilon}_{i,i+1}$) and mosaicing ($\hat{\epsilon}_{0,50}$) errors in pixels for 50 image pairs of the simulated video sequence-I. Bladder simulated video sequences without ($\sigma_{blurr} = 0$) and with ($\sigma_{blurr} = 2.5$) additional Gaussian blur (depicting defocus/refocus in cystoscopy [Chadebecq et al., 2012]) are used for quantifying the robustness of the methods. Mean registration time for image pairs with size of 512×512 pixels are also presented (CPU implementation time for images under no blur).

TV-L^1 methods led to more accurate results than the graph-cut method: their mosaicing error values $\hat{\epsilon}_{0,50}$ are 22.5, 19.24 and 16.81 pixels for the HAOF [Brox et al., 2004], the I-TV-L1

[Wedel et al., 2009b] and the correlation flow [Drulea and Nedevschi, 2013] methods respectively. Largest misalignment was noted for classical TV-L^1 method [Pock et al., 2007] as shown in Fig. 3.11 (b) ($\hat{\epsilon}_{0,50} \approx 62$ pixels). It is to be noted that the proposed method has the smallest global accumulated error when no blur is added to the images: $\hat{\epsilon}_{0,50} \approx 6$ pixels. In this test, since the mosaicing error is distributed over the sequence of 50 image pairs, it does not have perceptible visual misalignment at the loop closing (see Fig 3.11 (a) and (e)).

Among the 51 images of simulated video-sequence I, 7 images were randomly chosen for convolution with a Gaussian Kernel [Chadebecq et al., 2012]. This image blur was applied to test the robustness of the algorithms in the case of image focusing/de-focusing. As seen in Table 3.5, the mean registration error $\hat{\epsilon}_{i,i+1}$ was still the smallest (0.86 pixels) for the proposed method. However, the standard deviation of 0.92 for our method is slightly larger than that of the graph-cut method (0.8 pixels). An effect of robustness towards blur is seen on the mosaicing errors in the methods. Most of the bladder images are typically without strong de-focusing blur. For such images, the proposed method is by far the most accurate. For the pairs with at least one blurred images, the graph-cut method is the most robust in the sense that the registration accuracy is not overshooted by blur. This is confirmed by the large increase of standard deviation values up to 3.21, 1.30 and 1.21 pixels for the classical TV-L^1, I-TV-L^1 and HAOF respectively. In all cases, the I-TV-L^1 and the HAOF methods were among the less accurate methods, while TV-L^1 was the worst with 210.9 pixels of mosaicing error. Such large misalignment results lead to incoherent mosaic. In this case, subsequent map correction using bundle adjustment technique as proposed in Weibel et al. [Weibel et al., 2012b] becomes infeasible and time consuming. However, this is

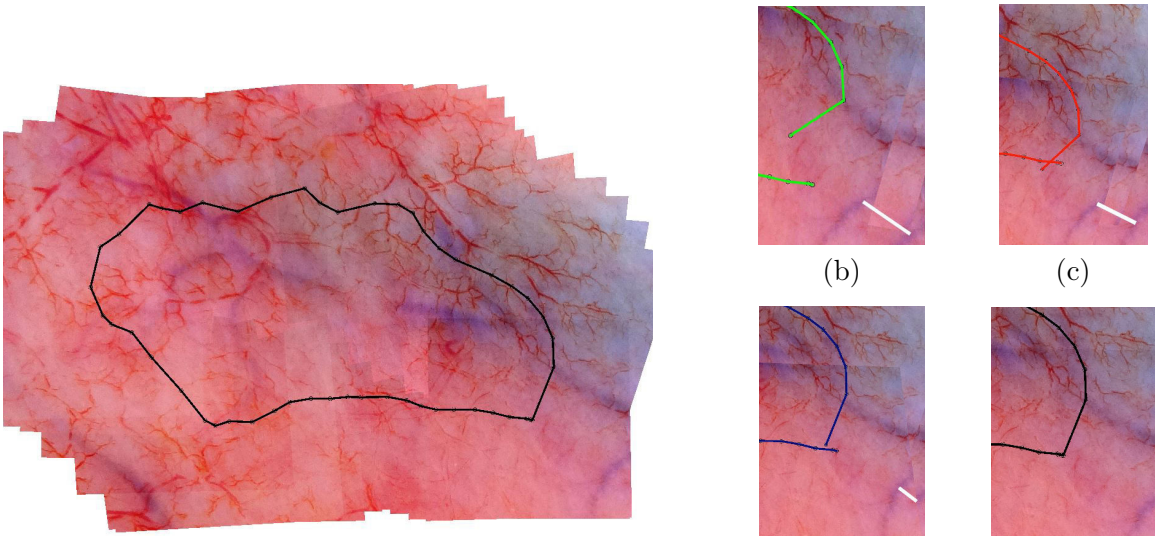


Figure 3.11: Mean mosaicing error evolution when placing the pixels of image I_i in the coordinate system of image I_0. (a) Pig bladder mosaic built with the proposed AOFW algorithm using simulated sequence I. The trajectory of the simulated image center path is shown in black. (b-e) Detailed view at the start and end of the loop (path closing) for the classical TV-L^1 (in green) [Pock et al., 2007], the graph-cut based method [Weibel et al., 2012b] (in red), the improved TV-L^1 [Wedel et al., 2009b] (in darkblue) and the proposed method (in black) respectively. Visual misalignment is also perceptible along the C-shaped vessel structure indicated by a white line. This error is visually imperceptible for the proposed method in (e).

Method	$\epsilon_{i,i+1}^{local}$ (in pixel)			$\epsilon_{0,49}^{global}$	$\bar{\mathbf{t}}$
	min	max	mean	(in pixel)	(in s)
Graph-cut method [Weibel et al., 2012b]	0.74	38.55	7.17	430.35	20
TV-L^1 [Pock et al., 2007]	0.11	37.9	4.66	270.65	7
HAOF [Brox et al., 2004]	0.09	19.23	3.62	201.75	9
I-TV-L^1 [Wedel et al., 2009b]	0.14	17.32	3.37	168.23	11
EPPM [Bao et al., 2014]	0.19	6.75	2.23	109.26	0.35 (GPU)
Corr-flow [Drulea and Nedevschi, 2013]	0.17	3.52	1.40	68.61	5
Proposed (AOFW)	**0.03**	**0.58**	**0.19**	**9.11**	**3**

Table 3.6: Errors obtained for the methods compared on simulated sequence-II. Mean registration time of image pairs with size 400×400 pixels for CPU implementation are also given unless GPU mentioned.

feasible for our proposed method, the correlation flow and the graph-cut based method.

However, the performances of the graph-cut based method and the proposed method are different in terms of computational speed. The reference graph-cut method and the proposed method were all implemented in C++ and run on 64-bit windows computer equipped with an Intel core 2 Quad at 2.83 GHz processor. From Table. 3.5, an average time of nearly 25 s was noted for registering image pairs with the graph-cut based method. This is the highest registration time among other reference methods. For our proposed method a lowest registration time of 2.5 s was recorded. The GPU implementation of the EPPM method recorded the least computational time of only 0.23 s per image pair.

Evaluation on simulated sequence-II (Rock image): In the simulated video-sequence II, the data consist of large homogeneous regions along with shadows in some images. Additionally, strong geometric transformations are used in this video simulation. It can be observed in Table 3.6 that the graph-cut based method is not robust enough to register all image pairs (a maximum local registration error of 39 pixels can be interpreted as a registration failure leading to a visually incoherent mosaic and a high global error of 430.35 pixels). The classical TV-L^1 [Pock et al., 2007] and the graph-cut based methods [Weibel et al., 2012b] are both not robust when image pairs are affected by large illumination variability. This sensitivity to illumination changes is due to the fact that these two methods minimize globally the sum of squared intensity differences, even if brightness constancy is not fulfilled. However, with the pre-processing step included in the improved-TV-L^1 [Wedel et al., 2009b] and a complementary gradient constancy term included in the cost function of HAOF [Brox et al., 2004], the maximum local error and the global error are reduced. The patch matching based approach (EPPM) [Bao et al., 2014] gave relatively robust results for image pairs with illumination changes but recorded larger errors in image pairs with large homogeneous regions. A maximum error of 6.75 pixels between image pairs was recorded for this method. The local errors between image pairs led to a large global registration error of 109.26 pixels in the image sequence II. Furthermore, using cross-correlation approach in the TV-L^1 framework [Drulea and Nedevschi, 2013], more promising results were obtained with mean local and global registration errors of only 1.40 and 68.61 pixels respectively. A closer look at the trajectory path in Fig. 3.12(b) reveals its robustness. For most of the image pairs, the trajectory of the correlation flow (in blue) is close to the ground truth trajectory (in green). The proposed AOFW method, however, is the most accurate among all the other methods. It records a global mosaicing error of only 9.11 pixels due to sub-pixel accuracy of the local registration errors in image pairs. In Fig. 3.12(b), the trajectory path of the proposed method (in magenta)

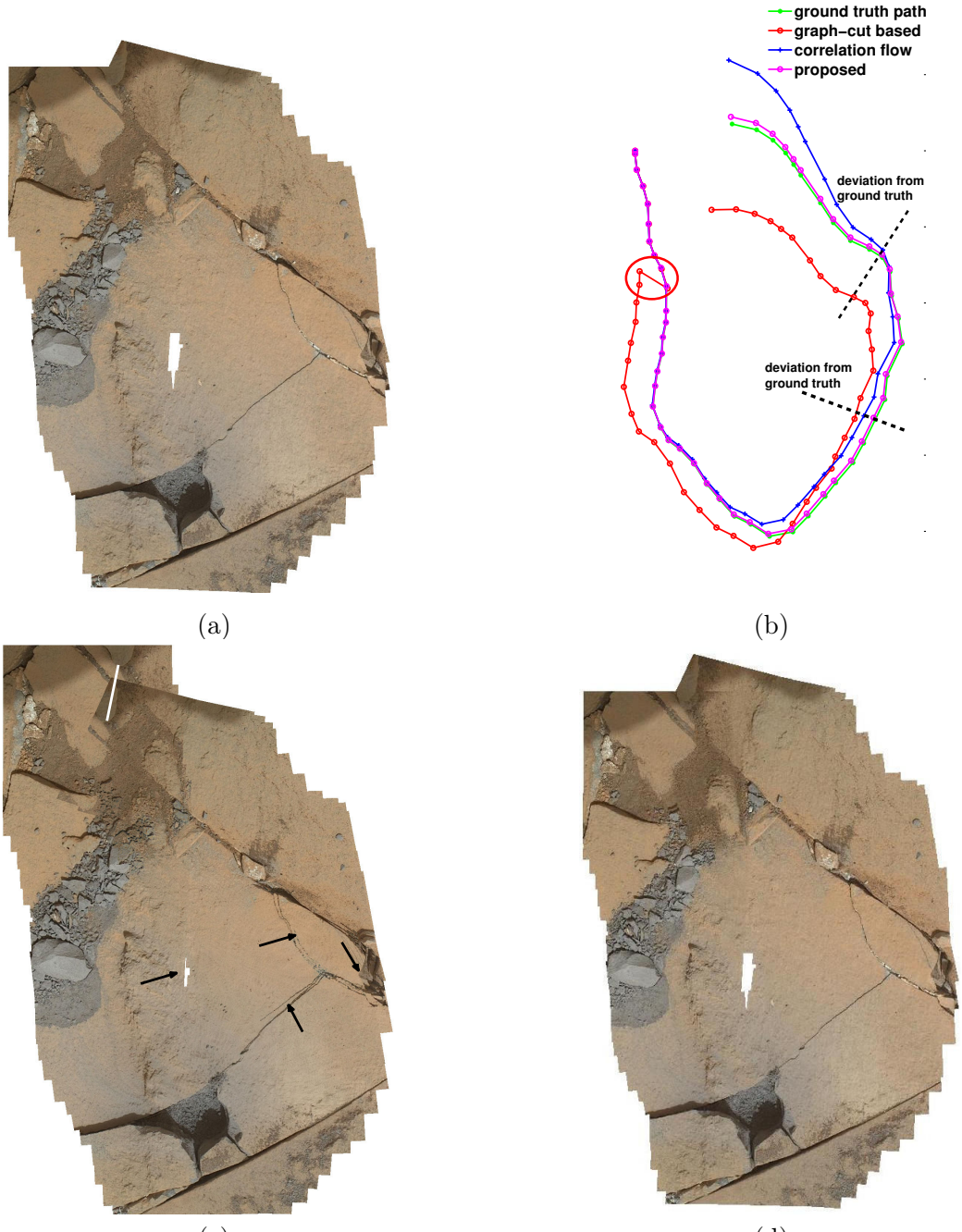
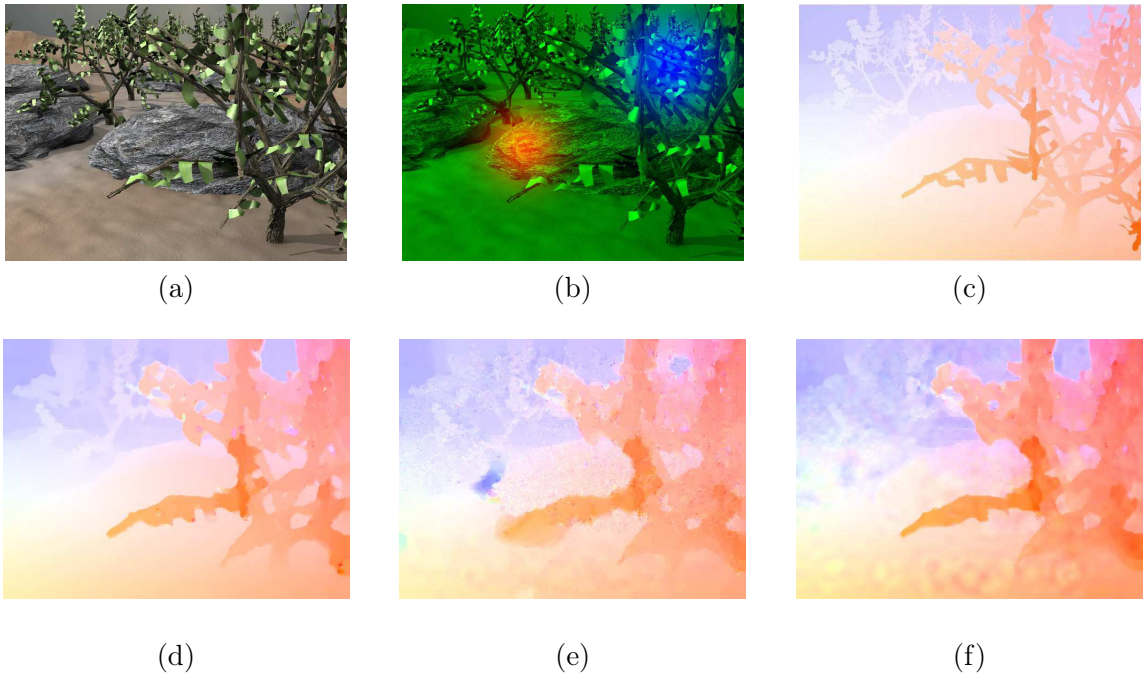


Figure 3.12: Experiments on simulated video-sequence II. a) Mosaic computed with the ground truth homographies. b) Path trajectory of the mosaics computed with the homographies obtained for the reference methods and the ground truth homographies. (c, d) Mosaics with the classical TV-L^1 approach and the proposed AOFW method respectively. The visual misalignments are indicated by arrows and a solid white line in mosaics (c) and (d). Comparing the shape and size of the white "holes" in the mosaic centres shows also that the shape of mosaic (d) is closer to the ground truth shape than that of the map in (c). Misalignment for each method can also be observed at each trajectory point in (b).



(a) (b) (c)

(d) (e) (f)

Figure 3.13: Visualization of the effect of strong illumination changes on on the Grove3 sequence of the Middlebury training dataset. (a) Original frame10 of Grove3. (b) Frame11 of Grove3 with strong simulated illumination changes [Drulea and Nedevschi, 2013]. (c) Ground truth flow field. (d) Optical flow obtained with the AOFW method on images without modified illumination. (e) Optical flow obtained with the AOFW method on images with the modified illumination (frame 11 as in (b)). (f) Optical flow obtained with the RFLOW method on images with modified illumination (frame 11 as in (b)).

is very close to the ground truth path in the whole mosaic, while the path computed with other methods deviates strongly after the registration of a given number of image pairs (see the doted line regions in Fig. 3.12.(b)). It is also evident from Fig. 3.12(d) that the mosaic obtained by the proposed method is visually coherent due to small accumulation of local errors. However, in Fig. 3.12(c), several misalignment are locally visible in mosaic regions which are indicated by solid black arrows and by a solid white line (accumulated error causing large visible misalignment between the first and the last frames).

The average computation time for image pair registration for this dataset is also presented in Table 3.6. It can be observed that the proposed method has simultaneously the best results in terms of accuracy and speed. While the graph-cut based method is the least robust (see the registration failure indicated by the red circle in Fig. 3.12.(b)) and also the most computational expensive algorithm for this dataset, the correlation flow method is second best in terms of accuracy and speed.

3.6.4 Robustness of algorithm against strong illumination changes

The proposed models (RFLOW in Chapter 2 and AOFW in this chapter) give accurate results for scenes with different textures and small illumination changes (such as small shadow regions). However, these methods have not been tested so far for strong illumination variations (local and/ global illumination changes). The second image I_2 of the Grove3 image pair of the Mid-

Method	Grove3	Grove3 (+illumination change)
Proposed model-I (RFLOW)	0.64/6.5	1.02/13.44
Proposed model-II (AOFW)	0.63/5.94	0.99/10.4

Table 3.7: Error quantification (EPE/AAE in pixels and degrees respectively) for the Grove 3 image pair with and without illumination change.

dlebury training dataset (known ground truth) is used to simulate non-uniform illumination. The illumination changes on the second image was simulated with the method used in [Drulea and Nedevschi, 2013]. The intensities of the red and the blue channels are separately treated producing a green image with two spots which are perceptible when comparing the non modified frame 10 of the the Grove3 sequence (see Fig. 3.13(a)) and the illumination modified frame 11 (given in Fig. 3.13(b)). The image is then normalized between 0 and 1. These changes lead to both local and global illumination difference. It can be visually verified from Figs. 3.13(e-f) that both the proposed models (RFLOW and AOFW) are affected largely due to such changes. Results obtained with the AOFW method for data without and with illumination changes are shown in Fig. 3.13(d) and Fig. 3.13(e) respectively. This highlights the effect of strong illumination changes on the proposed methods. It can be noticed that the optical flow field in Fig. 3.13 (e) with illumination is significantly degraded with edges being oversmoothed and visually observable errors around the circular bright spots. However, the flow field obtained using image pair without such strong illumination changes in Fig. 3.13 (d) is more close to the ground truth optical flow shown in Fig. 3.13 (c). Similarly, large and visually noticeable errors can be observed for RFLOW method when compared with the ground truth optical flow for image pair with such large illumination variation.

It is numerically confirmed in Table. 3.7 that the proposed methods are not suitable for strong illumination changes. Indeed, Table. 3.7 shows that the AEPE/AAE (pixel/degrees) significantly increase due to the simulated illumination changes. In short, the proposed models can be used without loss of accuracy in image pairs with small illumination changes like small shadow regions but they are not robust to strong illumination changes as the one simulated here on frame11 of Grove3. Both the error criteria AEPE and AAE increased almost by twice in case of both the proposed methods.

3.7 Main contributions and conclusion

Following summerizes the contributions of this chapter:

- Modeling of a data-cost on a Riesz wavelet scale-space which can handle the "flattening-out" problem prevalent at coarse pyramid levels in a multi-resolution energy minimization framework.
- Anisotropic regularization using a soft constraint on the smoothness term (use of both divergence free and curl free components) of the energy minimization function in the TV-L^1 framework. Such a modification enables to give correct flow vector orientations without constraining them strictly to the piecewise smooth constraint.
- Non-local weighted median filtering with embedded structure coherence measure in the classical bilateral filtering technique. Such a term has small but useful increase in accuracy.

- Tests on the Middlebury data set (images with very different textures and hidden textures) were done. The results suggest an improvement in optical flow accuracy in comparison to many baseline TV-L^1 methods.
- Experiments were conducted 1) on images acquired for a flattened pig-bladder on the metric table with controllable in-plane rotation and translations and 2) on simulated sequences with known homographies. For both experiments, the images were low textured and with small non uniformity in the illumination pattern.
- The proposed algorithm was quantitatively validated using both local errors (pairwise registration errors) and global errors (mosaicing errors) for the simulated low textured sequences.

In this chapter, the improvement of optical flow accuracy by preserving image structures at all levels of a multi-resolution pyramid framework was shown. Additionally, an anisotropic based regularization in the proposed model contributed to accurate optical flow results. As summarized above, the experiments performed in this chapter show the efficiency of the proposed model in handling both high and low textured image pairs/sequences. The proposed algorithms is able to handle small blur and small illumination changes such as small shadow regions. This case often arise in realistic scenes. However, a major limitation of the presented RFLOW (in Chapter 2) and AOFW (presented in this chapter) methods is that they are unable to handle strong illumination changes as presented in Section 3.6.4. An illumination invariant model is important and interesting for the framework of this thesis. The most important motivation of such an investigation is the ability to use the optical flow algorithm for modality independent cystoscopic image mosaicing. This objective has been achieved during the thesis duration and has been detailed in Chapter 4.

List of publications

[AGDB14] Sharib Ali, Ernest Galbrun, Christian Daul, François Guillemin and Walter Blondel "Anisotropic motion estimation on edge preserving Riesz wavelet: application to image mosaicing", Pattern Recognition, Volume 51, March 2016, Pages 425-442, http://dx.doi.org/10.1016/j.patcog.2015.09.021.

Chapter 4

Illumination invariant optical flow using neighborhood descriptors

Contents

4.1 Motivation: Robustness to illumination changes 93
4.1.1 Related work: Illumination robust methods 94
4.1.2 Aim of the chapter and its organization 95
4.2 Proposed optical flow approach (ROF-NND) 95
4.2.1 Neighborhood descriptors 95
4.2.2 Normalized neighborhood descriptor vectors 97
4.2.3 Illustration of effect of neighborhood descriptors 98
4.2.4 **NND** as data term in variational flow 99
4.2.5 Non-local filtering in flow regularization 100
4.2.6 Energy minimization 101
4.3 Coarse-to-fine optical flow approach of the ROF-NND algorithm . 103
4.4 Experiments on public datasets and algorithm benchmarking 104
4.4.1 Choice of algorithm parameters 105
4.4.2 Experiments on Middlebury benchmark 107
4.4.3 Experiments on KITTI benchmark 109
4.4.4 Experiments on MPI Sintel benchmark 110
4.4.5 Discussion on benchmarking 111
4.5 Experiments for illumination invariance 112
4.6 Experiments for large displacements (chosen $\alpha_{scale} = 0.7$) 114
4.7 Image mosaicing of low textured medical scenes 117
4.7.1 Datasets and evaluation criteria 117
4.7.2 Validation results 118
4.8 Main contributions and conclusion 120
List of publication 121

4.1 Motivation: Robustness to illumination changes

Chapters 2 and 3 have shown that the inherent flexibility of variational approaches used for optical flow field estimation enables the design of energy models that are able to handle texture variabilities, blur due to camera motion and small illumination changes. However, the previously proposed methods are sensitive to strong illumination changes. Such changes are unavoidable in different scenarios and occur for example as specular reflections (local illumination changes) in the endoscopic images of human organs like that of liver and stomach. In other medical or non-medical scenes, strong global illumination changes are either due to the environment in which the camera moves (e.g. as during underwater video acquisition) or corresponds to significant camera view-point changes between images (as in cystoscopy). Robust optical flow methods are needed for handling both of these local and global illumination changes which can be simultaneously strong. Additionally, in the medical field, an extrem illumination change occurs when passing from one modality to another. While multimodal image registration is a widespread problem in cardiology [Daul et al., 2009] or radiotherapy [Posada et al., 2007] for instance, in cystoscopy a same optical algorithm should ideally be able to register(with a constant parameter setting) either for white light data or data taken under fluorescence. As discussed in Chapter 1, these two modalities are widely used in cystoscopy.

To deal with strong local and global illumination changes observed in images due to various reasons, a total variational optical flow algorithm which is robust to illumination changes is proposed and validated in this chapter. The data-term of the proposed model is based on robust neighborhood descriptors constrained with a regularizer based on an accurate non-local median filtering technique within a coarse-to-fine energy minimization framework.

To illustrate the need of an illumination independent optical flow method we recall that for the Grove 3 image pair consisting of frames 10 and 11 (see the test in Section 3.6.4 of Chapter 3) the AEPE/AAE values were 1.02 $pixels$/13.44° and 0.99 $pixels$/10.44° for the accurate algorithms proposed in Chapter 2 (RFLOW, [Ali et al., 2014]) and Chapter 3 (AOFW, [Ali et al., 2015b]) respectively. These results are also visually represented in Fig. 4.1 shows the potential of a TV-L^1 algorithm specially conceived for strong illumination changes. For the same image pair as the one is Section 3.6.4 and for the same illumination change simulation in frame 11, the values of AEPE/AAE decreased to 0.58 $pixels$/6.05° when using the optical flow method described in this chapter. The errors were divided by a factor of 2 according the two optical flow accuracy criteria. The corresponding flow field given in Fig. 4.1 (d) can be visually compared to the ground truth optical flow sketched in Fig. 4.1 (c). It is also noticeable when comparing the flow

(a) (b) (c) (d) (e)

Figure 4.1: Visualization of the robustness of the proposed ROF-NND method against illumination changes. (a) Original frame10 of Grove3. (b) Frame11 of Grove3 with illumination. (c) Ground truth flow. (d) Result with ROF-NND before illumination changes ($AEPE/AAE = 0.58/6.05$). (e) Result of ROF-NND after illumination changes ($AEPE/AAE = 0.58/6.05$).

fields in Fig. 4.1 (e) (images with simulated illumination changes) and Fig. 4.1 (d) (same images without illumination changes), that the method proposed in this chapter give similar results for different illumination conditions of a scene. This in numerically confirmed by the very close AEPE/AAE values obtained by the algorithm without illumiantion changes (0.58 $pixels$/6.05°) and with illumination changes (0.59 $pixels$/6.15°).

4.1.1 Related work: Illumination robust methods

In the literature, self similarity patterns have been successfully used for handling scenes with illumination variations [Drulea and Nedevschi, 2013, Werlberger et al., 2010, Chen et al., 2013]. The technique in [Chen et al., 2013] is based on nearest neighbor field (NNF) using the PatchMatch based technique [Barnes et al., 2009] for the optical flow estimation which is followed by a motion segmentation step. Since NNF algorithms are not limited by the magnitude of displacement fields, they represent an efficient technique for handling large displacements. However, due to the non-convex energy formulation and integration of several steps like outlier rejection, multilabel graph-cut and then their fusion, this NNF based model [Chen et al., 2013] is computationally expensive. A fast and parallelizable energy minimization approach using zero-mean normalized cross-correlation (ZNCC) as a matching cost (data-term) was formulated in [Drulea and Nedevschi, 2013]. A similar approach using truncated ZNCC was also proposed by Werlberger et al. [Werlberger et al., 2010]. It has been shown in [Drulea and Nedevschi, 2013] that the ZNCC based matching cost is invariant to illumination changes in contrast to brightness constancy assumption (BCA) or gradient constancy assumption (GCA).

