Contributions au montage automatique de scènes complexes multi-vues en interaction avec l'environnement

Florent Lefèvre

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Contributions au montage automatique de scènes complexes multi-vues en interaction avec l'environnement

THÈSE

présentée et soutenue publiquement le 4 décembre 2019

pour l'obtention du

Doctorat de l'Université de Lorraine

(en Sciences, spécialité Automatique, Traitement du Signal et Génie Informatique)

par

Florent LEFEVRE

Composition du jury

Président :	Bertrand ROSE	Professeur, Université de Strasbourg
Rapporteurs :	Monique THONNAT	Directrice de Recherche INRIA, Sophia Antipolis
	Catherine ACHARD	Maître de Conférences HDR, Sorbonne université
Examinateurs :	Patrick CHARPENTIER	Professeur, Université de Lorraine (Directeur de thèse)
	Nicolas KROMMENACKER	Maître de Conférences, Université de Lorraine (Co-directeur de thèse)
	Vincent BOMBARDIER	Maître de Conférences, Université de Lorraine (Co-directeur de thèse)
Invité :	Bertrand PETAT	CEO CitizenCam

Centre de Recherche en Automatique de Nancy (CRAN) - UMR 7039
Campus Sciences, B.P. 70239, 54506 Vandoeuvre-lès-Nancy

Mis en page avec la classe thesul.

Remerciements

Je tiens à exprimer mes sincères remerciements à mes encadrant de thèse, Patrick Charpentier, Vincent Bombardier et Nicolas Krommenacker pour leurs conseils scientifiques, leurs critiques constructives et leur soutien tout au long de ce travail de thèse.

Je remercie ensuite Catherine Achard et Monique Thonnat d'avoir accepté d'être rapporteurs de ce travail et pour la qualité de leur rapport. Je remercie également Bertrand Rose pour avoir accepté de présider ce jury. Leurs questions et remarques, lors de la soutenance, ont permis de mettre en avant certains points à approfondir et d'ouvrir de nouvelles perspectives.

Je remercie également l'entreprise CitizenCam d'avoir rendu cette thèse possible. Je remercie particulièrement Bertrand Petat de m'avoir fait confiance dans ces nombreux projets. Mes remerciements s'adressent également à Guillaume pour m'avoir initié au Python, aux produits locaux ainsi qu'à de nouvelles recettes douteuses.

Je remercie tous les membres du CRAN, permanents, doctorants, post-doctorants ainsi que l'équipe administrative pour leur accueil au sein du département durant ma thèse. Je remercie tout particulièrement Sara, Fabian, Sorella Concetta, Chiara, Meri et Hang, pour les moments de détente que nous avons partagés ensemble.

Je remercie également les amis de Nancy et d'ailleurs qui ont toujours été là pour aller boire une bière (ou deux). Je remercie en particulier Jeanne, Jean-Guillaume, Pierre, Greg, Anaïs, Nicolas et les personnes déjà citées précédemment et qui sont devenus de proches amis.

Je remercie également Claire pour son amour et son accompagnement tout au long de cette thèse. Je remercie également sa famille de m'avoir accueilli à Nancy et ailleurs.

Enfin, je remercie ma famille pour leur soutien, leur confiance et leurs conseils.

Table des matières

Table des figures

Liste des tableaux

Introduction

Diffuser un évènement avec des vidéos est un moyen simple d'offrir une visibilité à une manifestation ou une organisation, car cela permet d'atteindre des spectateurs géographiquement distants. L'avènement d'Internet a permis à n'importe quelle collectivité d'avoir une visibilité locale, nationale, voire internationale. Il est, aujourd'hui, de plus en plus fréquent de voir un évènement diffusé sur Internet, que ce soit via les plateformes de réseaux sociaux ou de streaming. Une organisation (association, groupe sportif, conseil municipal, ...) peut ainsi facilement se faire connaître du public.

Cependant, l'enregistrement et la diffusion d'un évènement sont relativement coûteux. Ils requièrent la mobilisation d'une équipe de tournage (cameramen, preneurs de son, monteurs, assistants, ...) et d'un équipement spécifique (caméras, microphones, table de montage, enregistreur vidéo, ...) représentant un coût important. Dans le livre *Television Sport Production* [106], l'auteur estime que la retransmission d'un évènement sportif de taille moyenne, coûtait en 2005 environ 135 000 euros. Cette prestation comprenait trois jours de préparation et la mobilisation d'une vingtaine de personnes pour une diffusion de 2 h. De par le fait du coût d'une telle prestation, de nombreux évènements de petite envergure ne peuvent être diffusés.

La société CitizenCam [1] a pour objectif de rendre la captation de tout type d'évènements (concerts, manifestations sportives, conseils municipaux, etc) accessible économiquement. Pour ce faire, elle propose un système de captation et de diffusion à moindre coût. Le système est composé d'un ensemble de caméras pour filmer l'évènement sous différents angles. Un serveur est utilisé afin de permettre d'enregistrer et de diffuser les flux vidéos sur Internet. L'utilisation de matériels à faible coût (caméras IP de surveillance, câbles Ethernet, serveur sur mesure) permet de réduire les coûts d'une captation, tout en conservant une qualité d'image intéressante.

La simplicité d'utilisation du système proposé, fait qu'une seule personne est nécessaire pour la gestion de l'enregistrement. Sa tâche se concentre sur le démarrage et l'arrêt de l'enregistrement, ainsi qu'à un éventuel partitionnement de cet évènement (ordre du jour dans une réunion, mi-temps dans un match). Une fois l'enregistrement de l'évènement terminé, les flux vidéos sont disponibles sur Internet [2] et les spectateurs peuvent choisir le flux vidéo qu'ils souhaitent regarder. Ainsi, la transparence de la vie politique d'une commune est garantie en laissant aux utilisateurs, et non à un monteur, le choix des flux vidéo à visualiser.

Le système proposé par CitizenCam a cependant quelques limites. Lorsque l'on veut diffuser un évènement en direct, une bande passante importante est nécessaire afin d'émettre ou de recevoir les différents flux vidéo. L'envoi d'un flux vidéo contenant un assemblage des différentes vidéos (sous la forme de mosaïque) pourrait pallier ce problème. Cependant, il est difficile pour le spectateur d'assimiler les informations des différents flux et son expérience est ainsi dégradée. Lors d'une diffusion en différé, le spectateur doit changer de flux vidéo dès que l'action passe d'une caméra à une autre. Ce changement rend également l'expérience d'utilisation désagréable pour le spectateur lorsque les changements sont fréquents (évènements sportifs) ou lorsque plusieurs caméras sont présentes (réunions). Il est alors nécessaire de pouvoir proposer un flux vidéo monté afin de pallier ce problème. Le montage vidéo est le fait d'assembler une suite d'images de manière cohérente et agréable pour le spectateur. Ce montage, réalisé en direct ou en différé, est effectué par une ou plusieurs personnes : les monteurs. L'emploi d'une personne spécifiquement dédiée au montage est contraire à la volonté de CitizenCam de proposer une solution simple

1. Les travaux présentés dans ce mémoire de thèse ont été rendus possible grâce à une collaboration entre le CRAN (Centre de Recherche en Automatique de Nancy) et l'entreprise CitizenCam dans le cadre d'une CIFRE (Convention Industrielle de Formation par la Recherche).

2. Disponible sur citizencam.tv

d'utilisation et à bas coût. L'automatisation de la sélection de flux vidéo semble donc être une solution idéale pour palier ce problème. Une difficulté supplémentaire tient dans le fait que CitizenCam veut pouvoir proposer la captation et la diffusion de tout type d'évènements aussi différents qu'un conseil municipal (statique) et un match de basketball (dynamique).

De plus, l'installation des caméras diffèrent d'une captation à une autre. Dans la majorité des cas, les caméras sont installées et désinstallées à chaque captation. La méthode de montage automatique doit donc être polyvalente et adaptable au contexte.
Pour résumer, l'objectif des travaux de cette thèse est de proposer une méthodologie de sélection automatique de caméras, basée sur le contexte, permettant la diffusion en direct et en différé de tout type d'évènements et laissant la possibilité aux utilisateurs d'agir sur le flux monté.

Le chapitre 1 du mémoire présente un panorama des recherches conduites dans les domaines du montage automatique et de la sélection automatique de caméras. Nous commençons par décrire les interactions entre les différents composants d'un tel système, avant d'en étudier les fonctionnements et les limitations des méthodes existantes.

Dans le chapitre 2, nous présentons notre approche de montage automatique. Notre méthodologie s'appuie sur la prise en compte du contexte lors de l'élaboration d'un montage automatique. Nous postulons que la contextualisation du montage permet de répondre aux problématiques de généralisation et de personnalisation du contenu. Pour chaque application, nous partons des connaissances sur le contexte de captation afin d'identifier les différents points d'intérêt dans la scène. Cette étape nous permet d'étudier les différentes étapes à mettre en place, afin de permettre la sélection automatique de flux vidéo.

La méthodologie proposée est validée à travers deux applications différentes. Nous étudions, dans le chapitre 3, la diffusion de conseils municipaux et dans le chapitre 4, la diffusion de matchs de basketball. L'application de notre méthodologie nous permet d'identifier des manquements, en terme de traitements d'images, nécessitant de proposer de nouvelles méthodes.
Dans le cadre de la diffusion de conseils municipaux, l'absence d'informations provenant des microphones, ainsi que la distance entre les caméras et les locuteurs, nous conduisent à proposer une nouvelle méthode de détection de la prise de parole. Nous proposons également une nouvelle méthode d'identification des locuteurs ne nécessitant pas d'apprentissage.

Pour la diffusion de match de basketball, nous proposons une méthode de détection en temps-réel du jeu notable se basant sur l'exploitation d'une caméra azimutale. L'étude des connaissances sur le contexte nous permet également de proposer une méthode de détection des lancers-francs ainsi qu'une nouvelle méthode de suivi des joueurs.

Enfin, la conclusion synthétise les apports de nos travaux dans la réalisation d'un système de montage automatique. Les travaux de validation sont discutés et ouvrent sur de nombreuses perspectives.

Chapitre 1

Montage automatique

Sommaire

La diffusion d'un évènement permet, à n'importe quel spectateur, de vivre un évènement à distance comme s'il y était. De plus en plus d'organisations s'intéressent alors à la captation vidéo des évènements qu'ils organisent. Cependant, l'enregistrement et la diffusion d'un évènement par une équipe professionnelle est relativement coûteuse, limitant la diffusion des petits évènements. En effet, diffuser un évènement impose de mobiliser de nombreux équipements et un grand nombre de personnel, comme illustré figure 1.1. La captation d'un évènement nécessite l'utilisation de plusieurs caméras, ainsi que de cadreurs chargés de diriger les caméras en fonction de l'action se déroulant sur scène. Les flux vidéo sont envoyés vers une régie, où le réalisateur sélectionne le flux vidéo à diffuser en direct. Le réalisateur communique également avec les cameramen afin de donner des consignes de cadrage. Lors d'une diffusion en différé, un monteur se charge de rassembler les plans des différentes caméras afin de produire un flux vidéo monté. Des ingénieurs du son sont également présents pour contrôler les signaux audio provenant de divers microphones, utilisés pour capter la partie sonore de la scène. Enfin, il est fréquent qu'une production audiovisuelle emploie des statisticiens, des graphistes, des truquistes, afin de proposer des informations supplémentaires aux spectateurs.

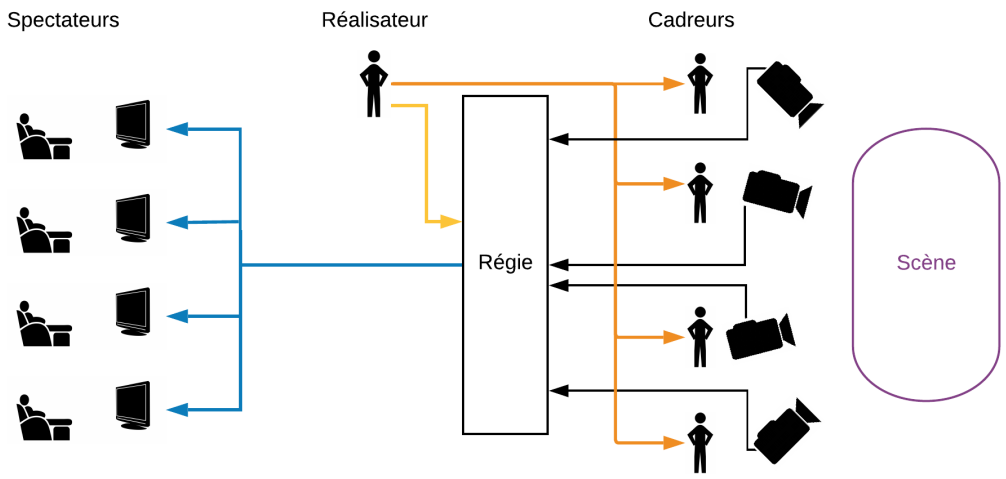


FIGURE 1.1 – Dispositif de captation vidéo traditionnel : les cadreurs orientent les caméras en fonction des consignes du régisseur. Ce dernier sélectionne également les flux vidéos à diffuser.

Le nombre important de personnels dans une équipe de production grève le prix d'une captation. Les évènements de grandes ampleurs ont l'audience nécessaire pour s'offrir une telle prestation, ce qui n'est pas le cas des évènements locaux ou régionaux. De ce fait, un grand nombre de petits évènements n'est pas diffusé. Il est donc nécessaire de réduire les coûts inhérents à la captation. L'automatisation de la chaîne de production est alors une solution pour faciliter la diffusion des petits évènements.

La sélection automatique de caméras est le fait de choisir de manière automatique la vue (cadrage et flux vidéo) la plus intéressante dans une scène multi-caméras. En d'autres termes, l'automatisation de la prise de vues a pour but de rendre la captation aussi vivante que celle réalisée par un humain tout en réduisant les coûts. Pour ce faire, les systèmes de montage automatique doivent agir à deux niveaux. Il est nécessaire de contrôler les caméras comme un cameraman le ferait et de sélectionner le flux vidéo à diffuser comme un réalisateur le ferait (cf. fig 1.2).

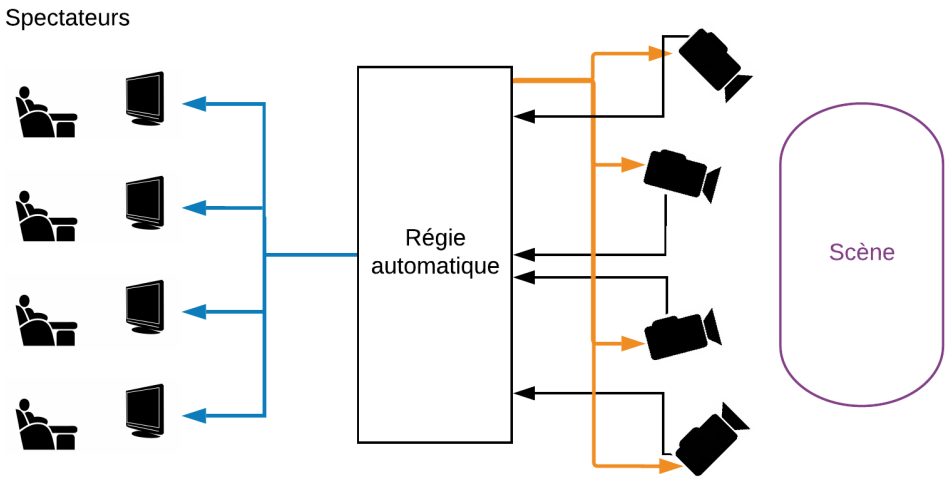


FIGURE 1.2 – Dispositif de captation vidéo automatique : les caméras sont contrôlées en fonction de l'objet d'intérêt dans la scène, et le flux vidéo, le plus intéressant pour les spectateurs, est sélectionné.

En 1995, Pinhanez et Bobick [111] proposent un système de captation d'émissions télévisuelles, où les caméras sont pilotées par un ordinateur et non par des cadreurs. Les caméras sont automatiquement orientées vers les zones d'intérêt dans la scène (tête du présentateur, mains, table, ...) sur l'ordre du réalisateur. La sélection du flux vidéo à diffuser est réalisée par le réalisateur, comme dans une production traditionnelle.

Ce premier système montre la possibilité de réduire le nombre d'opérateurs, pour une production d'émission de télévision, à une seule personne. Les progrès en vision par ordinateur et des capacités de calcul permettent aujourd'hui de sélectionner automatiquement le flux vidéo d'intérêt, afin de rendre la captation complètement autonome. Les montages automatiques ont été appliqués à de nombreux contextes comme dans l'enseignement, la surveillance ou encore la rediffusion d'évènements sportifs.

De manière générale, les systèmes de montage automatique visent à résoudre trois problèmes simultanément [24].

— Où se passe l'action à diffuser dans la scène ?
— Comment déplacer les caméras pour la filmer ?
— Quelle caméra doit-on diffuser ?

Afin de répondre à ces questions, les méthodes proposées dans l'état de l'art, sont régulièrement composées de trois étapes : la planification, le contrôle des caméras et la sélection du flux vidéo d'intérêt.

Le premier problème, la planification, consiste à extraire un certain nombre d'informations permettant la compréhension de la scène. Ces informations permettent de situer l'action dans la scène et ainsi de définir l'orientation des caméras pour permettre de la filmer. De plus, les informations sur le contenu de la scène sont utilisées afin de sélectionner la caméra présentant le meilleur contenu pour le spectateur à chaque instant.

Le contrôle des caméras permet de déplacer une caméra de son état initial jusqu'à la position

permettant au mieux de capter l'action. Ce déplacement doit être assez fluide pour ne pas gêner le spectateur. Le contrôle des caméras s'applique aux caméras réelles (Caméras PTZ : Pan Tilt Zoom : Rotation Inclinaison Zoom) mais également aux caméras virtuelles [3].

Enfin, lorsque plusieurs caméras observent l'action, la dernière étape des systèmes de montage consiste à sélectionner la caméra proposant la meilleure qualité d'image et présentant le meilleur angle de vue pour le spectateur.

Afin de pouvoir faciliter la mise en place d'un système de montage automatique, il est nécessaire de comprendre le fonctionnement de différentes étapes d'un tel système, ainsi que les relations qui les lient. Dans les parties suivantes de ce chapitre, nous faisons un état de l'art des différentes méthodes existantes, utilisées lors de la mise en place d'un système de montage automatique.

1.1 Planification du montage automatique

Le premier problème consiste à identifier ce que veulent voir les spectateurs. En d'autres termes, trouver les éléments importants dans la scène filmée, comme par exemple, localiser le porteur du ballon, identifier le locuteur, afin d'y focaliser l'attention du spectateur. Ces éléments importants de la scène dépendent du type d'évènement qui est diffusé. Les actions se déroulant lors d'un match de basketball sont différentes des actions lors d'une réunion. De plus, le placement des caméras par rapport à la scène implique l'utilisation de méthodes adaptées à l'angle de vue. Il est donc important de choisir les techniques d'identification de l'action adaptées à la situation.

Les méthodes de planification automatique se basent sur l'utilisation de capteurs afin de localiser où l'action se produit. Nous pouvons séparer ces capteurs en trois catégories distinctes [85] :

- les dispositifs de localisation embarqués, permettant de repérer l'action dans l'espace scénique à partir de données provenant des dispositifs (GPS, Infra rouge, ...) ;
- les microphones sont régulièrement utilisés dans les évènements où une ou plusieurs personnes parlent dans la scène ;
- les caméras qui permettent la localisation et/ou à l'identification de l'action dans la scène à partir d'une source vidéo.

Le choix d'un capteur est réalisé en fonction de l'objet ou de l'action recherché dans la scène et donc dépend du contexte d'utilisation. De plus, en fonction de la complexité de la scène, il peut être intéressant d'acquérir des informations de différents capteurs. Il est donc nécessaire de connaître les différents types de capteurs utilisés dans la littérature ainsi que leurs applications.

1.1.1 Approches basées sur les dispositifs de localisation embarqués

Les systèmes basés sur les dispositifs de localisation embarqués consistent à faire porter, aux personnes d'intérêt, un appareil émettant un signal électrique ou magnétique permettant de les localiser. Plusieurs récepteurs sont utilisés afin de déterminer la position de la personne portant le dispositif. Ces systèmes peuvent être utilisés à l'intérieur [72], mais les utilisations les plus fréquentes sont dans des évènements extérieurs [53, 46, 97].

3. Nous faisons ici référence aux caméras qui ré-échantillonnent les images provenant de vidéos du monde réel. Nous ne traitons pas le cas des caméras produisant des images dans un environnement virtuel.

Kameda et al. [72] utilisent un système de balises ultrasoniques afin de localiser précisément le professeur dans une salle de cours. Deux dispositifs sont positionnés sur les épaules de l'enseignant et quatre récepteurs sont positionnés au plafond permettant de le localiser avec une précision de 5 cm. Cependant, il est possible que la localisation échoue à cause d'occlusions ou de mouvements des épaules. Pour pallier ces problèmes, les auteurs utilisent également des méthodes basées sur l'audio et la vidéo pour assurer une localisation tout au long de l'évènement.

Dans [53, 46], les auteurs utilisent un système basé sur les radio-fréquences afin de pouvoir localiser les joueurs lors d'un match de football. L'installation requiert que chaque joueur porte un boîtier transpondeur sur lui et que différentes antennes soient installées dans le stade. Dans [97], l'utilisation de puce GPS permet de suivre les joueurs lors d'un match de football.

Mate et al. [94] utilisent les différents capteurs d'un smartphone afin de comprendre la scène et notamment la boussole magnétique, l'accéléromètre, le GPS. Ce grand nombre d'informations permet de réaliser un montage à partir de vidéo provenant de différents utilisateurs, quel que soit le contexte.

L'avantage principal de ces systèmes est la précision de la localisation que ce soit en intérieur ou en extérieur. Cependant, cette précision implique que les personnes d'intérêt portent un dispositif, qui peut être incommodant et/ou gênant. De plus, l'utilisation d'un tel système est relativement onéreux, ce qui se répercute sur le coût global de la production.

1.1.2 Approches basées sur les microphones

Dans le cas de réunions ou de conférences, les systèmes basés sur les microphones sont régulièrement utilisés pour localiser l'action dans la scène. En effet, lors de la diffusion ayant pour sujet un ou plusieurs locuteurs, des microphones sont utilisés afin de pouvoir entendre intelligiblement la personne en train de parler. Il est alors possible d'utiliser les microphones pour localiser des locuteurs.

Rui et al. [119] utilisent deux microphones afin de pouvoir localiser un étudiant posant une question dans une salle de classe. Le but de la localisation de source sonore est d'estimer l'angle de sorte que la caméra puisse pointer dans la bonne direction. La corrélation croisée entre les signaux issus de deux microphones permet d'estimer le délais temporel entre les deux captations et ainsi estimer la localisation du locuteur dans la pièce.

Comme seulement deux microphones sont utilisés, seul l'angle de rotation est calculé. L'augmentation du nombre de microphones permet une localisation plus précise du locuteur.

Dans [29], les auteurs proposent une méthode de détection de locuteurs en étudiant la corrélation entre le son et la vidéo. L'algorithme proposé fonctionne uniquement sur les scènes où un seul locuteur se trouve devant la caméra, et en plan rapproché. Dans [102], une série de microphones est ensuite utilisée afin de localiser plus précisément le locuteur. La source de son est estimée en comparant les différence de phase et d'intensité des différents microphones. Afin d'améliorer la localisation du locuteur, des caméras sont utilisées pour détecter des protagonistes dans l'image (détecteur de buste). L'association de ces deux types de capteurs améliore la robustesse du système en levant l'ambiguïté en cas d'occlusions.

Dans [82], les auteurs proposent une méthode pour localiser un locuteur dans une salle de réunion. Le système de captation est composé de 3 caméras et d'un ensemble de 16 microphones positionnés dans la salle sous la forme de deux cercles de 8 microphones. L'étude des signaux des différents microphones permet une localisation azimutale des sources sonores dans la pièce. La corrélation avec les images capturées par les caméras permet de localiser précisément le locuteur.

Les méthodes basées sur les microphones sont régulièrement utilisées dans des scènes inté-

rieures afin de localiser une ou plusieurs personnes prenant la parole. Toutefois, ces méthodes impliquent que le signal audio de chaque microphone soit identifiable. De plus, ces méthodes peuvent facilement être influencées par les sons extérieurs à la scène comme par exemple le public lors d'un match de basketball. C'est pourquoi de nombreuses méthodes se basant sur des microphones utilisent un système de vision pour lever les ambiguïtés.

1.1.3 Approches basées sur les caméras

Les approches basées sur les caméras consistent en l'extraction des informations à partir d'images. Un des avantages de l'utilisation des caméras est qu'aucun matériel supplémentaire n'est nécessaire. En effet, ces méthodes exploitent les images qui sont capturées par les caméras utilisées pour filmer la scène.

En 1998, Michael Bianchi [8] propose le système *AutoAuditorium* composé de 4 caméras. Trois de ces caméras sont fixes, dirigées vers la scène, où le professeur se trouve, vers le pupitre et vers le tableau. Enfin, la quatrième est une caméra PTZ qui suit automatiquement l'orateur. La sélection de la caméra d'intérêt est basée sur des règles heuristiques [9]. Une des limitations de ce système est le fait qu'aucune caméra ne filme l'audience. En effet, il est fréquent d'avoir une interaction entre les auditeurs et le professeur. Capturer ces interactions permet d'améliorer le confort utilisateur. Liu et al. [85] proposent également un système composé de quatre caméras. Trois d'entre elles servent à capturer l'enseignant et la présentation projetée. La dernière, quant à elle, filme l'audience. Deux microphones sont utilisés afin de localiser le locuteur dans la salle (professeur ou élève). Enfin certaines méthodes utilisent des techniques de suivi afin d'améliorer la reconnaissance du présentateur [95]. Vineet Gandhi [47] propose une méthode de sélection de plans pour la diffusion de pièce de théâtre en se basant sur l'identification des acteurs par les costumes et en utilisant la position de chaque acteur. Dans le cas de diffusion de cours, la plupart des méthodes localisent l'orateur afin de sélectionner la caméra. En effet, il est rare d'avoir plus d'un présentateur lors d'enseignements. De plus, le fond de la scène est souvent statique et ne présente que de faibles changements. L'estimation de la localisation du locuteur est alors souvent effectuée en utilisant une soustraction de fond [107], ou de différences de trames [144] utilisant une caméra fixe.

Dans le cadre des événements sportifs, la modélisation ou la soustraction de fonds est régulièrement utilisée pour localiser l'action.

Ariki et al. [4] suivent la position de la balle et des joueurs à partir de caméras fixes. Afin d'extraire ces positions, un modèle de fond est soustrait aux images des caméras, permettant ainsi de mettre en évidence les déplacements des objets. Un classifieur, basé sur la position de la balle et des joueurs, est utilisé afin de détecter des événements spécifiques dans un match de football, comme des pénaltys ou des coups-francs pour déterminer les paramètres de zoom pour les caméras virtuelles.

Santiago et al. [121] utilisent une modélisation floue du fond afin de pouvoir extraire la position des joueurs dans un match de handball. Carr et al. [21] utilisent une modélisation du fond par mélange de gaussienne afin de pouvoir extraire la position des joueurs, des arbitres et de la balle.

D'autres méthodes s'appuient sur le fait que des spectateurs sont physiquement présent, afin de trouver la source d'intérêt dans une scène. La direction du regard des spectateurs permet de localiser une personne d'intérêt dans une réunion [133], dans un film [67] ou bien une action d'intérêt dans un match de basketball [30].

Arev et al. [3] proposent une méthode de montage automatique basée sur l'orientation de différentes caméras manipulées par les spectateurs, réduisant ainsi la recherche de l'action à une

simple corrélation entre les orientations.

Les méthodes basées sur l'image sont les plus répandues dans la littératures car ces approches peuvent être mises en place dans des situations non contraintes. Un des avantages de ces méthodes est l'utilisation des caméras pour l'extraction des caractéristiques et la captation de l'évènement. De plus, ces méthodes sont facilement adaptables aux différents contextes. Cependant ces approches peuvent être imprécises ou avoir un coût calculatoire important. Enfin, un grand nombre de difficultés est rencontré, comme l'identification et le suivi de personnes en mouvement, la reconnaissance d'activité ou encore la gestion des occlusions.

L'étape de planification a pour but d'extraire les informations nécessaires pour le contrôle de caméras et la sélection du flux vidéo d'intérêt. Il est alors nécessaire de choisir des capteurs et des méthodes qui soient adaptés à l'évènement filmé, aux objets d'intérêt, ainsi qu'au type de diffusion visé. Les approches basées sur les caméras semblent être les plus adaptées aux différents contextes. Cependant, en fonction des objets d'intérêts et des types de diffusions, les approches basées sur les microphones et sur les dispositifs de localisation embarqués peuvent être nécessaires. Dans certaines situations, des informations peuvent être fournies manuellement, afin d'améliorer ou de remplacer l'étape de planification. Il peut s'agir d'informations connues à l'avance, comme par exemple la position des bancs des remplaçants dans un match de basketball [21], ou bien d'informations extraites par l'humain tout au long de l'évènement [20, 137]. L'utilisation de données fournies "à la main" permet d'obtenir des informations de manière précise, qu'un capteur ne pourrait proposer, mais requiert la présence d'au moins une personne pour la réalisation du montage.

1.2 Contrôle des caméras

La seconde étape dans la réalisation d'un montage automatique consiste à contrôler les caméras afin de déplacer la caméra de sa position initiale au point de vue défini lors de l'étape de planification. Ce déplacement, ou travelling, est effectué afin de mettre en évidence un objet dans la scène. Il peut s'agir d'une modification de l'angle de vue, permettant ainsi de suivre un déplacement, ou bien d'un travelling optique, ou zoom, c'est-à-dire de modifier la distance apparente entre les objets du champ et la perspective, permettant de mettre en évidence une partie de la scène. Ce déplacement doit s'effectuer en adoptant un mouvement agréable pour le spectateur. Le déplacement de la caméra doit être constant et à une vitesse adaptée afin de ne pas déstabiliser le spectateur [75].

Comme évoqué précédemment, le contrôle de caméra ne s'applique pas uniquement aux caméras robotiques. Il est possible de générer une nouvelle vue à partir d'images provenant de caméras fixes. Le principe est de modifier une image à partir d'informations géométriques ou photométriques. Il peut s'agir d'une simulation d'un mouvement dans une image fixe (effet Ken Burns) ou bien de la génération d'une nouvelle vue à partir d'un ensemble de caméras (Free Viewpoint camera [19]). Il est possible d'effectuer l'action de zoomer sur une vidéo (fig. 1.3), ou encore de créer une nouvelle vidéo, à partir d'un ensemble de flux vidéo [46].

Nous présenterons tout d'abord les deux types de contrôle de caméra : le contrôle de caméra réelle et le contrôle de caméra virtuelle. Puis, nous présenterons un certain nombre de règles

permettant de produire un contrôle de caméra agréable pour le spectateur.

1.2.1 Contrôle de caméra réelle

Le contrôle de caméra réelle, ou asservissement visuel, consiste à déplacer une caméra de sa position initiale à une position permettant de filmer l'objet d'intérêt. Lorsque des caméras réelles sont utilisées, il est nécessaire que ce déplacement soit réalisé en temps-réel afin de pouvoir filmer la totalité du déplacement. De plus, il est important que le déplacement de la caméra soit le plus proche possible du déplacement de la cible afin de la conserver dans le cadre de la caméra.

Bianchi et al. [8] utilisent, dans le système *autoauditorium*, une caméra grand-angle associée à une caméra PTZ. Les images de la caméra stationnaire sont utilisées pour suivre le déplacement du présentateur, permettant ainsi de contrôler la rotation, l'inclinaison et le zoom de la caméra robotique.

Rui et al. [119] utilisent également une caméra grand-angle afin de permettre le contrôle d'une caméra filmant l'enseignant. Comme la caméra utilisée ne permet pas un déplacement fluide, le déplacement des caméras est effectué uniquement lorsque l'enseignant sort du cadre, ou bien lorsque la caméra n'est pas diffusée. Le choix de l'orientation et du zoom de la caméra se base sur la position du rectangle encadrant de l'enseignant, détecté avec la caméra grand-angle.

Callemein et al. [16] s'intéressent à la captation d'émission de télé-réalité où plusieurs personnes se trouvent dans une maison. Plusieurs caméras fixes sont installées dans différentes pièces et sont utilisées afin de contrôler différentes caméras PTZ. Les informations sur la présence des protagonistes et l'orientation des têtes sont utilisées afin de contrôler les caméras. Afin d'éviter les déplacements fréquents de caméra, ils ne sont réalisés qu'une fois les protagonistes devenus stationnaires. Un ensemble de règles cinématographiques est utilisé afin de permettre un cadrage plaisant pour les spectateurs.

Le contrôle de caméra réelle est essentiellement utilisé pour les diffusions en direct car il doit être réalisé en temps réel. Ces systèmes sont souvent utilisés dans des scénarios d'enseignement, de présentation ou de réunion, permettant de focaliser l'attention sur les différents protagonistes. Cependant, ces systèmes sont difficiles à mettre en place dans des scènes dynamiques (sports) du fait du nombre important de joueurs et des déplacements et changements de direction rapides de ces derniers. En effet, afin d'obtenir une vidéo agréable à regarder, il est nécessaire que les mouvements des caméras soient précis et sans à-coups [35], ce qui n'est pas possible à partir des informations extraites de la scène.

1.2.2 Contrôle de caméra virtuelle

L'utilisation de caméras virtuelles dans un environnement réel consiste à retravailler les images obtenues par des caméras réelles. Dans le contexte de l'enseignement, Gleicher et al. [49] proposent de simuler des caméras virtuelles afin de proposer un cadrage adéquat à la rediffusion de cours. De nombreuses méthodes ont suivi cet exemple et il existe aujourd'hui des systèmes de caméra virtuelle dans deux domaines principaux : le sport et l'enseignement. Pour la majorité des méthodes, une caméra grand-angle est utilisée pour enregistrer l'évènement. Les images extraites de cette caméra sont ensuite, après l'enregistrement, utilisées afin de générer plusieurs caméras virtuelles.

Ariki et al [4] utilisent les évènements et les positions des joueurs afin de proposer le meilleur cadrage possible (fig. 1.3). Une caméra virtuelle est générée en prenant une sous-partie du flux vidéo d'origine. Le déplacement de la caméra virtuelle (fig. 1.3a) permet de suivre les joueurs

et la balle tout au long du match. En fonction des actions se déroulant, la vue présentée par la caméra virtuelle est agrandie ou rétrécie (1.3b), permettant le meilleur suivi de l'action.

(a) Panning

(b) Zooming

FIGURE 1.3 – Contrôle de caméras virtuelles proposé par Ariki et al. [4]

Chen et Carr [25] proposent de contrôler l'angle de rotation d'une caméra PTZ à partir de données extraites de la distribution des joueurs lors d'un match de basketball. Pour ce faire, une forêt d'arbres décisionnels est entraînée à partir de la position des joueurs sur le terrain et de l'angle d'une caméra manipulée par un opérateur. Les résultats obtenus sont proches de ceux fournis par un opérateur humain.

À cause des calculs nécessaires à la création de nouvelles vues, les méthodes utilisant des caméras virtuelles sont souvent utilisées pour des diffusions en différé. L'avantage est d'avoir une connaissance temporelle de la scène, permettant ainsi une meilleure compréhension des évènements s'y produisant. Les déplacements des caméras peuvent aussi prendre en compte les évènements futurs et ces méthodes ont souvent un rendu plus agréable pour le spectateur que les méthodes en temps réel. Un autre intérêt des caméras virtuelles est de pouvoir générer des flux vidéos différents à partir d'une seule vue. Par exemple, lors d'une rencontre sportive, il est possible de générer une caméra centrée sur chaque joueur, chaque équipe. Il est alors possible de proposer des angles de vues correspondant aux attentes des spectateurs, ce qui est impossible avec des caméras réelles.

1.2.3 Méthodes de contrôle

Que ce soit pour une caméra réelle ou virtuelle, certaines règles sont à prendre en compte afin d'assurer un cadrage optimal. Ces règles, issues du domaine du cinéma, permettent de produire un contenu agréable pour le spectateur. Le déplacement d'une caméra est utilisé afin de rendre un plan plus dynamique [46]. Cependant, un déplacement inadapté rend la visualisation peu agréable pour le spectateur.

De plus, les déplacements de caméras doivent être adaptés au type d'évènement : on ne peut pas filmer un évènement sportif et une pièce de théâtre de la même façon. Il existe néanmoins un certain nombre de règles communes pour différents types d'évènements. Par exemple, le fait de positionner les caméras à hauteur humaine [40] permet aux spectateurs de s'imaginer être présent dans la scène. Lorsque l'on cadre un personnage, il est souhaitable de laisser plus d'espace dans le cadre devant ses yeux que derrière la tête. Il en est de même dans le cas de sujet mobile : il est nécessaire de laisser plus d'espace dans la direction où il se déplace [14]. Dans le cas de scène dynamique, la règle de l'immobilité relative du cadre et du sujet mobile [14] stipule que, lors d'un plan, soit le sujet, soit le cadre doit être immobile afin que le spectateur ait le temps de comprendre la scène. Dans le cas où l'on filme un évènement sportif avec une seule caméra, il est fréquent que le cadre et les sujets se déplacent en même temps rendant la compréhension difficile pour le spectateur. Il est alors nécessaire que le cadre soit fixe au moment du tir et de se focaliser sur un seul objet (joueur ou ballon) lors des phases d'attaques.

Dans l'objectif d'obtenir un déplacement de caméras au rendu plus esthétique, de nombreux auteurs se sont intéressés à reproduire les mouvements réalisés par des caméramans professionnels. Certaines de ces méthodes sont basées sur l'étude des techniques employées [75], alors que d'autres [25] ont essayé d'apprendre un modèle permettant de relier le déplacement des protagonistes et le déplacement de la caméra. Doubek et al. [40] définissent trois critères calculés à chaque instant, pour le contrôle de la vue. Le premier, utilisé pour les plans larges, correspond à la proportion de l'image qu'un objet 2D occupe dans une vue. Cette mesure est diminuée si le sujet se trouve à la frontière de la vue. Le second critère se base sur la vitesse et la direction du sujet afin de sélectionner une caméra présentant le meilleur angle de vue sur la source d'intérêt en mouvement. Enfin, le dernier critère est basé sur la visibilité de la peau, permettant de proposer une vue rapprochée contenant la tête et les mains d'un protagoniste.

Le contrôle des caméras permet de proposer aux spectateurs la meilleure vue possible sur un évènement. Pour ce faire, différents types de caméras peuvent être utilisés, en fonction du style de diffusion, et des contraintes de la scène. Dans le cas de diffusions en direct, les caméras réelles sont préférées, tandis que les caméras virtuelles présentent un plus grand intérêt dans le cas de diffusions en différé. Le respect de règles cinématographiques permet de produire des déplacements de caméras, et des nouveaux plans de vue, avec un rendu esthétique. Les nouveaux flux vidéos générés présentant la scène sous différents angles, il est alors nécessaire de sélectionner la caméra correspondant aux attentes des spectateurs.

1.3 Sélection des flux vidéos d'intérêts

L'étape finale d'un montage automatique définit quelle caméra doit être diffusée à chaque instant. Pour ce faire, les méthodes se basent sur les connaissances extraites lors de l'étape de planification afin de savoir sur quelle caméra se déroule l'action que le spectateur souhaite voir. Lorsque plusieurs caméras filment le point d'intérêt dans la scène, il est nécessaire de sélectionner

la caméra présentant la meilleure vue sur la source d'intérêt. Enfin, dans le but de proposer un flux vidéo agréable à regarder, il est nécessaire de prendre en compte un certain nombre de règles. Pour résumer, l'objectif de l'étape de sélection est de choisir la caméra filmant l'action qui sera préférée par les spectateurs et de passer d'une vue à l'autre de manière esthétique.

Nous présenterons tout d'abord les deux approches de la littérature pour la sélection de la caméra d'intérêt : les approches basées sur des règles et les méthodes basées sur les données. Puis nous présenterons différentes règles permettant un changement de caméras esthétique.

1.3.1 Méthodes basées sur les règles

De nombreuses méthodes basent la sélection de caméras sur un ensemble de règles. Les changements de caméra sont effectués lorsque certaines activités sont détectées dans la scène. Ces activités peuvent se baser uniquement sur la vidéo, comme par exemple une personne entrant dans le champ de la caméra ou sur un événement audio, lorsqu'une personne prend la parole [80].

Dans le cadre de l'enseignement, Mukhopadhyay et al. [101] définissent des règles afin de passer d'une caméra suivant l'enseignant à une vue générale (plan situant l'action dans l'environnement et filmant l'écran de projection). Ainsi, lorsque l'enseignement change de diapositive, le plan d'ensemble de la pièce est proposé pendant 8 secondes afin de permettre aux étudiants de visualiser la diapositive. De plus, ils définissent une durée des plans minimale (3 secondes) et maximale (25 secondes). En effet, un plan trop court a tendance à perturber le spectateur. À l'inverse, un plan trop long nuit à l'attention de l'étudiant. Ainsi, si deux changements de diapositives se succèdent, alors la caméra générale est conservée. Dans le cas où une diapositive est présentée pendant plus de 25 secondes, des plans d'ensemble sont insérés pendant 5 secondes. Un algorithme heuristique a été proposé afin de pouvoir proposer une liste de décision du montage (Edit Decision List), c'est-à-dire une liste ordonnée des différents plans, satisfaisant les contraintes.