The method in [Mohamed et al., 2014] proposed an illumination robust approach by modifying the local directional pattern (MLDP) [Jabid et al., 2010] based on 8-bit binary feature descriptors as matching cost. Similar to [Sun et al., 2010], the author used a weighted median filtering approach for refining the flow vectors. However, the results given in [Mohamed et al., 2014] showed that the MLDP method is sensitive to large changes of illumination intensity and hue variations of colors. In the literature, the census transform (census-T) and the rank transform (rank-T) have been proven to be the most robust methods in case of illumination changes. However, their robustness towards illumination changes is achieved with a compromise in the accuracy [Hafner et al., 2013]. Both of these methods are based on the signatures obtained from the ordering of the intensity values in a patch of neighboring pixels. In both transforms, originally proposed by Zabih and Woodfill [Zabih and Woodfill, 1994], there is a loss of texture information when either the rank of the pixels (in rank-T) or a binary signature of the patch around the pixels (in census-T) are stored. Demetz et al. [Demetz et al., 2013] proposed to use a novel complete rank transform (CRT) which possess more local texture information and maintain the illumination invariance property of original rank or census transform. Classical patch-based data terms, such as the Census-T, fail in scale changes because of the strong changes in the local appearance. Ranftl et al. [Ranftl et al., 2014] proposed a scale-invariant census descriptor by sampling the radial stencils with different radii. A second-order total generalized regularization (TGV), orginally proposed by Bredies et al. [Bredies et al., 2010], was used along with non-local weights similar to the method used in [Werlberger et al., 2010, Drulea and Nedevschi, 2013]. It was shown that the TGV regularizer gives more accurate results than a TV-regularization. TGV regularization was also used by Demetz et al. [Demetz et al., 2015] confirming the increased accuracy of their previous CRT method [Demetz et al., 2013].

4.1.2 Aim of the chapter and its organization

This chapter describes an optical flow method which is robust towards illumination changes. It is based on a descriptor matching cost in a TV framework with weighted non-local regularization. The descriptors are based on the self-similarity measure of each pixel with respect to their neighboring pixels. The cost is built such that major part of the texture information around a pixel is retained. Such descriptors are computed at each level of a multi-resolution pyramid and then used in a coarse-to-fine energy minimization strategy to handle large displacements.

This chapter describes the use of neighborhood descriptors as a data-term in the TV scheme. These descriptors are created within a defined search-space demonstrating their self-similarity property. We successively explain 1) how the descriptors are computed, 2) the integration of these descriptors in the data-term, 3) the integration of a weighted non-local regularization in the primal-dual energy minimization framework, and 4) the use of three well-known publicly available optical flow datasets to benchmark our algorithm. Dedicated experiments have been conducted for demonstrating the robustness of the proposed method against illumination changes and accuracy of the proposed method for large displacements.

This chapter is organized as follows: Section 4.2 give the details on creating neighborhood descriptors, describes the data-term formulation, integrates the bilateral filtering technique into the regularizer and proposes a minimization scheme that is parallelizable due to the use of a primal-dual optimization approach. Section 4.3 gives an overview of the proposed optical flow algorithm referred to as robust optical flow using normalized neighborhood descriptors **ROF-NND**. In Section 4.4, we benchmark the **ROF-NND** algorithm with the publicly available Middlebury [Baker et al., 2011], KITTI [Geiger et al., 2013] and MPI Sintel [Butler et al., 2012] flow benchmarks. Section 4.5 demonstrates robustness tests of the algorithm to illumination changes and gives a comparison to some illumination invariant methods present in the literature. Section 4.6 is used to confirm the accuracy and robustness of the proposed algorithm with respect to large displacements. In Section 4.7, tests based on low textured and realistic (bladder and skin) phantom data with known ground truth are proposed to have a first indication about the appropriateness of the algorithm to be used in the frame of a a difficult image mosaicing application (endoscopy). Finally, Section 4.8 highlights the main contributions presented in this chapter.

4.2 Proposed optical flow approach (ROF-NND)

This section details the different aspects of the variational optical flow approach proposed in this thesis. After an introduction highlighting the interest of n-dimensional normalized neighborhood descriptors (**NND**) used as the data-term in an optical flow computation, we focus on three major parts of the method, namely, i) the linearization of the data-term based on n-dimensional **NND**, ii) the regularization of the flow field with classical bilateral filtering and iii) the first-order primal-dual approach for the minimization of the data-energy constrained with the non-local regularization.

4.2.1 Neighborhood descriptors

Data-terms in variational approaches are based on either classical BCA alone or GCA complementing the classical BCA. Such data-terms however do not allow for robust and accurate optical flow computation when strong illumination changes arise. The census-T and the rank-T are robust to illumination changes but usually suffer from loss of texture information in the matching

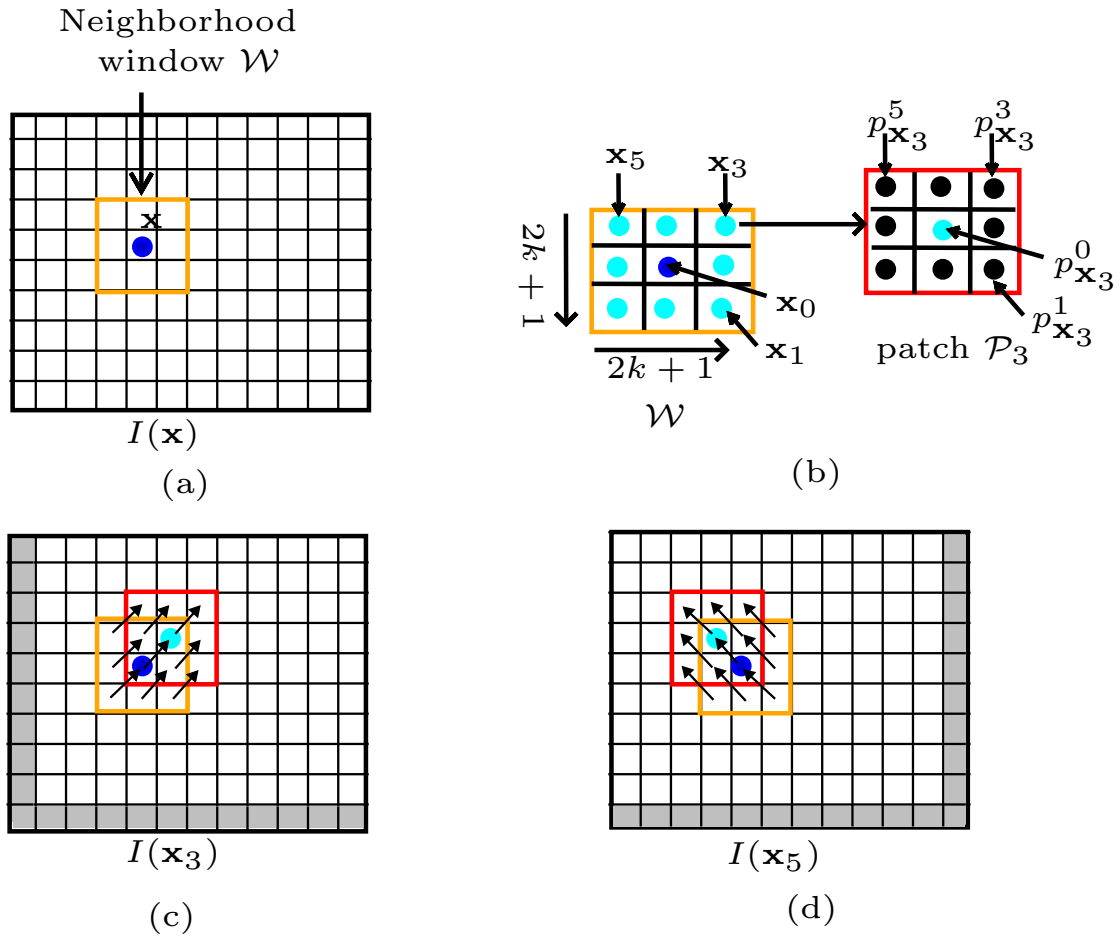


Figure 4.2: Definition of neighnorhood window $\mathcal{W}$, patches $\mathcal{P}_i$ and of their relative positions in image $I(\mathbf{x})$. (a) The dark blue dot gives the centre of window $\mathcal{W}$ delineated by the orange square. (b) $\mathbf{x}_i$ (light blue dots, $i \in [1, 8]$ in this example) are the neighbors of $\mathbf{x}$ in $\mathcal{W}$. A patch $\mathcal{P}_i$ (having the same size as $\mathcal{W}$) is centered on $\mathbf{x}_i$ and is represented by the red rectangle. The black dots (pixels $p^j_{\mathbf{x}_i}$) correspond to the neighbors of $\mathbf{x}_i$. (c) Illustration of the constant translation between corresponding pixels $\mathbf{x}_j$ and $\mathbf{p}^j_{\mathbf{x}_3}$ ($j \in [0, 8]$) of window $\mathcal{W}$ and patch $\mathcal{P}_3$ respectively. (d) Same "shift" representation for window $\mathcal{W}$ and patch $\mathcal{P}_5$.

cost. As a result, they are less accurate than many of accurate dense optical flow methods. Neighborhood descriptors (NDs) are used in order to reach a balance between robustness to illumination changes and optical flow accuracy. Such a descriptor holds self-similarity property of images.

Self-similarity of an image can be computed as the sum of square differences (SSD) between two patches of the same size. For a given pixel, the descriptor values are determined between a central patch (called neighborhood window) and patches located in the neighborhood of the pixel under treatments. All patches have the same size. Below diagrams are used to explain the steps used to establish the self-similarity property for building NDs.

NDs are vectors computed for a n-connected neighborhood centered at pixel $\mathbf{x}$. The principle of the neighborhood descriptors is illustrated in Fig. 4.2. As sketched in Fig. 4.2(a), a neighborhood window $\mathcal{W}$ is associated to each pixel $\mathbf{x}$. For simplicity, we have used a squared

window $\mathcal{W}$ of size $2k+1 \times 2k+1$ (*i.e.* with $(2k+1)^2$ pixels and a connectivity $n = \{(2k+1)^2-1\}$, $\forall k \geq 1$). So, for a given pixel $\mathbf{x}$ of image $I(\mathbf{x}) \in \Omega : \Omega \longrightarrow \mathbb{R}$, centered at window $\mathcal{W}$ (dark blue dot in Fig. 4.2(a)), there is an n-connectivity associated with it (represented by the light blue dots in Fig. 4.2(b) for $n = 8$). As an example, if $k = 1$, then $\mathcal{W}$ will be of size 3×3 and $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_8\}$ are the neighborhood pixels of $\mathbf{x}$ in $\mathcal{W}$ as shown in Fig. 4.2(b), where a patch $\mathcal{P}_i$ is associated to each light blue dot. These patches $\mathcal{P}_i$ are centered on the pixels $\mathbf{x}_i \in \mathcal{W}$. Fig. 4.2(b) (on the right) represents patch $\mathcal{P}_3$, centered on $\mathbf{x}_3$ which has 8 neighbors $\mathbf{p}^j_{\mathbf{x}_3}$ with $j \in [1, 8]$. The neighborhood windows $\mathcal{W}$ and patches P_i have the same size and contain an indentical n number of neighbors. $\{\mathbf{p}^1_{\mathbf{x}_i}, ... \ \mathbf{p}^j_{\mathbf{x}_i}, ..., \mathbf{p}^n_{\mathbf{x}_i}\}$ with $j \in [1, n]$ are the neighborhood pixels in patch $\mathcal{P}_i$. Figs. 4.2(c) and 4.2(d) illustrate the shift between window $\mathcal{W}$ and the two patches $\mathcal{P}_3$ and $\mathcal{P}_5$ respectively. Now, for computing the similarity measure $C_{\mathcal{P}_i}$ between the window $\mathcal{W}$ centered on $\mathbf{x}$ in image $I(\mathbf{x})$ and the shifted versions represented by patches $\mathcal{P}_i$ centered on $\mathbf{x}_i$ in $I(\mathbf{x})$ we estimate the sum of squared differences between the grey-levels of corresponding pixels in $\mathcal{W}$ and $\mathcal{P}_i$. This correspondence between neighbors of $\mathcal{W}$ and $\mathcal{P}_i$ is illustrated by the parallel arrows in Figs. 4.2.(c) and 4.2.(d). Mathematically, the similarity measure of window $\mathcal{W}$ and patch $\mathcal{P}_i$ can be written as:

$$C_{\mathcal{P}_i}(\mathbf{x}) = \sum_{j=0}^{n} \left(I(\mathbf{x}_j) - I(\mathbf{p}^j_{\mathbf{x}_i})\right)^2, \quad \text{with} \quad \mathbf{x}_j \in \mathcal{W} \text{ and } \mathbf{p}^j_{\mathbf{x}_i} \in \mathcal{P}_i, \tag{4.1}$$

where $\mathbf{x}_0 (= \mathbf{x})$ and $\mathbf{p}^0_{\mathbf{x}_i} (= \mathbf{x}_i)$ are the center pixels in the window $\mathcal{W}$ and patch $\mathcal{P}_i$ respectively. The $C_{\mathcal{P}_i}$ values can be physically interpreted as follows. It can be seen in Eq. (4.1) that each $C_{\mathcal{P}_i}$ relates to the mean grey-level variation in a direction defined by the pixels in $\mathcal{W}$ and $\mathcal{P}_i$ (centered at $\mathbf{x}$ and $\mathbf{x}_i = \mathbf{p}^0_{\mathbf{x}_i}$ respectively). When these grey-level variations correspond to the object edges or texture in different directions (*e.g.* 8 directions for $n = 8$ in Fig. 4.2), a set of $C_{\mathcal{P}_i}$-values can be seen as a local structure information measured at the pixel of interest $\mathbf{x}$.

4.2.2 Normalized neighborhood descriptor vectors

A non-negative monotonically decreasing Gaussian function, similar to the one used by Perona and Malik [Perona and Malik, 1990] for anisotropic diffusion, is used to normalize the neighborhood descriptors in the range $[0, 1]$:

$$C'_{\mathcal{P}_i}(\mathbf{x}) = \exp^{-\frac{C_{\mathcal{P}_i}(\mathbf{x})}{\sigma^2_{\mathbf{x}}}}, \tag{4.2}$$

where $\sigma^2_{\mathbf{x}}$ is the local measure of the mean change in grey value information in the 4-connected pixel neighborhood based on SSD between the neighborhood window $\mathcal{W}$ and a 4-shifted patches. We define this connectivity as $k = 0$ and with 4-possible patch shifts, *i.e.*, at $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} \in \mathcal{W}$, the mean of the local variation at a point $\mathbf{x}$ is computed as:

$$\sigma_{\mathbf{x}} = \sqrt{\left(\frac{1}{4}\sum_{i=1}^{4}\sum_{j=0}^{8}\left(I(\mathbf{x}_j) - I(\mathbf{p}^j_{\mathbf{x}_i})\right)^2\right)}, \tag{4.3}$$

with $\mathbf{x}_0$ and $\mathbf{p}^0_{\mathbf{x}_i}$ being the centers of the neighborhood window $\mathcal{W}_{3\times3}$ and the shifted patch $\mathcal{P}_i$ respectively. In Fig. 4.3, a 4-connected neighborhood of a 3×3 window $\mathcal{W}_{3\times3}$ centered at pixel $\mathbf{x}$ is shown in the light blue. These pixels corresponds to the 4 resulting patches result in 4-patches (patch $\mathcal{P}_3$ is shown for the pixel $\mathbf{x}_3$). This connectivity is fixed for the calculation of $\sigma_{\mathbf{x}}$ used

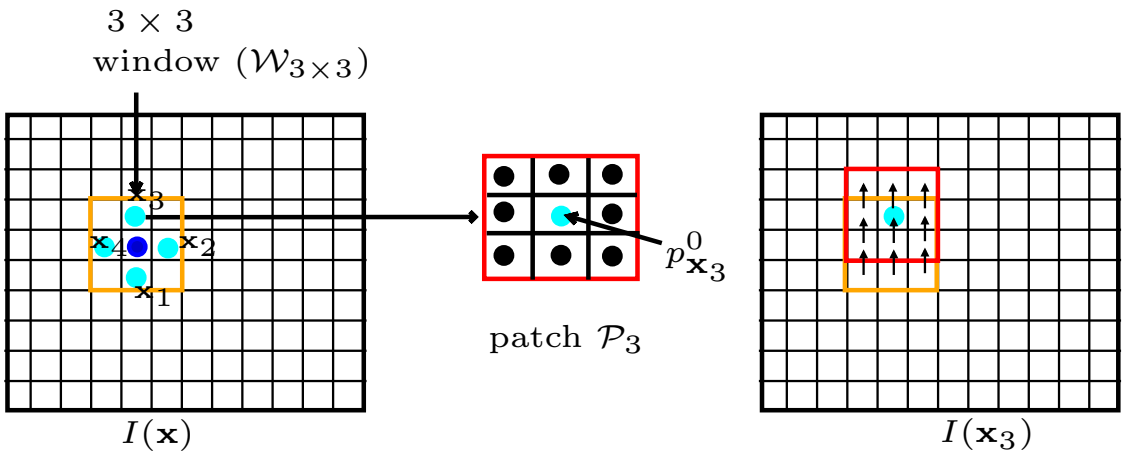


Figure 4.3: Pixel shifts in the 4-connected neighborhood in window $\mathcal{W}_{3\times3}$. On the left: 4-possible shifts in light blue dots $\{\mathbf{x}_1, .., \mathbf{x}_4\}$ around pixel $\mathbf{x}$ in $\mathcal{W}_{3\times3}$. On right: image $I(\mathbf{p}_{\mathbf{x}_3})$ formed with pixel shift on $\mathbf{x}_3 \in \mathcal{W}_{3\times3}$ forming a patch $\mathcal{P}_3$ (in red). The 8-pixels in $\in \mathcal{P}_3$ (black dots) corresponds to the pixel positions of neighborhood pixels in shifted image $I(\mathbf{x}_3)$ around $\mathbf{x}_3$ in original image $I(\mathbf{x})$.

in Eq. (4.2). The mean SSD between pixels for the fixed 3×3 window with the four patches is computed locally at each pixel of interest $\mathbf{x}$ by using Eq. (4.3). It is noticeable that the $\sigma_{\mathbf{x}}$ values are computed for only 4-connections in a fixed 3×3 window, *i.e.*, along horizontal and vertical directions of the images, while descriptors $C_{\mathcal{P}_i}$ are built with n number of connectivity usually greater than 4. In case of strong illumination changes or artifacts like specular reflections, an increase in connectivity of neighboring pixels can penalize largely the local structure or edge information. So, the local $\sigma_{\mathbf{x}}$ value computed with the four connected neighborhood is used for the normalization of the neighborhood descriptors $C_{\mathcal{P}_i}$, irrespective of their size of neighboring window $\mathcal{W}$. This will reduce the effect of strong illumination changes in the estimated $C_{\mathcal{P}_i}$s for a given window $\mathcal{W}$. Thus, the significance of using $\sigma_{\mathbf{x}}$ is that at the edge pixels in the proximity of $\mathbf{x}$, $\sigma_{\mathbf{x}}$ value is stronger and hence it will increase the response of the function $C'_{\mathcal{P}_i}(\mathbf{x})$ in Eq. (4.2). Consequently, a more contrasted edges will be favored over less contrasted edges.

The n-dimensional normalized neighborhood descriptor vector $\overrightarrow{\mathbf{NND}}(I, \mathbf{x})$ computed for pixel $\mathbf{x}$ of image $I(\mathbf{x})$ is thus given as:

$$\overrightarrow{\mathbf{NND}}(I, \mathbf{x}) = \left(C'_{\mathcal{P}_1}(\mathbf{x}), \ldots, C'_{\mathcal{P}_i}(\mathbf{x}), \ldots, C'_{\mathcal{P}_n}(\mathbf{x}) \right). \tag{4.4}$$

In Eq. (4.4) each value $C'_{\mathcal{P}_i}(\mathbf{x}), i \in [1, n]$ represents a component of the n-dimensional neighborhood descriptor vector $\overrightarrow{\mathbf{NND}}$.

4.2.3 Illustration of effect of neighborhood descriptors

The effect of **NND** is illustrated on a part of RubberWhale image pair of the Middlebury training dataset [Baker et al., 2011] having shadowed areas (in red rectangles of Fig. 4.4). Figs. 4.4 (b-c) give the neighborhood descriptor images respectively computed for the window and patch center pairs ($\mathbf{x}$, $\mathbf{x}_7$) and ($\mathbf{x}$, $\mathbf{x}_8$) following the neighborhood numbering convention in Fig. 4.2. Thus, the values the pixels $\mathbf{x}$ in the images in Fig. 4.4(b) and in Fig. 4.4 (c) are given respectively by the vector components NND(I, $\mathbf{x}$, 7) and NND(I, $\mathbf{x}$, 8) for $k = 1$ (i.e. Fig. 4.4 (b) and Fig. 4.4

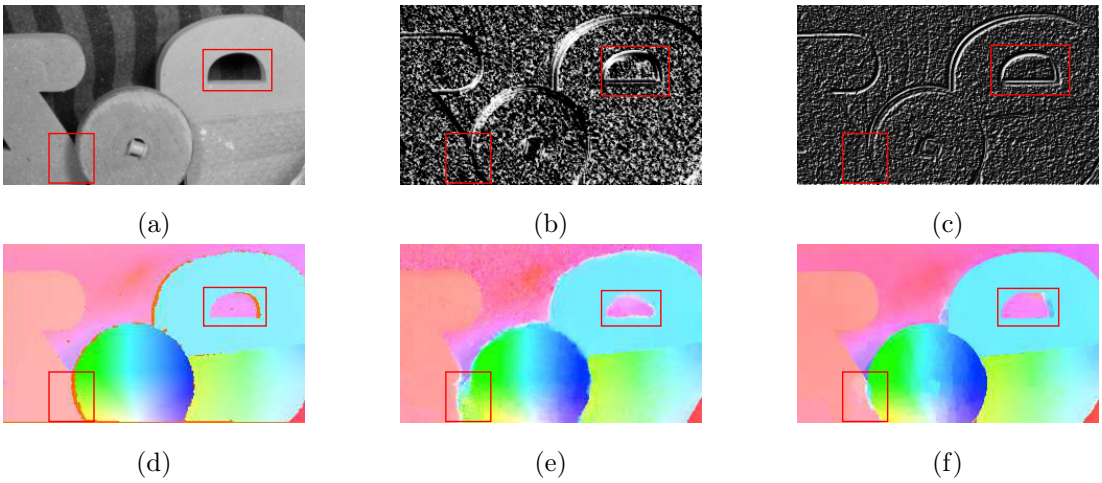


Figure 4.4: Illustration of the effect of **NND** on edge preservation and under illumination changes due to a shadow. (a) Orginal image with areas affected with shadows in red rectangles. (b) Image of $NND(I, \mathbf{x}, 7)$ (c) Image of $NND(I, \mathbf{x}, 8)$. (d) Ground truth flow field (classical flow color code used). (e) Result obtained with original image in Classical TV-L^1 framework [Pock et al., 2007]. (f) Result obtained with **NND** under similar implementation as the method used for (e).

(c) are NND component images). It can be observed in Figs. 4.4 (b-c) that the neighborhood descriptor possess significant information given under the form of strong gradient information and that each **NND** component has its own response according the structure (texture or object edge) orientation. A Gaussian function used to compute the **NND** decrease the sensitivity of the illumination change due to shadow in the direction of neighborhood pixel $\mathbf{x}_i$'s. It can be seen in the red rectangles of Figs. 4.4 (b-c) that the shadow has only a weak impact on these two **NND** components. As a consequence, as seen when comparing the content of the red rectangles in Fig. 4.4 (e) and in Fig. 4.4 (f) that the flow field obtained by including the **NND** in the data-cost is very close to the ground truth optical flow (see Fig. 4.4 (d)).

As also confirmed by the results in Sections 4.4, 4.5 and 4.6, the proposed neighborhood descriptors used in data-terms exhibit simultaneously several interesting properties (brightness, gradient and structure constancy) giving invariance to illumination changes and increased accuracy relative to many TV-L^1 based approaches in the literature. The advantage of **NND** over census-T can be explained as follows. The census-T is based on binary signatures giving for each pixel an information about the orientation of the gradient in a given neighborhood, whereas **NND** vectors include gradient information in several directions for each pixel. Unlike the census-T, **NND** vectors hold actually a structure information which leads to a detailed description of image regions around pixels of interest. Similarly, **NND** vectors include more detailed information than the Rank-T and the CRT methods. Indeed, such rank transform compute local description based on the comparison of a pixel with a surrounding neighborhood region. In contrary, **NND** vectors are determined by comparing two neighborhood regions (not only pixels) for all possible orientations in the local neighborhood (see Figure 4.2). This again leads to a richer set of information.

4.2.4 NND as data term in variational flow

The neighborhood descriptors are separately computed for the source (I_s) and target (I_t) images and are used in the data-term for optimizing the flow field, *i.e.* for superimposing the homologous

pixels of I_s and I_t. The coordinates of the superimposed pixels are $\mathbf{x}$ in source I_s and $\mathbf{x}+\mathbf{u_x}$ in target I_t, where $\mathbf{u_x}$ is the displacement vector at $\mathbf{x}$. Instead of using directly grey-level values, the proposed total variational scheme aims at minimizing the absolute difference of the corresponding n-dimensional vectors computed for $I_s(\mathbf{x})$ and $I_t(\mathbf{x}+\mathbf{u_x})$.

Let $\mathbf{u}=(u,v):\Omega\rightarrow\mathbb{R}^2$ be the flow field between I_s and I_t with their corresponding descriptor vectors $\overrightarrow{\mathbf{NND}}(I_s,\mathbf{x})$ and $\overrightarrow{\mathbf{NND}}(I_t,\mathbf{x}+\mathbf{u_x})$ respectively. In Eq. (4.5), the least absolute errors is computed between the corresponding components $NND(I_s,\mathbf{x},i)$ and $NND(I_t,\mathbf{x}+\mathbf{u_x},i)$ of the **NND**. As in Section 4.2.1, n (dimension of the **NND** vector) is also the number of pixels both in the neighborhood patch $\mathcal{P}$ centered on $\mathbf{x}$ in I_s and in the corresponding patch with pixels $\mathbf{x}+\mathbf{u_x}$ in image I_t.

$$E_{data}=\sum_{\mathbf{x}\in\Omega}\left\{\frac{1}{n}\sum_{i=1}^{n}\mid NND(I_s,\mathbf{x},i)-NND(I_t,\mathbf{x}+\mathbf{u_x},i)\mid\right\} \tag{4.5}$$

This gives a robust data-term which will suppress the outliers due to various imaging artifacts and scene ambiguities. This increased robustness is due to the presence of significant number of information present in each descriptor. A small mismatch in one descriptor will not penalize the whole data-term thus resulting in a more accurate and robust data-term than the classical approaches [Brox et al., 2004, Horn and Schunck, 1981, Werlberger et al., 2009]. Eq. (4.5) has to be linearized using first-order Taylor's series expansion to permit a convex formulation. After linearization, we can rewrite the energy associated with the data-term as:

$$\begin{aligned}E_{data}=\sum_{\mathbf{x}\in\Omega}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\mid NND(I_s,\mathbf{x},i)-NND(I_t,\mathbf{x}+\mathbf{u}^0_{\mathbf{x}},i)+\\ \nabla NND(I_t,\mathbf{x}+\mathbf{u}^0_{\mathbf{x}},i)(\mathbf{u_x}-\mathbf{u}^0_{\mathbf{x}})\mid\Bigg\},\end{aligned} \tag{4.6}$$

where $\mathbf{u}^0_{\mathbf{x}}$ is a close approximation of $\mathbf{u_x}$ and $\nabla=[\frac{\partial}{\partial x},\frac{\partial}{\partial y}]^T$.