Liu et al. [85] utilisent eux trois caméras pour enregistrer un cours. La première caméra est focalisée sur l'enseignant, la seconde est dirigée vers l'audience et est utilisée lorsqu'un étudiant prend la parole. Cette caméra est également utilisée pour montrer, de temps en temps, le public permettant d'obtenir un montage dynamique. Enfin, la dernière caméra propose un plan d'ensemble et est utilisée lorsque le suivi de l'enseignant échoue et qu'aucune personne de l'audience ne prend la parole. Un automate d'états à 3 états permet de sélectionner une des trois caméras utilisée pour l'enregistrement de cours. La transition d'un état à un autre prend en compte le changement de statut de la caméra ainsi que le temps maximal défini pour chaque caméra.

Dans le cas de déplacement d'une personne dans un bureau, Doubek et al. [40] définissent un ensemble de 4 règles basées sur les déplacements et les actions d'un sujet. Par exemple, lorsque le sujet est arrêté, le plan proposé est un plan rapproché ou bien une vue présentant ce que regarde le protagoniste. Un plan large est proposé lorsque le sujet est en déplacement. Enfin, l'écran d'ordinateur est sélectionné lorsque le sujet le regarde.

Ces méthodes permettent d'obtenir des montages proches de ceux réalisés par des équipes professionnelles. Cependant, il est nécessaire d'avoir une parfaite connaissance du type d'évènement filmé et de la façon dont un humain réaliserait ce montage. Les méthodes de sélection basées sur des règles sont souvent utilisées dans les scénarios ayant peu d'activités et peu d'acteurs. De plus, le montage est souvent peu agréable pour le spectateur : les plans peuvent être longs (peu d'actions) et les changements de plans non-opportuns (changements rapides et répétés de caméras). C'est pourquoi de nombreuses méthodes utilisent un ensemble de règles issues du domaine des productions audio-visuelles (règles cinématographiques) afin de rendre les montages plus attractifs. Liu et al. [85] ont interviewé cinq producteurs professionnels afin de créer des règles

comme : ne pas réaliser de coupes franches (jump cuts) ou encore définir une durée minimale à chaque plan. De même, dans le cas de déplacement d'une personne dans un bureau, Doubek et al. [40] définissent un ensemble de règles basées sur des règles cinématographiques comme par exemple, utiliser un plan large lorsque le sujet se déplace ou un plan rapproché lorsque le sujet devient immobile.

Les méthodes basées sur les règles nécessitent d'acquérir un grand nombre de connaissances sur l'évènement afin de choisir les conditions à mettre en place pour effectuer les transitions. Un des avantages de ces méthodes est qu'il est possible de les utiliser pour différents évènements d'un même type, en modifiant certaines conditions de transitions. Cependant, l'acquisition de ces connaissances et l'implémentation des règles est une étape longue à effectuer. De plus, le choix de ces règles est effectué par le créateur du système, ce qui peut ne pas refléter ce que veulent voir les spectateurs.

1.3.2 Méthodes basées sur les données

Afin de proposer une sélection automatique de caméras proche d'un montage manuel, des méthodes basées sur les données ont été proposées. Le principe est d'utiliser des données sur l'évènement à filmer, ainsi que des données provenant d'une équipe de production, afin d'entraîner un modèle capable de réaliser la sélection automatique de caméras.

Wang et al. [137] utilisent les caractéristiques de déplacement des caméras afin de choisir la meilleure caméra lors de la rediffusion de match de football. Le système étudié est composé d'une caméra principale proposant une vue large sur le terrain et de deux caméras, manipulées par des cadreurs, proposant une vue rapprochée sur les joueurs. La problématique de cet article est d'alterner entre la caméra principale et une des deux caméras rapprochées. Pour ce faire, la méthode se base uniquement sur la qualité des images afin de sélectionner une caméra ayant une image convenable pendant un certain temps. En effet, lors de mouvements rapides de caméras, l'image résultante est floue, la caméra ne sera alors pas sélectionnée par un réalisateur. Il est nécessaire d'attendre la fin du déplacement avant de changer de caméra afin d'obtenir un rendu esthétique [14]. Afin de déterminer si la vidéo est convenable pour la diffusion ou non, des modèles de Markov cachés (HMM) sont utilisés. Les informations sur les déplacements des caméras permettent de définir si les images d'une caméra sont convenables ou non.

Chen et al. [21] utilisent également un modèle de Markov caché afin de sélectionner la meilleure caméra virtuelle dans la production de vidéo de Basketball. Une des différences avec la méthode proposée par Wang et al. est est que la taille et la visibilité de l'objet à suivre sont également pris en compte. Ainsi, la qualité et le contenu de l'image sont pris en compte pour la sélection de caméras.

Plus récemment, Chen et al.[20] ont utilisé une approche basée sur les données en entraînant une forêt d'arbres décisionnels sur des données de suivi pour le hockey, et ont été capables de recommander la meilleure vue à un réalisateur humain. Des caractéristiques de bas niveau comme la visibilité du palet, la position des joueurs et l'orientation de caméras PTZ sont utilisées afin de proposer la meilleure vue.

Les méthodes basées sur les données utilisent des caractéristiques de bas-niveau afin de permettre la sélection de la caméra d'intérêt. Les informations extraites sur le jeu, ainsi que les données produites par les cadreurs et réalisateurs, permettent d'entraîner un modèle fournissant un montage proche des montages réalisés par l'humain. Cependant, il est nécessaire, afin de mettre en place une méthode basée sur les données, de disposer d'un jeu de données important.

De plus, la création de ce jeu de données nécessite la présence d'une équipe de captation professionnelle pour chaque évènement spécifique. Un nouvel apprentissage des modèles est nécessaire pour une utilisation dans un autre contexte.

L'étape de sélection est la dernière étape d'un système de montage automatique et permet aux spectateurs de pouvoir visualiser un évènement de manière agréable et compréhensible. Pour ce faire, un certain nombre de règles, souvent issues du domaine cinématographique sont à respecter. Pour garantir un montage agréable à regarder, les deux règles les plus récurrentes sont l'utilisation d'images de bonnes qualités (nettes et avec un déplacement fluide) et le changement régulier de plan afin d'éviter une monotonie dans le flux vidéo généré. Il est également nécessaire que le montage fourni respecte la narration de l'évènement. Pour ce faire, certaines règles cinématographiques doivent être utilisées, comme par exemple la règle des 180 ° qui consiste à sélectionner les caméras situées d'un même coté de l'axe de l'action [110]. Mais le plus important est de pouvoir proposer une caméra présentant l'action dans la scène. La sélection est donc fortement liée à l'étape de planification : afin de pouvoir choisir quelle caméra montrer aux spectateurs, il est nécessaire de pouvoir détecter certaines activités. Les actions produites sur scènes peuvent être reconnues de deux façons : en extrayant un certain nombre de caractéristiques sur la scène ou en s'appuyant sur l'expérience des cadreurs qui orientent les caméras pour filmer l'action.

1.4 Discussions

De nombreux systèmes de montage automatique ont été proposés pour la captation et la rediffusion de divers évènements. Nous synthétisons avec la table 1.1, les différents articles proposant des méthodes de montage automatique. Pour chaque méthode, le type d'évènement est spécifié, ainsi que les méthodes utilisées pour la planification, le contrôle et la sélection. De plus, le type de diffusion (en direct ou en différé), ainsi que la possibilité de personnaliser le montage ont été relevés.

Nous pouvons remarquer que des systèmes de sélection automatique de caméras ont été proposés pour différents types d'évènements dont les deux principales catégories sont le sport et l'enseignement. Les systèmes de montage automatique de la littérature réalisent, de manière générale, les trois étapes décrites dans ce chapitre, à savoir planifier, contrôler et sélectionner. Cependant, les méthodes misent en place dans la réalisation de ces différentes étapes, dépendent fortement du type d'évènement filmé ainsi que les types d'informations que les auteurs vont chercher à extraire. De ce fait, les méthodes proposées dans l'état de l'art sont rarement utilisables dans un autre contexte que celui initialement prévu.

La prise en compte des exigences utilisateurs n'est que rarement abordée dans les systèmes de montage automatique. Proposer la personnalisation aux utilisateurs nécessite de devoir extraire de la scène des informations supplémentaires que celles nécessaires à la réalisation d'un montage. Il en résulte un temps de calcul plus important qui restreint leur utilisation à la diffusion en différé. De plus, l'augmentation du coût calculatoire implique que les méthodes proposant la personnalisation restreignent le choix de personnalisation. En effet, l'utilisateur final n'a accès qu'à quelques règles permettant la personnalisation de contenu.

Enfin les méthodologies présentées fonctionnent soit en temps réel pour une diffusion en direct de l'évènement, soit avec un post-traitement pour une rediffusion. Cependant, les deux types de

diffusions présentent des avantages pour le public. Le temps réel permet aux spectateurs distants de suivre un événement pendant qu'il se déroule. De l'autre côté, la diffusion en différé permet une extraction d'un plus grand nombre d'informations, et notamment une compréhension temporelle, sur la scène. Cela permet de proposer des contrôles de caméras plus précis ainsi qu'une sélection de plans plus agréable pour le spectateur. De plus, il est possible dans le cas d'une diffusion en différé de proposer différents montages, répondant aux exigences des spectateurs.

Face à ces constats, notre travail vise à proposer une nouvelle méthodologie de montage automatique, prenant en compte le contexte, fonctionnant en direct et en différé et prenant en compte les préférences utilisateurs.

TABLE 1.1 – État de l'art des méthodes de sélection automatique de caméra. Cam : caméras, Mic : Microphones, DLE : Dispositifs de localisation embarqués, AM : Annotation Manuelle

Référence	Évènement	Planification	Contrôle	Sélection	Diffusion	Personnalisation
Pinhanez et al. [111]	Émission TV	Cam	Virtuelle	-	Différé	-
Bianchi [8]	Enseignement	Cam	Réelle	Règles	Différé	-
Mukhopadhyay et al. [101]	Enseignement	Cam	Réelle	Règles	Différé	-
Liu et al.[85]	Enseignement	Cam + Mic	Réelle	Règles	Direct	-
Kameda et al. [72]	Enseignement	Cam + Mic+ DLE	Réelle	Règles	Différé	-
Daigo et al. [30]	Basketball	Cam	Réelle	-	Différé	-
Doubek et al. [40]	Surveillance	Cam	Virtuelle	Règles	Direct	-
Rui et al. [119]	Enseignement	Cam + Mic	Réelle	Règles	Direct	-
Takemae et al.[133]	Réunion	Cam	Réelle	Règles	Différé	-
Ariki et al. [4]	Football	Cam	Virtuelle	Règles	Différé	X
Bocconi et al. [10]	Documentaire	AM	-	Règles	Différé	X
Kubicek et al. [80]	Réunion	Cam + Mic	Réelle	Règles	Différé	X
Wang et al. [137]	Football	Cam + Mic	-	Data	Différé	-
Kosmopoulos et al. [79]	Souvenirs	Cam	Réelle	Règles	Différé	X
Chen et al. [21]	Basketball	Cam	Virtuelle	Data	Différé	X
Daniyal et al. [32]	Surveillance	Cam	-	Data	Différé	-
Daniyal et al. [32]	Basketball	Cam	-	Data	Différé	-
Mavlankar et al. [95]	Enseignement	Cam	Virtuelle	-	Différé	X
Carr et al. [17]	Basketball	Cam	Hybride	-	Direct	-
Chen et al. [20]	Hockey	Cam + AM	-	Data	Différé	-
Gandhi [47]	Théâtre	Cam	Virtuelle	-	Différé	-
Hulens et al. [64]	Enseignement	Cam	Réelle	-	Direct	-
Gaddam et al. [46]	Football	Cam + AM	Virtuelle	-	Direct	-
Yus et al. [145]	Rowing race	Cam	Réelle	-	Direct	-
Callemein et al. [16]	Emission TV	Cam	Réelle	Règles	Différé	-
Mate [94]	Évènements sociaux	Cam + Mic + DLE	-	Règles	Différé	X

Chapitre 2

Méthodologie générique de montage automatique

Sommaire

2.1 Introduction

Nous avons présenté dans le chapitre 1 les différentes méthodes utilisées dans les systèmes de montage automatique. La majorité des méthodes proposées dans l'état de l'art se concentre sur un contexte d'application spécifique. En effet, chaque méthode s'intéresse à réaliser un montage automatique adapté à un type d'évènement particulier et à une installation (nombre de caméras, localisation des caméras, dispositifs de détection) particulière. De ce fait, les méthodes existantes ne sont pas aisément transposables ou adaptables à d'autres contextes. Il est souvent nécessaire de refaire toute une étude lorsque l'on souhaite diffuser un nouvel évènement.

Trois étapes sont nécessaires dans la réalisation d'un montage automatique : la planification, le contrôle et la sélection [24]. De ce fait, lorsque l'on souhaite mettre en place un système de montage automatique, il est nécessaire de réfléchir à trois questions :

— que veut voir le spectateur ?
— où doivent être orientées les caméras ?
— quel flux renvoyer au spectateur ?

Le concepteur d'un système de montage automatique doit choisir les méthodes à mettre en place afin de répondre à ces trois questions. Ce choix se base alors sur les connaissances de l'évènement capté et sur les besoins du client.

L'objectif principal de ce chapitre est de concevoir une méthodologie permettant de simplifier la mise en place d'un système de montage automatique, quel que soit le contexte d'application (Sport/Enseignement/Théâtre), le type de caméra utilisé (virtuelle ou réelle), et la nature de la diffusion (directe ou différée), à partir des connaissances sur le contexte d'application.

Pour ce faire, nous proposons dans ce chapitre de réaliser une analyse fonctionnelle des systèmes de montage automatique pour en proposer une architecture générique. Une étude des systèmes existants est menée dans l'objectif d'identifier leurs différences et ainsi prendre en compte les sources de ces différences dans notre méthodologie. Cette étude nous permet d'identifier le contexte comme étant le point d'origine de chaque méthode de montage automatique. Nous intégrons alors les connaissances sur le contexte d'application dans l'architecture proposée, permettant ainsi de s'adapter à chaque évènement. Nous proposons ainsi une modélisation du contexte afin de guider la sélection des sources d'intérêts dans la scène et la sélection des méthodes à mettre en place pour l'élaboration d'un système de montage automatique. Enfin, nous expliquons comment la prise en compte du contexte permet la diffusion en direct d'un évènement et la diffusion en différé des flux vidéos répondant aux desiderata des spectateurs.

2.2 Architecture d'un système de montage automatique

Afin de proposer une méthodologie générique de montage automatique, il est nécessaire de comprendre comment est structuré un tel système et quelles sont les différentes interactions entre les étapes d'un montage automatique. Nous proposons alors de modéliser un système de montage automatique grâce à la méthode SADT (Structured Analysis and Design Technics) [92]. Cette méthode permet une description graphique d'un système par analyse fonctionnelle descendante, c'est-à-dire une analyse partant d'une vue générale d'un système pour aller vers une vue détaillée des actions qui le composent. Cette modélisation nous permet de comprendre les interactions entre les différentes parties d'un système.

2.2.1 Analyse fonctionnelle des systèmes de montage

Les différents systèmes de montage automatique de la littérature reposent sur la réalisation de trois étapes successives afin de produire des flux vidéo montés [24] : la planification, le contrôle et la sélection. La figure 2.1 présente une vue générale des interactions entre chacune de ces étapes. L'étape de planification interprète des données provenant de capteurs afin de fournir des

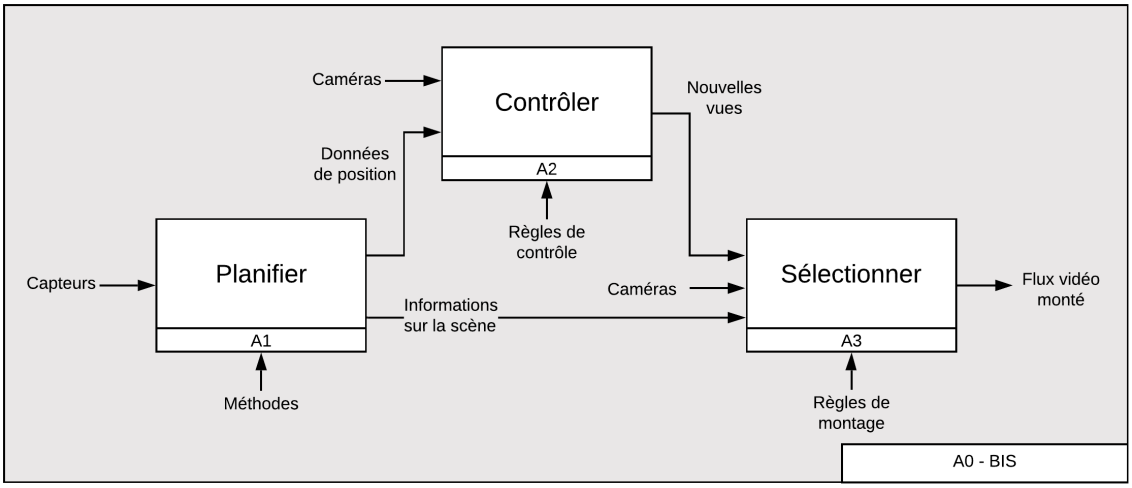


FIGURE 2.1 – Interactions entre les étapes d'un système de montage automatique (vue A0-bis).

informations de localisation, de trajectoire ou d'identification de la source d'intérêt aux étapes de contrôle et de sélection. Ces informations sont utilisées afin de fournir de nouvelles vues sur la source d'intérêt lors de l'étape de contrôle de caméras (virtuelles ou réelles). Ces informations permettent également, dans la troisième étape, de sélectionner la caméra représentant au mieux l'action. Afin de mieux comprendre les différentes opérations réalisées, nous proposons une analyse fonctionnelle de plus bas niveau de chaque étape.

Étape de planification

L'étape de planification a pour objectif d'extraire des informations sur le centre d'attention dans la scène en utilisant différents capteurs et différents algorithmes pour être ensuite exploitées par les étapes de contrôle et de sélection. Il s'agit donc d'une étape clef dans la réalisation d'un système de sélection automatique de caméras.

De manière générale, l'étape de planification vise à générer des informations sur la localisation, la trajectoire ou l'identité de la source d'intérêt (fig 2.2). Pour ce faire, trois mécanismes peuvent être mis en oeuvre.

Localiser la source d'intérêt permet d'obtenir la position de l'objet à filmer dans la scène. Différents capteurs et méthodes peuvent être mis en place afin d'obtenir la position d'une source d'intérêt. Les informations de localisation ou de suivi obtenues aux instants précédents peuvent être utilisées afin d'améliorer la détection à l'instant t. Cette étape produit des informations sur la position des sources d'intérêt qui peuvent être utilisées directement dans les étapes de contrôle et de sélection. Cependant ces informations de localisation sont régulièrement utilisées dans les autres étapes de la planification.

Le suivi d'une source d'intérêt permet de s'assurer que la source considérée est toujours la même au cours du temps afin de s'assurer de la stabilité de la détection. De plus, le suivi permet

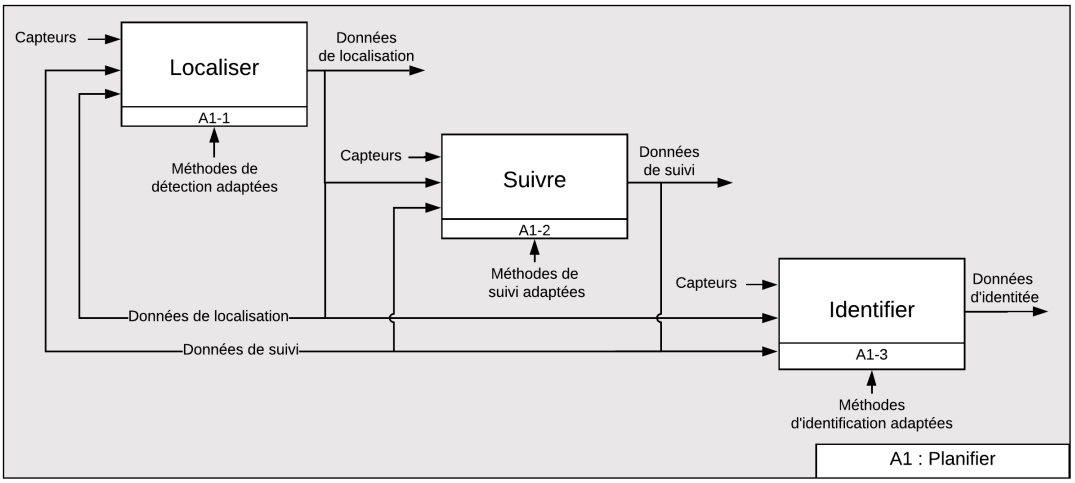


FIGURE 2.2 – Étape de planification (vue A1) d'un système de montage automatique.

de connaître la trajectoire de la source d'intérêt au cours du temps. La prédiction de trajectoire aide à la localisation en anticipant la zone où se trouvera l'objet d'intérêt aux instants suivants. La prédiction de trajectoire est également utilisée pour le contrôle et la sélection de caméras : anticiper les déplacements permet de proposer un contrôle et une sélection plus fluides.