4.2.5 Non-local filtering in flow regularization

The data-term in the total variational approach is attached to an optical flow regularizer [Pock et al., 2007, Brox et al., 2004, Horn and Schunck, 1981]. The role of the regularizer is to enforce a continuous optical flow in image regions with homogeneous pixel values, while preserving flow field discontinuities at edge pixels. The discontinuities in the flow field can be robustly preserved when scene information are incorporated in the regularizer. This technique is well established as non-local regularization [Drulea and Nedevschi, 2013, Werlberger et al., 2010, Sun et al., 2010, Ranftl et al., 2014, Li and Osher, 2009]. The regularization term in our total variational scheme uses an approach similar to that proposed in [Sun et al., 2010]:

$$E_s(\mathbf{u})=\sum_{\mathbf{x}\in\Omega,\,\forall\mathbf{x}'\neq\mathbf{x}\in\mathcal{N}_{\mathbf{x}}}\sum w^{\mathbf{x}'}_{\mathbf{x}}\mid\mathbf{u_x}-\mathbf{u_{x'}}\mid. \tag{4.7}$$

In Eq. (4.7), $w^{\mathbf{x}'}_{\mathbf{x}}$ is the weight assigned to each pixel $\mathbf{x}'$ in the neighborhood $\mathcal{N}_{\mathbf{x}}$ of pixels centered on $\mathbf{x}$. The weight $w^{\mathbf{x}'}_{\mathbf{x}}$ is defined as the correlation entity based on (i) the spatial distance and (ii) the color distance of the center pixel $\mathbf{x}$ with the neighborhood pixels $\mathbf{x}'\in\mathcal{N}_{\mathbf{x}}$:

$$w^{\mathbf{x}'}_{\mathbf{x}}=e^{-|\mathbf{x}-\mathbf{x}'|^2/2\beta_1^2}.e^{-|I(\mathbf{x})-I(\mathbf{x}')|^2/2\beta_2^2}, \tag{4.8}$$

where $\{\beta_1, \beta_2\}$ are normalization factors and $I(\mathbf{x})$ is the color vector in the CIE Lab color space. Edges are preserved since for further away pixels ($\mid \mathbf{x} - \mathbf{x}' \mid$ large) and for pixels which are not in the same region as the pixel of interest $\mathbf{x}$ (large color differences $\mid I(\mathbf{x}) - I(\mathbf{x}') \mid$) the value of weight $w_x^{x'}$ will be small, thus preventing from smoothing of the edge pixels.

4.2.6 Energy minimization

The error in the data-term in Eq. (4.6) is constrained with the smoothness term of Eq. (4.7) and has to be minimized in order to obtain the displacement vectors between image pairs (I_s, I_t). The Energy $E(\mathbf{u})$ to be minimized can be represented by following classical equation:

$$E(\mathbf{u}) = \sum_{\mathbf{x}\in\Omega} \lambda E_{data}(\mathbf{u}) + E_s(\mathbf{u}) \tag{4.9}$$

In Eq. (4.9), the minimization of $E(\mathbf{u})$ leads to the flow field $\mathbf{u}$ between source I_s and target I_t and the weight λ sets the trade-off between the data and the regularization term (setting the value of λ is discussed at end of Section 4.3). The minimization of the energy function $E(\mathbf{u})$ using a primal-dual model in convex optimization has been well established in the literature of variational optical flow [Pock et al., 2007, Zach et al., 2007, Chambolle, 2004]. However, some key steps relating to the minimization of $E(\mathbf{u})$ in Eq. (4.9) are discussed below which adapts such a primal-dual optimization for the proposed energy formulation incorporating non-local regularization.

With the positive scalar weights $w_{\mathbf{x}}^{\mathbf{x}'}$ in the neighborhood $\mathcal{N}_{\mathbf{x}}$, the non-local regularizer $E_s(\mathbf{u})$ in Eq. (4.7) is convex in $\mathbf{u}$. Since the flow field at each pixel $\mathbf{u}_{\mathbf{x}}$ has to be evaluated for all other pixels in the given neighborhood $\mathcal{N}_{\mathbf{x}}$, we introduce a linear operator $K : \mathbb{R}^{2.|\Omega|} \rightarrow \mathbb{R}^{2.|\Omega||\mathcal{N}_{\mathbf{x}}|}$ defined as [Drulea and Nedevschi, 2013]:

$$K\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 - \mathbf{u}_{1,1} & \dots & \dots & \mathbf{u}_{|\Omega|} - \mathbf{u}_{1,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{u}_1 - \mathbf{u}_{m,n} & \dots & \dots & \mathbf{u}_{|\Omega|} - \mathbf{u}_{m,n} \end{pmatrix}, \tag{4.10}$$

where $\{\mathbf{u}_{1,1}, \dots, \mathbf{u}_{m,n}\} \in \mathcal{N}_{\mathbf{x}}(\mathcal{N}_{\mathbf{x}} \rightarrow \mathbb{R}^{m\times n})$ represents the neighborhood flow vectors around each pixel of flow field $\mathbf{u}$ and subscript $\{1, \dots, \mid \Omega \mid\}$ in $\mathbf{u}$ is the array of pixel indexes of the estimated flow field $\mathbf{u}$ itself. The smoothness term can now be represented as a function of $K\mathbf{u}$ since the smoothness spans in all the neighboring pixels so $E_s(\mathbf{u})$ can be re-written as $E_s(K\mathbf{u})$, such that:

$$E_s(K\mathbf{u}) = \sum_{\mathbf{x}\in\Omega} \sum_{\mathbf{x}'\in\mathcal{N}_{\mathbf{x}}} \mid w_{\mathbf{x}}^{\mathbf{x}'}(\mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}'}) \mid . \tag{4.11}$$

Thus, the minimization problem of Eq. (4.9) can be redefined as:

$$\min_{\mathbf{u}} \{\lambda E_{data}(\mathbf{u}) + E_s(K\mathbf{u})\} \tag{4.12}$$

with E_{data} and E_s as the convex functions. Due to the presence of the linear operator K, the minimization of Eq. (4.12) is not trivial. We therefore have used a duality based approach [Zach et al., 2007] and reformulated Eq. (4.12) into a saddle-point min-max problem as below:

$$\min_{\mathbf{u}} \max_{\mathbf{q}} \{\lambda E_{data}(\mathbf{u}) + \langle K\mathbf{u}, \mathbf{q}\rangle - E_s^*(\mathbf{q})\}, \tag{4.13}$$

where E_s^* is the convex conjugate of the convex function E_s, $\mathbf{q} \in \mathbb{R}^{2|\Omega||\mathcal{N}_\mathbf{x}|}$ is the dual variable which lies in the closed convex set $\mathcal{S}_\mathbf{q}$ and $\langle . \rangle$ is the dot product. Eq. (4.13) is convex in $\mathbf{u}$ and concave in dual variable $\mathbf{q}$. The convex set $\mathcal{S}_\mathbf{q}$ is defined as non-negative entries of the bilateral filter weights in Eq. (4.8):

$$\mathcal{S}_\mathbf{q} = \{w_\mathbf{x}^{\mathbf{x}'}, \forall \mathbf{x} \in \Omega, \forall \mathbf{x}' \in \mathcal{N}\} \tag{4.14}$$

The relaxed proximal point algorithm (PPA) [Chambolle and Pock, 2011, Rockafellar, 1976, Gu et al., 2014] is used in order to solve the saddle-problem in Eq (4.13). Let τ and α be the positive scalars such that $\tau\alpha = \| K^T K \|$ and $(\hat{\mathbf{u}}, \hat{\mathbf{q}})$ be the initial estimates (usually set to zero), then the iteration scheme can be divided into following steps [Gu et al., 2014]:

I. PPA step:
The primal-variable solution is given by:

$$\tilde{\mathbf{u}}^k = \arg\min_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \left\{ \lambda E_{data}(\mathbf{u}) + \frac{\tau}{2} \| \mathbf{u} - [\mathbf{u}^k - \frac{1}{\tau} K^T \mathbf{q}^k] \|^2 \right\}, \tag{4.15}$$

where $\mathcal{S}_\mathbf{u}$ is the convex set of $\mathbf{u}$, $\tilde{\mathbf{u}}^k$ is the current estimate, $\mathbf{u}^k = \hat{\mathbf{u}}$ is an initial estimate, $\mathbf{u}$ is the update values (usually null at the beginning of the algorithm) and K^T is the adjoint linear operator. And the dual-variable estimate $\tilde{\mathbf{q}}^k$ is given by:

$$\tilde{\mathbf{q}}^k = \arg\min_{\mathbf{q}\in\mathcal{S}_\mathbf{q}} \left\{ \lambda E_s^*(\mathbf{q}) + \frac{\alpha}{2} \| \mathbf{q} - [\mathbf{q}^k + \frac{1}{\alpha} K(2\tilde{\mathbf{u}}^k - \mathbf{u}^k)] \|^2 \right\} \tag{4.16}$$

II. Relaxation step (update):
This step is used to generate the new iterate $\mathbf{u}^{k+1}$:

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \gamma(\mathbf{u}^k - \tilde{\mathbf{u}}^k). \tag{4.17}$$

In our case, $\gamma = 2$ in Eq. (4.17) so, $\mathbf{u}^{k+1} = 2\tilde{\mathbf{u}}^k - \mathbf{u}^k$.

Now, in order to minimize Eqs. (4.15) and (4.16) which are convex, first order optimality condition is used (*i.e.* $\frac{\partial \tilde{u}^k}{\partial \mathbf{u}} \geq 0$ and $\frac{\partial \tilde{p}^k}{\partial \mathbf{q}} \geq 0$). This results in following equations for Eq. (4.15) and Eq. (4.16) respectively:

$$\lambda \frac{\partial E_{data}}{\partial \mathbf{u}} + \tau(\mathbf{u} - \mathbf{u}^k) + K^T \mathbf{q}^k \geq 0 \tag{4.18}$$

$$\alpha(\mathbf{q} - \mathbf{q}^k) - K\mathbf{u}^{k+1} \geq 0. \tag{4.19}$$

Eq. (4.18) being linear, it can be solved by pointwise iterations for $\tilde{\mathbf{u}}^k$. Note that, we solve Eq. (4.18) for $\mathbf{u}$ which is actually the current value that has be determined. $\mathbf{u}$ can be represented by $\tilde{\mathbf{u}}^k$ which is used to update $\mathbf{u}^{k+1}$ in Eq. (4.17). Since the estimation of $\mathbf{u}$ (refer Eq. (4.18)) is dependent on the dual variable $\mathbf{q}$, it is important to solve Eq. (4.19) for $\mathbf{q}$. The dual variable $\mathbf{q}$ lies in the closed convex set $\mathcal{S}_\mathbf{q}$ so for each point there exists a unique solution for $\mathbf{q}$ in $\mathcal{S}_\mathbf{q}$ (*i.e.* closest to the optimal solution of $\mathbf{q}$ in Euclidean sense). The point-wise projection Pr. of the gradient solution in Eq. (4.19) onto the convex set $\mathcal{S}_\mathbf{q}$ is:

$$\mathbf{q}^{k+1} = \text{Pr.}_{\mathcal{S}_\mathbf{q}}\{\sigma K u^{k+1} + \mathbf{q}^k\}, \tag{4.20}$$

where $\sigma = \frac{1}{\alpha}$ and the positive scalars τ in Eq. (4.18) and σ in Eq. (4.20) take the same value: 0.22.

Algorithm 2: Optical flow estimation using n-dimensional **NND** descriptors

Input : Image pairs (I_s, I_t)
Output : Optical flow field $\mathbf{u}$
Parameters : 8-connected neighborhood ($\mathcal{N}_8$), $\alpha_{scale} = 0.7$, $N_{warps} = 3$, $N_{iter.} = 30$, $\lambda = 90$, $\epsilon = 0.0001$, $\tau = \sigma = 0.22$, $s = 0.5$, $\Delta = 0.001$
Initialization : initial flow field at coarsest level $j = N_{scales}$:$\tilde{\mathbf{u}}^j = 0$

Using α_{scale}, downsample I_s^1 and I_t^1 into N_{scales};
Using Eqs. (4.1)-(4.4), compute **NND** vectors of all pixels $\mathbf{x}$ of I_s^1 and I_t^1 all pyramid levels (i.e. $j = 1 : N_{scales}$);
for *scale* $j = N_{scales}$ **to** 1 **do**
  for $l = 1$ **to** N_{warps} **do**
    Warp image I_s^j with flow field $\tilde{\mathbf{u}}^j$ with a step-size s;
    Use Eq. (4.6) to compute the data-term energy obtained for the neighborhood descriptors at level j.;
    Compute the weights of Eq. (4.8) using the CIE-Lab representation of I_s^j and I_i^j;
    m=1;
    do
      Primal-dual energy minimization of $\mathsf{E}(\mathbf{u}^j)$ using Eqs. (4.18) and (4.20) with step-widths of τ and σ respectively;
      m=m+1;
    while $((m \leq N_{iter})$ *and* $(\| \boldsymbol{u}^j - \boldsymbol{u}^0 \|^2 > \Delta))$;
  if $j \neq 1$ **then**
    $\mathbf{u}^{j-1} = median filter\left(\uparrow \circ \frac{\mathbf{u}^j}{\alpha_{scale}}\right)$;
    $\tilde{\mathbf{u}}^{j-1} = \mathbf{u}^{j-1}$;
return optical flow $(\mathbf{u} = \mathbf{u}^1)$;

4.3 Coarse-to-fine optical flow approach of the ROF-NND algorithm

A global overview on the coarse to fine approach is presented in Algorithm 2. The source and target images, I_s and I_t, are dowsampled from level (scale) $j = 1$ to $j = N_{scales}$. The value of the downsampling factor α_{scale} is set to 0.7. The number of downsampled scales are such that the height and width of images I_s^j and I_t^j remains greater or equal to 16 pixels at a given scale j. The normalized n-dimensional vectors $\overrightarrow{\mathbf{NND}}(I_s, \mathbf{x})$ and $\overrightarrow{\mathbf{NND}}(I_t, \mathbf{x})$ are computed for all pixels $\mathbf{x}$ of source image I_s and target image I_t (see Eqs. (4.1)-(4.4) in Section 4.2.1) for all the scales. At scale j, the source and target images and their corresponding **NND** vectors are represented as I_s^j, I_t^j, $\overrightarrow{\mathbf{NND}}^j(I_s, \mathbf{x})$ and $\overrightarrow{\mathbf{NND}}^j(I_t, \mathbf{x})$ respectively.

The energy minimization using the primal-dual approach in convex optimization is done in a coarse-to-fine strategy. The start of this minimization (coarsest level) is done with the initialization of $\tilde{\mathbf{u}}^{Nscales} = 0$ (*i.e.* primal solution of Eq. (4.18)). Similarly, the dual variable p in Eq. (4.20) is also set to zero for this level. The flow field at finer levels $\mathbf{u}^{j-1}$ are computed as the up-scaled version of the previous scale j giving an initial flow fields to the finer pyramid levels. This flow field obtained at each level j ($\mathbf{u}^j$) is then used to warp the source image I_s^j and superimpose it on target I_t^j. The data-terms at each level j is computed using Eq. (4.6). Also, at each level j, a non-local weight $w_{\mathbf{x}}^{\mathbf{x}'}$ using Eq. (4.8) is computed exploiting the CIE-Lab representation of I_s^j and I_i^j. This is done to perform a non-local regularization of the flow field

in the primal-dual energy minimization scheme.

The formulated L^1 total variation energy in Eq. (4.9) is solved in N_{iter} iterations using the primal-dual approach in convex optimization presented in Section 4.2.6. The flow field $\mathbf{u}^j$ updated in this way is then used to warp I_s^j on target I_t^j and the L^1 variation energy in Eq. (4.9) is solved in the third warp to obtain the final flow field $\mathbf{u}^j$ at level j. This flow field is then upsampled to level $j-1$ by the re-scaling factor α_{scale}. A median filter of size $[3,3]$ is applied during each upscaling procedure to smooth the obtained flow field which will result in minimization of interpolation error.

The parameter values in Algorithm 1 are the following: an 8-connected neighborhood ($k=1$ and $n=8$) and patches $\mathcal{P}$ with a size of 3×3 were used to compute the 8-dimensional neighborhood vectors required for the data-term in Eq. (4.6). The energy of the non-local regularizer in Eq. (4.7) was dependent on the neighborhood weights $w_{\mathbf{x}}^{\mathbf{x}'}$ having following parameter values in Eq. (4.8): $\beta_1=5$ and $\beta_2=7$. These values were set experimentally on Middlebury dataset giving better accuracy than at other tested values. The relative importance of the data-term with respect to the regularizer is adjusted with $\lambda=90$. The value of parameters τ in Eq. (4.18) and σ in Eq. (4.20) were both set to 0.22 for solving iteratively the L^1-variation energy in Eq. (4.9) using the primal-dual approach presented in Section 4.2.6. It is worth noting that these parameter values were kept constant for all the tests presented in Sections 4.4, 4.5, 4.6 and 4.7 involving different image datasets. The method of parameter adjustment (*i.e.* the choice of the parameter values) is detailed in Section 4.4.1.

4.4 Experiments on public datasets and algorithm benchmarking

Three reference datasets were used to benchmark the proposed algorithm. These publicly available datasets include 1) the Middlebury flow dataset [Baker et al., 2011], 2) the KITTI flow dataset [Geiger et al., 2013] and 3) the MPI Sintel dataset [Butler et al., 2012]. These databases cover a large variability in scene (like illumination conditions), image quality (sharp or blurred textures) and in displacement (small/large flow vector magnitudes).

Three quantitative measures [Baker et al., 2011, Geiger et al., 2013] of the optical flow quality were employed both to adjust the parameters listed in the algorithm overview (Section 4.3) and to benchmark the algorithm. The first and the second error measures are the average angular error (AAE) in degrees and the average end-point error (AEPE) computed with all pixels. The third measure corresponds to the percentage of bad pixels (% BP) in the non-occluded areas (Out-Noc). The Middlebury flow database benchmark is based on both the AAE and AEPE values for all pixels. While, the KITTI flow benchmark is based on % BP in the non-occluded areas (Out-Noc) at 3 pixels AEPE threshold (% BP3). The MPI Sintel benchmark consists of clean and final pass sets, the latter being more challenging. Algorithms are evaluated for this dataset based on the overall AEPE in the whole test sequence for each pass separately.

The aim of this chapter is to highlight the robustness of the proposed **ROF-NND** method for small displacements, large displacements, different textures and illumination varying scenes. To do so, it was compared with two types of algorithms, namely: baseline TV-L^1 algorithms suited for small displacement [Zach et al., 2007, Brox et al., 2004, Wedel et al., 2009b] and TV-L^1 algorithms that are robust to illumination changes and that have convincing performance for large displacements [Drulea and Nedevschi, 2013, Mohamed et al., 2014, Demetz et al., 2013, Ranftl et al., 2014, Xu et al., 2012]. However, our main focus is on illumination robustness of the compared methods.

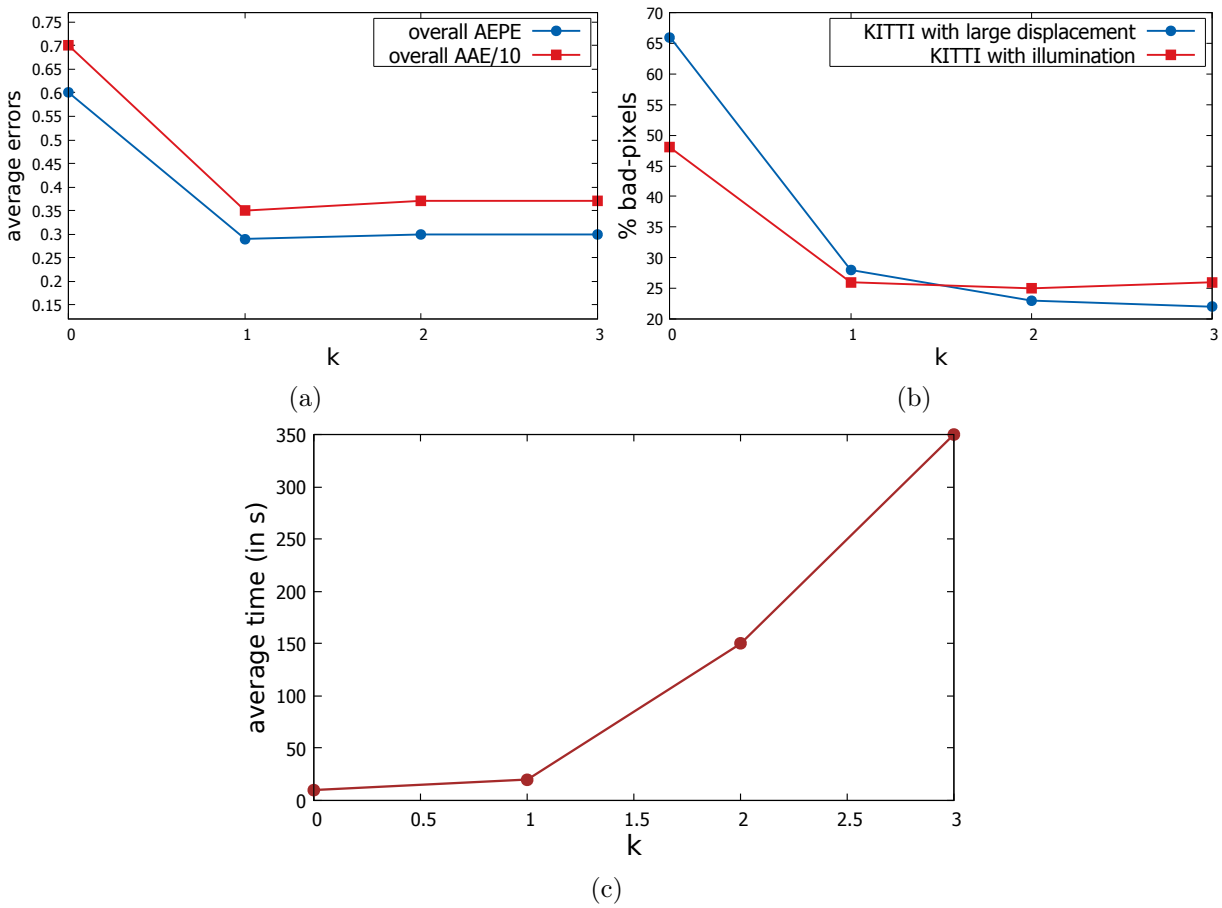


Figure 4.5: Tests for setting the optimal size of patch $\mathcal{P}$. (a) Average errors on Middlebury training dataset for varying k. The AAE values in degrees are divided by 10 for representing both errors (AEPE and AAE) on a common range of values (for visualization purpose). (b) % of bad-pixels at AEPE threshold of 3 pixels for the KITTI training dataset with illumination changes (in red) and large displacement (in blue). (c) Average optical flow computation time on the KITTI for increasing k.

4.4.1 Choice of algorithm parameters

This section presents the method used to adjust three important parameters of the proposed algorithm, namely the size of neighborhood window $\mathcal{W}$ and that of patches $\mathcal{P}$ (both determined by the same k, refer Section 4.2.1) leading to a robust data-term, the λ-parameter (relative importance between the data-term and the regularizer) for ensuring accurate flow and the down-sampling factor α_{scale} in the multiresolution approach enabling computation of optical flow for large displacements.

The Middlebury data-base consists of various image types (synthetic images, hidden textures in real scenes, indoor and outdoor scenes, etc.) combining moderate optical flow displacements and large texture variations in image pairs. The texture variability of these images is particularly useful for adjusting the λ-parameter in order to achieve optical flow accuracy for a large range of texture variations. Using eight training image pairs (with known ground truth) of this dataset we experimentally verified that with $\lambda \in [90-120]$ the AEPE and AAE errors were the least and close to respectively 0.30 pixels and 3.54°. This accuracy in a large weight interval shows that

the λ parameter in our algorithm can be easily set constant for a large number of image types (this will be confirmed in the results of Sections 4.4.2, 4.4.3 and 4.4.4 in which very different image types are treated with $\lambda = 90$).

Among the parameters of the data-term to be set, the size of the neighborhood patch $\mathcal{P}$ and the window $\mathcal{W}$ has the greatest impact on the robustness of the method. The size of $\mathcal{P}$ and $\mathcal{W}$ depend on k since an area of $(2k+1) \times (2k+1)$ pixels leads to **NND** vector with a dimension of $[(2k+1)^2 - 1]$ when k is > 0. For the particular case of $k = 0$, a four-connected neighborhood is considered (i.e. 4 dimensional **NND** vectors). In order to determine the optimal value of parameter k, tests were performed on the Middlebury training dataset (with synthetic and hidden texture images), on the KITTI training dataset with illumination changes (image pairs#11, #15, #44, #74) and on the large displacement image pairs (#117, #144, #147, #181) of the same dataset. It can be observed from Fig. 4.5 (a) for the Middlebury dataset, when k lies in $[0, 3]$, the smallest AEPE and AAE errors are obtained for $k = 1$, *i.e.* for 8-dimensional **NND** and a patch size of 3×3. The errors are the highest for a 4-connected neighborhood ($k = 0$). For the KITTI dataset (refer to Fig. 4.5(b)), in the image pairs with illumination changes the % of bad-pixels does not change significantly from $k = 1$ to $k = 3$. However, for large displacements in the same dataset, the % of bad-pixels is 5% less with $k = 2$ and 7% less for $k = 3$ in comparison to the same error when k is set to 1. But, this accuracy is obtained at the expense of high computational time as shown by the plot in Fig. 4.5 (c). The computational time grows significantly with the increase of the value of k (refer to Fig. 4.5(c)). With the above observations, it becomes clear that choosing an 8-connected neighborhood (*i.e.* $k = 1$) leads to the best compromise between accuracy and computational time.

In the proposed multi-resolution approach, the down-sampling factor α_{scale} has to be chosen such that large displacements can be treated while maintaining the optical flow errors in acceptable limits. Fig. 4.6 illustrates the impact of the value of $\alpha_{scale} \in [0.5, 0.9]$ on the optical flow computation speed and accuracy using one of the large displacement image pairs of the KITTI

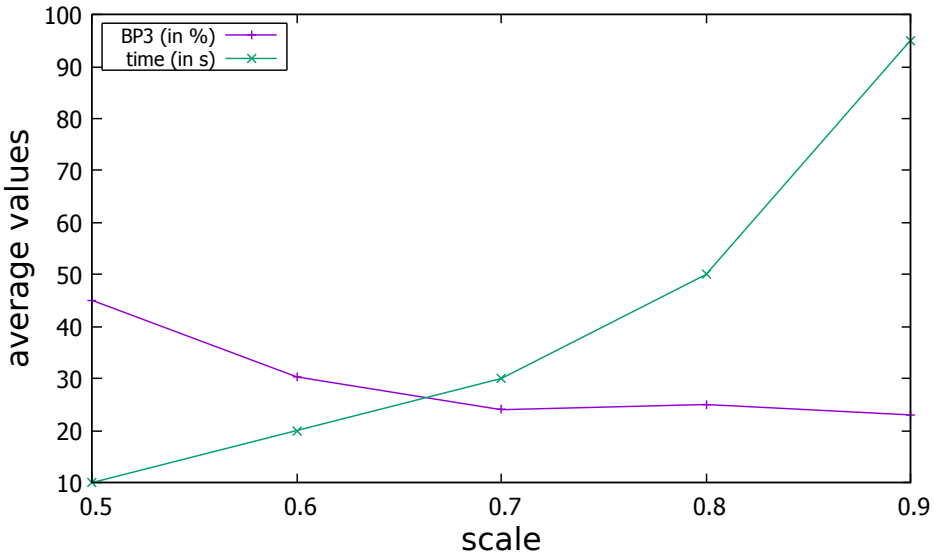
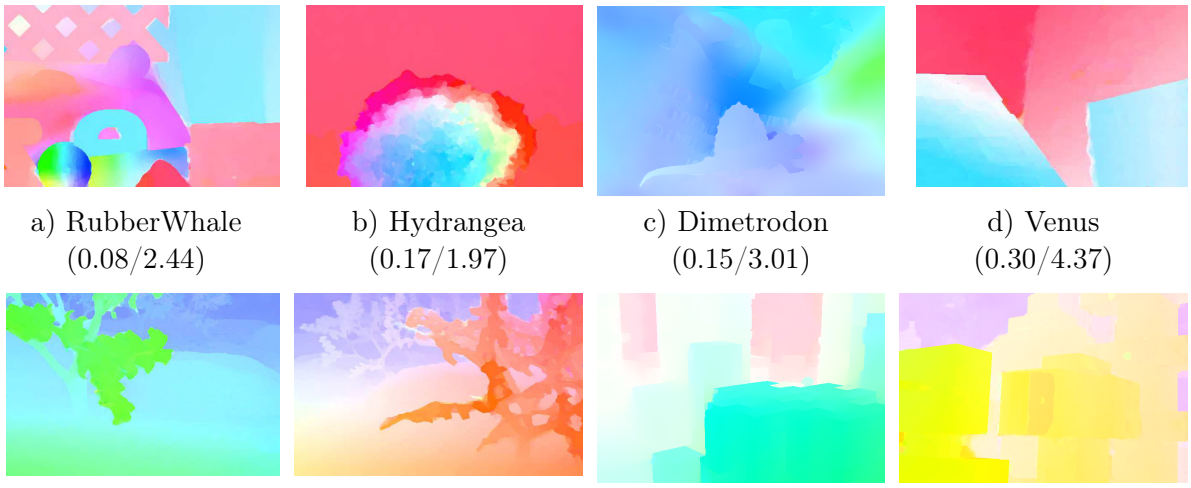


Figure 4.6: Percentage of average bad-pixels at AEPE threshold of 3 pixels (BP3) and average computation time as a function of the scale-factor (α_{scale}). The values are given for a combined MATLAB/C- implementation of the proposed algorithm (**ROF-NND** with $k = 1$) tested on large displacement image pairs (#117, #144, #147 and #181) of the KITTI training dataset.



a) RubberWhale (0.08/2.44) b) Hydrangea (0.17/1.97) c) Dimetrodon (0.15/3.01) d) Venus (0.30/4.37)

e) Grove 2 (0.15/1.98) f) Grove 3 (0.58/6.05) g) Urban 2 (0.36/3.27) h) Urban 3 (0.47/3.51)

Figure 4.7: Optical flow results (in flow color code) obtained with the proposed method (**ROF-NND**) for the training image pairs of the Middlebury dataset along with their corresponding AEPE/AAE (pixels/°). Average overall AEPE/AAE on this training sequence are 0.28 pixels/3.32°.

dataset (average values are given for pair numbers 117, 144, 147 and 181). For $\alpha_{scale} = 0.6$, the average percentage of bad pixels is over 30%. This error grows significantly when α_{scale} decreases. However, for $\alpha_{scale} = 0.7$ this percentage value is 24.17 % and remains almost constant when α_{scale} increases. This experiment shows that $\alpha_{scale} = 0.7$ leads to the best compromise between the number of down-sampled images in the coarse-to-fine approach and the percentage of erroneous pixels. When α_{scale} increases, the computation time grows exponentially while the improvement of the percentage of erroneous pixels remain weak. The values plotted in Fig. 4.6 show that, when passing from $\alpha_{scale} = 0.9$ to $\alpha_{scale} = 0.7$, the computation times is divided by 3 while the percentage of bad pixels only increases by 1%.

4.4.2 Experiments on Middlebury benchmark

The Middlebury dataset [Baker et al., 2011] consists of 8 image pairs having known ground truth (GT) in the training dataset and 8 image pairs without available flow field ground truth in the test dataset (evaluated online). These set include data with hidden texture, synthetic image and stereo pairs. Since this dataset was built for the estimation of small motion fields associated to varying texture conditions, it is interesting for robustness evaluation of the proposed algorithm for small displacements.

A visual representation of the results on 8 images of the Middlebury training dataset is shown in Fig. 4.7. The algorithm is robust to both for hidden texture dataset (first row of images in Fig. 4.7) and for synthetic dataset (second row of images). Both AEPE and AAE are given in bracket below their color-code flow representation in Fig. 4.7. The small errors obtained for a standard dataset are a first indication of the accuracy of the proposed method.

An online benchmarking of the proposed algorithm was done with the Middlebury test dataset using the parameters given in Section 4.4.1. The performances of the proposed algorithm are compared to those of other approaches chosen to cover a large variety of TV-L^1

a) Army b) Mequon c) Schefflera d) Wooden

e) Grove f) Urban g) Yosemite h) Teddy

g) Backyard h) Basketball h) Dumptruck h) Evergreen

Figure 4.8: Optical flow results (in flow color code) obtained with the proposed method (**ROF-NND**) for the Middlebury test dataset (hidden ground truth).

Method	**Overall AEPE**	**Overall AAE**	**Runtime (in s)**
MDP-flow2 [Xu et al., 2012]	0.25	2.44	342
corr-flow [Drulea and Nedevschi, 2013]	0.31	3.01	290
Classic-NL [Sun et al., 2014]	0.32	2.90	972
MLDP [Mohamed et al., 2014]	0.35	3.21	165
OFH [Zimmer et al., 2011]	0.36	3.40	620
WPB [Werlberger et al., 2010]	0.38	3.8	20
ROF-NND (our)	0.39	3.81	10
Aniso. Huber-L1 [Werlberger et al., 2009]	0.40	4.22	2 (GPU)
HS [Sun et al., 2014]	0.41	3.92	486
CRTflow [Demetz et al., 2013]	0.43	4.46	13
Simpleflow [Tao et al., 2012]	0.47	3.12	1.7 (GPU)
TV-L1-improved [Wedel et al., 2009b]	0.54	4.17	2.9 (GPU)
HAOF [Brox et al., 2004]	0.57	4.54	18
EPPM [Bao et al., 2014]	0.62	-	0.2 (GPU)

Table 4.1: Comparison of state-of-the-art methods with the proposed **ROF-NND** method with the overall AEPE (in pixels) and AAE (in degrees) on the Middlebury optical flow benchmark [Baker et al., 2011]. Runtimes are provided for the Urban image pair under CPU implementation unless mentioned as GPU. Average ranking is provided online at `http://vision.middlebury.edu/flow/eval/results-e1.php`.

Method	BP3[%]		AEPE [px]		t (in s)
	Noc.	Occ.	Noc.	Occ.	
NLTGV-SC [Ranftl et al., 2014]	5.93	11.96	1.6	3.8	16 (GPU)
MLDP [Mohamed et al., 2014]	8.67	18.78	2.4	6.7	160
CRTflow [Demetz et al., 2013]	9.43	18.72	2.7	6.5	18 (GPU)
ROF-NND (k=2)	10.44	21.23	2.5	6.5	50
C-NL [Sun et al., 2014]	10.49	20.64	2.8	7.2	888
C-NL-fast [Sun et al., 2014]	12.36	22.28	3.1	7.9	174
ROF-NND (k=1)	12.43	22.69	3.3	8	20
EPPM [Bao et al., 2014]	12.75	23.55	2.5	9.2	0.25 (GPU)
HS [Sun et al., 2010, Sun et al., 2014]	14.75	24.11	4.0	9.0	18 (GPU)
DB-TV-L1 [Zach et al., 2007]	30.87	39.25	7.9	14.6	16
HAOF [Brox et al., 2004]	35.87	43.46	11.1	18.3	16.2

Table 4.2: KITTI flow benchmark comparition for the proposed and existing reference state-of-the-art methods (without incorporating stereo-matching or epipolar geometry). Average runtimes (**t**) are given for CPU implementation unless mentioned. For details see also http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=flow. Noc represents errors evaluated for non-occluded regions and occ represents errors evaluated for all image pixels including occlusion. The % of bad pixels and the AEPE for the pixels at AEPE threshold of 3 pixels are given.

implementations. The overall average errors of the proposed algorithm (**ROF-NND**) and the chosen reference methods for 8-image pairs (available online) is given in Table 4.1. The MDP-flow2 method [Xu et al., 2012] is on top of this chart with the least error among all the methods in comparison. However, this accuracy is obtained at the expense of the computational time (342 s is required to compute the flow field in Urban image pair). Most of the top performing algorithms (corr-flow [Drulea and Nedevschi, 2013], Classic-NL [Sun et al., 2014], MLDP [Mohamed et al., 2014] and OFH [Zimmer et al., 2011]) in Table 4.1 have high accuracy but are with very high computational time. The proposed method (ROF-**NND**) gives a competitive result which is similar to that of the non-local approach of Werlberger et al. [Werlberger et al., 2010] while leading to a much smaller computational time (of only 10 s) in comparision to the most accurate algorithms. ROF-**NND** is more accurate than improved TV-L^1 approaches such as Aniso. Huber-L1 [Werlberger et al., 2009], TV-L1-improved [Wedel et al., 2009b] and HAOF [Brox et al., 2004]. Additionally, the proposed method also outperforms CRTflow method [Demetz et al., 2013] which is based on complete rank transform.

4.4.3 Experiments on KITTI benchmark

The KITTI optical flow test dataset [Geiger et al., 2013] consists of 194 color image pair obtained with a wide-view camera fixed on a moving vehicle. These image pairs have challenging characteristics of an outdoor scene like with illumination variability, large perspective changes, repeated texture patterns and with large displacements. The results obtained by the proposed optical flow algorithm for the outdoor scene were compared to those of other methods based on total variational approach and particularly suited for large displacements.

As seen in Table 4.2, the NLTGV-SC method [Ranftl et al., 2014] which uses a second-order variant of the TGV based non-local regularization has the least BP3 percentage and AEPE:

Method	AEPE all [px]	s0-10 [px]	s10-40 [px]	t [s]
MLDP [Mohamed et al., 2014]	8.287	1.312	5.122	NA
MDP-flow2 [Xu et al., 2012]	8.445	1.420	5.449	547
EPPM [Tao et al., 2012]	8.377	1.834	4.955	NA
NLTGV-SC [Ranftl et al., 2014]	8.746	1.587	4.780	NA
Classic-NL [Sun et al., 2010]	9.153	1.113	4.496	888
ROF-NND (k=1)	9.286	1.221	4.700	50
NLTV-SC [Ranftl et al., 2014]	9.855	1.202	4.757	NA
HS [Sun et al., 2014]	9.610	1.882	5.335	156
Classic++ [Sun et al., 2010]	9.959	1.403	5.098	510
Classic-NL-fast [Sun et al., 2010]	10.088	1.092	4.7	174
Aniso-Huber-L1 [Werlberger et al., 2009]	11.927	1.155	7.966	3.2 (GPU)
SimpleFlow [Tao et al., 2012]	13.364	1.475	9.582	2.9 (GPU)

Table 4.3: Performance on MPI Sintel benchmark (`(http://sintel.is.tue.mpg.de/results)`). Only methods with variational approach implementation are shown here for final pass dataset. The column “s0-10" and “s10-40" represents the AEPE over regions with flow vector magnitudes ranging in $[0, 10]$ pixels and $[10, 40]$ pixels. Average runtime (t) is given for CPU implementation unless mentioned. NA stands for not applicable.

respectively 5.93% and 1.6 pixels for non-occluded image regions (Noc) and 11.96% and 3.8 pixels for image regions with occlusion (occ). This method uses a scale invariant census transform based approach. However, the NLTV-SC method using a first-order TV approach was mentioned in [Ranftl et al., 2014] to have higher percentage of BP3 of 9.19 %. Comparing the results of the latter method (NLTV-SC) with the proposed **ROF-NND** method in this chapter is more relevant since both of these methods are based on TV-regularization. However, the results given for the NLTV-SC method give an idea of the performance of our method relative to TGV variants. The MLDP [Mohamed et al., 2014] and CRTflow [Demetz et al., 2013] methods exhibit accurate results for the KITTI data-set. The proposed method with, $k = 2$, with only 10.44% of BP3 for non-occluded pixels (see Table 4.2) competes well with both of these methods. Compared to the the proposed approach, the C-NL and C-NL-fast methods [Sun et al., 2010, Sun et al., 2014] lead to higher computational time and without improving the flow field accuracy (these methods have larger % of BP3 and AEPE). The fast edge preserving method [Bao et al., 2014] using plane-fitting for flow field refinement gives close to real-time performance on GPU while having an acceptable accuracy which is better than that of the classical HS implementation of Sun et al. [Sun et al., 2010]. The baseline TV-L^1 approaches are the least accurate ones with % of bad pixels above 30%.

4.4.4 Experiments on MPI Sintel benchmark

The MPI Sintel dataset [Butler et al., 2012] consists of different image sequences. For our comparison we used the final pass sequences to benchmark the proposed algorithm on this complicated data. The 12 final pass sequences consist of image pairs with large displacements, specular reflections, blur due to camera defocus/refocus, motion blur and atmospheric effects such as fog and smoke. In Table 4.3, three criteria were used to evaluate the algorithms: i) the overall average-end-point errors computed with all flow field vectors of the dataset (AEPE all), ii) the

overall AEPE of pixels with displacement vectors in the $[0, 10]$ pixel interval (s0-10) and in the $[10, 40]$ pixel interval (s10-40) and iii) the run time.

For all of these three criteria, the results in Table 4.3 show that the proposed method has better performance than the recent TV-L^1 baseline methods [Werlberger et al., 2009, Sun et al., 2010, Tao et al., 2012, Sun et al., 2014]. It can be noted in the Table 4.3 that the second-order variant of TGV (NLTGV-SC method [Ranftl et al., 2014]) is not the most accurate method in the comparison as in the KITTI benchmark. Moreover, for s0-10 (AEPE for small displacements), NLTGV-SC method has a large error of 1.587 pixles which is more than many other methods in the comparison Table 4.3. The NLTV-SC approach has slightly higher errors than than the proposed method on this dataset. MLDP has the smallest overall AEPE of 8.287 pixels while Classic-NL-fast has the smallest AEPE for s0-10 (1.092) and Classic-NL has the least for s10-40 (1.113 pixels). The smallest overall AEPE of 8.287 pixels, the smallest AEPE for "small" displacements (s0-10) of 1.092 pixels and the least AEPE for "large" displacements (s10-40) of 1.113 pixels were recorded for the MLDP, the Classic-NL-fast and the Classic-NL methods respectively. The proposed method has no criterion with the smallest value, but when comparing its s0-10 (1.221 pixels) and s10-40 (4.7 pixels) errors to those of other methods it exhibits often the best results globally. The SimpleFlow approach [Tao et al., 2012] recorded the highest errors on this dataset with a value of 13.364 pixels for the overall AEPE.

4.4.5 Discussion on benchmarking

The proposed algorithm has consistently performed well in comparison to the state-of-the-art methods on all the three publicly available benchmark datasets [Baker et al., 2011, Geiger et al., 2013, Butler et al., 2012]. On the Middlebury test dataset, the proposed method has an overall AEPE of 0.39 pixels and an overall AAE of 3.81 pixels which are smaller than many baseline methods [Brox et al., 2004, Wedel et al., 2009b, Werlberger et al., 2009, Sun et al., 2014, Demetz et al., 2013]. On the same dataset other methods are more accurate [Xu et al., 2012, Drulea and Nedevschi, 2013, Zimmer et al., 2011, Werlberger et al., 2010] than ours, but this performance is reached at the expense of computation time which is by far higher than that of the proposed approach.

On the KITTI data-set, recent top-level methods (MLDP [Mohamed et al., 2014], CRT-flow [Demetz et al., 2013] and NLTGV-SC [Ranftl et al., 2014]) were more accurate than the proposed method, but this accuracy is again obtained at the expence of computation speed: the MLDP method ([Mohamed et al., 2014], 160s) is 8 times slower than our ROF-NDD method with $k = 1$ (20 s), while the CRTflow ([Demetz et al., 2013], 18s) and the NLTGV-SC ([Ranftl et al., 2014], 16 s) methods require a GPU implentation to reach the CPU time of our CPU algorithm version. However, with $k = 2$, our method has the accuracy very close to the CRTflow method [Demetz et al., 2013] but with an increased computational time of 50 s with a CPU implementation. It is to be noted that the CRTflow method [Demetz et al., 2013] has much higher overall AEPE and overall AAE on Middlebury dataset than the proposed method. Moreover, the proposed method again performed better in comparison to the baseline TV-L1 methods [Werlberger et al., 2009, Brox et al., 2004, Sun et al., 2010, Bao et al., 2014].

For the KITTI images, the NLTGV-SC method [Ranftl et al., 2014] based on the second order variant of the TGV regularization was the most accurate one. But, on MPI sintel dataset, this method was not the top performing method. The overall AEPE of the NLTGV-SC method is 8.746 pixels. However, the same value obtained for its TV version (NLTV-SC) is 9.855 pixels while the proposed method has an overall AEPE of 9.286 pixels only. MLDP [Mohamed et al., 2014] and MDP-flow2 [Xu et al., 2012] methods are the most accurate on the Sintel dataset. The

(a) (b) (c) (d)

Figure 4.9: Simulated illumination changes on the second image of the Grove2 image pair. a) Original image I_{in} b) I_{out} with an additive term $a = 30$, c) I_{out} with multiplicative $m = 1.8$ and d) I_{out} with $\gamma = 3.5$.

proposed method gave more accurate results compared to many of other methods including the baseline methods [Werlberger et al., 2009, Sun et al., 2010, Sun et al., 2014, Tao et al., 2012].

From the benchmarking experiments performed on three standard datasets, we can conclude that our algorithm can efficiently handle illumination variations and scenes with large blur, local deformations and low textures with no compromise in computational speed (only 10 s for Urban image pair of Middlebury dataset, 30 s for KITTI dataset and 50 s for MPI Sintel dataset).

4.5 Experiments for illumination invariance

An evaluation of the proposed **ROF-NDD** method has been done with the image pair Grove2 of the Middlebury training dataset to compare its behavior to that of other self-similarity based approaches under changing illumination conditions. In Eq. (4.21), I^i_{in} stands for the second image of the Grove pair, with $i = R$, G or B representing the red, green or blue color channel. Eq. (4.21) is used to compute an image I^i_{out} simulating illumination changes as proposed in reference illumination robust optical flow contributions [Mohamed et al., 2014, Hafner et al., 2013]. Parameter $a > 0$, $m > 0$ and $\gamma > 0$ correspond to an additive term, a multiplicative factor and a gamma correction respectively. The same parameter values are used for all channels.

$$I^i_{out} = uint8\left(255.\left(\frac{m.I^i_{in} + a}{255}\right)^{\gamma}\right), \tag{4.21}$$

In Eq. (4.21), $uint8$ represents the function converting a real value to an unsigned byte integer and the values of the channels of image I^i_{in} and the simulated image I^i_{out} are all in the [0, 255] interval. The flow field between the pair consisting of the non-modified image and the illumination modified image is computed for different values of parameters m, a and γ. Modified images I^i_{out} are shown in Fig. 4.9. Severe scene illumination changes can be observed in Figs. 4.9(b-d). However, the strongest change can be seen for the gamma correction ($\gamma = 3.5$) in Fig. 4.9(d).

Fig. 4.10 gives the error quantification of the proposed method compared to some recent illumination invariant dense optical flow methods [Drulea and Nedevschi, 2013, Mohamed et al., 2014, Hafner et al., 2013]. The AEPE and AAE are plotted in Fig. 4.10 for each value of the additive term a, the multiplicative term m and the gamma correction γ. It can be observed that for the additive term (refer to Figs. 4.10(a-b)), all the methods are robust to such illumination changes which lead to a very small deviation of the errors when the value of a increases. However, such robustness is not observed for the MLDP method [Mohamed et al., 2014] when the multiplicative term m varies as shown in Figs. 4.10 (c-d). For this method, a large deviation can be

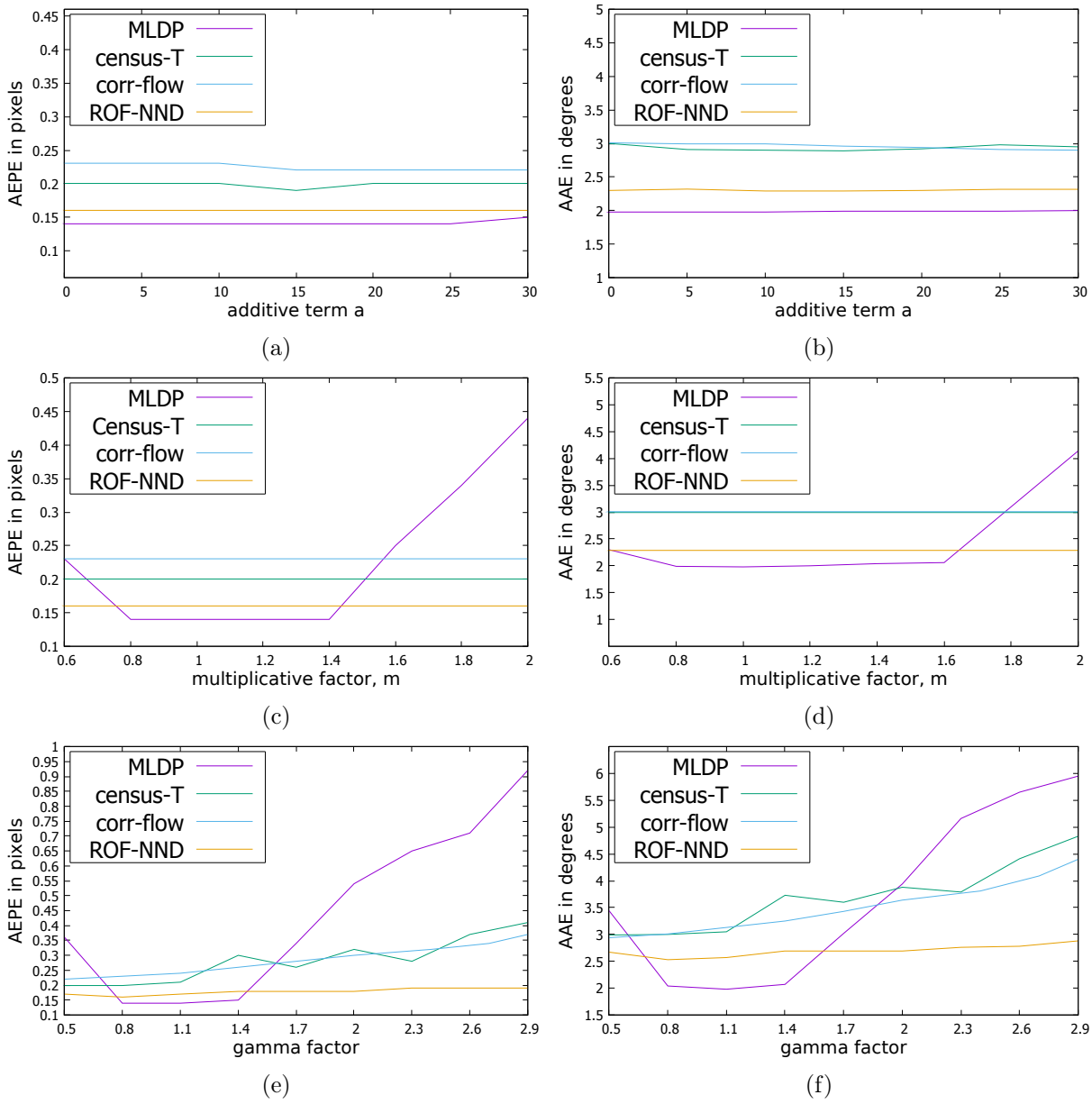


Figure 4.10: Illustration of the effect of illumination changes (by varying a, m and γ) on AEPE and AAE for the MLDP method [Mohamed et al., 2014], the census transform [Hafner et al., 2013], the correlation flow [Drulea and Nedevschi, 2013] and the proposed method (ROF-**NND**).

noted for both the AEPE and AAE criteria at $m = \{0.6, 1.6, 1.8, 2.0\}$. The accuracy of proposed method, the census transform [Hafner et al., 2013] and the correlation flow approach [Drulea and Nedevschi, 2013] is not affected by the tested changes in m. However, for all methods, a strong impact of the gamma correction can be noticed for both the AEPE value in Fig. 4.10 (e) and the AAE criterion in Fig. 4.10 (f). While the proposed method has the least deviation in both the AEPE and AAE errors, the MLDP method has the highest deviation except for $\gamma > 1.4$ and $\gamma < 0.8$. For $\gamma = 2.9$, the AEPE and AAE errors of MLDP method reached almost 1 pixel and nearly 6 degrees respectively. The correlation flow and census-T methods are by far less affected

Method	Seq.#11	Seq.#15	Seq.#44	Seq.#74	Avg.
MLDP [Mohamed et al., 2014]	15.09% (2.87)	10.22% (2.72)	8.88% (1.85)	49.87% (14.85)	21.02% (5.57)
ROF-NND ($k = 1$)	19.09% (4.76)	10.54% (1.96)	16.60% (3.67)	54.41% (16.33)	25.16% (6.68)
corr-flow [Drulea and Nedevschi, 2013]	18.72% (3.40)	12.98% (2.32)	16.88% (5.63)	54.65% (16.60)	25.81% (6.98)
census-T [Hafner et al., 2013]	19.83% (5.06)	15.03% (3.41)	24.30% (7.96)	51.10% (15.14)	27.57% (7.89)
OFH [Zimmer et al., 2011]	24.32% (6.48)	18.34% (3.63)	11.17% (2.44)	57.40% (17.25)	27.81% (7.45)
SRB [Sun et al., 2014]	27.83% (6.43)	18.93% (4.05)	14.66% (2.44)	57.36% (17.36)	29.69% (7.54)
BW [Bruhn et al., 2005]	20.54% (3.62)	36.85% (6.67)	22.38% (3.16)	67.22% (23.77)	36.75% (9.305)
MDP-flow2 [Xu et al., 2012] ($\lambda = 6$)	23.69% (4.47)	53.85% (18.13)	26.38% (3.74)	67.30% (23.77)	42.80% (14.85)
HS [Horn and Schunck, 1981]	25.98% (6.79)	49.57% (7.95)	34.18% (4.61)	79.57% (21.55)	47.32% (10.23)
WPB [Werlberger et al., 2010]	39.25% (18.75)	60.50% (17.63)	40.85% (5.88)	87.02% (24.09)	56.90% (16.60)

Table 4.4: Percentage of bad pixels and AEPE value (in brackets), at AEPE threshold of 3 pixels, for the state-of-the-art methods and the proposed **ROF-NND** method. The results are given for the non-occluded ground truth of the four KITTI training image sequences (**#11, #15, #44 and #74**) which include illumination changes.

by gamma changes than the MLDP approach, but still remain more affected than the proposed method.

The KITTI training data-set was used in order to examine the robustness of the proposed method exhibiting illumination changes in a real scenes. This set of images are reference data for validating the robustness against illumination changes [2]. Table 4.4 quantifies the accuracy (at a AEPE threshold of 3 pixels is classically used to determine both percentage of bad pixels and overall AEPE values) of numerous state-of-the-art algorithms and allows for a comparison with the proposed **ROF-NDD** method. It can be observed in Table 4.4 that the MLDP method (with an average of 21.03% of bad pixels) and the proposed ROF-NDD approach (25.16% of bad pixels in average) gave the most and the second most accurate results respectively for the illumination changes of the KITTI dataset sequences. Correlation flow [Drulea and Nedevschi, 2013] has an accuracy very close the proposed method, while the base-line methods [Horn and Schunck, 1981] and [Sun et al., 2014] recorded higher average values for the percentage of bad pixel (47.32% and 29.7% respectively). MLD-flow2 method [Xu et al., 2012], which has the smallest error in the Middlebury dataset, has also a much higher percentage of average BP3 and average AEPE (42.80%, 14.85 pixels) compared to the proposed method (25.16%, 6.68 pixels).