Enfin, l'identification d'une source d'intérêt permet, lorsque plusieurs sources d'intérêts sont sur scène, de choisir la caméra présentant la source d'intérêt que le spectateur veut voir. Par exemple, lors d'un évènement sportif où plusieurs joueurs sont sur le terrain, l'identification permet de contrôler les caméras vers un joueur particulier et de sélectionner la caméra filmant avec le meilleur cadre ce joueur. L'identification des sources d'intérêt est donc un élément important pour la personnalisation des flux vidéo.

Étape de contrôle

Les caractéristiques extraites de la scène, lors de l'étape de planification, peuvent ensuite être utilisées afin de diriger les caméras vers la source d'intérêt dans la scène. L'objectif de l'étape de contrôle est, en effet, de déplacer les caméras de sorte que la source d'intérêt soit toujours dans son cadre. L'étape de contrôle doit donc déplacer une caméra de sa position actuelle à la position optimale pour filmer l'objet d'intérêt. Cette étape est illustrée dans la figure 2.3.

Les informations produites lors de l'étape de planification sont liées aux repères des capteurs utilisés. Afin de pouvoir les utiliser, les coordonnées de la source d'intérêt dans le repère de la caméra doivent être calculées afin d'obtenir la position que la caméra doit atteindre. En d'autres termes, la première étape est de trouver l'orientation que la caméra doit avoir afin de cadrer la position de la source d'intérêt. De plus, il est nécessaire de connaître la position initiale de la caméra afin de pouvoir définir le déplacement de la caméra. D'autres informations peuvent être utilisées afin d'en améliorer le contrôle. Les informations sur la trajectoire de la source d'intérêt sont particulièrement intéressantes car elles permettent d'anticiper le mouvement de la source d'intérêt à cours terme et ainsi d'éviter les à-coups lors du contrôle. De plus, dans le cas d'une diffusion en différé, il est possible de connaître les positions futures de la source d'intérêt permettant d'offrir un déplacement fluide des caméras.

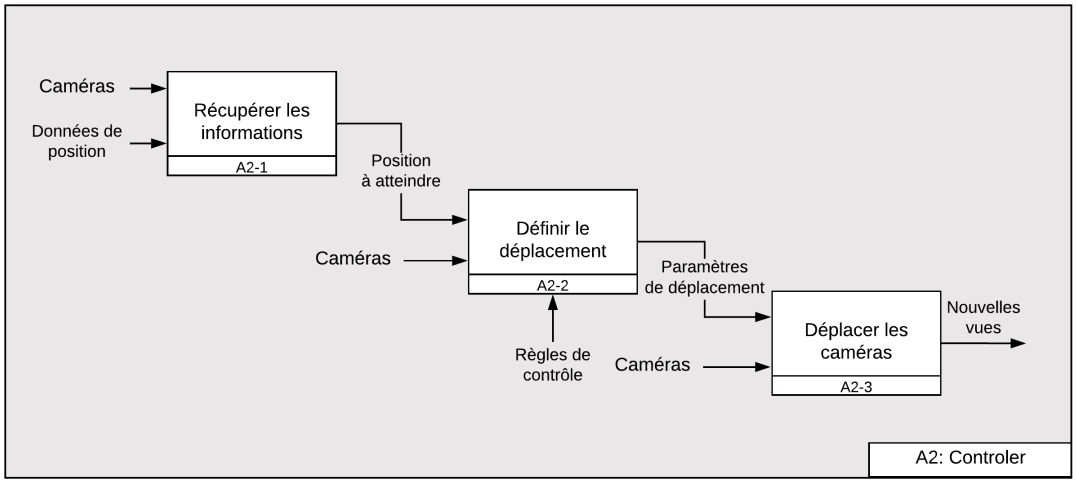


FIGURE 2.3 – Étape de contrôle (vue A2) d'un système de montage automatique.

Une fois les coordonnées dans l'espace caméra obtenues, la seconde étape est de savoir comment déplacer la caméra afin d'atteindre sa position optimale. Les informations sur la position à l'origine et sur la position à atteindre sont utilisées pour calculer les paramètres de déplacement (Rotation, Inclinaison, Zoom) de chaque caméra. Afin d'offrir un flux vidéo agréable pour le spectateur, un certain nombre de règles cinématographiques peuvent être prises en compte, comme par exemple le fait que le mouvement réalisé ne soit ni trop rapide, ni trop lent.

Enfin, la dernière étape du contrôle est le déplacement en lui-même. Ce déplacement produit de nouvelles images qui sont ensuite exploitées lors de l'étape de sélection.

Étape de sélection

La dernière étape d'un montage automatique consiste à sélectionner la caméra d'intérêt. Le choix des différents plans doit permettre aux spectateurs de pouvoir comprendre ce qu'il se passe dans la scène. De plus, il est nécessaire que les transitions entre les différents plans soient faites à un moment opportun, permettant de proposer un montage agréable à regarder pour le spectateur. La figure 2.4 présente les différents modules qui composent l'étape de sélection d'un système de montage automatique.

Le premier point est de définir la source d'intérêt à diffuser à chaque instant. En effet, différentes sources d'intérêts peuvent être extraites durant un évènement permettant de garantir une diffusion agréable pour le spectateur. Par exemple, lors de la retransmission de conférence, il est possible que l'intérêt change du conférencier aux diapositives étant affichées. Ce changement de source d'intérêt permet à l'utilisateur de mieux assimiler les informations transmises par le conférencier. Le choix de la source d'intérêt repose sur les informations obtenues lors de la planification, aux vues sélectionnées précédemment et à un certain nombre de règles de montage défini lors de la mise en place du système.

Lorsque plusieurs caméras filment la source d'intérêt, il est nécessaire de définir la meilleure caméra filmant l'objet d'intérêt. Le déplacement des caméras ayant modifié le point de vue, l'analyse des images provenant des différentes caméras est requise afin de s'assurer de la qualité et du cadrage des images.

Enfin, le dernier point est la sélection de la caméra à diffuser à l'instant t. Cette sélection

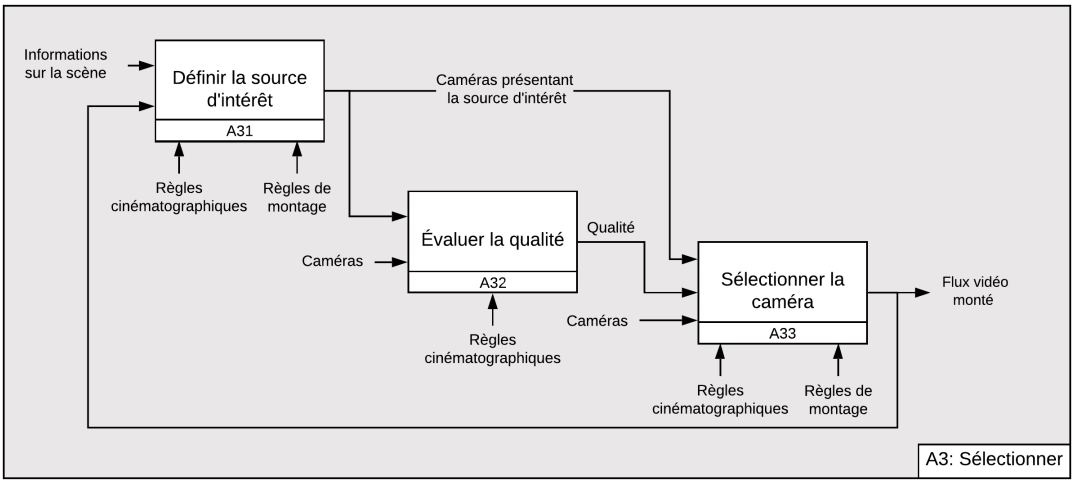


FIGURE 2.4 – Étape de sélection (vue A3) d'un système de montage automatique.

repose sur la qualité visuelle des flux vidéo présentant la source d'intérêt ainsi qu'un certain nombre de règles cinématographiques propres à l'évènement filmé. De plus, certaines d'entre elles influent sur les transitions entre les différentes sources d'intérêt. Par exemple, il est important de veiller à conserver une luminosité constante entre les caméras ou à respecter la règle des 180°.

2.2.2 Impact du contexte sur les systèmes existants

Nous avons proposé une architecture générique de système de montage automatique grâce à une analyse SADT. Cette architecture est obtenue à partir des connaissances sur les différentes méthodes de sélection automatique de caméras. Elle prend en compte les différents traitements réalisables dans un système de montage automatique. Cependant, en fonction de l'utilisation des systèmes, toutes les étapes ne sont pas réalisées et les méthodes mises en place sont différentes. C'est pourquoi nous proposons de relever les trois principales sources de différence, à savoir le types de diffusion, le type d'évènement et l'installation des caméras, afin de pouvoir par la suite les prendre en compte dans la réalisation d'un système de montage autonome.

Type de diffusion

Un des premiers aspects qui différencie les méthodes de la littérature est le type de diffusion qui est visé. En effet, pour proposer un montage automatique pour une diffusion en direct, il est nécessaire que le système mis en place fonctionne en temps réel, ce qui a un impact dans les différentes étapes le composant.

Les méthodes et capteurs mis en place pour la planification doivent extraire les informations sur les sources d'intérêts dans un temps compatible avec la vidéo. La majorité des systèmes existants utilisent alors des caractéristiques de bas niveau. De plus, il est nécessaire que les caméras soient orientées vers la source d'intérêt. Les caméras réelles sont régulièrement utilisées dans ce cas car il est nécessaire que leurs contrôles soient réalisés en temps réel. Les caméras virtuelles nécessitant plus de temps de calcul, leur utilisation est rare dans ce cas. Enfin, les diffusions en direct impliquent qu'un seul flux soit disponible. Il est alors nécessaire que ce flux

contienne la source d'intérêt répondant aux attentes du plus grand nombre de spectateurs.

Lors d'une diffusion en différé, les méthodes mises en place lors de la planification ne sont pas contraintes temporellement. En effet, la différence de temps entre la captation et la diffusion permet l'utilisation de méthodes ne fonctionnant pas en temps réel. Le fait que la planification ne soit pas contrainte par le temps rend possible d'extraire plus d'informations sur les sources d'intérêt. De plus, la connaissance de l'évènement dans sa totalité permet d'avoir des informations temporelles sur les actions réalisées. Le fait qu'un plus grand nombre d'informations soit disponible, rend possible un meilleur contrôle de caméras virtuelles. Il en va de même pour la sélection des flux vidéo où les transitions entre les plans peuvent être effectuées en fonction du déroulement des différentes actions.

Type d'évènement

Le type d'évènement détermine les méthodes mises en œuvre dans un montage automatique : on ne filme pas tous les évènements de la même façon. Les méthodes mises en place pour la planification, le contrôle ou la sélection doivent alors être adaptées à l'évènement filmé.

En effet, les méthodes utilisées lors de la planification ne seront pas les mêmes que l'on filme un évènement sportif ou une réunion. Par exemple, le suivi de personne ne sera pas nécessaire dans les scénarios avec peu d'action. Dans le cadre des réunions, les différentes personnes participantes ne se déplacent pas et il n'est alors pas nécessaire de mettre en place une méthode de suivi. De la même manière, il n'est pas nécessaire de mettre en place une méthode d'identification dans le cas où une seule personne est sur scène. Le contrôle et la sélection de caméras sont également impactés par le type d'évènement filmé. Dans certains cas, les déplacements de caméra doivent être rapides, comme par exemple lors d'évènements sportifs. Dans d'autres cas, les déplacements de caméras sont plus lents comme dans le cas de l'enseignement. Il en va de même lors de la sélection des caméras, où les changements sont plus fréquents du fait de l'activité dans la scène.

Installation des caméras

L'installation des caméras, que ce soit leurs nombres ou leurs positions, est un élément important dans la configuration d'un système de montage autonome. Les méthodes mises en place lors de la planification dépendent de la façon dont les caméras sont installées. Par exemple, dans [30], les auteurs exploitent le fait d'avoir une caméra orientée vers les spectateurs afin de trouver la source d'intérêt dans un match de basket, en fonction de l'orientation des spectateurs. Dans [17], aucune caméra n'est dirigée vers le public, mais deux caméras azimutales sont disponibles et exploitées afin de trouver où est la source d'intérêt dans le match. Les étapes de contrôle et de sélection sont également dépendantes de l'installation des caméras. Le nombre de caméras filmant un évènement a une influence sur la façon d'orienter les caméras et sur les règles à mettre en place pour la sélection de flux vidéo d'intérêt. Par exemple, lors d'une captation de réunion, si une caméra est orientée vers chaque participant, le contrôle de caméra ne sera pas nécessaire et seule l'étape de sélection permettra de montrer aux spectateurs la caméra où l'action se passe. Dans le cas où une seule caméra est utilisée pour filmer une scène, le contrôle est primordial afin de proposer aux spectateurs le meilleur angle de vue possible et la sélection de flux vidéo n'est pas nécessaire.

Le type d'évènement, le type de diffusion et l'installation des caméras influent sur les éléments mis en place dans un système de montage automatique. Connaître le contexte d'application est nécessaire pour la mise en place d'un système de montage automatique. Une des causes principales de variabilité est les sources d'intérêts qui sont l'élément de base des méthodes mises en place. En effet, elle doit être choisie en fonction du type de diffusion et du type d'évènement. De plus, l'installation des caméras influe sur les méthodes à mettre en oeuvre pour extraire cette source d'intérêt, ainsi que sur les méthodes permettant d'avoir, à tout moment, une caméra orientée vers la source d'intérêt dans la scène. La connaissance de la source d'intérêt à extraire permet de choisir les méthodes de planification à mettre en place (algorithmes et capteurs), ainsi que la façon de contrôler la caméra et de sélectionner le flux vidéo d'intérêt. Il est alors nécessaire, dans le but de proposer une méthodologie générique de mise en place d'un système de sélection automatique de caméras, d'intégrer la connaissance du contexte.

2.2.3 Vers une généralisation du montage automatique

Les différences des systèmes de montage automatique, présentées dans la section 2.2.2, étant liées au contexte d'application, il est nécessaire de l'intégrer à l'architecture de montage automatique pour proposer une méthode générique. Les connaissances sur le contexte permettent de choisir les méthodes, les règles et les sources d'intérêts qui seront utilisées lors de la mise en place d'un système de montage automatique. Nous proposons donc d'intégrer ces connaissances du contexte dans l'architecture proposée afin de permettre de fournir une méthodologie générique de montage automatique.

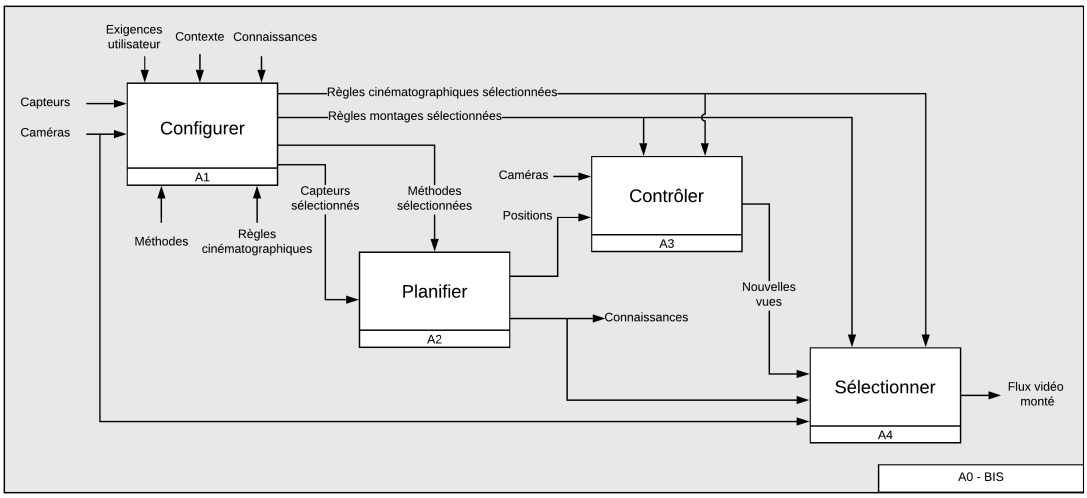


FIGURE 2.5 – Proposition de méthodologie générique de montage automatique (vue A0-bis)

La figure 2.5 présente notre proposition de méthodologie générique de montage automatique. L'ajout de l'étape **configurer** permet d'adapter le système de montage automatique à chaque situation. Cette étape a pour rôle de définir les paramètres des différentes étapes du système à partir des informations du contexte. Son premier objectif est d'identifier les sources d'intérêts dans la scène. La connaissance des sources d'intérêts permet ensuite de choisir les méthodes pour extraire les informations sur ces dernières, ainsi que pour contrôler et sélectionner les caméras, tout en respectant les contraintes de la captation.

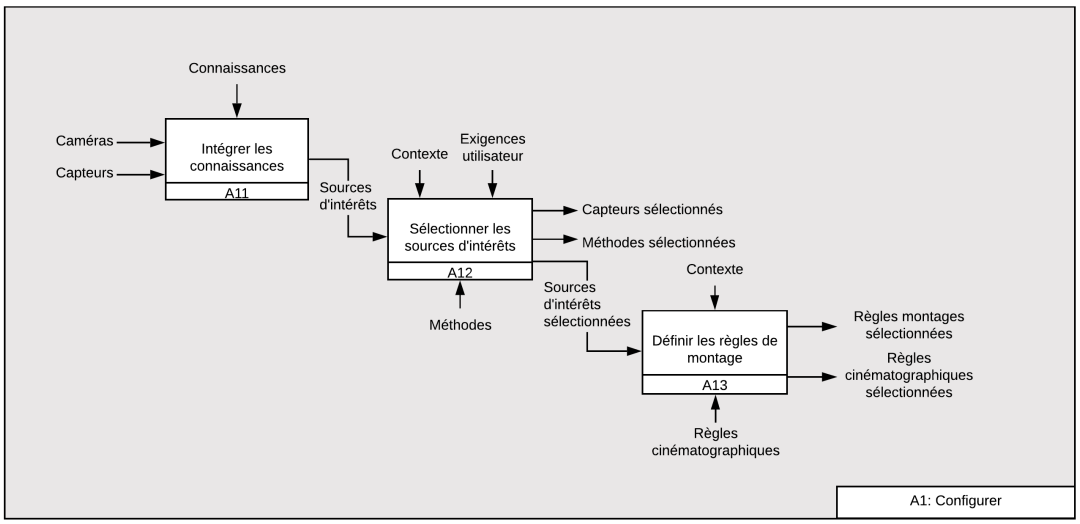


FIGURE 2.6 – Étape de configuration (vue A1) d'un système de montage automatique.

La figure 2.6 présente l'analyse de plus bas niveau de la configuration qui se découpe en trois étapes. La première consiste à intégrer un certain nombre de connaissances sur l'évènement que l'on souhaite diffuser. Ces connaissances peuvent venir de différents sources : experts, textes, base de données, ... L'intégration de ces connaissances permet de pouvoir identifier les différentes sources d'intérêts, ainsi que leurs attributs, intervenant lors de l'évènement. En fonction du type de diffusion et des exigences des utilisateurs, une ou plusieurs sources d'intérêts sont sélectionnées. Cela permet de choisir les capteurs et les méthodes à mettre en place afin d'extraire leurs caractéristiques dans la scène. Enfin, en fonction des sources d'intérêts sélectionnées et du contexte de l'évènement, des règles de montage et cinématographiques sont définies.

2.3 Intégration des connaissances pour une généralisation

La contextualisation du montage automatique permet d'adapter les systèmes à chaque situation. Les sources d'intérêt étant un des éléments clefs dans le choix des méthodes des différentes étapes, il est nécessaire de mettre en œuvre une méthodologie pour identifier et sélectionner ces sources.