4.6 Experiments for large displacements (chosen $\alpha_{scale} = 0.7$)

Classically, in order to handle large displacements in the scene, many algorithms (like variational approaches) use coarse-to-fine approaches. However, the accuracy of numerous algorithms still

[2] http://www.dagm.de/symposien/special-sessions/

Method	Seq.#117	Seq.#144	Seq.#147	Seq.#181	Avg.
ROF-NND (k=2)	11.69% (2.36)	25.84% (6.07)	7.72% (1.29)	43.76% (19.04)	22.25% (7.19)
MLDP [Mohamed et al., 2014]	13.80% (4.19)	28.02% (6.93)	4.66% (1.05)	46.81% (21.03)	23.32% (8.3)
ROF-NND (k=1)	10.17% (2.21)	27.15% (6.16)	8.02% (1.36)	48.05% (22.28)	23.35% (8.00)
OFH [Zimmer et al., 2011]	9.09% (2.17)	29.62% (6.77)	8.03% (1.98)	52.32% (23.46)	24.76% (8.60)
SRB [Sun et al., 2014]	18.11% (5.28)	39.55% (9.33)	7.55% (1.74)	56.51% (22.88)	30.43% (9.80)
corr-flow [Drulea and Nedevschi, 2013]	14.05% (2.64)	38.75% (8.68)	10.22% (1.60)	63.18% (31.31)	31.55% (11.05)
BW [Bruhn et al., 2005]	22.25% (4.23)	35.01% (8.17)	10.07% (2.20)	59.05% (22.58)	31.60% (11.62)
census-T [Hafner et al., 2013]	20.52% (9.82)	36.29% (7.71)	6.78% (0.95)	65.55% (31.26)	32.29% (12.43)
MDP-flow2 [Xu et al., 2012] ($\lambda = 6$)	35.78% (11.04)	28.79% (5.52)	18.04% (6.30)	59.34% (24.49)	35.48% (11.83)
HS [Horn and Schunck, 1981]	37.82% (9.77)	41.30% (7.32)	18.52% (3.38)	65.77% (23.40)	40.85% (10.96)
WPB [Werlberger et al., 2010]	41.23% (9.18)	41.53% (8.94)	25.92% (4.43)	68.27% (25.96)	44.24% (12.13)

Table 4.5: Large displacements tests on four training image sequences of the KITTI dataset (pair number 117, 144, 147 and 181) with non-occluded ground truth results. Two criteria at error threshold of 3 pixels (percentage of bad pixels and the AEPE value given in brackets), are used to evaluate state-of-the-art methods and the proposed **ROF-NND** method. See also http://www.dagm.de/symposien/special-sessions/ for such large displacements tests).

largely depends on their ability to handle scene variabilities at different scales. In the proposed approach, neighborhood descriptors were used so that the energy minimization is not only based on pixels value taken individually or locally in small image regions, but rather depends on extended neighborhood pixels information. Unlike, the census or rank based approaches where signatures are used for a local neighborhood, the neighborhood descriptors used in our approach are based on patch based similarity functions which are measured as patch distances between the window of a pixel of interest and all the patches obtained with all possible connectivity of this centered pixel (refer to Eq. (4.1)). This means that the descriptors hold the complete information of the pixel of interest with respect to its surrounding pixels. Moreover, a monotonically decreasing function is used further in Eq. (4.2)) to make the descriptors robust to outliers (like due to strong illumination variabilities).

Large displacement tests are performed with the KITTI data-set in order to observe the usefulness of the accuracy of the proposed descriptor as the data-term constrained with non-local regularization in the energy minimization scheme. The patch-size or neighborhood window (*i.e.* value of k) make a significant role in obtaining improved accuracy due to following reasons: 1) increased patch-size will be able to handle occlusions efficiently and 2) in large displacements, constructing patches at each pyramid level will assist in handling largely displaced pixels. Our experiments showed that increasing the scale factor $\alpha_{scale} > 0.7$ did not change the result much for large displacement sequences. The large displacement KITTI training dataset classically used in the literature allow for a comparison of the proposed approach with some well-known reference

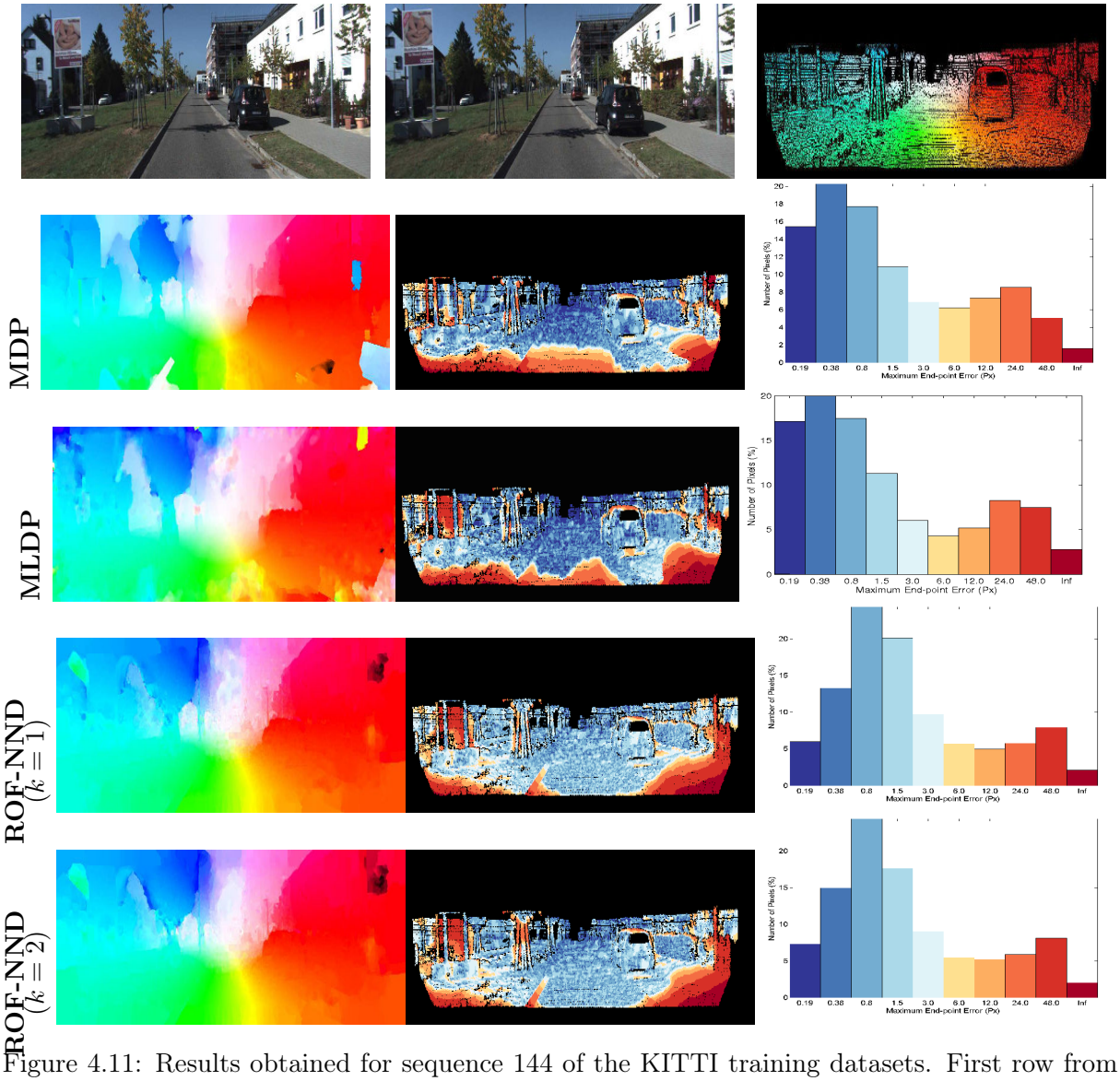


Figure 4.11: Results obtained for sequence 144 of the KITTI training datasets. First row from the top: two images for which the optical flow has to be determined with known ground truth flow on the right. Second row: results obtained for the MDP-flow2 method [Xu et al., 2012] (on the right: flow field with its usual color code, in the middle: end point error image with small and large values in blue and red respectively, on the right: bar chart of the end point errors in pixels). Third row : same results for the MLDP method [Mohamed et al., 2014]. Fourth row: results for the proposed **ROF-NND** method with $k = 1$. Fifth row : results for the **ROF-NND** method with $k = 2$.

methods given in Table 4.5.

As can be seen in Table 4.5, the proposed **ROF-NND** method with $k = 2$ (i.e. with **NND** vectors of dimension 24) led to the best results in the comparison tests (the average bad-pixel percentage is 22.5% at an AEPE threshold of 3 pixels). However, increasing k also increases exponentially the computation time. It is noticeable in Table 4.5 that the proposed algorithm remains competitive when diminishing the neighborhood size to bring it to the value adjusted in Section 4.4.1 ($k = 1$) and systematically used in most of the tests with **ROF-NND** method.

With $k = 1$, the average bad-pixel percentage of 23.35% put our method in the third place of the ranking given in Table 4.5. This accuracy is close to that of the MLDP method [Mohamed et al., 2014] resulting in an average percentage of BP3 of 23.32 %.

Visual results are presented in Fig. 4.11. It can be observed in the error images that the MDP-flow2 approach [Xu et al., 2012] (MDP in Fig. 4.11) and the MLDP method [Mohamed et al., 2014] have numerous pixels in red indicating large AEPE errors. The proposed method (**ROF-NND**) has much lower errors. The error diminishing can be observed in the AEPE image of the last (bottom) row including more blue pixels with weak errors. In the AEPE bar charts given in the last column for the four methods the blue and red colors again represent small and larger values respectively. It can be noticed that with the proposed approach (two last rows on the bottom) the number of occurrences of orange to red pixels (high end point errors) is lower than the MDP-flow2 approach [Xu et al., 2012] (second row from the top) the and MLDP method [Mohamed et al., 2014] (third row).

4.7 Image mosaicing of low textured medical scenes

In previous sections, it has been shown that the ROF-NND method is robust to illumination changes and gives competitive results when compared with different state-of-the-art methods. However, in order to achieve the objective of this thesis, it is important to validate the robustness and accuracy of the ROF-NND method for low textured medical scenes as well. To do so, we have simulated two epithelial surface acquisitions (a bladder video and a skin video, both with known ground truth transformations between the images). These test data were built in a similar way as the simulated test sequences used in Section 3.6.3 of Chapter 3. The results have been compared with the algorithms developed in this thesis for low textured bladder scenes (*i.e.* the RFLOW and the AOFW methods). Additionally, we have also included the results on these sequences for the graph-cut based method adapted to bladder image mosaicing [Weibel et al., 2012b].

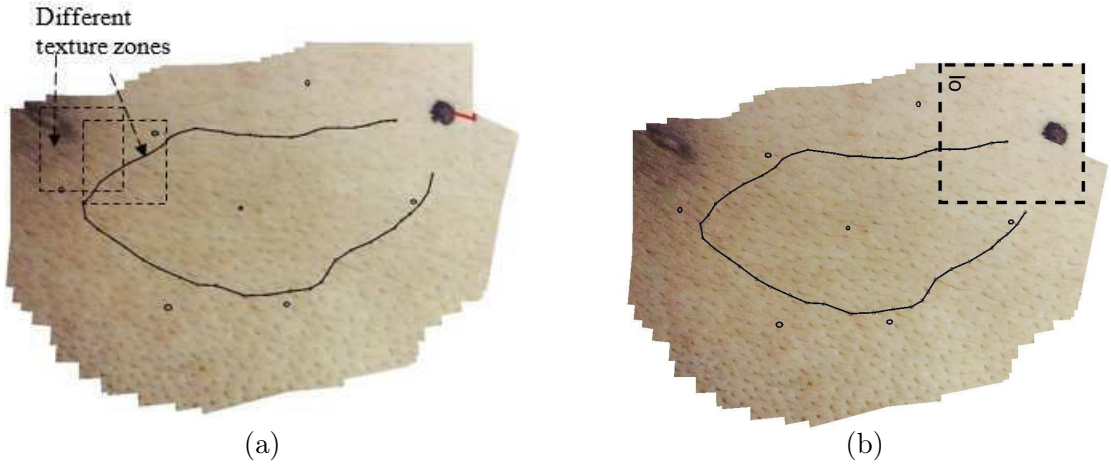
4.7.1 Datasets and evaluation criteria

High resolution images of 1) a human skin surface and 2) the inner surface of an excised pig bladder were first acquired. A dataset with 50 images in each (data-I for skin and data-II for pig bladder) were extracted. A subimage I_0 was first chosen to define a reference coordinate system from the high resolution images. Then, the images I_1 to I_{49} were extracted with known global transformation $H^{true}_{0,i}$, computed using simulated local transformations $H^{true}_{i,i+1}$ superimposing exactly image I_{i+1} on I_i. The intervals of the randomly chosen parameter values of the homographies are shown in Table 4.6.

The ROF-NDD optical flow method is used to determine the dense homologous point corre-

Dataset	θ	s_x, s_y	f_x, f_y	h_1, h_2	$\sqrt{t_x^2 + t_y^2}$
Data-I	$\pm\ 10^0$	0.90-1.10	0.90-1.10	$\pm\ 10^{-5}$	70
Data-II	$\pm\ 5^0$	0.95-1.05	0.95-1.05	$\pm\ 10^{-5}$	50

Table 4.6: $H^{true}_{i,i+1}$ homography parameter intervals used for computing the displacements between consecutive images. θ, $\{s_x, s_y\}$, $\{f_x, f_y\}$, $\{t_x, t_y\}$ and $\{h_1, h_2\}$ are the in-plane rotation, shear, scale, 2D translation and perspective parameters respectively.



(a) (b)

Figure 4.12: Human skin data-I mosaics. 900×1400 pixels mosaic was obtained with I_0 being the first image. (a) Mosaic with RFLOW method. A visual misalignment is shown between the first frame and the last frame with a red line. (b) Mosaic obtained with the ROF-NND method..

spondence to estimated the local homographies $H^{est}_{i,i+1}$ between images I_{i+1} and I_i. Similar to the Chapter 3, two criteria (see Eq. (4.22)) were used for comparing the proposed method with some recent methods in bladder epithelium mosaicing. The local error $\epsilon^{local}_{i,i+1}$ gives the registration accuracy when superimposing pixels p of images I_i and I_{i+1}, whereas $\epsilon^{global}_{0,49}$ is the global (mosaicing) error when placing image I_{49} in the coordinate system of I_0. Recalling the formulas for local and global image registration errors, we have:

$$\begin{aligned} \epsilon^{local}_{i,i+1} &= \frac{1}{N} \sum_{p \in I_i \cap I_{i+1}} \| H^{true}_{i,i+1} p - H^{est}_{i,i+1} p \|^2 \\ \epsilon^{global}_{0,49} &= \frac{1}{N} \sum_{p \in I_0 \cap I_{49}} \| H^{true}_{0,49} p - H^{est}_{0,49} p \|^2 . \end{aligned} \tag{4.22}$$

4.7.2 Validation results

Table 4.7 and Table 4.8 give the results for the human skin epithelium (data-I) and the pig bladder epithelium (data-II) respectively. Accurate methods that are adapted for image mosaicing of scenes with large texture variability like that in bladder scene have been used used for comparative

Method	$\epsilon^{local}_{i,i+1}$ (in pixel)			$\epsilon^{global}_{0,49}$	$\bar{\mathbf{t}}$
	min	max	mean	(in pixel)	(in s)
Graph-cut method [Weibel et al., 2012b]	0.18	4.72	0.84	36	20
RFlow method [Ali et al., 2014]	0.15	2.23	0.70	30	4
AOFW	**0.04**	1.32	0.37	9.1	5
ROF-NND	**0.04**	**0.21**	**0.18**	**3.1**	**3**

Table 4.7: Registration and mosaicing results obtained for datset "data I" (human skin epithelium).

Method	$\epsilon_{i,i+1}^{local}$ (in pixel)			$\epsilon_{0,49}^{global}$ (in pixel)	$\bar{\mathbf{t}}$ (in s)
	min	max	mean		
Graph-cut method [Weibel et al., 2012b]	0.32	6.8	3.01	11	48
RFlow method [Ali et al., 2014]	**0.02**	1.56	0.33	7.5	**3**
AOFW	0.03	**1.06**	0.21	**4.4**	5
ROF-NND	0.03	1.23	**0.17**	4.8	4

Table 4.8: Registration and mosaicing results for dataset "data-II" (pig bladder phantom).

analysis. From the Table 4.7, it can be observed that in case of low textured human skin epithelium sequence, the graph-cut based method and RFLOW method that are adapted for image registration of bladder epithelium [Weibel et al., 2012b, Ali et al., 2014] lead to large global registration errors of 36 pixels and 30 pixels respectively. The smallest error of 3.1 pixels was obtained for the proposed method in this chapter (ROF-NND). It was observed that, for the image pairs with difference in texture zones (refer to Fig. 4.12), all the methods except ROF-NND method gave large errors with a magnitude of local registration error greater than 1 pixels. A valid reason for large errors comes from the fact that an abrupt change in illumination was observed between the image pairs indicated in Fig. 4.12. Among these methods, the largest error was observed for the graph-cut based method (4.72 pixels).

In Table 4.8, for the ROF-NND method and the AOFW method shows very competitive results. The graph-cut based method still being the least accurate method among the compared methods, AOFW gave a global error of only 4.4 pixels. The pig-bladder sequence (data-II) has the presence of textures (blood vessels) in them and with only small change in illumination (as can be observed in first few frames of Fig. 4.13), both the RFLOW and the AOFW methods perform well on this data.

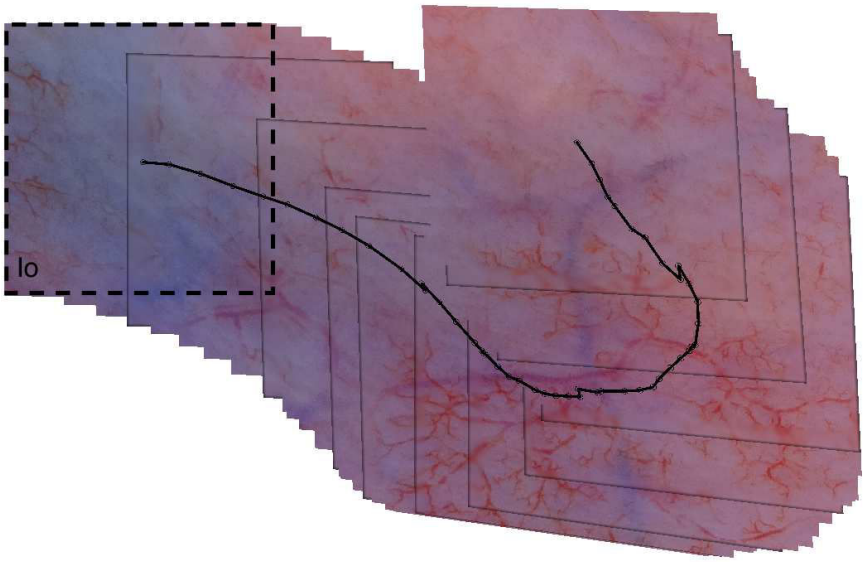


Figure 4.13: Results for dataset "data II": pig bladder mosaics built with the homographies $H_{i,i+1}^{est}$ estimated with the proposed ROF-NDD method. This mosaic has a size of 900 × 1500 pixels.

In all the above results, graph-cut based method has the highest average time for paired image registration while the proposed methods are comparatively computationally efficient (almost 5 times in skin epithelium and almost 15 times bladder epithelium).

4.8 Main contributions and conclusion

The main contributions of this chapter are:

- The modeling of a data-term based on self-similarity neighborhood descriptors for improving the robustness of optical flow determination under changing illumination conditions.
- The use of a non-local weighted regularization implemented in the TV-L^1 energy minimization framework. The non-trivial optimization step is simplified using a linear operator.
- The algorithm benchmarking on three standard public datasets used for optical flow estimation.
- The comparison of the proposed method with many state-of-the-art methods (both TV-L^1 baseline methods and illumination robust methods).
- The validation of the robustness of the method against illumination changes is done. The method is also compared to illumination robust methods in the literature.
- Since, the method builds its descriptors with the patches in its neighborhood pixels, extending the patch size makes the algorithm more accurate than other state-of-the-art methods. In other words, the occlusion problem is tackled efficiently in extended neighborhood patches. Such an experiment was presented in the Section 4.6.
- Testing the appropriateness of the method on weakly contrasted phantom data. The result obtained for low-textured medical test sequences with known ground truths are a first indication that the algorithm can be used for image mosaicing of different textured scenes including illumination changes.

In this chapter, an algorithm robust to illumination changes (both for strong and weak illumination changes with either local and/or global changes) have been proposed. This robustness is due to the fact that self-similarity descriptors, built with the neighborhood pixels in the proximity of the pixel of interest, hold strong edge information which is invariant to the illumination changes. The components of the descriptor vectors relate to the orientation of the objects /textures and to the magnitude of their edges. Embedding the descriptor vectors in the data-term contributes to the algorithm robustness towards illumination changes. Moreover, the use of the weighted non-local median filtering in the framework of the primal-dual energy minimization has enabled the ROF-NND algorithm to be more accurate than many state-of-the-art methods in literature. The illustration of the proposed method on different textured scenes including low texture medical scenes confirmed the usability of this proposed method for various scenes, especially for the scenes with strong illumination changes. These results have to be confirmed in Chapter 5 which is dedicated to complicated real scenes (various endoscopic scenes and other medical and non medical scenes).

List of publications

[ADGA] Sharib Ali, Christian Daul, Ernest Galbrun and Walter Blondel "Illumination invariant optical flow using neighborhood descriptors," Computer Vision and Image Understanding, Available online 17 December 2015, http://dx.doi.org/10.1016/j.cviu.2015.12.003.

Chapter 5

Image mosaicing in endoscopy and of other complicated scenes

Contents

5.1 Motivation: Fast, accurate and robust image mosaicing of various complicated real data sequences . . . 122
5.2 Point correspondence estimation for for complicated scenes . . . 123
5.3 Endoscopic dataset . . . 124
5.4 Qualitative mosaicing results on patient data . . . 125
5.4.1 White light cystoscopy . . . 125
5.4.2 Fluorescence cystoscopy . . . 131
5.4.3 Gastroscopy . . . 133
5.4.4 Laparoscopy . . . 134
5.5 Quality mosaics for other scenes . . . 134
5.5.1 Dermoscopy . . . 134
5.5.2 Underwater scenes . . . 135
5.5.3 Video mosaic of the Mars surface . . . 135
5.6 Main contributions and conclusion . . . 137
List of publication . . . 137

5.1 Motivation: Fast, accurate and robust image mosaicing of various complicated real data sequences

The robustness and accuracy of image registration algorithms largely depend on the characteristics of the scenes for which image sequences are acquired. Important scene characteristics are related to the texture and the illumination conditions. A severe variation in texture and/or illumination is often observed in medical scenes, especially in human organs such as on the internal bladder walls (cystoscopy), the liver (laparoscopy) and the stomach (gastroscopy). Notably, strong intra- and inter-patient texture variability is present in cystoscopic and dermoscopic image sequences (see Figs. 5.1 (a-d)). Images which are affected by large specular reflections are observed in gastroscopic and laparoscopic images as shown in Figs. 5.1 (e-f). In such video-sequences, weak textures combined with changing illumination conditions need the development

of robust and dedicated image registration techniques as discussed in Chapter 1. Additionally, in other non-medical scenes like underwater scenes or space exploration scenes, image sequences have similar complications making the optical flow algorithms more challenging. For example, for underwater scenes one major problem that is observed is the presence of large areas with repeated textures and non-uniform illumination. Such data, shown in Fig. 5.1 (a), causes severe correspondence ambiguities when feature extraction and matching approaches are used for homologous point estimation. These ambiguities do not guarantee robust image mosaicing. Moreover, due to viewpoint changes, depth changes in different regions of an image (refer to Fig. 5.1 (b)) and different lighting conditions in standard modalities (refer to Fig. 5.1 (c-d)) strong illumination differences are often observed in scenes and leads to high registration errors when no robust algorithms are used.

The motivation of this chapter is to investigate the advantages of the optical flow approaches proposed in Chapters 2 (RFLOW), 3 (AOFW) and 4 (ROF-NDD) which comply different scene types, image modality and illumination variability. Though this chapter is focuses on endoscopic scenes in general it also present examples of other complicated scenes such as underwater and space exploration scene. Applying the proposed algorithms to a large variety of scene aims at showing their robustness and generality.

5.2 Point correspondence estimation for for complicated scenes

The optical flow accuracy of the methods proposed for scenes with weak textures, image blur, pure in-plane rotations and/or small (local and global) illumination changes were discussed in Chapters 2 (RFLOW method) and 3 (AOFW approach). Additionally, the illumination invariant optical flow method introduced in Chapter 4 (ROF-NDD algorithm) was able to compute accurate optical flow fields, even for large displacements in varying texture conditions. Each of the three proposed algorithms has its own advantages and suits especially for a particular set of scenes.

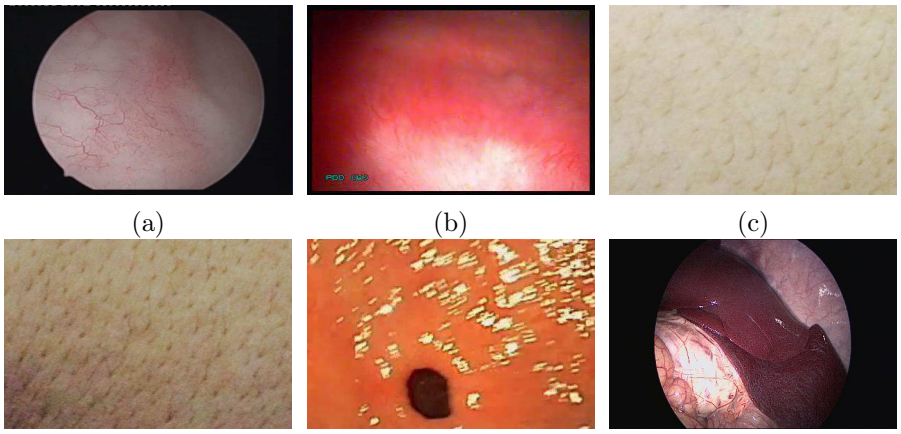
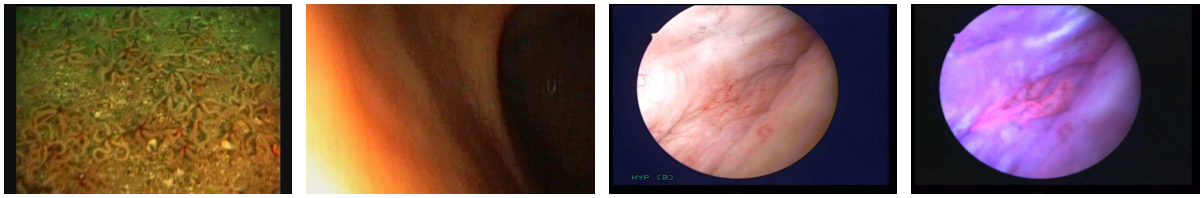


Figure 5.1: Strong scene variability inside and between image modalities. (a-b) Intra-patient texture and illumination variability in cystoscopy for images acquired with a rigid (a) and flexible (b) cystoscope repectively. (c-d) Inter-patient illumination variability in dermoscopy. (e-f) Strong specular reflections due to moistness of regions around the organs (stomach and liver respectively).



(a) (b) (c) (d)

Figure 5.2: Illustration of repetitive patterns and of scenes with illumination variability. (a) Non-uniform illumination and repeated texture in underwater scene. The illumination changes occur notably due to reflection of light from moving water during image acquisition. (b) Non-uniform illumination due to the organ depth and view-point. (c-d) Illumination changes in bimodality cystoscopic imaging: white light modality (WL) in (b) and fluorescence modality (FL) in (c).