Afin d'identifier les différentes sources d'intérêts dans n'importe quel contexte, nous capitalisons les connaissances à partir de la littérature. Ces connaissances sont exprimées dans l'état de l'art, dans un langage naturel parfois imprécis et subjectif. Nous entreprenons une formalisation objective de ces connaissances en utilisant la méthodologie NIAM (Natural language Information Analysis Method) [105] associée au formalisme ORM (Object Role Modelling)[51].

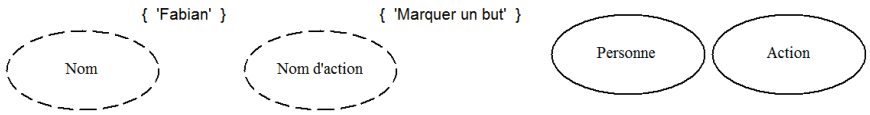
La méthode NIAM/ORM est basée sur l'acquisition de connaissances à partir d'un énoncé en langage naturel et sa modélisation comme modèle conceptuel. Afin d'assurer la cohérence de ce modèle, elle permet d'ajouter des contraintes (unicité, totalité,...) sur les objets et leurs relations. Le modèle de connaissances NIAM/ORM ainsi obtenu peut enfin être soumis pour validation par un expert après transcription en langage naturel. L'un des avantages de NIAM est la méthodologie de mise en œuvre. De plus, le modèle NIAM peut être facilement traduit en langage OWL (Web Ontology Language) [62] ou dans le formalisme UML (Unified Modelling

Language) [52].

2.3.1 Méthodologie NIAM/ORM

L'objectif premier de la méthodologie NIAM/ORM est de permettre l'analyse des connaissances relatives à un domaine délimité (appelé Univers de Discours ou Univers d'Intérêt). L'approche de modélisation de la méthode repose sur trois axiomes pour assurer la qualité du modèle obtenu [2]. Le premier consiste à énoncer l'équivalence sémantique entre les énoncés d'un fait en langage naturel et un ensemble de phrases élémentaires. Une phrase élémentaire est une phrase qui ne peut être décomposée en phrases plus courtes sans perte de sémantique. Elle représente un seul fait, et est de la forme "sujet verbe complément". La décomposition de l'énoncé de la sorte permet de minimiser l'ambiguïté de l'utilisation du langage naturel. Un sujet peut être représenté par deux types d'objets dans le formalisme NIAM/ORM. Les Types d'Objets Lexicaux (LOT) représentent les objets du monde réel et les Types d'Objets NOn Lexicaux (NOLOT) représentent les objets abstraits.

Prenons l'exemple de la phrase "Fabian marque un but", il est possible de reformuler la phrase de manière à éviter toute ambiguïté. Dans ce cas, la phrase devient : "La Personne qui a un Nom "Fabian" réalise l'action désignée par le Nom d'action "marquer un but". Les objets "Fabian" et "marquer un but" sont les objets du monde réel (LOT) et les objets "Personne" et "Action" sont les objets abstraits (NOLOT). Ces deux concepts sont illustrés Figure 2.7 avec le formalisme ORM.



(a) LOT (b) NOLOT

FIGURE 2.7 – Notions sur le formalisme NIAM/ORM

La seconde étape de la modélisation consiste à mettre en évidence les "idées" et les "ponts de dénominations". Les idées représentent les liens entre les objets abstraits. Les ponts de dénominations représentent les liens entre les objets abstraits et les objets qu'ils représentent dans le monde réel.

La phrase "Une Personne réalise une Action" élaborée à partir de la phrase précédente représente une idée. Chaque idée est porteuse de l'information contenue dans le modèle conceptuel. En reprenant la phrase de départ reformulée, deux autres phrases sont présentes : "une Personne a un Nom" et "une Action a un Nom d'Action". Ces deux phrases représentent des ponts de dénomination et ne portent aucune information. Elles servent juste à représenter la partie réelle de l'Univers d'Intérêt dans la base de données finale.

Enfin, l'ajout de contraintes sur les objets composant une idée doit être effectué afin de s'assurer de la cohérence entre la connaissance modélisée et les faits observés. À la phrase " Une personne a un nom", nous pouvons ajouter des contraintes d'unicité et de totalité afin de supprimer tout équivoque (Fig. 2.8) : "Toute personne a un et un seul nom" et "un Nom est d'une ou plusieurs personnes". La figure 2.9 présente la modélisation des connaissances exprimées dans la phrase "Fabian marque un but".

Dans cette section, nous avons présenté les bases de la méthode NIAM/ORM permettant l'acquisition et la modélisation des connaissances. Dans les sections suivantes, nous allons nous

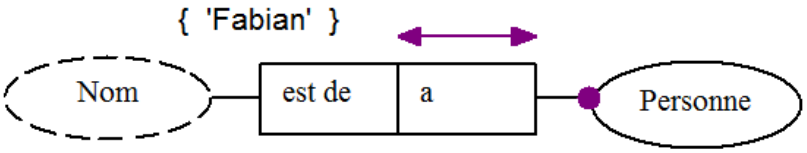


FIGURE 2.8 – Pont de dénomination avec contraintes

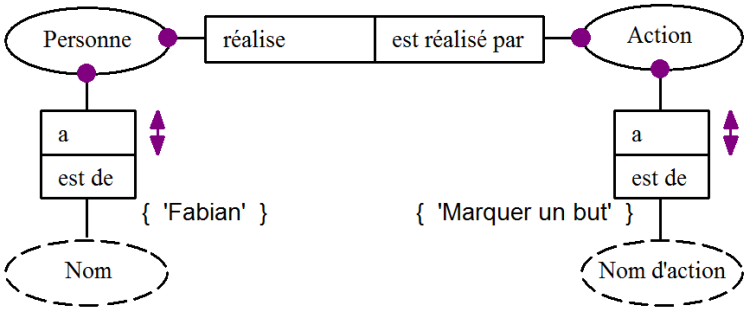


FIGURE 2.9 – Modélisation ORM de la phrase "Fabian marque un but"

appuyer sur les expériences de l'état de l'art pour acquérir des connaissances afin de pouvoir proposer une modélisation générique des sources d'intérêts.

2.3.2 Acquisition de connaissances

Afin de pouvoir proposer une méthode qui soit adaptable à toute sorte de contexte, il est nécessaire de prendre en compte un grand nombre de connaissances. Du fait qu'il n'existe aucun texte, aucune base de données ou encore d'expert traitant du montage vidéo de manière générique, il est nécessaire de regrouper les informations de différentes sources. Nous avons, dans le chapitre 1, analysé les différentes méthodes de montage automatique de caméras de l'état de l'art. Nous présentons ici quatre scénarios différents, issus de cet état de l'art.

Pour la diffusion d'émissions de cuisine, Pinhannez et al. [111, 112] extraient des éléments importants du scénario pour contrôler les caméras. Par exemple, dans l'expression "le chef enveloppe le poulet avec un sac en plastique", les auteurs relèvent différentes informations. Tout d'abord l'acteur, c'est-à-dire la personne qui exécute l'action : le chef. Ils extraient aussi l'action, les opérations effectuées par le chef : envelopper. Des instruments ou des ingrédients sont également extraits de la phrase afin de faciliter la détection d'événements ou le contrôle de la caméra. Comme l'émission est scénarisée, il est possible de connaître les informations sur le lieu et l'orientation du chef. L'utilisation d'un modèle approximatif du monde ("approximate world model"), c'est-à-dire une description grossière des éléments d'une scène, permet d'utiliser des méthodes de traitement d'images spécifiques à la situation et au contexte.

Dans [72], Kameda et al. expliquent le processus d'un cours magistral. "Le professeur se tient devant ses élèves et marche librement. L'enseignant donne une conférence et peut utiliser un

diaporama préparé à l'avance, et peut aussi utiliser le tableau blanc pour écrire les formules. Les étudiants peuvent poser des questions, mais seule une personne peut parler en même temps". De ces phrases, nous pourrions extraire les différentes sources d'intérêt. Tout d'abord, nous avons deux catégories de personnes : l'enseignant et les élèves. L'enseignant peut effectuer différentes actions pendant la leçon. Il "marche" sur scène, parle et peut utiliser le diaporama. Les élèves réalise une seule action :"poser des questions".

Dans le contexte de la diffusion du football, Ariki et al. [4] utilisent les connaissances sur la scène afin de pouvoir proposer une méthode permettant de reconnaître les différents événements d'un match de football. Par exemple, si le ballon se trouve sur le coin et que la distance entre le ballon et la position moyenne des joueurs est intermédiaire, l'événement est reconnu comme un coup de pied de coin (corner). Dans cet exemple, nous pouvons également séparer les actions qui sont exécutées dans la scène des personnes les exécutant. Pour l'action "tirer un coup de pied de coin", les objets qui participent à cette action sont le ballon et les joueurs.

Dans le scénario de surveillance, même si toutes les méthodes permettent la détection des personnes, la source d'intérêt est souvent une action. Par exemple, Daniyal et al. [32] utilisent l'événement "objet étant sur la route" comme source d'intérêt pour un jeu de données contenant des piétons. Dans un scénario de surveillance aéroportuaire, trois événements sont utilisés : lorsqu'une personne court, lorsqu'un objet n'entre pas dans l'ascenseur alors que la porte de l'ascenseur ouverte et lorsqu'une personne marche dans une direction opposée à la direction autorisée.

A partir de ces quatre exemples, nous pouvons remarquer que deux concepts apparaissent dans la littérature : les personnes d'intérêt (POI) et les actions d'intérêt (AOI). L'étude des différents cas de la littérature permet de pouvoir proposer une modélisation adaptable à chaque situation. Nous avons, dans le tableau 2.1, recensé les différentes sources d'intérêt (POI et AOI) des publications présentées dans le chapitre 1, en fonction du type d'évènement filmé.

Nous pouvons remarquer que pour chaque article, au moins une POI et une AOI est présente. Dans le cas des événements sportifs, les sources d'intérêt les plus fréquentes sont : les joueurs [21, 32, 20, 46, 116, 108], l'équipe [21] ou les évènements se produisant durant la rencontre [30, 4, 137, 21, 116]. Dans le cas des conférences, réunions de représentations théâtrales, le point d'intérêt le plus fréquent est le conférencier : la personne qui parle sur scène [8, 85, 72, 133, 95, 107, 64]. Dans certains articles, une autre source d'intérêt est quelqu'un qui pose une question dans la salle [85]. Dans une application de surveillance, le point clé est l'action qui est effectuée dans la scène [40, 32]. Il peut être nécessaire de s'intéresser à la personne qui a fait l'action lorsqu'il est nécessaire de la suivre [96].

Pour certaines méthodes seule la personne ou l'action d'intérêt est indiquée. Cependant, la POI ou l'AOI non mentionnée est facilement identifiable. Par exemple, dans le cas où seule une POI est spécifiée ([17, 145, 7, 47, 16]), l'action d'intérêt correspondante est généralement le déplacement de cette dernière. Dans le cas où la POI n'est pas mentionnée, l'intérêt se porte essentiellement sur l'action et non sur la ou les personnes la réalisant. Par exemple, dans [137] les auteurs s'intéressent aux actions réalisées par les joueurs, mais l'identité des POI n'est pas pertinente pour le montage.

TABLE 2.1 – Différents points d'intérêt des systèmes de montage automatique de la littérature

Référence	Personne d'intérêt	Action d'intérêt
	Enseignements	
[8]	La personne sur scène	Une diapositive est projetée
[85]	Une personne	Pose une question
[72]	Une personne	Parle, change de slide, écrit au tableau
[95]	L'enseignant	se déplace
[64]	L'enseignant	ses mouvements
	Réunion	
[133]	Le locuteur et le destinataire	Parler
[80]	Des personnes	Ses évènements
	Surveillance	
[32]	Des personnes	Piéton sur la route, Personne qui cours
[40]	Une personne	se déplace, regarde
[96]	Des personnes	-
	Sports	
[30]	-	Jeu notable
[4]	La balle et les joueurs,	Les évènements
[137]	-	Attaques, Fautes et autres
[21]	Position des joueurs	Évènements
[32]	Joueurs	Tir au panier, Haute activité
[17]	Les joueurs	-
[20]	Le palet, des joueurs	-
[46]	La balle, un joueur, des joueurs	Mouvements
[145]	Un bateau	-
[7]	Un nageur	-
	Autres	
[111]	Le chef	Parle, mélange, emballe,
[10]	Des personnes	Parlent d'un sujet, avec une attitude
[79]	Un visiteur	se déplace dans un parc
[47]	Les acteurs	-
[16]	Des personnes	-
[94]	indéfini	indéfini

2.3.3 Modélisation des connaissances

La section précédente a mis en avant le fait que les sources d'intérêts dans le contexte du montage automatique pouvaient être séparées en deux concepts distincts, les personnes d'intérêts (POI) et les actions d'intérêts (AOI), quel que soit le type d'évènement filmé. La phrase "Une personne réalise une action" est commune à tous les systèmes de montage automatique.

Concernant les actions, nous pouvons relever un certain nombre de connaissances valables

pour toutes les AOI.

- Toutes les actions décrites dans les différentes expériences de l'état de l'art ont un type, ou un nom d'action, permettant de les identifier (un lancer-franc, une prise de parole, un changement de diapositive, ...) ;
- Toutes les actions possèdent un certain nombre de caractéristiques qui leur sont propres. Ces caractéristiques permettent d'identifier ces actions. Il peut s'agir d'informations visuelles, textuelles, auditives, ... ;
- Toute action a une existence temporelle : une action a une date de début et une date de fin. Cette information temporelle associée aux caractéristiques et au type d'une action permet d'identifier de manière unique une action au cours d'un évènement.
- Toute action est réalisée par une ou plusieurs personnes.

Des connaissances communes à chaque POI sont également identifiables.

- Tout individu est identifié de manière unique par un identifiant. Cet identifiant peut être généré à l'apparition d'un individu dans la scène, ou bien connu à l'avance comme un numéro de dossard dans une équipe ou un numéro d'identifiant dans une base de données ;
- Une Personne d'intérêt possède également un nom ;
- Toute personne a un rôle dans la scène, comme par exemple : joueur, enseignant, acteur, guitariste.

Enfin, une personne d'intérêt peut être un individu seul, ou un groupe d'individus. Un groupe est composé d'un ensemble d'individus ayant un ou plusieurs points communs. Il peut s'agir d'une équipe sportive, un groupe politique, l'ensemble des étudiants, etc. Un groupe est alors défini par un nom de groupe et par un ensemble de caractéristiques qui lui sont propres (couleur de maillot, emplacement dans la scène, ...).

La figure 2.10, présente la modélisation ORM obtenue à partir des connaissances de l'état de l'art. Le modèle présenté permet d'aider à la mise en place d'une méthode de montage automatique. L'instanciation de ce modèle permet de mettre en avant les différentes sources d'intérêts présentes dans la scène et ainsi sélectionner la ou les sources d'intérêts les plus pertinentes pour la réalisation d'un montage automatique.

Afin de faciliter le choix des caractéristiques et des méthodes à mettre en place afin d'extraire les informations de la scène, nous proposons de séparer les attributs des objets d'intérêts en différentes classes. Au sens de la systémique, un objet d'intérêt peut être représenté dans un référentiel triadique [83]. Un objet peut posséder des attributs de temps, d'espace et/ou de formes. Une personne d'intérêt possède par exemple des attributs d'espace (position dans la scène), de forme (identité, couleur, numéro, ...) et de temps (présence/absence dans la scène). Il en va de même pour les actions d'intérêts. Par exemple, le déplacement d'un objet est la transformation de ses attributs en terme d'espace et de temps.

La figure 2.11 présente une modélisation générique, dont les caractéristiques des objets d'intérêts sont décomposées en fonction de la nature de leurs attributs.
Exprimer les caractéristiques d'un objet dans ces trois catégories aide au choix et à la mise en place des méthodes d'extraction des caractéristiques de la scène [11]. Les attributs des POI étant essentiellement de forme et d'espace, des méthodes d'identification et de détection devront être mises en places. Ceux des AOI concernant essentiellement le temps et l'espace, les méthodes de détection et de suivi seront choisies pour l'extraction des caractéristiques.

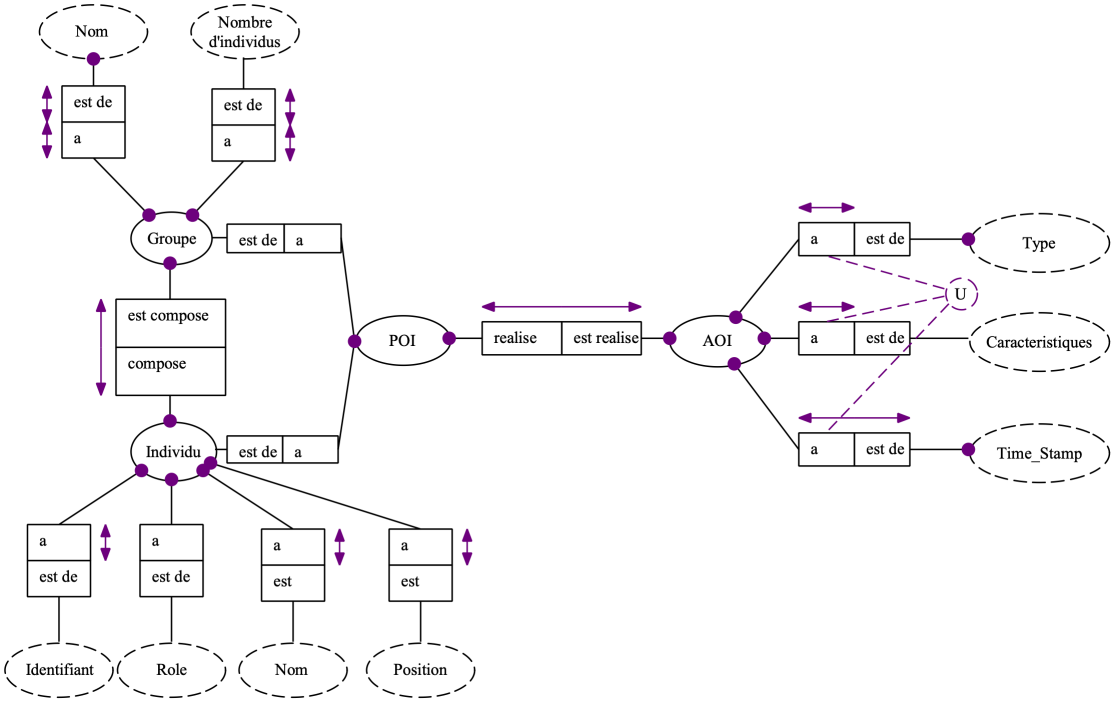


FIGURE 2.10 – Modélisation générique des sources d'intérêt pour la réalisation d'un montage automatique

2.4 Mise en place d'un système de montage automatique

Lors de la mise en place d'un système de montage automatique, il est nécessaire de sélectionner la source d'intérêt adaptée au type de diffusion. Nous proposons d'identifier deux catégories de sources d'intérêt. Les sources d'intérêts principales, c'est-à-dire celles répondant à l'attente du plus grand nombre de personne, dont la détection est nécessaire pour une diffusion en direct. La seconde catégorie consiste en toutes les autres sources d'intérêt, ainsi qu'en leurs caractéristiques, dont l'extraction n'est pas possible en temps réel. Les connaissances sur ces sources d'intérêt permettent de proposer de nouveaux montages.

Cette section précise également en quoi le mode de diffusion influe sur les méthodes à mettre en place pour extraire les sources d'intérêts.

2.4.1 Cas d'une diffusion en direct

Diffuser un évènement en direct permet à n'importe quel spectateur géographiquement distant de pouvoir suivre en évènement, en même temps qu'il se déroule. La diffusion de vidéos nécessitant une grande bande passante, il est souvent d'usage de ne diffuser qu'un seul flux, afin de garantir la qualité de celui-ci. Il est alors nécessaire de réaliser un montage automatique pour permettre la diffusion d'un évènement.