In the preliminary results given in Chapter 2, the first proposed method (RFLOW) efficiently preserves the image textures and is robust to noise and small illumination changes. However, the optical flow accuracy is strongly affected in case of in-plane rotations. The AOFW method was proposed in Chapter 3 to deal with the in-plane rotations. A curl operator along with the divergence was used to recover accurately the actual displacements for in-plane rotations. Additionally due to the use of the multiresolution approach with Riesz wavelet basis filters, weak textured images were efficiently handled. The "flattening-out" problem at coarser levels of the image pyramids was thus resolved. The experiments conducted for images with low and repeated texture, image blur and small illumination changes in Chapter 3 demonstrated the usability of AOFW method for complicated scenes like bladder and other weak textured image sequences. However, both the RFLOW and AOFW methods were not suitable for strong illumination variations often arising when changing the modality in cystoscopy and when specular reflections occur as in gastroscopic or laparaoscopic images. An algorithm (ROF-NND) which is robust towards such strong illumination changes was proposed and validated in Chapter 4. The ROF-NND was validated for obtaining accurate optical flow fields for various scene types on publically available datasets. Additionally, the competitiveness of the proposed ROF-NND method was validated against the other proposed methods (RFLOW and AOFW) for robust and accurate image registration on sparsely textured dermoscopic image sequence and bladder image sequence.

5.3 Endoscopic dataset

Patient video-sequences under WL cystoscopy were acquired by Prof. François Guillemin at the comprehensive oncology center of Nancy (Institut de cancérologie de Lorraine). These data include image sequences acquired by both rigid and flexible cystoscopes. The age group of the patients under observation (pre- and post- diagnosis procedure) were between 45-70 years including both male and female patients. Patient data under FL modality was provided by Centre Hospitalier Vaudois at Lausanne, Switzerland. The sequences were acquired of 40-50 year old man (smoker). Gastroscopic videos sequences were acquired by Prof. Dominique Lamarque who is a gastroentorologist at the Ambroise Paré hospital from Boulogne Billacourt. The objective of such image acquisition is to observe the inflammations around the pyloric antrum region of the stomach. Another, endoscopic video-sequence was provided by Christophe Doignon from the Icube laboratory. This video was acquired with a laparoscope at the Institut de Recherche Contre les Cancers de l'Appareil Digestif (IRCAD) in Strasbourg, France.

5.4 Qualitative mosaicing results on patient data

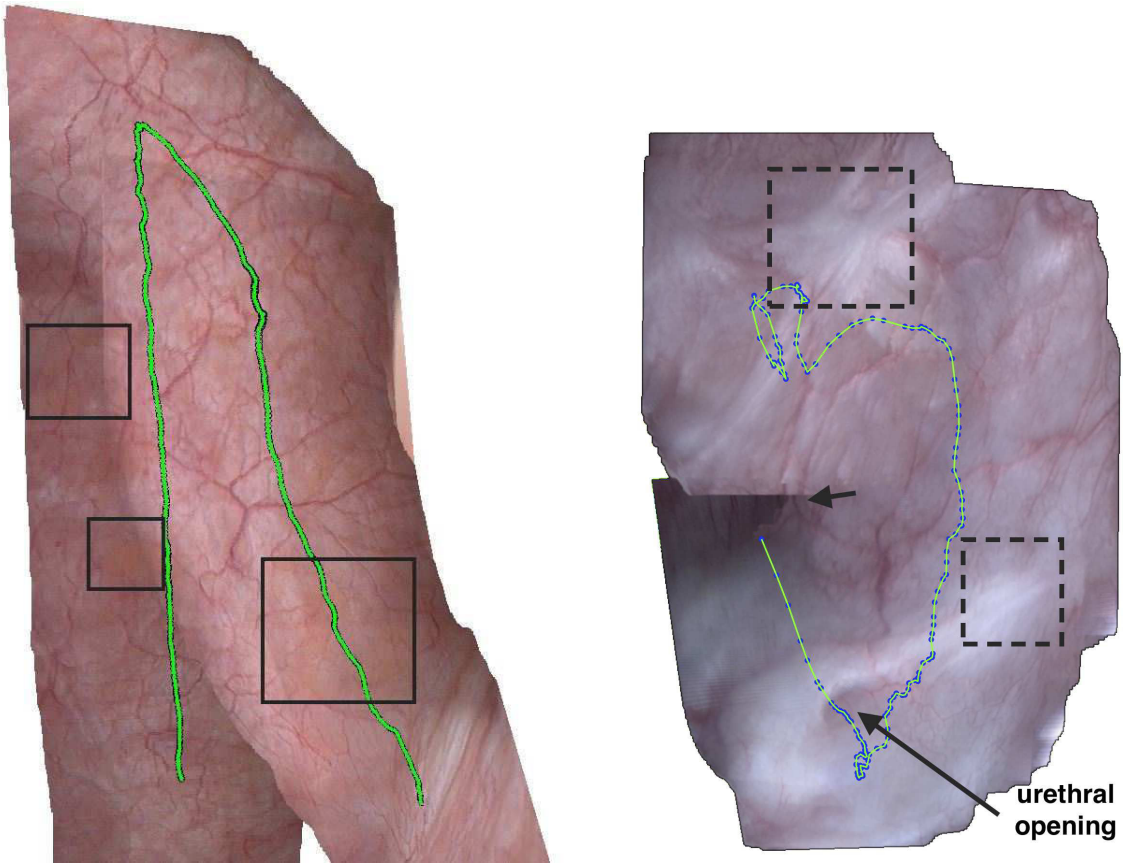
In this section, we show qualitative results of our proposed algorithms on various endoscopic dataset. While cystoscopic data being the main focus of this thesis, various mosaics for both inter- and intra-patient is presented, both for white light modality and fluorescence modality have been presented. Being able to robustly mosaic other complicated endoscopic data also validates the wide applicability of the designed algorithms. This section will also give a clear insight to the reader about the advantages and limitations of the proposed algorithms in this thesis.

5.4.1 White light cystoscopy

Efficient preoperative diagnosis requires the inner bladder wall to be fully scanned by clinicians. But, the small field of view of cystocopes cause a major obstruction to this objective since it is not easy for clinicians to mentally reconstruct the bladder with only a video seen on a screen and to decide whether the complete organ was scanned or not. Enlarging the field of view through image mosaicing can help the clinician to scan the bladder while diminishing the risk of non observed bladder parts [Hamadou et al., 2009, Behrens et al., 2009, Weibel et al., 2012b, Bergen et al., 2013a, Ali et al., 2015a, Weibel et al., 2010]. The methods proposed in this thesis are designed for the building of accurate mosaics without compromise in speed, thus making the second diagnosis possible in the examination room while the patient is dressing up. On one hand, the state of scares or multi-focal cancerous lesions have to be observed globally. This is not possible in single image of video-sequences as it shows only a part of region of interests. On the other hand, comparing mosaics built with videos acquired from a patient facilitates lesion follow-up and examination traceability. Such mosaics can be interpreted even after some weeks or months after the acquisition, while lesions evolution assessment with direct use of video-sequences is not possible. This is the reason why bladder video-sequences are most often not archived.

Bladder mosaicing in presence of distributed texture

Two mosaics of patient cystoscopic data are presented in Fig. 5.3. Fig. 5.3(a) shows a loop trajectory video-sequence with scars delineated by solid lined black rectangles. These scars can be seen as the slight hue changes of the bladder epithelium. In some image pairs, simultaneously affected by image blur and acquired from viewpoints with large endoscope displacements, the registration was very challenging. However, for these image pairs, a robust and very accurate registration was possible with the RFLOW and the AOFW optical flow methods. According to the image quality (presence of well spread and contrasted textures or images with weak textures or blur), the automated switching method described in [Ali et al., 2013b] was used to select either the proposed optical flow approach (RFLOW in Fig. 5.3 (a) and Fig. 5.3 (b)) or a feature based approach (SURF, [Bay et al., 2008]) for determining the correspondence between homologous points. A very accurate and visually coherent mosaic was built as shown in Fig. 5.3 (a) where no texture discontinuities are perceptible between overlapping and non consecutive image pairs in the video-sequence. Indeed, due to accumulating registration errors, texture discontinuities are mainly visible for such image pairs. Here, the registration errors were constantly small enough so that the textures remain coherent (i.e. without discontinuities) in all mosaic parts. It is worth noticing that the mosaicing error (i.e. the registration error accumulation) increases drastically for each inaccurate registration. Using the optical flow method for less textured images instead the feature based method allowed us to obtain this coherent mosaic. An advantage of such



(a) (b)

Figure 5.3: Patient cystoscopic mosaics. (a) First image mosaic of 500 images constructed with the RFLOW algorithm. Scars can be seen on the bladder wall, some of them are shown in black rectangles. (b) Second bladder map of an urethral opening region (200 image pairs). This mosaic was built with the AOFW method. The low textured areas with dashed rectangles represent healed scar regions. A black arrow at loop closing in (b) shows a small misalignment between the vessels.

switching between optical flow and feature based methods for homologous point estimation was a gain in mean registration time down to 0.55 s per frame while preserving the accuracy (with only the optical flow method the mean registration time of an image pair was 3.5 s). In this example due to well spread texture present in almost all images, only 20 image pairs were registered using our optical flow model while the remaining image pairs were superimposed with the feature based method. However, with a feature based approach, the failure risk of the registration of one of these 20 image pair is very high and impedes the mosaic construction.

Fig. 5.3(b) shows the mosaicing results for a more complicated video-sequence including an urethral opening region. Large homogeneous regions (*i.e.* with weak textures) can be seen in the dashed rectangles. The corresponding white scars in Fig. 5.3(b) are due to the healing of tumors and other lesions done by a process called "fulguration" (*i.e.* burning of the cells). In these image pairs, the SURF method was able to find enough features for homography computation in only 5 image pairs out of 200. A feature based approach alone was unable to deal with such sequences. Conclusively, mosaicing for longer sequence of this cystoscopy image sequence was solely possible because of the robustness of the AOFW method. A very small global misalignment can be seen

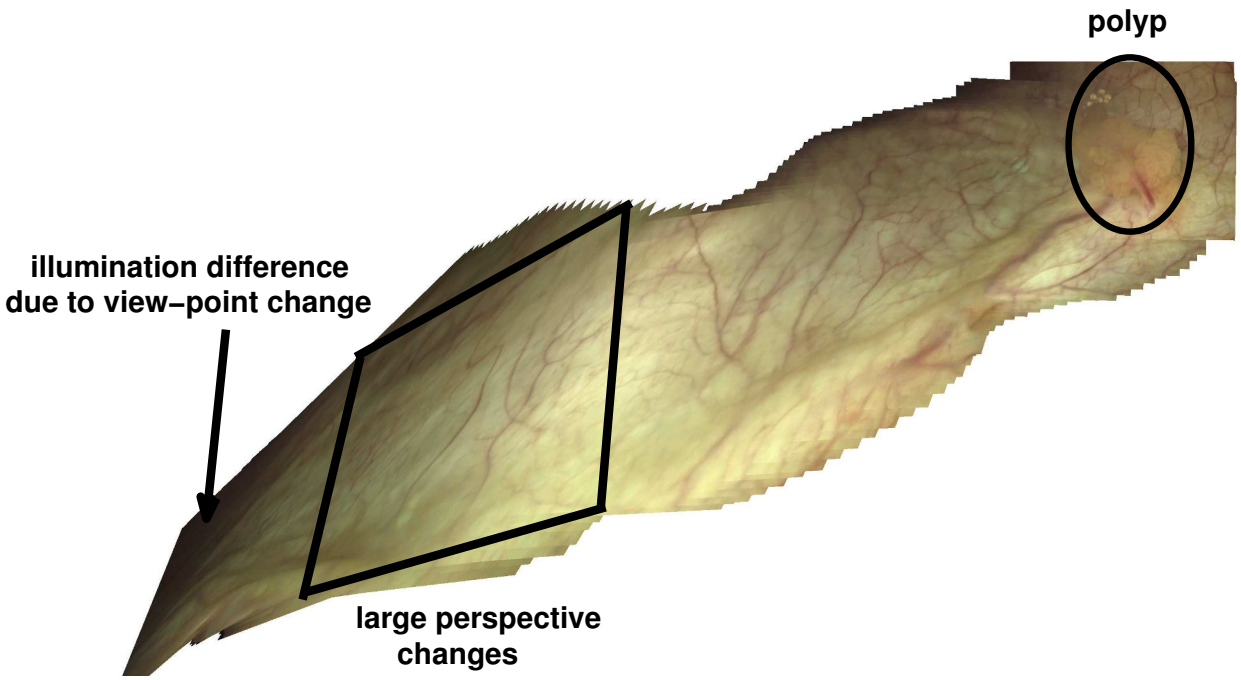


Figure 5.4: Mosaic using 500 frames of patient data. Texture and illumination variability is persistent. Illumination changes are due to view-point changes. A polyp is shown in circular black region. Additionally, scale- and perspective- changes are observed when moving from right to left.

at the loop closing spotted by the arrow in the Fig. 5.3(b). This mosaic was obtained with a mean image registration time of 1 second per frame.

Fig. 5.4 gives an idea of texture and illumination variability in intra-patient data. It can be observed that due to strong view point changes during the camera motion, changes in contrast (refer to the area pointed by the arrow in Fig. 5.4) and geometrical transformations (refer to the quadrangle in Fig. 5.4) are stronger. A polyp indicated by the black ellipse. The mosaic was obtained using AOFW with a mean registration time of 3 s per frame.

Image mosaicing in presence of low texture and pure/large in-plane rotations

A third mosaic built with a video-sequence of 255 images is shown in Fig. 5.5. It demonstrates in a qualitative way, the robustness and accuracy of the AOFW method. This is an important test sequence because it involves large in-plane rotations as can be seen at regions with trajectory loops. One criterion for visual evaluation of the registration accuracy is the continuity of the vessels at the points indicated by the black arrows at image pair interfaces in the non-blended mosaic of Fig. 5.5. Indeed, when passing from one image palced in the mosaic coordinate to another the no vessel discontinuity must be observed for ensuring the visual coherence of the panorama. In Fig. 1.5 no blending or color correction were done so that the image borders remain perceptible. The time required to obtain this mosaic was approximately 3 minutes (*i.e.* stitching time of ≈ 0.6 s/frame).

A major advantage is that computationally expensive and rigorous bundle adjustment algorithm was not required for obtaining this realistic mosaic. In case of inaccurate mosaicing, texture misalignment have to be corrected *i.e.* time consuming global map correction algorithms

have to be applied as described in [Weibel et al., 2012b, Weibel et al., 2012a].

Image mosaicing in presence of large specular reflections

Mosaics of video-sequences obtained under WL modality with a flexible cystoscope are shown in Figs. 5.6 (a-b). It can be observed that the image pairs used to build these mosaics are affected by strong specular reflections. The RFLOW method and the AOFW method worked only for first few image pairs of these sequences. As visible in Fig. 5.6 (a), the first sequence is not only affected by strong specular reflections, but is also with very large homogeneous regions. In these scene conditions, it was only possible to compute the optical flow between the image pairs using the illumination robust ROF-NND method. Similarly, for the sequence in Fig. 5.6 (b) only the ROF-NDD method was able to compute the optical flow for the strong specular reflections at the beginning of the sequence (on the right of the image in Fig. 5.6 (b)) lead thus to a complete mosaic. The other approaches were not able to compute a mosaic for this sequence.

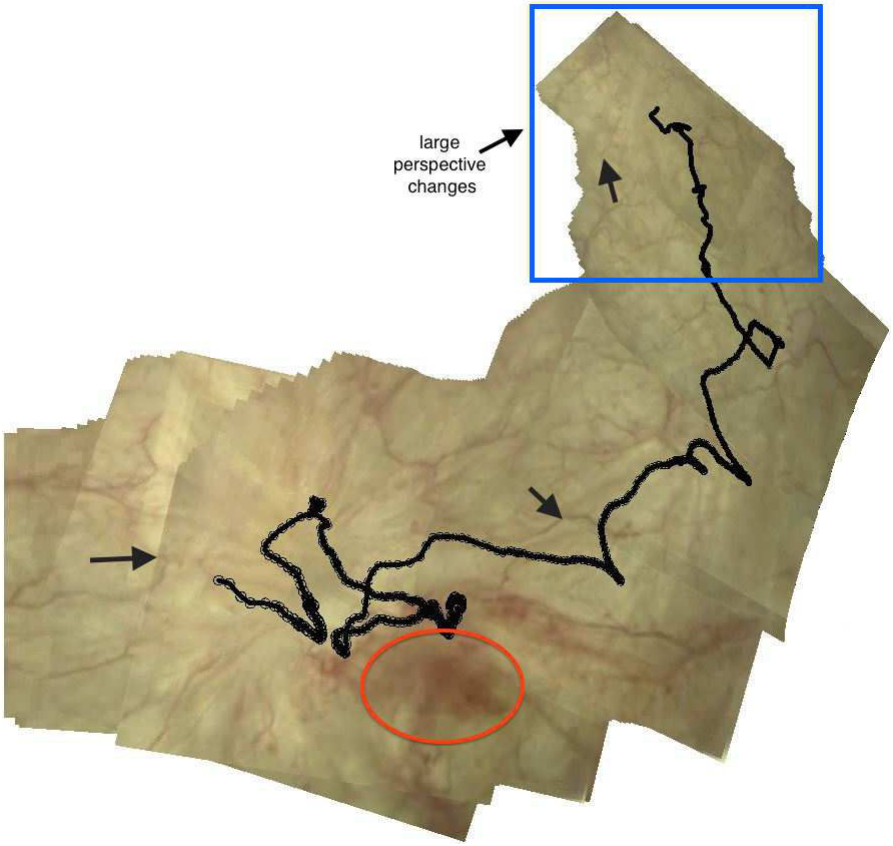
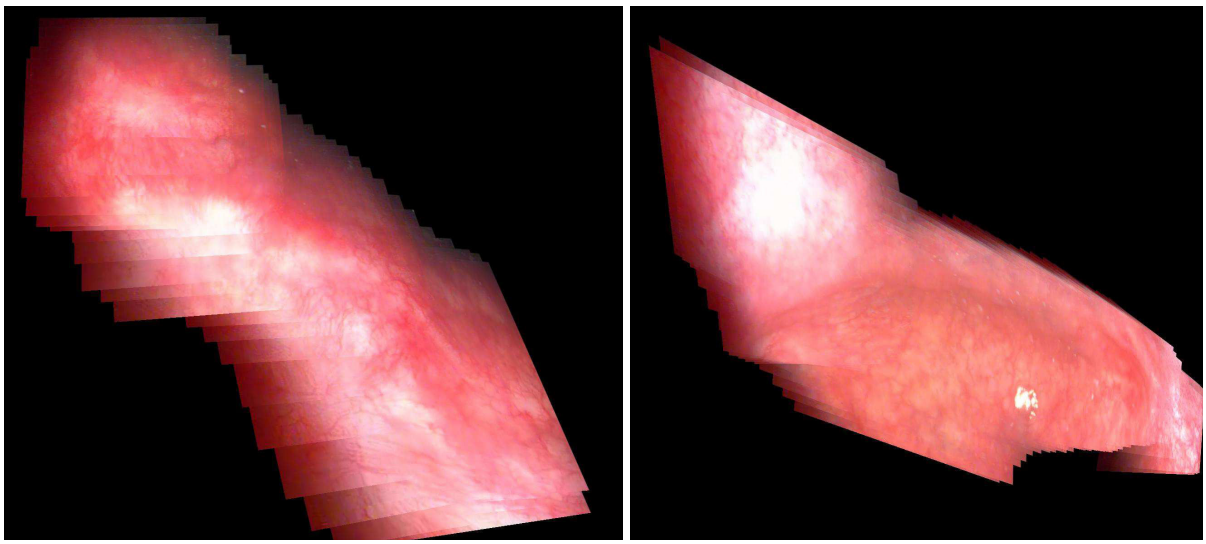


Figure 5.5: Bladder mosaic with strong in-plane rotations and perspective changes. The acquisition of the cystoscopic video-sequence was done after transurethral resection of a bladder tumor. The red circle represents the resected region with blood stains. The black line represents the reconstructed trajectory of the cystoscope projected onto the mosaicing plane. The arrows indicate vessel continuity points (for qualitative/visual estimate of global registration errors) at 30th, 175th and 225th image pairs respectively (left to right).



(a) (b)

Figure 5.6: Mosaic of patient data obtained with flexible cystoscope. (a) Strong specular reflections in all the image pairs additionally with large displacement and strong perspective changes. (b) Strong specular reflection along with non-planar organ surface showing an air bubble in the cavity.

Image mosaicing in presence of large displacements

The textures of the WL mosaic shown in Fig. 5.7 are visually different from those of the video-sequences previously evaluated and discussed. No texture discontinuity is perceptible in this mosaic, mostly at places with tissue structures (e.g. polyps). Weak and homogeneous textures

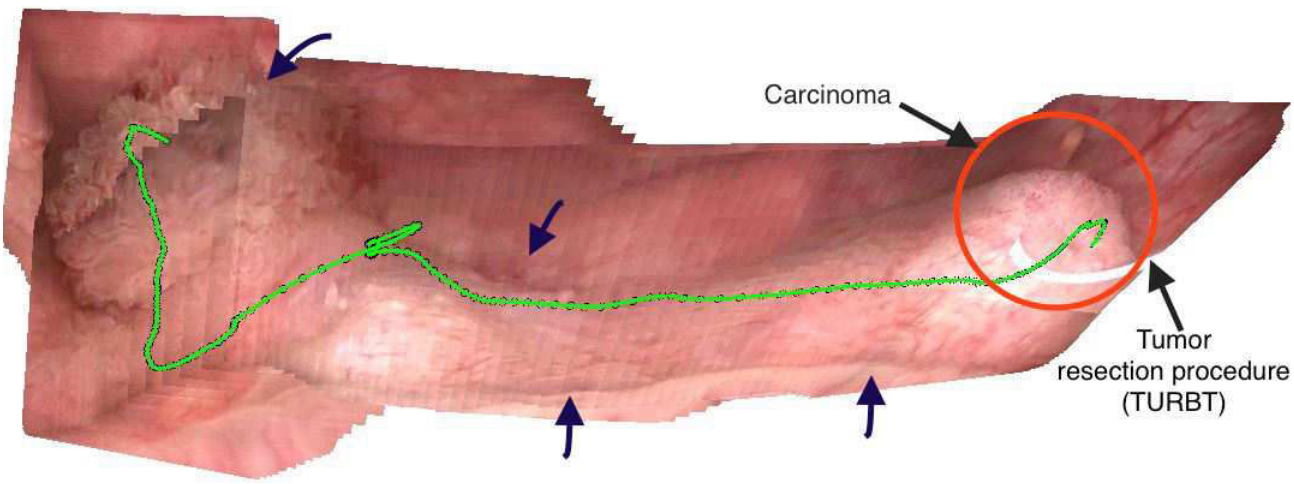


Figure 5.7: Mosaic of every 10^{th} frame of patient data. Image mosaic showing the trans-urethral resection (surgical) procedure. White looped wire can be observed at the end (on the right) of the mosaic which is being used for removing the bladder polyp in extended FOV bladder mosaic. Green lines in mosaics represent the reconstructed camera trajectory

are observed in many image pairs at the mosaic sequence. The optical flow computation and the registration of image pairs were possible with the AOFW and the ROF-NND methods. Large misalignments were observed for few image pairs having large in-plane rotations in this sequence. The accuracy of the ROF-NDD optical flow method is visually validated at the places where blue arrows indicate the continuity and alignment of structures in the mosaic in Fig. 5.7. The green curve represents the trajectory of the registered image centers (*i.e.*, ideally, the projection of the 3D endoscope trajectory into the mosaic plane). In the wide field of view mosaic of Fig. 5.7, it is possible to observe simultaneously a polyp (at the left) and the resection of another one (at the right). This mosaic used every 10^{th} frame in the sequence. Images pairs (I_1, I_{i+10}) involve displacements of more than 20 pixels. Due to their coarse-to-fine energy minimization strategy, all the proposed techniques proposed in this thesis (RFLOW, AOFW and ROF-NDD) were able to efficiently handle these large displacements.

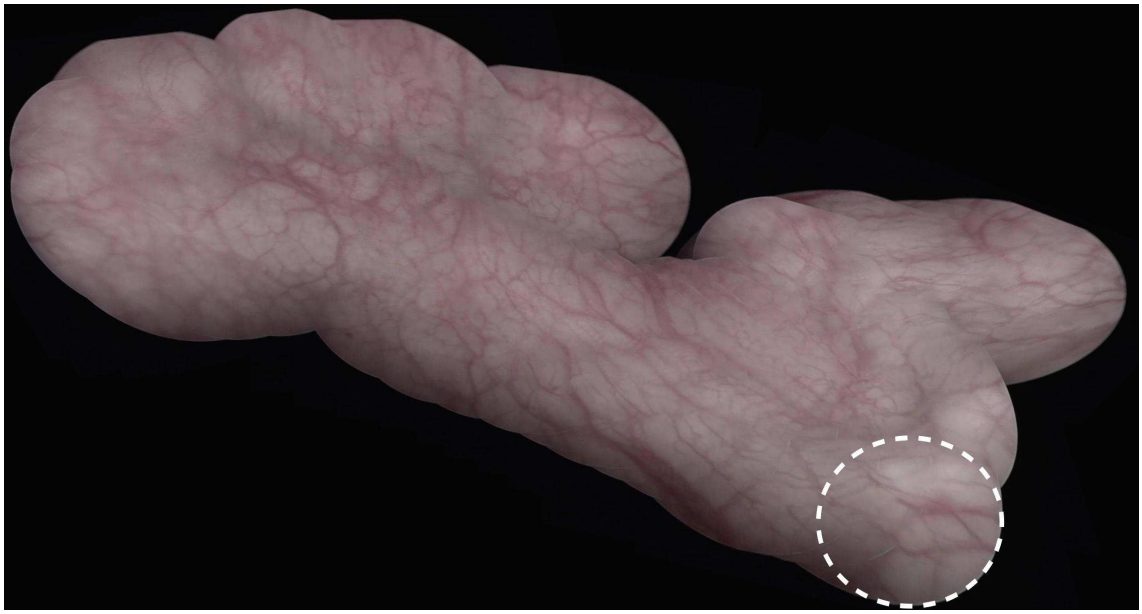
Mosaicing of large image sequences

In the literature, often short sequences are used because the registration errors generated by these algorithms accumulate themselves. Such accumulating errors lead to visually incoherent mosaics when the sequences are large. Such incoherence is is visually disturbing, especially at the loop closings. However, we have already shown in our experiments in Chapter 3 and Chapter 4 that our proposed methods (the AOFW and the ROF-NND) gives very accurate paired image registrations leading to negligible mosaicing (global) errors. But, to minimize accumulating errors it would be important to place a bladder mosaicing algorithm in optimal conditions by controlling the acquisition conditions: the organ scanning procedure should notably avoiding very large perspective changes. However, this is not possible all the time for urologists or surgeons.

The white circle in Fig. 5.8 (a) gives the position of the first image in the mosaic and represents the field of view of the cystoscope. It can be noticed that this field of view was significantly increased. Large perspective changes can be seen at the mosaic end (image which is both over the white circle and completely on the right of the mosaic). At this closing loop place, no texture discontinuities are perceptible even for this rather long image sequence. This means that the accumulating registration errors (leading mosaicing errors) remained weak in all places of the mosaic, even without global panorama correction.

A second visually coherent mosaic is given in Fig. 5.8 (b). The first image is again the one which is both on the bottom and on the right in the mosaic, while the last one is on the top and the left in a mosaic region were the image size becomes smaller. This size change is due to fact that the distance between endoscope's distal tip and the surface constantly changed during the acquisition (the first image is at original data resolution). It can be observed at the end of the mosaic in Fig. 5.8 (b) that even for large scale changes the mosaic remains visually coherent.

Providing a solution ensuring a complete bladder scan is important, for instance to be sure to miss no tissue with lesions. However, computing a unique mosaic of the complete bladder is not of interest since the resolution of the images degrades itself along the mosaicing process due to perspective changes, scale changes and other geometrical transformations. This leads to mosaics with large regions in which the image resolution is inappropriate (too weak) or in which the textures and structures are too distorted. Moreover, computing successfully a mosaic of a complete bladder is very uncertain (or impossible), whatever the algorithm (notably due to the continuous transformation changes).



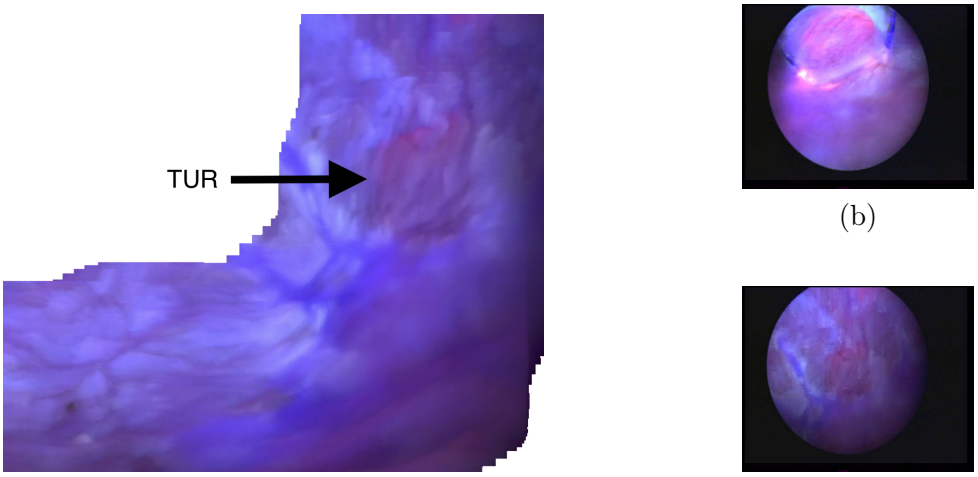
(a)

(b)

Figure 5.8: Mosaic built with large sequences. (a) Inner bladder wall mosaic using the ROF-NND method. It uses 900 frames corresponding to 35 s of cystoscopic video data. The white circle represents the FOV of a video frame. (b) Video image sequence corresponding to 25 seconds of the same video sequence used in (a) and consisting of 618 frames. A maximal displacement of 30 pixels between the image pairs occurs in this video-sequence extract.

5.4.2 Fluorescence cystoscopy

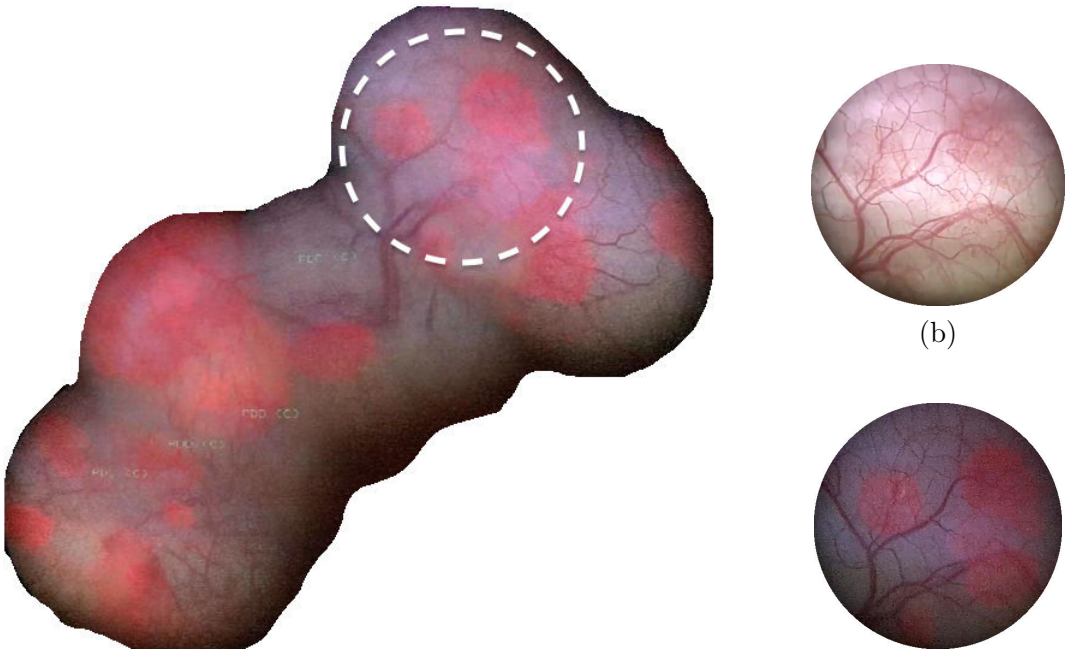
Fluorescence cystoscopy detects hard-to-find bladder tumors that might not be visible by standard WL cystoscopy. These blue-light techniques improves tumor detection that assists an



(a) (b) (c)

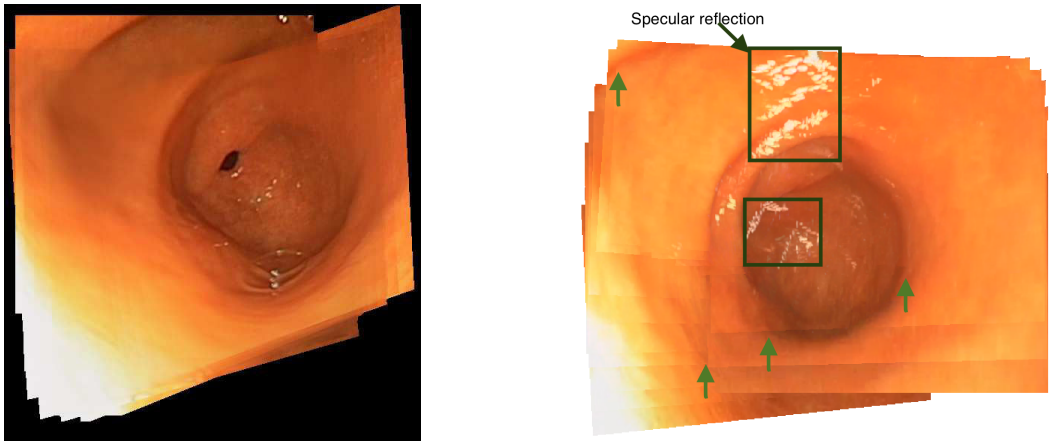
Figure 5.9: Mosaicing tests under FL modality. (a) Patient data mosaic built with 100 image pairs. (b, c) Bladder region in original small FOV image before and after transurethral resection of bladder tumor (TURBT). The corresponding bladder region in mosaic a) is indicated by an arrow.

improved resection and thereby reduce residual tumors. The interest of combining WL and FL mosaics was discussed in [Hernández-Mier et al., 2006]. Figure 5.9 (a) shows a mosaic of 100 image pairs under FL modality. In this mosaic the mark of a transurethral resection (TUR)



(a) (b) (c)

Figure 5.10: Second mosaic with FL data. White circle represents the FOV of the cystoscope and corresponding image under WL and FL are shown respectively in (a) and (b).



(a) (b)

Figure 5.11: Gastroscopy image mosaics of the pyloric antrum region. (a) Image mosaic without strong specular reflections. (b) Mosaic with 70 images with strong specular reflections. Regions in the mosaic having large specular reflections are in the black rectangles and the green arrows mark the structure continuity which demonstrates the quality of image alignment.

region is pointed by the arrow. The corresponding region in the small FOV images is shown in Figs. 5.9 (b) and (c) for before and after this resection (at different view-points). During this procedure tissues with lesions are removed by burning of cancerous cells. The mosaic shown here was obtained with the ROF-NDD method, however the AOFW algorithm provides a mosaic with a comparable visual quality.

Another mosaic in Fig. 5.10 (a) shows how the cancerous cells which are often not clearly visible under WL are more apparent under FL modality. The red spots in this mosaic corresponds to cancerous parts in the bladder. On the left of this mosaic an image under WL modality is shown in Fig. 5.10 (b) and its counter part FL image in Fig. 5.10 (c).

5.4.3 Gastroscopy

In gastroscopy, an endoscope is used to scan the inner surface of the stomach. Inflammations in the pyloric antrum region (aperture region corresponding to the junction of the stomach and the intestine) may be the early signs of potential lesions which can notably degenerate into cancer. A unique image only shows a part of the antrum, whereas having large field of views of such scenes (as that in Figs. 5.11(a-b)) can greatly improve the diagnosis and treatment traceability through mosaic screening and archiving. These images are characterized by the presence of large homogeneous regions, specular reflections and shadows along the duodenum opening as can be observed in the mosaic shown in Fig. 5.11(a).

Most of the image pairs in this sequence gave a great challenge for baseline approaches [Weibel et al., 2012b, Pock et al., 2007, Wedel et al., 2009b, Brox et al., 2004, Drulea and Nedevschi, 2013] to robustly register weak textured (mostly homogeneous) images, also affected by specular reflections indicated by rectangles in Fig. 5.11(b). The ROF-NND algorithm was robust enough to register all the image pairs required for building a mosaic of the pyloric antrum region as shown in Fig. 5.11(b). Image borders, which remain visible since no blending was used, show that there is effectively no structure discontinuity, as required for visually coherent mosaics.

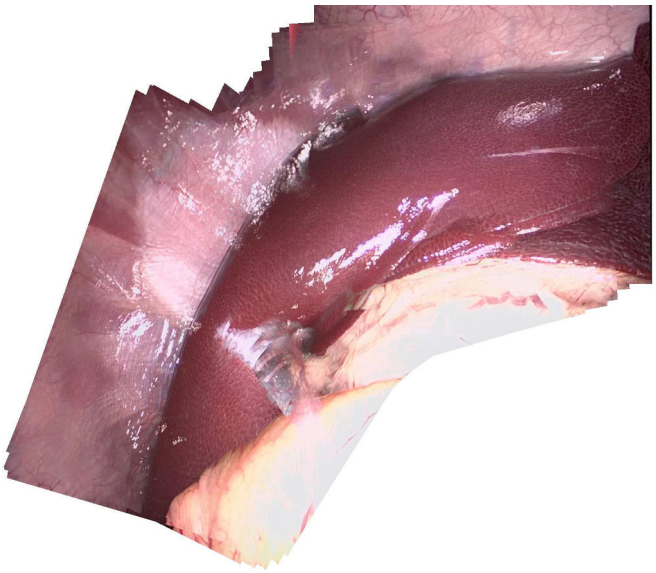


Figure 5.12: Stitching of 100 frames (every 10^{th} frame of the sequence) extracted from a laparoscopic video sequence of the region around the liver. Large specular reflections and brightness changes can be observed in image pairs. The first mosaic image is located at the top right of the map corner.

5.4.4 Laparoscopy

Laparoscopic surgery or keyhole surgery is a modern and popular minimally invasive surgical (MIS) procedure which is performed via small incisions with a size from 0.5 up to 1.5 cm. During such MIS, a laparoscope is inserted in the abdominal cavity through the opening and visualizes the area of interest. However, due to small sizes of the openings, the displacement of the endoscopes are limited. Usually, the video-sequences obtained by a laparoscope has large in-plane rotations and small displacements. As observed in Fig. 5.12, large specular reflections are often visible due to organ's surface optical properties. Additionally, inhomogeneous illumination is observed since the center of the image are more lightened than their periphery (vignetting effect). Stitching of 100 frames with the ROF-NND method shows the robustness of the algorithm to large illumination changes between the frames. Such a wide FOV mosaics can assist the surgeons in the MIS procedures. It can be observed that the majority of image pairs are almost homogeneous (*i.e.* with very less structures in them).

5.5 Quality mosaics for other scenes

5.5.1 Dermoscopy

Fig. 5.13 gives a representative result in terms of skin mosaics. This result was obtained for a video-sequence of the left facial part of a human with some neck region on a patient. It can be observed that skin pores act as structure information for the RFLOW algorithm. The RFLOW method is robust enough to lead to visually coherent mosaic of such image sequences. Even though the geometrical information visualized in the images is not quasi-planar as in the region including the border between the cheek and neck, the RFLOW method robustly register the image pairs giving visually coherent mosaic in Fig. 5.13.

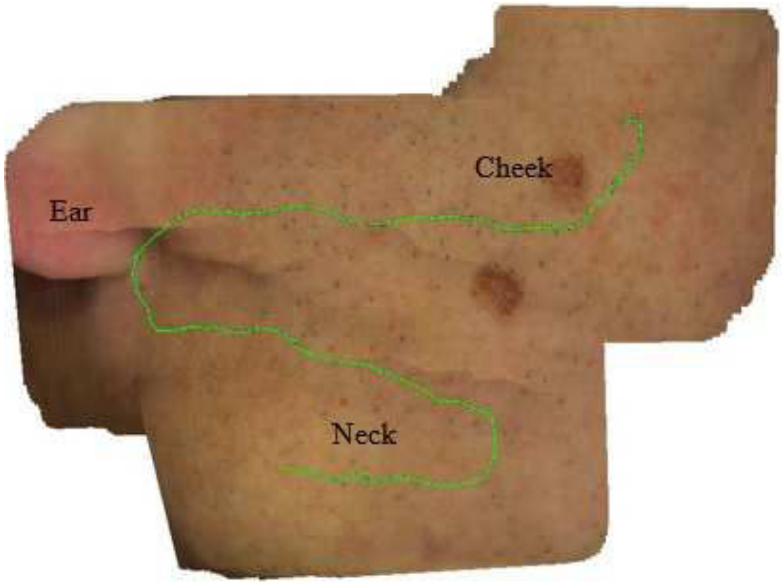


Figure 5.13: Human data mosaic of the face and the neck region. The green line represents the camera trajectory.

5.5.2 Underwater scenes

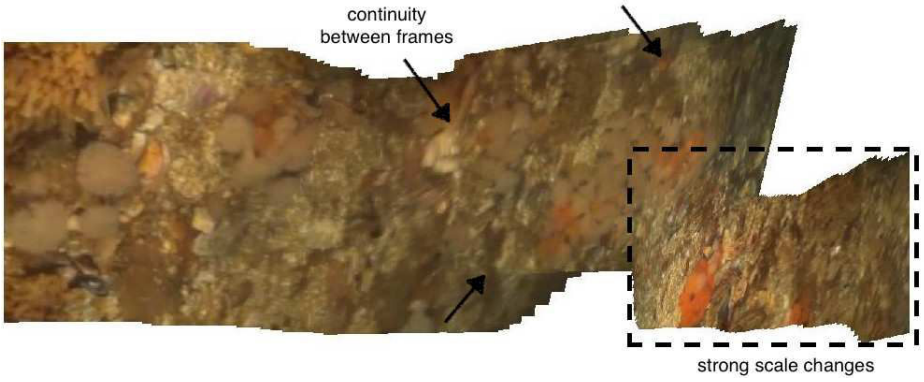
Mosaicing of underwater data is challenging since these scenes are affected by strong variability in illumination (caused by strong reflectance), presence of large number of repeated patterns, large image regions without texture, presence of blur (due to both camera motion and tidal motion of water) and moving objects into the field of view.

In Fig. 5.14 (a), blurred images can be observed at the left side of the constructed panorama. This is due to strong tidal currents persistent in this video sequence [Michel J. et al., 2011]. Strong scale changes can be noticed mainly in the right side of the mosaic indicated by a dashed black rectangular region. Another example for a seabed sequence is presented in Fig. 5.14.(b) which visualizes large areas with repeated patterns and with non-uniform illumination. Strong perspective changes are seen at the end of this mosaic indicated by a dashed black rectangle. Since it was shown in Chapter 3 that the AOFW method can efficiently handle repetitive structures and patterns in them, perspective and scale changes and small illumination changes, this optical flow method was used to build the mosaics of the underwater scenes.

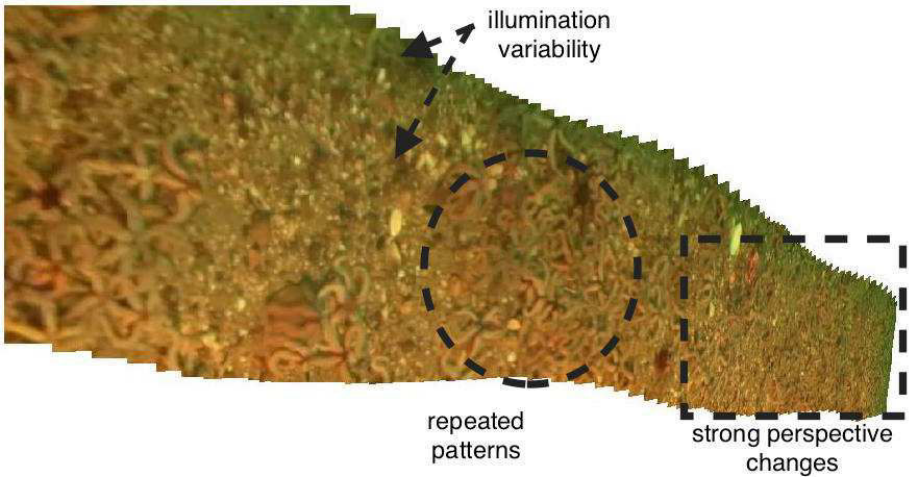
Both of the mosaics in Figs. 5.14 (a-b) are visually coherent since there is no texture discontinuity in them which is a good indicator signifying accurate alignment of the images. Texture continuity can be observed in locations marked by black arrows in Fig. 5.14 (a). These mosaicing results thus provides a clear idea about the registration robustness of the AOFW algorithm under such variations of scene characteristics and illumination conditions including also strong scale (Fig. 5.14 (a)) or perspective (Fig. 5.14 (b)) changes.

5.5.3 Video mosaic of the Mars surface

A mosaicing result for a last scene type is shown in Fig. 5.15. As seen in the figure, the images of this video sequence (surface of Mars) has large homogeneous texture distribution. The proposed methods (RFLOW, AOFW and ROF-NND) robustly registered the 179 images required to build the mosaic (in practice, only each 10^{th} image of the available video-sequence was used



(a)



(b)

Figure 5.14: Underwater mosaics computed for video-sequences described in [Michel J. et al., 2011]. a) Seabed mosaic visualizing sessile fauna, hermit crabs, and horse mussels embedded at the surface of the sediment. b) Mosaic with repeated patterns of brittle stars strewing the seabed (in circle). Blur is perceptible in the left and central mosaic parts in (a). This blur is due to strong tidal currents. Large illumination variability can be observed.

Figure 5.15: Panorama of the landing site of the NASA's curiosity rover. The black line represents the trajectory of the camera motion. This mosaic was obtained with the AOFW method.

for constructing the mosaic since the camera motion is small for this video). All the proposed algorithms in this thesis were able to establish dense correspondence for all the image pairs for this sequence.

5.6 Main contributions and conclusion

From the point of view of the various applications which can benefit from the optical flow methods proposed in this thesis we can mention following contributions:

- In the specific case of bladder mosaicing, it was shown that the proposed algorithms are robust and accurate enough to construct maps of large organ parts. Such mosaics of large bladder areas can potentially be used to represent the bladder in several panoramas.
- Depending upon the image quality and strong changes in transformation parameters, proposed algorithms in this thesis proved to be robust to such changes.
- Without code optimization or GPU parallelization, the bladder mosaics can be constructed in several minutes allowing a second diagnosis few time after the examination itself which can be also used for efficient data archiving for follow-up and examination traceability.
- It was verified that the ROF-NDD algorithm effectively allowed for the mosaicing of both WL ans FL cystoscopic data with constant parameter settings.
- It was shown that it is possible to construct mosaicis for very different and numerous scene types. The algorithm to be used according the scene type was intelligibly identified.

Thus in this chapter, a wide application of the proposed algorithms have been identified and several examples in context to both endoscopic and other scene data are provided. Robustness of the proposed AOFW and ROF-NND algorithms that were demonstrated through public datasets and simulated sequences have been finally validated on the challenging real data sequences.

List of publications

[AFDB15] Sharib Ali, Khuram Faraz, Christian Daul and Walter Blondel “Optical flow with structure information for epithelial image mosaicing," 37^{th} Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, August 2015.

[ADGA+15] Sharib Ali, Christian Daul, Ernest Galbrun, Marine Amouroux, François Guillemin and Walter Blondel “Robust bladder image registration by redefining data-term in total variational approach," In Proceedings of SPIE, Medical Imaging, Image Processing Conference (SPIE), pp. 94131H–12, Orlando, USA, Febraury 2015.

Conclusion and perspectives

The major contribution presented in this thesis concerns the development of optical flow methods that can robustly give dense point-to-point correspondences between image frames despite of their strong texture variability, poor information (weakly pronounced or repeated textures), lighting artifacts (e.g. specular reflections), blur and/or large displacements (i.e. strong in-plane rotations, large perspective changes and/or scale variations). Such a dense estimation of homologous pixels representation between images can be used for very accurate pairwise image registration. In the framework of cystoscopic image mosaicing, since the images obtained are assumed to be rigid and planar, we have used homography for stitching of these image pairs. In order to make the methods more general, homography assumption have not been embedded in the optical flow estimation. Also because even in cystoscopic image pairs for some cases both the planarity and the rigidity assumptions do not hold. So, these cases can be dealt easily with out modifying the whole pipeline but rather estimating parameters with higher Degrees of Freedom (DOFs) or just by warping pixel-to-pixel between the image pairs for obtaining sub-maps.

From the medical point of view, the optical flow algorithms proposed in this thesis are an important step in the whole image mosaicing process required for constructing extended field of views. The mosaics which can be constructed due to the proposed approaches improve the diagnosis and examination traceability, especially in cystoscopy on which this thesis focuses. However, the thesis shows also the potential of the proposed algorithms for other endoscopic applications like gastroscopy. In the specific the case of cystoscopic image mosaicing, the experiments conducted in this thesis have shown the robustness of the optical flow algorithms, even for a strong variabilty in terms of inter- and intra- patient image characteristics. This shows the usability potential of proposed methods which can handle a large variety of acquisition (endoscopy trajectory, illumination quality) and scene (texture variability and quality) conditions. Registering images with the best possible compromise between robustness, accuracy and computation speed is one of the main issues for the bladder image mosaicing. In this thesis, this issue was successfully addressed using new dense optical flow assessment methods.

From the scientific point of view, the optical flow robustness and computational speed have been improved by performing an iterative first-order primal-dual energy minimization using coarse-to-fine multiresolution approach. A novel structure constancy assumption was used in the RFLOW algorithm to make the optical flow determination robust to small illumination changes. It was shown that integrating such structure constancy assumption in the data-term improves the flow field quality in both medical and non-medical data in comparison to baseline TV-l^1 methods. Later, a deeper insight for structure preservation was done by treating images at all pyramid levels with Riesz basis filters. Such a pyramid type attenuates the "flattening-out" effect at coarse pyramid levels. This guaranteed an accurate initialization of flow fields when passing from coarse to finer pyramid levels. Additionally, the accuracy of the flow field has been improved by a modified regularization term so that i) only appropriate pixels (i.e. pixels without texture) contribute to the flow field smoothness (computation of adaptive weights for

anisotropic diffusion) and ii) pure in-plane rotation vector fields were not treated as translation vector fields. A new weighted median filtering algorithm has also been proposed to minimize outliers in the motion field. Tests on the different reference databases have highlighted that, in comparison to existing baseline optical flow methods, this proposed method (referred to as AOFW) can handle images with weak textures, shadows and/or blur. In particular scenes (like medical and weak textured Mars scenes), and in comparison to methods specifically developed for the optical flow determination in weakly textured images, the results obtained have shown that the AOFW algorithm exhibits clearly the best compromise between robustness, accuracy and computation speed. These promising results were obtained on realistic simulated video sequence data and confirmed on real video data. While having improved accuracy and robustness of the TV-L^1 approaches, the proposed algorithm guarantees for feasible mosaics that can be improved by bundle adjustment techniques even in presence of large blur due to focus/de-focus of the cystoscope's camera. At minimum, such accuracy is important since accumulation of large registration errors impedes even the global map correction.

The dense point correspondence given by the optical flow is used to compute the homography geometrically linking the images. The underlying assumption is that the surfaces viewed in the images are quasi-planar for medical scenes and planar for other scenes under observation. However, in presence of strong illumination changes between the images (e.g. due to viewpoint changes, modality changes or specular reflections), the pixel values can be sufficiently affected to lead to inaccurate flow vectors, even when the AOFW method is used. A method robust to such strong illumination variability has also been proposed in this thesis (ROF-NDD approach). It has been shown that using i) self similarity neighborhood descriptors as vector representation of the data-term in a total variational framework and ii) a non-local regularization based on bilateral filtering can lead to very accurate flow fields even in presence of large illumination differences between images. The proposed algorithm has been validated on three well-known flow benchmark datasets. Dedicated tests for accuracy and robustness check of the algorithm were done on complicated scene conditions including both texture variability and illumination changes. It has been shown that the method is robust in all these scene conditions and gives improved accuracy compared to the baseline variational methods. It is also competitive with regard to the most recent and accurate methods dedicated to illumination changes, while being most often with less computational cost. The application of the ROF-NDD algorithm to the mosaicing of weak textured endoscopic datasets also validates the robustness and the accuracy of the proposed optical flow method.

Perspectives.

The inherent flexibility of the energy model in variational approaches allow for several improvements of the proposed optical flow algorithms. One interesting improvement could be to incorporate information of the scene geometry in the energy function. An affine invariant hypothesis was incorporated in the discrete energy minimization using graph-cuts in [Weibel et al., 2012b] for bladder image registration. Such incorporation as an additional constraint can minimize the abrupt divergence of the flow field vectors.

A first-order total variational energy minimization technique has been used in this thesis. One future development could be to use a second-order variant of the total generalized variation (TGV) approach. Indeed, recent literature [Ranftl et al., 2014, Demetz et al., 2015] has shown that, for a same energy to be minimized the second order variant of the TGV minimization approach shows, in comparison to the first-order TV approach, a great improvement in flow field

estimation.

This work is mainly focused on the accuracy and robustness of optical determination and image registration. However, in order to complete the image mosaicing pipeline, another important future work could also be towards using the variational approach proposed in this thesis to improve the visual coherence of mosaics (in terms of texture and color discontinuities). It will notably allow for providing blended mosaics with best possible contrasted textures.

The processing time is one crucial criterion when considering the urological context. Currently, the proposed algorithms can build mosaics of large bladder parts in some few minutes (this is a significant speed improvement in comparison to the graph-cut or mutual information based algorithms dedicated to bladder image mosaicing). Urologists or surgeons can thus scan different bladder parts and the proposed mosaicing methods provide a bladder representation on several mosaics. However, this procedure can be simplified in terms of practical use by visualizing in real-time the mosaics of the bladder parts on a screen during the cystoscopy itself.

The current codes of the proposed algorithms relies on own C functions. However, numerous processing steps are also based on the OpenCV library. The next work to be done for enabling a straightforward code optimization would be to re-implement all OpenCV functions in own C code. Since the proposed TV-L^1 approaches can be parallelized, a fully dedicated GPU implementation of this C-code would lead to a significantly improved computation time and give near to real time performance to the algorithms presented in this thesis.

Bibliography

[Adal et al., 2013] Adal, K., Ali, S., Sidibé, D., Karnowski, T., Chaum, E., and Mériaudeau, F. (2013). Automated detection of microaneurysms using robust blob descriptors. In *Proceeding SPIE*, volume 8670, pages 86700N–86700N–7.

[Agarwala et al., 2004] Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., and Cohen, M. (2004). Interactive digital photomontage. *ACM Transactions on Graphics (Proceeding SIGGRAPH)*, 23(3):294–302.

[Ali et al., 2013a] Ali, S., Blondel, W., and Daul, C. (2013a). Tv-l1 based fast and robust mosaicing of cystoscopic images. In *XXIVe Colloque GRETSI Traitement du Signal & des Images, GRETSI 2013*, page CDROM.

[Ali et al., 2014] Ali, S., Daul, C., and Blondel, W. (2014). Robust and accurate optical flow estimation for weak texture and varying illumination condition: Application to cystoscopy. In *4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 140–145.