Étant donné que l'on s'adresse à un nombre important de spectateurs, il est essentiel que la source d'intérêt (SOI) sélectionnée, soit celle que le plus grand nombre veut voir. De ce fait, le choix de cette SOI doit être fait de la manière la plus neutre que possible. Elle peut être l'enseignant dans un contexte d'enseignement, le locuteur pour une réunion ou un ensemble d'acteur dans une pièce de théâtre. La source d'intérêt sélectionnée doit être la personne ou

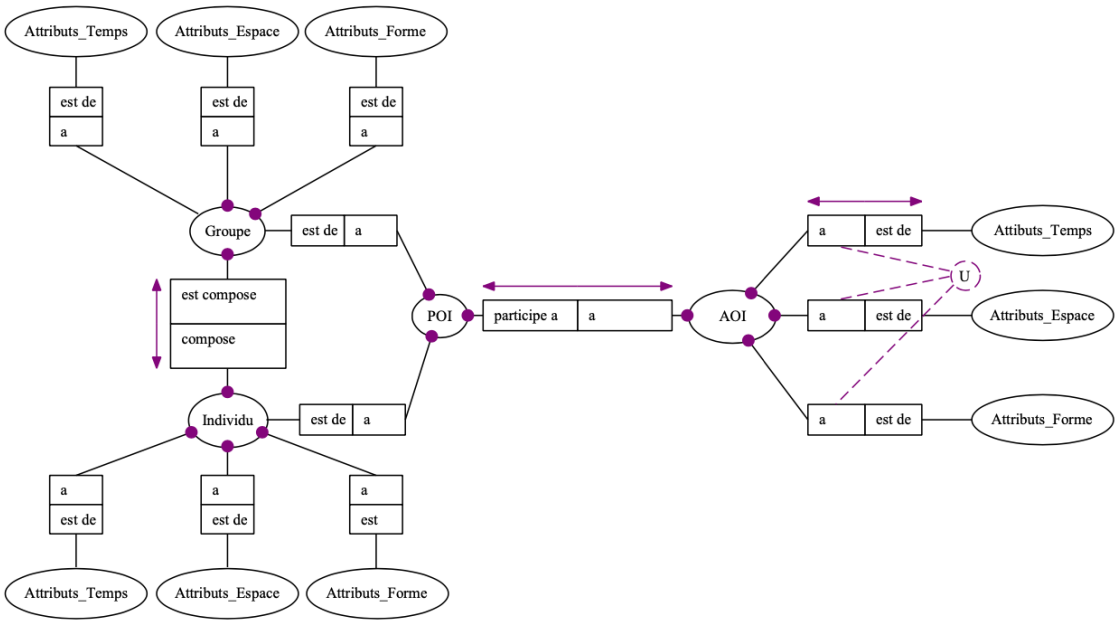


FIGURE 2.11 – Modélisation systémique des sources d'intérêt pour la réalisation d'un montage automatique

l'action que les spectateurs regarderaient s'ils étaient sur place.

Un autre point important lors d'une diffusion live est que les calculs effectués lors de l'étape de planification doivent être réalisés en un temps inférieur ou égale au temps réel vidéo. Le temps de calcul par image doit être inférieur à 40 millisecondes pour une vidéo à 25 images par seconde ou 16 millisecondes pour une vidéo à 60 images par seconde. Il est donc nécessaire de choisir des méthodes respectant cet objectif de temps de calcul.

Ainsi, certaines méthodes fonctionnant en temps réel, se concentrent sur des aspects particuliers du montage automatique. L'utilisation d'une seule caméra permet de se focaliser sur le contrôle de caméra et non sur la sélection du meilleur flux vidéo [17, 64, 46, 145] . D'autres méthodes utilisent des informations de bas niveaux et des règles basiques pour la sélection de caméras [85, 40, 119].

Lors d'une diffusion en direct, deux types d'informations sont générés. Tout d'abord le flux vidéo monté qui permet aux spectateurs distants de visualiser un évènement en direct. Mais également les méta-données, c'est-à-dire toutes les informations extraites ou fournies par le diffuseur, permettant la génération du montage. Ces informations peuvent présenter un intérêt pour une diffusion ultérieure et notamment pour les diffusions en différé. Il est donc important de stocker les informations extraites pendant cette première diffusion, afin d'éviter de devoir les extraire de nouveau.

2.4.2 Cas d'une diffusion en différé et personnalisée

La personnalisation d'un montage consiste en la sélection des flux vidéos en fonction des desiderata des spectateurs. Afin de permettre aux utilisateurs de voir un évènement de la façon qu'ils souhaitent, il est nécessaire d'augmenter le nombre de sources d'intérêts à extraire lors de planification, comme nous pouvons le voir dans le tableau 2.2. En effet, les spectateurs peuvent avoir différentes envies lorsqu'ils regardent un évènement. Les sources d'intérêts peuvent être

différentes pour chaque spectateur. Augmenter le nombre de sources d'intérêts à détecter permet de pouvoir proposer un montage répondant aux exigences des différents spectateurs.

TABLE 2.2 – Les sources d'intérêt pour la personnalisation des systèmes de montage automatique de la littérature

Référence	Personne d'intérêt	Action d'intérêt
[95]	L'enseignant, un tableau	Actions de l'enseignant
[80]	Une personne, une diapositive	Actions des personnes
[4]	La balle et les joueurs,	Différents évènements
[46]	La balle, un/des joueur(s)	Mouvements
[20]	La balle, des joueurs	
[10]	Des personnes	Parler d'un certain sujet
[79]	Un visiteur	Se déplacer dans un parc
[94]	Une personne, un objet, ...	

La modélisation des connaissances sur un événement met en avant un grand nombre de sources d'intérêts potentielles. Certaines sont utilisées pour la réalisation d'une diffusion en direct, les autres peuvent être utilisées pour la personnalisation des flux vidéos. Ces sources d'intérêt peuvent être des actions, des personnes ou des attributs d'action ou de personne, dont la détection en temps-réel est difficile. Dans le cas d'une diffusion différé, le temps de calcul ne représente pas une contrainte. En effet, la diffusion ne devant pas être en simultané avec l'enregistrement, il est possible de calculer un grand nombre de ces caractéristiques après l'événement.

De plus, la diffusion en différé permet d'avoir une segmentation temporelle de l'évènement. Un segment temporel peut être un point de l'ordre du jour dans une réunion, une mi-temps dans une rencontre sportive, une chanson lors d'un concert ou encore une scène dans une pièce de théâtre. La sélection d'un segment temporel par le spectateur permet d'obtenir un montage uniquement sur la partie qui l'intéresse. Il est de même possible de se focaliser uniquement sur la réalisation d'un type d'action d'intérêt. Par exemple, si un spectateur veut voir tous les buts marqués dans un match de football, on ne peut les proposer qu'à la fin de la rencontre. Le flux vidéo produit est alors un résumé chronologique des buts marqués durant le match. Dans le cas où plusieurs caméras filment ces buts, alors le choix de la caméra est réalisé uniquement sur la durée des buts et non sur le match en entier. Il est également envisageable de proposer une personnalisation avancée en proposant aux utilisateurs de sélectionner une action, ainsi que les personnes d'intérêt l'ayant faite. Ainsi, le spectateur peut visionner uniquement les buts marqués par son équipe préférée, ou par son joueur préféré.

Un certain nombre de connaissances peuvent également être connues en dehors de l'événement et notamment les caractéristiques d'une POI. Une personne par exemple a un nom, un prénom, un sexe, un rôle dans l'événement, une position dans la scène. Elle peut appartenir à un groupe qui lui-même a des caractéristiques. Ces informations peuvent être connues avant un évènement et peuvent être stockées dans une base de données. Ainsi, l'identification d'une personne dans la scène permet d'obtenir facilement un grand nombre d'informations, offrant alors un plus grand choix de personnalisation.

L'extraction d'information permet alors la personnalisation de contenu. En laissant à l'utilisateur l'accès aux expressions sémantiques des sources d'intérêt acquises avant, pendant ou après l'événement rend la visualisation d'un événement extrêmement personnalisable.

2.5 Discussions

La mise en place d'un système de sélection automatique de caméras nécessite une connaissance parfaite de l'évènement que l'on souhaite diffuser. Dans ce chapitre, nous avons proposé une méthodologie de conception de système de montage automatique se basant sur le contexte. L'analyse des systèmes de la littérature nous permet de proposer une architecture générique, instanciable à chaque situation. L'apport tient aussi dans la modélisation et la synthèse des informations nécessaires au paramétrage du système de montage automatique. Ainsi, l'adaptation d'un système au contexte d'application est guidée par l'acquisition et l'exploitation des connaissances sur un évènement. L'identification des sources d'intérêts à extraire permet de guider la mise en place des méthodes aux différents niveaux d'un système de montage automatique.

Pour résumer, la méthodologie proposée présente de nombreux avantages.

— L'indépendance au contexte d'application permet de pouvoir adapter un système de montage automatique à chaque type d'évènement ;
— L'utilisation du formalisme NIAM/ORM permet d'identifier facilement les points d'intérêt d'un système multi-caméras, notamment grâce à l'utilisation des deux concepts : "Personne d'intérêt" et "Action d'intérêt" ;
— La connaissance des points d'intérêts aide à la mise en place des méthodes utilisées à l'étape de la planification. Une connaissance approfondie de la scène permet de sélectionner la méthode la plus appropriée pour extraire les caractéristiques nécessaires à la sélection de la caméra.
— La modélisation d'un événement permet de réfléchir aux différents points d'intérêt pouvant exister dans un événement. Cela présente l'avantage de pouvoir réfléchir aux différents éléments qu'un spectateur aimerait voir et ainsi de proposer une personnalisation des flux vidéo.

Nous proposons, dans les sections suivantes, l'application de notre méthodologie à deux types d'évènements différents. Nous nous intéressons à la mise en place d'un système de montage automatique dans le contexte d'application des conseils municipaux dans le chapitre 3. le chapitre 4 vise l'application au contexte de la diffusion de rencontre de basketball. Pour chaque cas d'application, une étude des sources d'intérêts est présentée, permettant de déterminer les méthodes de traitements d'image à mettre en place pour une diffusion en direct et en différé.

Chapitre 3

Montage automatique pour la diffusion de conseils municipaux

Sommaire

3.1 Introduction

De nombreuses réunions ont lieu tous les jours. Une réunion peut être définie comme étant un rassemblement de personnes dont l'objectif est de débattre sur un ou plusieurs sujets. Dans le cas où une réunion doit être visible par un certain nombre de personnes distantes, il est nécessaire de diffuser l'évènement afin que les personnes concernées puissent obtenir les informations échangées durant le débat.

Dans ce chapitre, nous nous intéresserons au cas spécifique des conseils municipaux. Ce sont des assemblées de personnes, élues, se réunissant régulièrement afin de délibérer sur les affaires d'une commune. Ces événements, portant sur la vie d'une ville, intéressent les concitoyens souhaitant être informés de l'avenir de la commune. Il est alors important que ces conseils puissent être vus par le plus grand monde.

De nombreuses communes françaises ont fait appel à des sociétés de production, afin de diffuser des conseils en direct ou en différé, permettant ainsi au citoyen ne pouvant assister aux conseils de pouvoir s'informer. Cependant, la majorité des systèmes utilisés rencontrent divers problèmes. Le fait que ces systèmes soient opérés par l'humain implique que ces solutions sont relativement coûteuses. L'équipe de production induit un coût pouvant rendre ces systèmes inaccessibles pour les petites communes. De plus, le spectateur est tributaire des choix du réalisateur. Cela peut aller à l'encontre du principe de transparence que souhaitent les communes.

Certaines communes ont fait le choix, afin de palier ces problèmes, d'utiliser un seul plan large du conseil, dans le but de réduire les coûts de la captation et d'en permettre la neutralité. Cependant, le grand nombre d'élus présents dans un conseil ne permet pas à l'utilisateur de pouvoir porter son attention sur une personne particulière. Ce choix de réalisation est propice aux occultations et offre une visibilité réduite de chaque individu. Enfin, leur visualisation n'est pas toujours agréable pour le spectateur, qui ne regarde alors qu'une partie de ce conseil.

La solution de CitizenCam est de proposer une captation multi vues des conseils municipaux en installant plusieurs caméras filmant l'intégralité des élus présents. Le citoyen peut alors, après le conseil municipal, choisir la caméra d'intérêt qu'il souhaite. Il peut s'agir de la caméra où une personne parle, ou encore de se focaliser sur un élu particulier, pour analyser ses réactions, ses interventions sur un sujet donné.

Cette solution, répond aux problèmes des situations actuelles. En automatisant la captation, son coût diminue. En utilisant plusieurs caméras, il est alors possible de voir précisément chaque élu. Enfin, le spectateur peut choisir le flux vidéo qu'il veut voir et non un flux choisi par un monteur. En revanche, le choix de la caméra peut rendre la visualisation de l'événement difficile pour le spectateur. Si celui-ci souhaite voir les personnes qui prennent la parole, il doit chercher dans chaque flux vidéo, où se trouve la personne qui s'exprime, à chaque changement. Ces actions rendent la visualisation peut confortable pour l'utilisateur final. De plus, le grand nombre de flux vidéo proposé ne permet pas une diffusion en live à moins qu'un monteur soit présent, rendant la diffusion onéreuse.

Ces constats montrent qu'il est nécessaire de mettre en place une solution de montage automatique dans le contexte des conseils municipaux. Pour ce faire, nous commençons par mettre en pratique notre méthodologie afin de déterminer les sources d'intérêts dans ce contexte. La connaissance de l'action d'intérêt principale, nécessaire pour la diffusion en direct, nous conduit à proposer une nouvelle méthode de détection de locuteur. Afin de proposer la personnalisation des flux vidéos, l'identification des participants au conseil municipal s'avère nécessaire. Nous présentons ainsi une nouvelle méthode d'identification des locuteurs basée sur l'utilisation de la technologie Visible Light Communication.

3.2 Modélisation du contexte d'un conseil municipal

La première étape de notre méthodologie est l'acquisition des connaissances relatives à l'évènement, dans notre cas, les conseils municipaux. En France, les conseils municipaux sont définis dans le Code général des collectivités territoriales (CGCT). Il s'agit donc d'une source intéressante d'informations sur leur réglementation. Cependant, ces textes ne traitent pas du déroulement des séances. L'expertise acquise par CitizenCam lors de captations de conseils municipaux permet de pallier ce manque de connaissances.

Nous cherchons, dans un premier temps, à acquérir des connaissances sur la composition du conseil municipal pour identifier les personnes d'intérêt potentielles ainsi que leurs attributs. Puis, nous nous intéressons à l'étude des actions d'intérêts qui peuvent avoir lieu lors d'un conseil municipal.

3.2.1 Définition des personnes d'intérêts (POI)

Un conseil municipal est une assemblée de personnes délibérantes sur les affaires d'une commune. Concernant sa composition, l'article L2121-1 du CGCT dit que "Le corps municipal de chaque commune se compose du conseil municipal, du maire et d'un ou plusieurs adjoints". Les personnes participant à un conseil municipal ont donc un rôle qui peut être "maire", "adjoints" ou "conseiller municipal". Ces personnes sont élues et le Code électoral nous apprend que lors de l'élection, "les candidats peuvent se présenter de façon isolée ou groupée." (article L255-3).

Une POI peut donc être un élu seul. Un élu est défini par un certain nombre d'attributs, à savoir : un nom et un prénom, un genre (Forme), souvent une place attitrée (Espace) (cf figure 3.1). La connaissance de la place d'un conseiller permet alors une identification indirecte de celui-ci. De plus, les élus peuvent être présent ou non (Temps). Enfin, les élus appartiennent de manière générale à différents groupes.

FIGURE 3.1 – conseil municipal de la ville de Palaiseau : les élus ont une place attitrée, identifiée par leurs noms ou leurs rôles.

Au sein d'un conseil municipal, des groupes de différents types existent. Tout d'abord, un élu peut appartenir à un groupe politique, un ensemble de personnes, uni autour d'une philosophie ou une idéologie commune, identifié par un nom de groupe. D'autre part, il est fréquent de trouver dans les conseils municipaux la présence d'au moins deux groupes d'élus : la majorité

et l'opposition. Le groupe de la majorité comprend les élus faisant partie du groupe politique ayant obtenu la majorité des scrutins lors de l'élection. Il s'agit de manière générale du groupe auquel appartient le maire. Le groupe nommé opposition est composé des autres élus du conseil municipal. Un groupe d'élu est composé d'au moins deux élus, à moins que le règlement intérieur de la ville ne retienne un chiffre supérieur. Les élus appartenant à un même groupe sont souvent regroupés lors d'un conseil municipal. Les différents groupes possèdent donc plusieurs attributs : un nom, un type, une place et un nombre d'élus le composant.

Pour résumer, lors d'un conseil municipal, un certain nombre d'élus sont présents. Un élu peut être identifié par son nom, localisé par sa place dans le cas où elles sont définies. Les élus peuvent avoir différents rôles qui sont maires, adjoint au maire ou conseiller. Enfin, les élus possèdent un et un seul genre. D'autre part, les élus appartiennent à des groupes qui peuvent être de deux sortes : les groupes politiques et les groupes d'élus. Ces groupes, identifiés par un nom de groupe, sont composés d'un certain nombre d'élus ayant des places proches.

3.2.2 Définition de l'action d'intérêt (AOI)

Avant la tenue d'une réunion, le maire d'une commune convoque les différents élus en leurs communicant la liste des "questions portées à l'ordre du jour" (article L2121-10 du CGCT). Lors du conseil municipal, ces différents points sont débattus. Les actions "débattre d'un point de l'ordre du jour" permettent de segmenter temporellement les captations de conseils municipaux.

Pour chaque point de l'ordre du jour, les élus peuvent exprimer leur opinion. Il est donc intéressant pour le spectateur de pouvoir suivre un élu en train de s'exprimer. L'action "prendre la parole" apparaît donc comme l'action d'intérêt principale lors des conseils municipaux. Le maire dirige le débat et donne la parole aux différents intervenants tout au long du conseil municipal. Il veille à ce que les personnes souhaitant s'exprimer le fasse les uns après les autres.

Il est fréquent qu'un vote sur les décisions à prendre conclut les débats sur un point de l'ordre du jour. Ces votes peuvent être réalisés selon trois modes de scrutin :

— le scrutin ordinaire, à main levée
— le scrutin public, par bulletin écrit ou appel nominal
— le scrutin secret.

La détection de l'action d'intérêt "voter" permettrait d'offrir aux spectateurs le montage automatique d'une diffusion personnalisée en visualisant par exemple un résumé des décisions prises lors d'un conseil municipal.

Pour résumer, peu d'actions différentes ont lieu lors d'un conseil municipal. Un conseil est constitué de différents points de l'ordre du jour, pendant lesquels les élus prennent la parole et, dans certains cas, réalisent un vote. Ces actions ont donc des attributs de temps, caractérisés par une date de début et de fin. Ces actions comportent des noms et un certain nombre de caractéristiques les définissant (attributs de forme). Les actions d'intérêts dans le cadre des conseils municipaux ne présentent pas d'attributs d'espace.

3.2.3 Prise en compte du contexte

Les connaissances extraites du code général des collectivités territoriales, ainsi que nos connaissances sur le déroulement des conseils municipaux, nous ont permis d'instancier le modèle NIAM/ORM, comme nous pouvons le voir dans la figure 3.2.

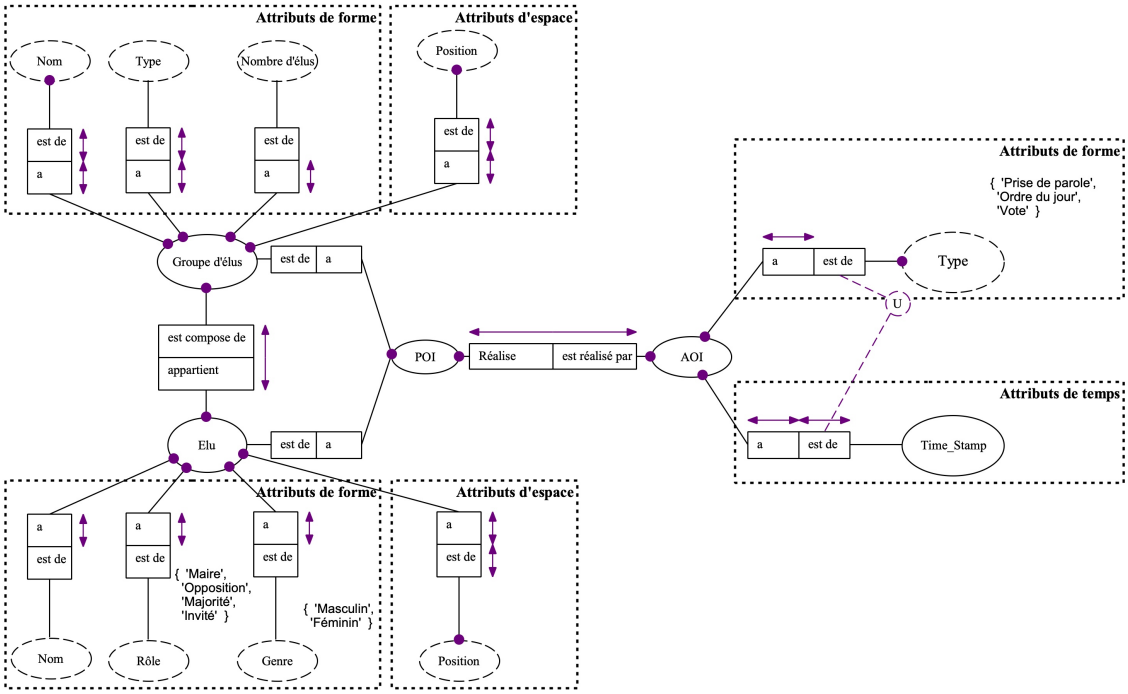


FIGURE 3.2 – Modélisation NIAM/ORM d'un conseil municipal

Un conseil municipal est une réunion d'élus débâtant sur différents points de la vie d'une commune. La diffusion des conseils municipaux permet aux citoyens d'être informés facilement des décisions prises.