[Ali et al., 2015a] Ali, S., Daul, C., Galbrun, E., Amouroux, M., Guillemin, F., and Blondel, W. (2015a). Robust bladder image registration by redefining data-term in total variational approach. In *Proceeding SPIE*, volume 9413, pages 94131H–94131H–12.

[Ali et al., 2015b] Ali, S., Daul, C., Galbrun, E., Guillemin, F., and Blondel, W. (2015b). Anisotropic motion estimation on edge preserving riesz wavelets for robust video mosaicing. *Pattern Recognition*, pages 1–20.

[Ali et al., 2013b] Ali, S., Daul, C., Weibel, T., and Blondel, W. (2013b). Fast mosaicing of cystoscopic images from dense correspondence: combined SURF and TV-L1 optical flow method. In *20th IEEE International Conference on Image Processing, (ICIP)*, pages 1291–1295.

[Alvarez et al., 1999] Alvarez, L., Esclarin, J., Lefebure, M., and Sanchez, J. (1999). A pde model for computing the optical flow. In *Proceeding XVI Congreso de Ecuaciones Diferenciales y Aplicaciones*, pages 1349–1356.

[Alvarez et al., 2000] Alvarez, L., Weickert, J., and Sánchez, J. (2000). Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision (IJCV)*, 39(1):41–56.

[Aubert et al., 1999] Aubert, G., Deriche, R., and Kornprobst, P. (1999). Computing optical flow via variational techniques. *SIAM Journal on Applied Mathematics*, 60:156–182.

[Baker et al., 2003] Baker, S., Gross, R., Ishikawa, T., and Matthews, I. (2003). Lucas-kanade 20 years on: A unifying framework: Part 2. *International Journal of Computer Vision (IJCV)*, 56:221–255.

[Baker et al., 2011] Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., and Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*, 92:1–31.

[Bao et al., 2014] Bao, L., Yang, Q., and Jin, H. (2014). Fast edge-preserving patchmatch for large displacement optical flow. *IEEE Transactions on Image Processing (TIP)*, 23(12):4996–5006.

[Barnes et al., 2009] Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proceeding SIGGRAPH)*, 28(3).

[Barreto et al., 2009] Barreto, J., Roquette, J., Sturm, P., and Fonseca, F. (2009). Automatic Camera Calibration Applied to Medical Endoscopy. In *BMVC 2009 - 20th British Machine Vision Conference*, pages 1–10. The British Machine Vision Association (BMVA).

[Barron et al., 1994] Barron, J. L., Fleet, D. J., and Beauchemin, S. S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision (IJCV)*, 12:43–77.

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110:346–359.

[Behrens et al., 2011] Behrens, A., Bommes, M., Stehle, T., Gross, S., Leonhardt, S., and Aach, T. (2011). Real-time image composition of bladder mosaics in fluorescence endoscopy. *Computer Science - R&D*, 26:51–64.

[Behrens et al., 2010] Behrens, A., Guski, M., Stehle, T., Gross, S., and Aach, T. (2010). Intensity based multi-scale blending for panoramic images in fluorescence endoscopy. In *International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pages 1305–1308.

[Behrens et al., 2009] Behrens, A., Stehle, T., Gross, S., and Aach, T. (2009). Local and global panoramic imaging for fluorescence bladder endoscopy. In *IEEE International Conference Engineering in Medicine and Biology Society (EMBC)*, pages 6990–6993.

[Bellmunt et al., 2014] Bellmunt, J., Orsola, A., Leow, J. J., Wiegel, T., De Santis, M., and Horwich, A. (2014). Bladder cancer: ESMO Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology*, 25(suppl 3):iii40–iii48.

[Bergen et al., 2013a] Bergen, T., Wittenberg, T., Münzenmayer, C., Chen, C. C. G., and Hager, G. D. (2013a). A graph-based approach for local and global panorama imaging in cystoscopy. In *Proceeding SPIE, Medical Imaging: Image Processing Conference*, volume 8671, pages 86711K–86711K–7.

[Bergen et al., 2013b] Bergen, T., Wittenberg, T., Münzenmayer, C., Chen, C. C. G., and Hager, G. D. (2013b). A graph-based approach for local and global panorama imaging in cystoscopy. In *Proceeding SPIE*, volume 8671, pages 86711K–86711K–7.

[Berger et al., 2013] Berger, T., Hastreiter, P., Münzenmayer, C., Buchfelder, M., and Wittenberg, T. (2013). Image stitching of sphenoid sinuses from monocular endoscopic views. In *12. Jahrestagung der Deutschen Gesellschaft für Computer-und Roboterassistierte Chirurgie, November 28-30, 2013, Innsbruck, Austria*, pages 226–229.

[Black and Anandan, 1996] Black, M. J. and Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding (CVIU)*, 63(1):75 – 104.

[Bredies et al., 2010] Bredies, K., Kunisch, K., and Pock, T. (2010). Total generalized variation. *SIAM Journal on Imgaging Science (SIIMS)*, 3(3):492–526.

[Brown and Lowe, 2007] Brown, M. and Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision (IJCV)*, 74:59–73.

[Brox et al., 2004] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024, pages 25–36.

[Brox and Malik, 2011] Brox, T. and Malik, J. (2011). Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):500–513.

[Bruhn et al., 2005] Bruhn, A., Weickert, J., and Schnörr, C. (2005). Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision (IJCV)*, 61(3):211–231.

[Burt and Adelson, 1983] Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540.

[Butler et al., 2012] Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag.

[Chadebecq et al., 2012] Chadebecq, F., Tilmant, C., and Bartoli, A. (2012). Measuring the size of neoplasia in colonoscopy usingdepth-from-defocus. In *34th IEEE International Conference Engineering in Medicine and Biology Society (EMBC)*.

[Chambolle, 2004] Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97.

[Chambolle and Pock, 2011] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

[Chen et al., 2013] Chen, Z., Jin, H., Lin, Z., Cohen, S., and Wu, Y. (2013). Large displacement optical flow from nearest neighbor fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Danielsson, 1980] Danielsson, P. E. (1980). Euclidean distance mapping. *Computer Graphics and Image Processing*, 14(3):227–248.

[Daul et al., 2009] Daul, C., Lopez-Hernandez, J., Wolf, D., Karcher, G., and Ethévenot, G. (2009). 3-D multimodal cardiac data superimposition using 2-D image registration and 3-D reconstruction from multiple views. *Image and Vision Computing (IVC)*, 27(6):790 – 802.

[Davison et al., 2007] Davison, A. J., Reid, I. D., Molton, N. D., and Stasse, O. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067.

[Demetz et al., 2013] Demetz, O., Hafner, D., and Weickert, J. (2013). The complete rank transform: A tool for accurate and morphologically invariant matching of structures. In *24th British Machine Vision Conference (BMVC)*. BMVA Press.

[Demetz et al., 2015] Demetz, O., Hafner, D., and Weickert, J. (2015). Morphologically invariant matching of structures with the complete rank transform. *International Journal of Computer Vision (IJCV)*, 113(3):220–232.

[Drulea and Nedevschi, 2013] Drulea, M. and Nedevschi, S. (2013). Motion estimation using the correlation transform. *IEEE Transactions on Image Processing (TIP)*, 22(8):3260–3270.

[Fan et al., 2010] Fan, Y., Meng, M.-H., and Li, B. (2010). 3D reconstruction of wireless capsule endoscopy images. In *IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 5149–5152.

[Ferlay et al., 2013] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F. (2013). Cancer incidence and mortality worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer. Available from: http://globocan.iarc.fr.

[Frangi et al., 1998] Frangi, F., Niessen, J., Vincken, L., and Viergever, A. (1998). Multiscale vessel enhancement filtering. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1496, pages 130–137. Springer-Verlag.

[Geiger et al., 2013] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237.

[Grasa et al., 2009] Grasa, O. G., Civera, J., Guemes, A., Munoz, V., and Montiel, J. M. M. (2009). EFK monocular SLAM 3D modeling, measuring and augmented reality from endoscope image sequences. In *5th workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery, held in conjunction with MICCAI*, pages 1–8.

[Gu et al., 2014] Gu, G., He, B., and Yuan, X. (2014). Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach. *Computational Optimization and Applications*, 59(1-2):135–161.

[Hafner et al., 2013] Hafner, D., Demetz, O., and Weickert, J. (2013). Why is the census transform good for robust optic flow computation? In *Scale Space and Variational Methods in Computer Vision*, volume 7893 of *Lecture Notes in Computer Science (LNCS)*, pages 210–221. Springer Berlin Heidelberg.

[Hamadou et al., 2009] Hamadou, A. B., Soussen, C., Blondel, W., Daul, C., and Wolf, D. (2009). Comparative study of image registration techniques for bladder video-endoscopy. In *European Conferences on Biomedical Optics*, volume 737118, pages 1–7.

[Hartley and Zisserman, 2003] Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition.

[Hernández-Mier et al., 2006] Hernández-Mier, Y., Blondel, W., Daul, C., Wolf, D., and Bourg-Heckly, G. (2006). 2-D panoramas from cystoscopic image sequences and potential application to fluorescence imaging. In *IFAC conference on Modeling and Control in Biomedical Systems,*, volume 6, pages 291–296.

[Hernandez-Mier et al., 2010] Hernandez-Mier, Y., Blondel, W., Daul, C., Wolf, D., and Guillemin, F. (2010). Fast construction of panoramic images for cystoscopic exploration. *Computerized Medical Imagaging and Graphics (CMIG)*, 34:579–592.

[Horn and Schunck, 1981] Horn, B. and Schunck, G. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

[Igarashi et al., 2009] Igarashi, T., Zenbustsu, S., Yamanishi, T., and Naya, Y. (2009). Computer-based endoscopic image-processing technology for endourology and laparoscopic surgery. *Journal of endourology / Endourological Society*, 16(6):533–543.

[Jabid et al., 2010] Jabid, T., Kabir, M. H., and Chae, O. (2010). Local directional pattern (ldp) - a robust image descriptor for object recognition. In *7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 482–487.

[Kim and Pollefeys, 2008] Kim, S. J. and Pollefeys, M. (2008). Robust radiometric calibration and vignetting correction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (PAMI)*, 30(4):562–576.

[Koppel et al., 2007] Koppel, D., Chen, C.-I., Wang, Y.-F., Lee, H., Gu, J., Poirson, A., and Wolters, R. (2007). Toward automated model building from video in computer-assisted diagnoses in colonoscopy. In *Proceeding SPIE*, volume 6509, pages 65091L–65091L–9.

[Li and Osher, 2009] Li, Y. and Osher, S. (2009). A new median formula with applications to PDE based denoising. *Communications in Mathematical Sciences*, 7(3):741–753.

[Lindeberg, 1994] Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers.

[Liu et al., 2011] Liu, C., Yuen, J., and Torralba, A. (2011). SIFT Flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994.

[Liu et al., 2003] Liu, H., Chellappa, R., and Rosenfeld, A. (2003). Fast two-frame multiscale dense optical flow estimation using discrete wavelet filters. *Journal of the Optical Society of America (JOSA)*, 20:1505 – 1515.

[Liu et al., 2015] Liu, J., Wang, B., Hu, W., sun, P., Li, J., Duan, H., and Si, J. (2015). Global and local panoramic views for gastroscopy: An assisted method of gastroscopic lesion surveillance. *Biomedical Engineering, IEEE Transactions on*, PP(99):1–1.

[Lorenz et al., 1997] Lorenz, C., Carlsen, I.-C., Buzug, T. M., Fassnacht, C., and Weese, J. (1997). Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images. In *Proceeding of the First Joint Conference on Computer Vision*, pages 233–242.

[Lourenço et al., 2014] Lourenço, M., Stoyanov, D., and Barreto, J. P. (2014). Visual odometry in stereo endoscopy by using pearl to handle partial scene deformation. In *9th International Workshop on Augmented Environments for Computer-Assisted Interventions Held in Conjunction with MICCAI*, pages 33–40.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679.

[Mahmoud et al., 2012] Mahmoud, N., Nicolau, S., Keshk, A., Ahmad, M. A., Soler, L., and Marescaux, J. (2012). Fast 3d structure from motion with missing points from registration of partial reconstructions. In *Conference on Articulated Motion and Deformable Objects site (AMDO)*, pages 173–183.

[Maier-Hein et al., 2014] Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.-L., Clancy, N. T., Elson, D. S., Haase, S., Heim, E., Hornegger, J., Jannin, P., Kenngott, H., Kilgus, T., Müller-Stich, B., Oladokun, D., Röhl, S., Dos Santos, T. R., Schlemmer, H.-P., Seitel, A., Speidel, S., Wagner, M., and Stoyanov, D. (2014). Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction. *IEEE Transactions on Medical Imaging*, 33:1913–30.

[Maier-Hein et al., 2013] Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., and Stoyanov, D. (2013). Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17:974–996.

[Mallat, 1989] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 11:674–693.

[Malti et al., 2012] Malti, A., Bartoli, A., and Collins, T. (2012). Template-based conformal shape-from-motion-and-shading for laparoscopy. In *Information Processing in Computer-Assisted Interventions - Third International Conference, IPCAI*, pages 1–10.

[Marquardt, 1963] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441.

[Marzotto et al., 2004] Marzotto, R., Fusiello, A., and Murino, V. (2004). High resolution video mosaicing with global alignment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–692–I–698 Vol.1.

[Maurin et al., 2009] Maurin, B., Doignon, C., de Mathelin, M., and Gangi, A. (2009). A fast and automatic stereotactic registration with a single CT-slice. *Computer Vision and Image Understanding (CVIU)*, 113(8):878–890.

[Michel J. et al., 2011] Michel J., K., Martin J, A., Simon, J., David N, T., David K. A., B., Andrew S., B., Jan G., H., Hermanni, K., Nicholas V. C., P., and Raffaelli, D. G. (2011). *Marine Ecology: Processes, systems and impacts*. Oxford University Press.

[Miranda-Luna et al., 2004] Miranda-Luna, R., Blondel, W., Daul, C., Hernandez-Mier, Y., Posada, R., and Wolf, D. (2004). A simplified method of endoscopic image distortion correction based on grey level registration. In *IEEE International Conference on Image Processing (ICIP)*, pages 3383–3386.

[Miranda-Luna et al., 2008] Miranda-Luna, R., Daul, C., Blondel, W., Hernandez-Mier, Y., Wolf, D., and Guillemin, F. (2008). Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm. *IEEE Transactions on Biomedical Engineering (TBME)*, 55:541–553.

[Mohamed et al., 2014] Mohamed, M., Rashwan, H., Mertsching, B., Garcia, M., and Puig, D. (2014). Illumination-robust optical flow using a local directional pattern. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(9):1499–1508.

[Moreau, 1965] Moreau, J. J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la S. M. F.*, 93:273–299.

[Nagel and Enkelmann, 1986] Nagel, H.-H. and Enkelmann, W. (1986). An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(5):565–593.

[Papenberg et al., 2006] Papenberg, N., Bruhn, A., Brox, T., Didas, S., and Weickert, J. (2006). Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision (IJCV)*, 67(2):141–158.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(7):629–639.

[Pezaro et al., 2012] Pezaro, C., Liew, M., and Davis, I. (2012). Urothelial cancers: using biology to improve outcomes. *Expert Review of Anticancer Therapy*, 12(1):87–98.

[Pock et al., 2007] Pock, T., Urschler, M., Zach, C., Beichel, R., and Bischof, H. (2007). A duality based algorithm for TV-L1-optical-flow image registration. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 10:511–518.

[Posada et al., 2007] Posada, R., Daul, C., Wolf, D., and Aletti, P. (2007). Towards a noninvasive intracranial tumor irradiation using 3D optical imaging and multimodal data registration. *International Journal of Biomedical Imaging*, 2007:1–14.

[Ranftl et al., 2014] Ranftl, R., Bredies, K., and Pock, T. (2014). Non-local total generalized variation for optical flow estimation. In *European Conference on Computer Vision (ECCV)*, volume 8689 of *Lecture Notes in Computer Science (LNCS)*, pages 439–454. Springer International Publishing.

[Rashwan et al., 2013] Rashwan, H., Mohamed, M., García, M., Mertsching, B., and Puig, D. (2013). *Illumination Robust Optical Flow Model Based on Histogram of Oriented Gradients*, volume 8142 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

[Rockafellar, 1976] Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization (SICON)*, 14(5):877–898.

[Rousseeuw and Leroy, 1987] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA.

[Rudin et al., 1992] Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Journal of Physics D*, 60:259–268.

[Saito and Toriwaki, 1994] Saito, T. and Toriwaki, J.-I. (1994). New algorithms for euclidean distance transformation of an n-dimensional digitized picture with applications. *Pattern Recognition*, 27(11):1551 – 1565.

[Scharstein and Szeliski, 1998] Scharstein, D. and Szeliski, R. (1998). Stereo matching with nonlinear diffusion. *International Journal of Computer Vision (IJCV)*, 28:155–174.

[Schnörr and Sprengel, 1994] Schnörr, C. and Sprengel, R. (1994). A nonlinear regularization approach to early vision. *Biological Cybernetics*, 72(2):141–149.

[Schuster et al., 2012] Schuster, M., Bergen, T., Reiter, M., Münzenmayer, C., Friedl, S., and Wittenberg, T. (2012). Laryngoscopic image stitching for view enhancement and documentation - first experiences. In *Biomedizinsche Technik. Biomedical engineering*.

[Shevchenko et al., 2012] Shevchenko, N., Fallert, J., Stepp, H., Sahli, H., Karl, A., and Lueth, T. (2012). A high resolution bladder wall map: Feasibility study. In *IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 5761–5764.

[Simoncelli et al., 1992] Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J. (1992). Shiftable multi-scale transforms. *IEEE Transactions on Informations Theory*, 38(2).

[Snavely et al., 2006] Snavely, N., Seitz, S. M., and Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835–846.

[Soper et al., 2012] Soper, T., Porter, M., and Seibel, E. J. (2012). Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *Biomedical Engineering, IEEE Transactions on (TBME)*, 59(6):1670–1680.

[Stein, 2004] Stein, F. (2004). Efficient computation of optical flow using the census transform. In *Pattern Recognition*, volume 3175 of *Lecture Notes in Computer Science*, pages 79–86. Springer Berlin Heidelberg.

[Sun et al., 2010] Sun, D., Roth, S., and Black, M. J. (2010). Secrets of optical flow estimation and their principles. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439.

[Sun et al., 2014] Sun, D., Roth, S., and Black, M. J. (2014). A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision (IJCV)*, 106(2):115–137.

[Suter, 1994] Suter, D. (1994). Motion estimation and vector splines. In *Computer Vision and Pattern Recognition, 1994. Proceedings (CVPR), IEEE Computer Society Conference on*, pages 939–942.

[Szeliski, 2006] Szeliski, R. (2006). Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–104.

[Tao et al., 2012] Tao, M., Bai, J., Kohli, P., and Paris, S. (2012). Simpleflow: A non-iterative, sublinear optical flow algorithm. *Computer Graphics Forum*, 31:345–353.

[Thorsten et al., 2002] Thorsten, T., Hellward, B., and Peter, M. (2002). *Three-Dimensional Endoscopy*. Springer Netherlands.

[Tikhonov, 1963] Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Doklady*, volume 4, pages 1035–1038.

[Tistarelli, 1996] Tistarelli, M. (1996). Multiple constraints to compute optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (PAMI)*, 18(12):1243–1250.

[Totz et al., 2012] Totz, J., Fujii, K., Mountney, P., and Yang, G.-Z. (2012). Enhanced visualisation for minimally invasive surgery. *International Journal of Computer Assisted Radiology and Surgery*, 7(3):423–432.

[Triggs et al., 2000] Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (2000). Bundle adjustment — a modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer Berlin Heidelberg.

[Tschumperlé and Deriche, 2003] Tschumperlé, D. and Deriche, R. (2003). Vector-valued image regularization with PDEs: A common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 506–517.

[Unser et al., 2011] Unser, M., Chenouard, N., and Ville, D. V. D. (2011). Steerable pyramids and tight wavelet frames in $L_2(\mathtt{R}^d)$. *IEEE Transactions on Image Processing (TIP)*, 20(10):2705–2721.

[Uras et al., 1988] Uras, S., Girosi, F., Verri, A., and Torre, V. (1988). A computational approach to motion perception. *Biological Cybernetics*, 60(2):79–87.

[Uyttendaele et al., 2001] Uyttendaele, M., Eden, A., and Szeliski, R. (2001). Eliminating ghosting and exposure artifacts in image mosaics. In *Computer Vision and Pattern Recognition (CVPR)*, pages 509–516. IEEE Computer Society.

[Vercauteren, 2008] Vercauteren, T. (2008). *Image registration and mosaicing for dynamic In vivo fibered confocal microscopy*. PhD thesis, École Nationale Supérieure des Mines de Paris.

[Vogel et al., 2013] Vogel, C., Roth, S., and Schindler, K. (2013). An evaluation of data costs for optical flow. In *Pattern Recognition*, volume 8142 of *Lecture Notes in Computer Science*, pages 343–353. Springer Berlin Heidelberg.

[Wedel et al., 2009a] Wedel, A., Cremers, D., Pock, T., and Bischof, H. (2009a). Structure- and motion-adaptive regularization for high accuracy optic flow. In *International Conference on Computer Vision (ICCV)*.

[Wedel et al., 2009b] Wedel, A., Pock, T., , C., Bischof, H., and Cremers, D. (2009b). An improved algorithm for TV-L1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, Lecture Notes in Computer Science, pages 23–45. Springer Berlin Heidelberg.

[Weibel et al., 2012a] Weibel, T., Daul, C., Wolf, D., and Rösch, R. (2012a). Contrast-enhancing seam detection and blending using graph cuts. In *21st International Conference on Pattern Recognition (ICPR)*, pages 2732–2735.

[Weibel et al., 2010] Weibel, T., Daul, C., Wolf, D., Rösch, R., and Ben-Hamadou, A. (2010). Endoscopic bladder image registration using sparse graph cuts. In *17th IEEE International Conference on Image Processing, (ICIP)*, pages 157–160.

[Weibel et al., 2012b] Weibel, T., Daul, C., Wolf, D., Rösch, R., and Guillemin, F. (2012b). Graph based construction of textured large field of view mosaics for bladder cancer diagnosis. *Pattern Recognition*, 45(12):4138–4150.

[Weickert et al., 2006] Weickert, J., Bruhn, A., Brox, T., and Papenberg, N. (2006). A survey on variational optic flow methods for small displacements. In *Mathematical Models for Registration and Applications to Medical Imaging*, volume 10, pages 103–136. Springer Berlin Heidelberg.

[Weickert and Schnörr, 2001] Weickert, J. and Schnörr, C. (2001). A theoretical framework for convex regularizers in pde-based computation of image motion. *International Journal of Computer Vision (IJCV)*, 45(3):245–264.

[Weinzaepfel et al., 2013] Weinzaepfel, P., Revaud, J., Harchaoui, Z., and Schmid, C. (2013). DeepFlow: Large displacement optical flow with deep matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1385–1392.

[Werlberger et al., 2010] Werlberger, M., Pock, T., and Bischof, H. (2010). Motion estimation with non-local total variation regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2464–2471.

[Werlberger et al., 2009] Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., and Bischof, H. (2009). Anisotropic Huber-L1 optical flow. In *20th British Machine Vision Conference (BMVC)*.

[Xu et al., 2012] Xu, L., Jia, J., and Matsushita, Y. (2012). Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(9):1744–1757.

[Yi et al., 2013] Yi, S., Xie, J., Mui, P., and Leighton, J. A. (2013). Achieving real-time capsule endoscopy (ce) video visualization through panoramic imaging. In *Proceeding SPIE*, volume 8656, pages 86560I–86560I–7.

[Zabih and Woodfill, 1994] Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision (ECCV)*, volume 801 of *Lecture Notes in Computer Science*, pages 151–158. Springer Berlin Heidelberg.

[Zach et al., 2007] Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for real-time TV-L1 optical flow. In *Annual Sympsium of the German Association Pattern Recognition (DAGM)*, pages 214–223.

[Zhang et al., 2003] Zhang, L., Curless, B., Hertzmann, A., and Seitz, S. M. (2003). Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In *International Conference on Computer Vision (ICCV)*, volume 2, page 618.

[Zimmer et al., 2011] Zimmer, H., Bruhn, A., and Weickert, J. (2011). Optic flow in harmony. *International Journal of Computer Vision (IJCV)*, 93(3):368–388.

Résumé

La cystoscopie est l'examen de référence pour le diagnostic et le traitement du cancer de la vessie. Le champ de vue (CdV) réduit des endoscopes complique le diagnostic et le suivi des lésions. Les mosaïques d'images sont une solution à ce problème car elles visualisent des CdV étendus. Toutefois, pour la vessie, le mosaïque d'images est un véritable défi à cause du faible contraste dans les images, des textures peu prononcées, de la variabilité intra- et inter-patient et des changements d'illumination dans les séquences. Ce défi est également à relever dans d'autres modalités endoscopiques ou dans des scènes non médicales comme les vidéos sous-marines. Dans cette thèse, une énergie variationnelle totale a d'abord été minimisée à l'aide d'un algorithme primal-dual du premier ordre pour obtenir un flot optique fournissant une correspondance dense et précise entre les points homologues des paires d'images. Les correspondances sont ensuite utilisées pour déterminer les paramètres des transformations requises pour le placement des images dans le repère global de la mosaïque. Les méthodes proposées pour l'estimation du flot optique dense incluent un terme d'attache aux données qui minimise le nombre des vecteurs aberrants et un terme de régularisation conçu pour préserver les discontinuités du champ de vecteurs. Un algorithme de flot optique qui est robuste vis-à-vis de changements d'illumination importants (et utilisable pour différentes modalités) a également été développé dans ce contexte. La précision et la robustesse des méthodes de recalage proposées ont été testées sur des jeux de données (de flot optique) publiquement accessibles et sur des fantômes de vessies et de la peau. Des résultats sur des données patients acquises avec des cystoscopes rigides et flexibles, en lumière blanche ou en fluorescence, montrent la robustesse des algorithmes proposés. Ces résultats sont complétés par ceux obtenus pour d'autres séquences endoscopiques réelles de dermatoscopie, de scène sous-marine et de données d'exploration spatiale.

Mots-clés: approches variationnelles totales, flot optique, constance de structure, descripteurs de voisinages, régularisation anisotropique, optimisation convexe, recalage d'images, mosaïquage d'images endoscopiques.

Abstract

Cystoscopy is the reference procedure for the diagnosis and treatment of bladder cancer. The small field of view (FOV) of endoscopes makes both the diagnosis and follow-up of lesions difficult. Image mosaics are a solution to this problem since they visualize large FOVs of the bladder scene. However, due to low contrast, weak texture, inter- and intra-patient texture variability and illumination changes in these image sequences, the task of image mosaicing becomes challenging. This is also a major concern in other endoscopic data and non-medical scenes like underwater videos. In this thesis, a total variational energy has been first minimized using a first-order primal-dual algorithm in convex optimization to obtain optical flow vector fields giving a dense and accurate correspondence between homologous points of the image pairs. The correspondences are then used to obtain transformation parameters for registering the images to one global mosaic coordinate system. The proposed methods for dense optical flow estimation include a data-term which is modeled to minimize at most the outliers and a regularizer which is designed to preserve at their best the flow field discontinuities. An optical flow algorithm, which is robust to strong illumination changes (and which suits to different modalities), has also been developed in this framework. The registration accuracy and robustness of the proposed methods are tested on both publicly available datasets for optical flow estimation and on simulated bladder and skin phantoms. Results on patient data acquired with rigid and flexible cystoscopes under the white light and the fluorescence modality show the robustness of the proposed approaches. These results are also complemented with those of other real endoscopic data, dermoscopic sequences, underwater scenes and space exploration data.

Keywords: Total variational approach, optical flow, structure constancy, neighborhood descriptors, anisotropic regularization, convex optimization, image registration, endoscopic image mosaicing.