Cette information peut être obtenue en temps réel, pendant que le conseil municipal a lieu. Il est alors nécessaire que le montage proposé soit celui souhaité par le plus grand nombre. Lors d'une réunion, l'intérêt est porté sur la personne en train de s'exprimer. L'action d'intérêt lors d'un conseil municipal est donc la prise de parole d'un élu. Il est alors nécessaire de mettre en place une méthode de détection des locuteurs, fonctionnant en temps réel, afin de sélectionner automatiquement la caméra où une personne s'exprime.

Dans le cas où le conseil municipal est diffusé en différé, il est nécessaire d'offrir une possibilité de personnalisation du montage à l'utilisateur final afin de garantir la transparence demandée par les communes.

Le découpage du conseil municipal en points de l'ordre du jour permet d'en segmenter temporellement le déroulement. Les spectateurs ont alors la possibilité de visualiser les périodes du conseil municipal qui les intéresse. Il est ainsi possible de proposer des résumés répondant aux attentes de chaque spectateur.

D'autre part, un spectateur peut s'intéresser à une personne ou à un groupe de personnes. De nombreuses informations sont attachées à chaque locuteur. L'extraction des informations sur ces personnes rend possible une personnalisation avancée.

Les informations sur les locuteurs (noms, fonctions au sein du conseil municipal, genre, etc) rendent possible la génération d'un flux ciblant un élu, ou un groupe d'élus partageant un même attribut, particulier. Un spectateur peut, par exemple, avoir envie de voir les prises de parole d'un adjoint au maire délégué à la jeunesse et aux sports. Il peut également avoir envie de regarder

les prises de parole d'un élu particulier sur un point précis de l'ordre du jour.

La notion de groupe est également importante pour la personnalisation. Un électeur peut s'intéresser aux membres d'un groupe (politique ou non). Les informations sur les groupes, ainsi que les informations sur les élus sont, sauf changement durant le conseil, connues à l'avance. Ainsi, il est possible, à partir de l'identité d'une personne, d'obtenir un grand nombre d'informations. L'identification des différents élus lors d'un conseil municipal permet alors d'enrichir le montage, et de pouvoir proposer automatiquement à l'utilisateur une sélection de flux vidéo répondant à ses attentes.

L'intégration des connaissances dans le cadre des conseils municipaux nous a permis d'identifier différentes sources d'intérêts. La source d'intérêt principale est l'action d'intérêt "prise de parole" qui est rattaché à la personne d'intérêt "élu". La détection et la localisation de l'AOI "prise de parole" rend possible la diffusion d'un conseil municipal en direct. D'autre part, l'identification des locuteurs permet d'obtenir les caractéristiques connus des élus (nom, rôle, appartenance à un groupe,...) et ainsi rend possible la personnalisation des flux vidéo.

3.3 Détection de l'AOI "prise de parole"

Afin de proposer une diffusion en direct d'un conseil municipal, il est nécessaire de pouvoir détecter l'action d'intérêt que la majorité des personnes veut voir. Cette source d'intérêt est, dans le cas de réunion et plus particulièrement des conseils municipaux, la prise de parole.

Nous étudions dans une première partie les méthodes de la littérature permettant la détection de locuteur dans une scène. Les limites des méthodes existantes par rapport à notre contexte nous mènent à proposer une nouvelle méthode, basée sur la détection de microphone actif. Enfin, nous validons notre méthode sur des séquences vidéos réelles, extraites de conseils municipaux de la ville de Villers-Lès-Nancy

3.3.1 État de l'art des méthodes de détection de locuteur

La détection de locuteurs a été abordée dans diverses situations. L'application la plus récurrente est celle des systèmes de vidéo conférence. Deux groupes de personnes, géographiquement distants, discutent grâce à une liaison audio et vidéo. Une des problématiques est de localiser la personne en train de parler, afin de focaliser l'attention de la vidéo sur cette dernière. Dans le contexte des montages automatiques, l'application la plus fréquente est la localisation d'un conférencier pour une retransmission de conférence. La localisation du locuteur dans la scène permet de pouvoir en proposer le meilleur cadrage possible. D'autres applications, comme la discussion entre deux personnes ou la séparation des locuteurs pour l'indexation ont également été étudiées.

On trouve dans la littérature trois approches pour la détection de locuteur. La première utilise les signaux audio provenant de microphones pour détecter la prise de parole. La seconde approche est d'utiliser les informations vidéo afin de localiser le locuteur. Enfin, la dernière couple l'analyse des signaux audio et vidéo pour résoudre le problème.

Méthodes de détection de locuteurs basées sur l'audio

L'objectif des méthodes basées sur les microphones est d'étudier les signaux provenant des microphones, afin de déterminer la position d'une source de bruit, et plus particulièrement des différents locuteurs dans la scène.

La majorité des méthodes se base sur la différence des temps d'arrivée (time difference of arrival - TDOA) [5] en analysant les signaux captés par différents microphones disposés dans la pièce, une fois synchronisés.

Rui et al. [119] utilisent deux microphones afin de pouvoir localiser un étudiant posant une question dans une salle de classe. La corrélation croisée entre les deux signaux émis par les microphones permet d'estimer le délai temporel entre les deux captations et ainsi localiser le locuteur dans la pièce.

L'utilisation de deux microphones ne permet qu'une localisation approximative. En effet, seule la direction de la source sonore est obtenue. Aucune information sur la distance entre le locuteur et les microphones ne peut être calculée. L'augmentation du nombre de microphones permet de pouvoir localiser plus précisément le locuteur [38].

Ainsi, des ensembles de microphones, pouvant être installés en ligne [142, 89, 88] ou en cercle [100], autorisent une localisation spatiale précise des locuteurs.

Les méthodes basées sur l'utilisation de l'audio sont à ce jour celles offrant les meilleures performances en terme d'efficacité. Toutefois, ces méthodes impliquent que le signal audio de chaque microphone soit exploitable et que les différents signaux soient parfaitement synchronisés. De plus, ces méthodes sont sensibles aux bruits dans la scène. C'est pourquoi de nombreuses méthodes se basant sur des microphones utilisent un système de vision pour lever certaines ambiguïtés.

Méthodes de détection de locuteurs basées sur la vidéo

De nombreuses études ont été menées afin d'utiliser uniquement les pistes vidéos pour la localisation. Du fait, des applications (vidéo conférences, enregistrement de cours magistraux, ...) des caméras sont déjà utilisées pour la retransmission. Ainsi, l'utilisation de la vidéo seule permet de réduire les coûts.

De nombreux travaux existent sur l'analyse du mouvement des lèvres. Ces travaux peuvent avoir comme objectif la reconnaissance vocale [28], la détection de l'activité vocale [127, 123, 65] ou la lecture sur les lèvres [130]. Ces méthodes s'appuient généralement sur la détection de visage et l'utilisation des contours actifs, afin de modéliser les lèvres. Il est alors nécessaire que les images analysées soient suffisamment proches des visages.

Certains travaux se sont intéressés à la détection de locuteurs à partir de vues larges de la scène. La direction des regards des participants peut permettre d'identifier où est le locuteur [48]. En effet, lors d'une réunion, les différents participants ont de manière générale, le visage orienté vers la personne qui parle.

La prise de parole peut également être détectée en étudiant les mouvements des différents protagonistes. Quek et al. [114] ont étudié les relations entre le mouvement des personnes et la prise de parole. Leur étude révèle qu'il existe une corrélation entre les deux. Ces travaux ont été repris dans [135] où deux personnes discutent face à face. L'analyse des mouvements permet de détecter efficacement le locuteur. L'avantage de cette méthode est qu'il n'est pas nécessaire d'avoir une vue de face des protagonistes, contrairement aux méthodes se basant sur l'analyse du mouvement des lèvres. Cependant, le nombre de personnes présentes dans la scène doit être relativement faible.

Les méthodes basées sur l'analyse visuelle des locuteurs nécessitent d'avoir une vue rapprochée, de face des potentiels locuteurs. De plus, ces méthodes sont sensibles à la luminosité, à l'orientation des visages, aux occultations, ... Ainsi, ces méthodes sont peu efficaces dans les scénarios où plusieurs personnes sont présentes. De plus, ces méthodes ne permettent pas de différencier les locuteurs principaux, de deux personnes discutant entre elles, en aparté.

Méthodes de détection de locuteurs basées sur le couplage de l'audio et la vidéo

Du fait de la présence de caméras et de microphones pour capter une scène, il est normal que des méthodes de détection de locuteurs se basent sur les informations audiovisuelles.

Dans [29], les auteurs proposent une méthode résolvant une partie des problèmes de la section 3.3.1, en étudiant la corrélation entre le son et la vidéo. L'algorithme proposé fonctionne uniquement sur les scènes où un seul locuteur se trouve devant la caméra, et en plan rapproché.

Dans [102], une série de microphones est ensuite utilisée afin de localiser le locuteur. La source du son est estimée en utilisant les différences de phase et d'intensité des différents microphones, ainsi que la théorie de Dempster-Shafer [125]. Afin d'améliorer la localisation, des caméras sont utilisées afin de détecter des protagonistes dans l'image (détection de torse). L'association de ces deux types de capteurs améliore la robustesse du système en levant l'ambiguïté en cas d'occlusions.

AV16.3 [82] est une plateforme de recherche constituée de 2 disques de 8 microphones ainsi que de 3 caméras. Les caméras sont utilisées pour localiser les éventuels locuteurs dans la pièce. Les microphones permettent de localiser le ou les locuteurs dans la pièce, en s'appuyant sur les données extraites des vidéos.

Le projet Smart Room [15] s'intéresse à une application de vidéo conférence. 5 caméras sont utilisées pour localiser les locuteurs autour de la table. 16 microphones sont employés afin de localiser et d'identifier le locuteur actif. L'identification est réalisée en analysant les transformées de Fourrier à court terme des phrases enregistrées et en les comparant avec des modèles spécifiques à chaque locuteur.

D'Arca et al. [33] s'intéressent à la localisation d'une source de son en mouvement. La méthode proposée se base sur l'étude de la corrélation entre le mouvement dans la vidéo (vitesse et accélération des flux optique) et les informations audio (Mel-Frequency Cepstral Coefficient MFCC) permettant de suivre une source sonore en prenant en compte les éventuelles occlusions.

Les méthodes exploitant les informations visuelles et auditives sont celles présentant les meilleurs résultats. Les informations visuelles permettent de localiser les éventuelles sources de bruits et les informations auditives permettent d'identifier la position du locuteur. Ces méthodes nécessitent cependant un grand nombre de matériels rendant leur utilisation onéreuse.

Dans notre contexte, plusieurs caméras filment les différents locuteurs, et chaque participant possède un microphone devant lui. Cependant, les systèmes audio mis en place, ne permettent pas d'isoler individuellement les signaux audio de chaque microphone. En effet, une seule source sonore est disponible pour tous les microphones, impliquant que nous ne pouvons pas utiliser de méthodes basées sur l'utilisation des microphones. Le cadre utilisé par chaque caméra filmant entre 3 et 10 personnes, la qualité visuelle ne permet pas d'utiliser des méthodes basées sur l'extraction du contour des lèvres. Il serait, en outre, impossible de différencier le locuteur principal de deux personnes discutant entre elles. Les méthodes se basant sur les mouvements des différents protagonistes permettent de localiser les différents locuteurs dans des scènes avec peu de personne. Ces méthodes sont donc non adaptées au contexte à cause du nombre important d'élus lors des conseils municipaux.

En conclusion, les méthodes existantes ne donnent pas satisfaction ou ne sont pas adaptables à notre contexte. Nous proposons donc une nouvelle méthode de détection de locuteur répondant à nos besoins.

3.3.2 Détection visuelle de microphones actifs

La plupart des communes Françaises sont équipées de systèmes de microphone où une LED s'allume au niveau de la collerette du microphone comme illustré figure 3.3. L'allumage de cette LED signifie que le microphone devient actif et il est alors possible, pour le conseiller de pouvoir s'exprimer. La détection de la collerette des microphones, et notamment la détection de l'état du microphone, permet de détecter la prise de parole sans avoir besoin de modifier ou d'intervenir sur le système déjà existant.

FIGURE 3.3 – Présence de microphones disposant d'une lumière pour signaler la prise de parole dans différents conseils municipaux

D'une certaine manière, la détection de LED de microphone se rapproche des problématiques de détection de feux de circulations. Nous cherchons des informations lumineuses dans une vidéo.

[132] proposent une méthode de détection de feux de circulation en utilisant un seuillage des couleurs dans le domaine HSV. La recherche des centres des sources lumineuses en utilisant un masque Gaussien et en vérifiant les zones candidates avec une carte des poids d'existence (Existence Weight Map). Cette carte permet de mettre en évidence le fait que la position d'une lumière dans une trame est pratiquement identique dans la trame suivante. Pour ce faire, la partie haute de l'image est découpée en MxN blocs. Pour chacun de ces blocs, un poids de possibilité d'existence est associé. Si un feu est détecté dans une trame précédente à l'endroit m,n : $(m, n, t-1)$, il y a une forte probabilité qu'il apparaisse à l'endroit (m, n, t). Le bloc supérieur $(m, n+1, t)$ à la seconde plus forte probabilité. Enfin les blocs latéraux $(m \pm 1, n, t)$ et $(m \pm 1, n+1, t)$ ont la troisième plus forte probabilité.

Le système est associé au système de guidage de la voiture. La détection de feu est ainsi déclenchée 300 mètres avant un croisement. Le système permet de détecter le feu à 90 mètres, avec une justesse de 80%.

[27] proposent une méthode de détection de feu dans une scène fixe. Pour ce faire l'illu-

mination de la scène est évaluée à partir d'une image d'arrière-plan [63]. Cette évaluation de l'illumination est ensuite utilisée pour corriger les images, grâce à un opérateur flou. Une opération morphologique floue est conjointement utilisée pour éliminer les bruits et pour obtenir toutes les formes circulaires possibles. Les informations de teintes, d'intensités et de formes sont ensuite réunies afin d'extraire les zones lumineuses correspondantes aux feux. Les informations spatiales des feux tricolores et les informations temporelles sont également utilisées afin de confirmer la présence et couleur du feu de circulation détecté.

Description de l'algorithme

Ce paragraphe décrit le fonctionnement général de l'algorithme que nous avons développé en Python. La plupart des fonctions utilisées proviennent de la bibliothèque OpenCV. L'organigramme 3.4 représente les différentes étapes détaillées ensuite.
Afin de pouvoir utiliser l'algorithme dans toutes les situations, nous avons introduit une étape d'initialisation du système permettant de sélectionner manuellement les positions des microphones, les limites des zones de recherche, les seuillages HSV et ainsi entraîner un arbre de décision.

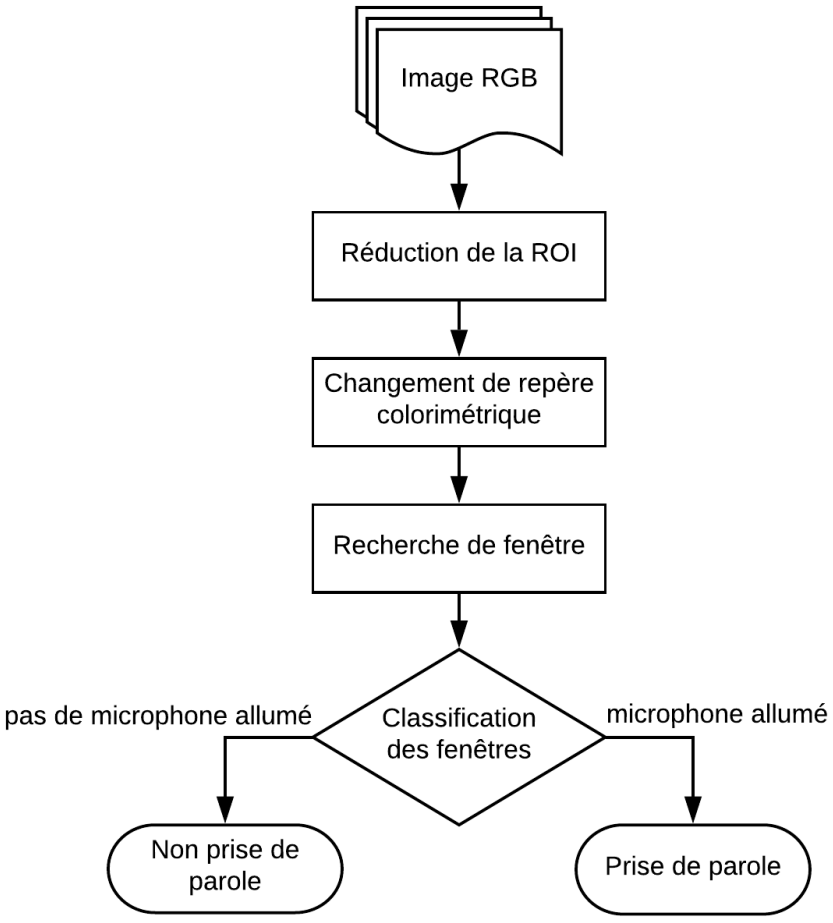


FIGURE 3.4 – Organigramme de l'algorithme proposé pour la détection de microphones actifs

Définition de la zone de recherche : Les microphones se trouvant sur la table, devant les éventuels locuteurs, il est donc opportun de focaliser la recherche sur une zone comprise entre la table et la tête des locuteurs. Analyser la présence des microphones dans une zone réduite permet de réduire le temps de calculs, ainsi que de diminuer le risque de fausses détections. Comme les microphones sont manipulés par les élus pendant la séance, la zone de recherche doit être assez large pour pallier ces déplacements. La figure 3.5 présente la zone de recherche définie pour la vidéo "vue1" (cf Table 3.2). Dans cette situation, les microphones des deux élus latéraux ont été éloignés durant le conseil municipal. Ils sont cependant replacés par les élus avant une prise de parole.
Cette zone est pour l'instant définie de manière manuelle, cependant cette étape peut être remplacée par la mise en place d'un système de détection de buste afin de cibler de manière automatique la zone de recherche et ainsi de chercher les microphones uniquement aux endroits où un locuteur potentiel se trouve.

FIGURE 3.5 – Exemple de zone de recherche (en bleu) pour la détection de microphones actifs

Seuillage des couleurs : Nous utilisons le domaine de couleur HSV pour analyser la luminance des microphones. Les microphones actifs émettant de la lumière, il est intéressant de rechercher les zones ayant une forte luminosité. Le repère HSV et notamment la composante V permet de mettre en évidence les zones de forte luminance dans l'image. Un filtrage sur les composantes S et V est ainsi effectué pour obtenir des zones candidates, comme le montre l'équation :

$$C_{x,y} = \begin{cases} 1 \quad si \quad ((S_{x,y} \geq S1 \cap S_{x,y} \leq S2) \cap (V_{x,y} \geq V1) \cap V_{x,y} \leq V2)) \\ 0 \qquad\qquad\qquad\qquad sinon \end{cases} \tag{3.1}$$

Où C correspond à l'image seuillée, $S_{x,y}$ est la valeur en saturation et $V_{x,y}$ est la valeur de l'intensité du pixel (x,y). Les quatre seuils (S1,S2,V1 et V2) sont définis de manière empirique.

Nous pouvons, dans la figure 3.6, voir le résultat du seuillage effectué.

FIGURE 3.6 – Seuillage des couleurs dans le domaine HSV

Analyse en composante connexe : Afin de pouvoir étudier les zones, il est nécessaire de labelliser chaque zone comprenant une forte luminosité. Pour ce faire nous réalisons une analyse en composante connexe afin de pouvoir analyser par la suite si une zone contient un microphone allumé.

Vérification de la présence d'un microphone : Il existe cependant dans les images que nous traitons un grand nombre de perturbations : lumière produite par un smartphone, reflets, etc. Il est alors nécessaire de discriminer ces émissions lumineuses de celles que nous recherchons. Pour ce faire, nous utilisons un classificateur basé sur un arbre de décision [115] et une technique de fenêtres glissantes pour séparer les zones qui contiennent un microphone actif des autres. Pour chaque image, une fenêtre de taille 19x19 (voir 3.3.2), est balayée dans les régions candidates. Les caractéristiques calculées (voir 3.3.2) dans chaque emplacement de fenêtre sont ensuite testées à l'aide du classificateur. L'arbre de décision est généré pendant l'étape d'initialisation de notre système. Ce type de classificateur a été choisi pour sa rapidité de calcul permettant de respecter la contrainte de temps réel.

Sélection de la taille de la fenêtre

La sélection de la taille de la fenêtre pour le calcul des caractéristiques est une étape importante pour la caractérisation de l'état du microphone. Afin de déterminer la meilleure taille à utiliser, trois fenêtres ont été testées : une petite fenêtre de 3x3 pixels, une fenêtre moyenne de 9x9 pixels et une grande de 19x19 pixels, comme le montre la fig. 3.7. Le rapport pixel/millimètre obtenus après calibration est de 0,9.

Afin de valider le choix de la taille de la fenêtre, une base de données d'images spécifique a été créée. Pour chaque taille de fenêtre, 500 images de microphones actifs et 500 images d'autres parties de la scène ont été extraites de chacune des quatre vidéos détaillées dans la table 3.2. Les arbres de décisions ont été entraînés à l'aide des caractéristiques présentées dans la section

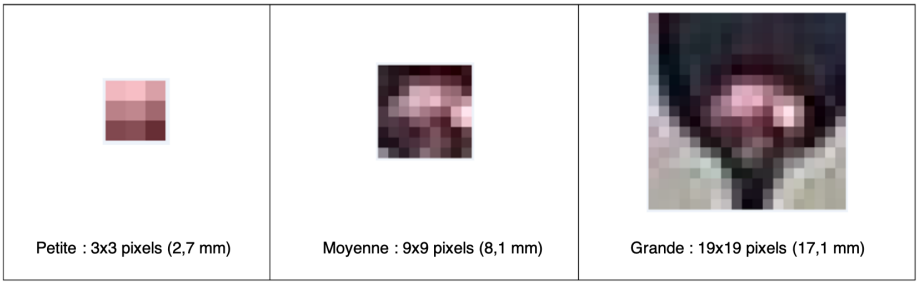


FIGURE 3.7 – Les tailles des trois fenêtres testées

3.3.2, extraites de ces 4000 images. Nous utilisons une validation croisée en k-fold[77] avec k=3 afin d'éviter le sur-apprentissage. Chaque classificateur formé a ensuite été testé sur chaque vidéo, présentée en détail dans la table 3.2. La table 3.1 résume les résultats obtenus en terme de justesse, c'est-à-dire la proportion des prédictions correctes effectuées par le classificateur.

TABLE 3.1 – Influence de la taille des fenêtres sur la classification

Taille :	3x3	9x9	19x19
Vue 1	100%	100%	100%
Vue 2	97.9%	99.6%	100%
Vue 4	88.6%	95.6%	98.2%
Vue 6	87.7%	96.8%	98.1%

L'utilisation d'une petite fenêtre ne permet pas de séparer efficacement les deux classes dans chaque situation. En effet, sa taille est limite pour représenter toute la structure du microphone, ce qui peut expliquer les résultats obtenus avec les images des vues 4 et 6. Les fenêtres moyennes et grandes permettent une meilleure séparation de ces classes. Néanmoins, les fenêtres de taille 19x19 présentent de meilleurs résultats dans le cas de vues de biais (vue 4 et 6) que des fenêtres de taille moyenne. C'est pourquoi nous utilisons des fenêtres de taille 19x19 pour le contrôle de présence du microphone.

TABLE 3.2 – Caractéristiques des vidéos utilisées

Nom	Nombre d'image	Zone de recherche	Surface (pixel)	Rapport pxl/mm	Propriétés
Vue 1	9 957	1800 x 340	345	0.71	Face, gros plan
Vue 2	22 459	830 x 230	190	0.23	Face, plan large
Vue 4	15 737	1000 x 300	120	0.21	Biais, plan large
Vue 6	32 508	1000 x 220	210	0.30	Biais, plan large

Sélection des caractéristiques

Une des principales problématiques de la détection de la LED des microphones est de trouver les mesures permettant de pouvoir la caractériser de manière efficace. Afin de trouver ces caractéristiques, nous avons décidé d'étudier un grand nombre de mesures statistiques dans des repères colorimétriques différents (RGB, HSV [69] et CIE L*a*b* [66]). Ces repères permettent

une distinction simple de la chrominance ((HS) et (AB)) d'une image avec la luminance : composante V (Valeur) du domaine HSV et la composante L (Luminance) du domaine LAB.

Nous avons donc pour chacune de ces composantes (3 composantes par repère) sélectionné des mesures statistiques afin d'en déterminer les plus pertinentes pour la séparation des microphones allumés du reste de l'image. Nous avons donc calculé les mesures suivantes pour les histogrammes de chaque composante.

- La moyenne de l'histogramme
- Le mode, la valeur la plus représentée dans l'histogramme
- La variance et l'écart type, permettant de caractériser l'éloignement par rapport à la moyenne
- Le moment d'ordre 3 (skewness), mesurant l'asymétrie de la distribution
- La racine troisième du skewness
- Le moment d'ordre 4 (kurtosis), mesurant l'aplatissement de la distribution
- La racine quatrième du kurtosis
- Le Khi-2 [81]

Afin de pouvoir déterminer les caractéristiques à utiliser pour la détection de microphone, des méthodes de sélection de paramètres sont utilisées. Ces méthodes nous permettent de réduire le nombre de caractéristiques à calculer pour ne garder que celles qui ont un fort pouvoir discriminant. Nous pouvons identifier 3 types de caractéristiques.

- Les caractéristiques complémentaires, c'est-à-dire celles qui combinées entre elles permettent une meilleure différenciation des classes.
- Les caractéristiques redondantes, celles qui apportent des informations identiques.
- Les caractéristiques antagonistes qui apportent des informations contradictoires quant à la séparation des classes.

La suppression des caractéristiques antagonistes permet d'obtenir des taux de reconnaissance plus élevés et la suppression des redondantes permet de diminuer les temps de calculs.

Nous avons extrait deux images par seconde de nos vidéos de test (2 images par seconde) et sélectionné des fenêtres de taille 19 x 19 pixels entourant les microphones, ainsi que des images provenant de zone de même taille ne comprenant pas de microphone actif.

Chacune de ces fenêtres réduites a ensuite été annotée "positives" ou "négatives". Disposant ainsi de jeux de données suffisant mais pas trop grand, nous avons pu utiliser les méthodes de type wrapper [78] suivantes :

- ReliefF [76].
- SFS : Sequential Feature Selection [139].
- SBS : Sequential Backward Selection [93].
- SFFS : Sequential Forward Floating Selection[113].
- SBFS : Sequential Backward Floating Selection[113].

Pour chaque jeu de données, nous avons, grâce aux algorithmes de sélection, extrait un sous-ensemble des 10 caractéristiques les plus discriminantes lors de la séparation des classes "microphones allumés" (fenêtre positive) et "rebut" (fenêtre négative). Les pourcentages d'apparition de chaque caractéristique sont présentés figure 3.8.

Nous avons remarqué que les caractéristiques dans le domaine HSV, notamment les informations sur la teinte (Hue), sont particulièrement importantes dans la caractérisation des microphones allumés. Nous avons sélectionné les 7 caractéristiques récurrentes pour chaque jeu de

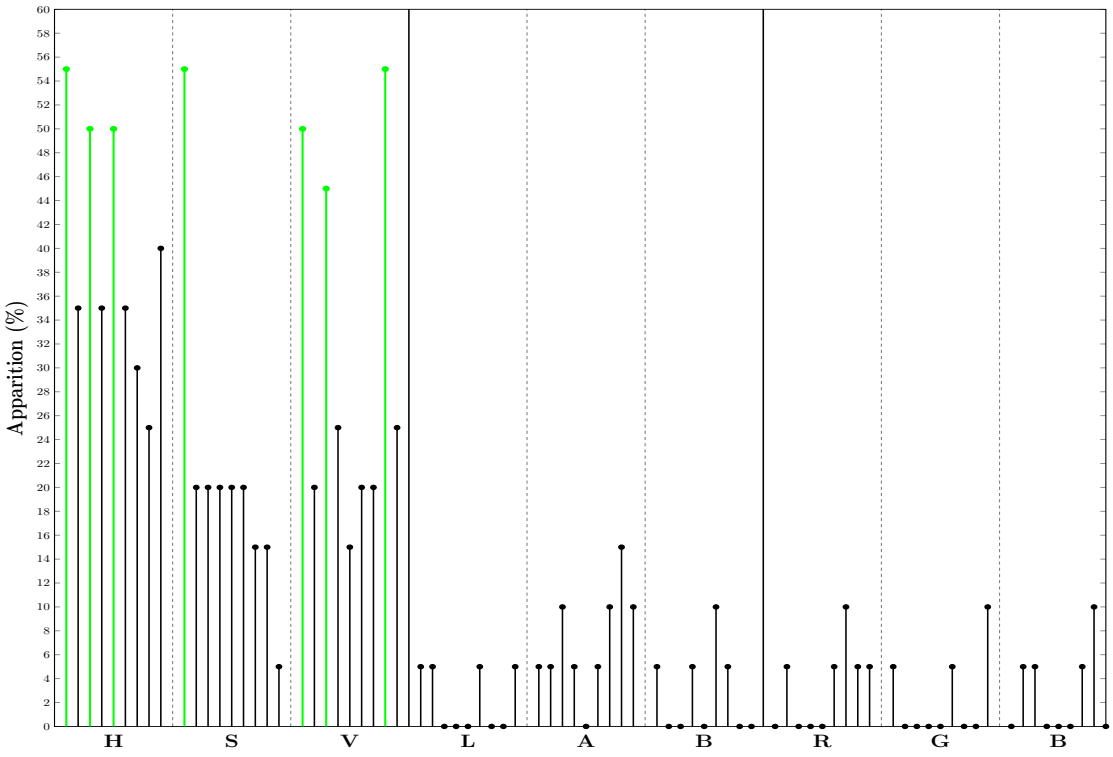


FIGURE 3.8 – Pourcentage d'apparition des caractéristique des 5 algorithmes de sélection de caractéristiques, appliqués aux 4 jeux de données. Les caractéristiques sélectionnées sont représentée en vert.

données : moyenne, variance et moment d'ordre 3 dans la composante H, moyenne dans la composante S, ainsi que la moyenne, la variance et la racine quatrième de Kurtosis de la composante V.

3.3.3 Résultats

L'évaluation de cette méthode a été effectuée sur un ordinateur disposant d'un processeur Intel Core I7-5557U cadencé à 3.1 Ghz et de 8 GO de mémoire RAM.

L'évaluation de la méthode a été réalisée à partir de vidéos d'un conseil municipal tourné à Villers-Les-Nancy en France. La table 3.2 présente les différentes vidéos utilisées. Les images originales ont une taille de 1920x1080 pixels pour la vidéo "Vue 1" et 1080x720 pixels pour les autres. Comme les vidéos 2, 4 et 6 ont des caractéristiques similaires, nous pouvons les regrouper sous le nom DB2. La taille de la zone correspondant à la région d'intérêt est définie manuellement lors de l'étape d'initialisation. Les temps de calcul sont vraiment influencés par la définition de ces zones. Différents cadrages ont été utilisés pour la captation de ces vidéos. La première, "Vue 1", est une caméra dirigée vers le maire, avec une résolution plus élevée. Les autres caméras filment les conseillers et ont une résolution inférieure. C'est pourquoi nous utilisons deux arbres de décision différents dont l'apprentissage est effectué à l'étape d'initialisation. Le premier est formé à partir de 500 images de la caméra 1 et est ensuite appliquée sur l'ensemble de la séquence de "Vue 1". Le seconde est formé à partir de 500 images de la caméra 2 et appliqué sur les 3 autres séquences : vue 2, 4 et 6 (DB2).

La table 3.3 résume les résultats obtenus lors de l'utilisation de l'algorithme avec les vidéos tests. Les résultats sont exprimés en terme de précision (nombre de détections correctes par rapport au nombre de détections totales), de rappel (nombre de détections correctes par rapport au nombre réel de détections) et de justesse (nombre de détections correctes (vrai positif) et non détections correctes (faux positif) par rapport au nombre total de fenêtres).

Le temps de traitement par image (T.T.I.) exprime le temps moyen, en millisecondes, nécessaire pour détecter la présence d'un microphone actif dans une image. Ce temps dépend de la zone de recherche définie lors de l'initialisation.

TABLE 3.3 – Résultats obtenus par la méthode proposée.

Nom	Précision	Rappel	Justesse	T.T.I. (ms)
Vue 1	100 %	97.65 %	98.20%	33
Vue 2	98.47 %	94.27%	99.18%	19
Vue 4	99.72 %	100 %	99.93%	12
Vue 6	100 %	99.63%	99.89%	27

Les résultats obtenus montrent les performances de la méthode proposée sur les vidéos sélectionnées. La précision et la justesse sont d'environ 99% et le rappel d'environ 98%. Ces résultats confirment l'efficacité de la sélection des caractéristiques présentées dans la partie 3.3.2. Le temps de traitement par image montre la possibilité de fonctionnement en temps réel, grâce à l'utilisation d'une zone de recherche adaptée. De plus, les résultats obtenus sur la base de données DB2 montrent qu'il est possible de générer un modèle général, à partir d'une caméra, et traiter plusieurs flux provenant d'autres caméras présentant des réglages différents (grossissement, orientation).

La majorité des faux positifs obtenus sont dus à la lumière les perturbations (réflexions, smartphones,....). Un apprentissage plus long pourrait peut-être réduire ces erreurs. Les faux négatifs sont causés par l'occultation totale du microphone (lorsque le locuteur tient le microphone au niveau de la LED). Les résultats présentés ne tiennent pas compte de la dynamique de la séquence vidéo. Dans le cadre de la sélection de caméras, les détections et les pertes de détection inférieures à 300 ms sont ignorées. Cette temporisation, choisie empiriquement, permet de lisser les détections. Ainsi, les occlusions et les faux positifs de faible durée ne sont pas ressentis par l'utilisateur.

La méthode de détection de locuteur proposée permet une diffusion des conseils municipaux en temps réel. À chaque fois qu'une nouvelle personne parle, nous pouvons changer le flux relayé afin de montrer ce nouvel orateur. Afin d'améliorer le flux vidéo produit, il serait intéressant d'indiquer aux spectateurs qui est la personne en train de parler. Il est alors nécessaire de pouvoir identifier les locuteurs.

3.4 Identification des POI "locuteurs"

Afin de pouvoir proposer la personnalisation des flux vidéo montés, il est nécessaire d'extraire plus d'informations sur la scène. L'extraction des connaissances sur le conseil municipal nous a permis de mettre en avant de nombreuses informations disponibles sur les personnes d'intérêt telles que son nom, son rôle, son appartenance politique, etc. Ces informations pouvant être

stockées dans une base de données, il est nécessaire de faire le lien entre ces informations et la personne en train de parler. L'identification des locuteurs est donc une nécessité pour la personnalisation des flux vidéos.

3.4.1 État de l'art en identification de personne

Différentes méthodes ont été proposées dans la littérature pour l'identification de personnes dans une scène. Les approches biométriques sont les méthodes ayant suscité le plus de recherche ces dernières années. Parmi elles, la reconnaissance faciale [148] et la reconnaissance vocale [39, 55] sont les deux approches les plus adaptées dans le contexte des réunions.

La reconnaissance faciale consiste à identifier une personne dans une image à partir des caractéristiques morphométriques de son visage. Un processus de reconnaissance faciale est généralement composé de quatre étapes [73] : la détection de visage, les pré-traitements, l'extraction de caractéristiques et la reconnaissance faciale. Différentes méthodes ont été proposées pour la reconnaissance faciale. On peut citer par exemple celle se basant sur l'analyse en composante principale [103, 136], celles basées sur les séparateurs à vaste marge (SVM) [57], ou encore les méthodes basées sur les réseaux de neurones [73]. Parmi ces méthodes, celles basées sur les réseaux de neurones sont à ce jour celles qui présentent les meilleures performances [73]. Une des problématiques des méthodes de reconnaissance faciale est qu'il est nécessaire d'avoir une vue rapprochée des individus afin de pouvoir procéder aux identifications. De plus, il est nécessaire, pour le fonctionnement de ces systèmes, de construire une base de données contenant un grand nombre d'images de chaque personne à identifier afin de servir de base de comparaison ou d'apprentissage pour les modèles.

Les méthodes de reconnaissance vocale reposent sur les caractéristiques de la voix pour identifier les locuteurs. Ces méthodes se basent donc sur l'analyse de signaux audio provenant de microphones [55]. Les caractéristiques acoustiques les plus souvent utilisées sont les Mel-Frequency Cepstral Coefficients (MFCC)[98] et les Codages Prédictifs Linéaires (LPC)[60] .

Cependant, tout comme pour la reconnaissance faciale, il est nécessaire de réunir un grand nombre d'échantillons pour chaque personne à identifier.

Pour chaque approche de reconnaissance biométrique, il est nécessaire de collecter un grand nombre de données. Une base de données doit donc être constituée, afin d'identifier chaque élu. Ce recueil d'informations est nécessaire pour chaque nouvelle personne, rendant l'installation d'un système de reconnaissance biométrique long et coûteux.

Dans de nombreux conseils municipaux, des chevalets sont présents devant les élus indiquant le nom des personnes, comme nous pouvons le voir figure 3.9. Un système de reconnaissance de caractères (OCR)[41] pourrait permettre l'identification des individus présents derrière chaque chevalet. Plusieurs limitations telles que les occlusions, les surexpositions, les rotations de chevalets, comme illustrées figure 3.9b, rendent les techniques d'OCR peut fiables.

Les méthodes existantes montrent ainsi leurs limites dans notre contexte. Il est alors nécessaire de réfléchir à de nouvelles méthodes adaptées à notre contexte. Nous avons dans la partie 3.3 de ce chapitre, proposé une méthode pour détecter les locuteurs, se basant sur la détection de la LED des microphones. Cette information lumineuse nous informe sur la prise de parole d'un élu, mais peut également nous servir de support de transmission d'informations [31]. Le microphone devient alors un composant actif du système en communiquant un identifiant, propre à chaque

(a) (b)

FIGURE 3.9 – Identification de personnes par informations textuelles

locuteur, qui est capté par les caméras. Ceci permet une vérification de la détection ainsi que l'identification de la personne en train de parler.

3.4.2 Communication par lumière visible pour l'identification de locuteurs

Les communications optiques sans fil exploitent les spectres infrarouges et visibles des ondes électromagnétiques comme vecteur de communication. La communication par lumière visible (VLC), a attiré l'attention au cours de la dernière décennie grâce aux progrès réalisés dans la recherche sur les diodes électroluminescentes (LED).
Le principe de la VLC est de moduler l'intensité lumineuse par un signal d'information. Les commutations entre état allumé et éteint (On-Off Keying - OOK) des LEDs à haute vitesse permettent de transmettre de l'information sans que les personnes présentent dans la pièce ne perçoivent ces oscillations. La lumière modulée est détectée côté récepteur par des photo-diodes qui peuvent transformer l'intensité lumineuse en un courant électrique proportionnel.

Étant donné que de nombreux appareils comme les smartphones intègrent des caméras et des lampes flash, de nombreux chercheurs ont concentré leurs études sur la communication en utilisant ces composants pour produire des émetteurs-récepteurs à faible coût. De nombreuses applications de communication optique par caméra (OCC) ont été développées pour la technologie des véhicules tels que les communications véhicule à véhicule (V2V) et infrastructure à véhicule (I2V) ainsi que les applications de positionnement [120, 1].

Techniques de modulation pour la communication optique par caméra

Contrairement aux systèmes utilisant des photo-diodes, les caméras grand public ont une fréquence d'échantillonnage faible. Dans la majorité des cas, ces caméras ont une fréquence d'acquisition inférieure à 60 images par seconde. Une oscillation en basse fréquence des Leds serait perçue par les personnes présentes dans la pièce comme un scintillement, pouvant être dérangeant. Il est alors nécessaire d'émettre les messages avec une vitesse d'oscillation supérieure à 100Hz [61] afin de ne pas déranger les personnes présentes dans la scène, ainsi que les spectateurs. Différents types de modulation ont été mis en place pour la communication optique par caméra : la modulation par écran, la modulation sur-échantillonnée et la modulation sous-échantillonnée [86].

Modulation par écran
Cette technique consiste à coder l'information sur des plans en deux dimensions présents sur l'image capturée par la caméra. Les informations codées peuvent être visibles ou non par l'être

humain. La transmission par écran visible code le flux de données en codes visuels 2D successifs, de type QR code, affichés sur un écran [13]. En ce qui concerne le non-visible, les informations sont le plus souvent dispersées dans l'image et parfois dans le temps. Ceci est fait pour minimiser la perception d'artefacts visibles dans les images/animations causée par des informations intégrées [104].
Ces méthodes de modulation sont utilisées le plus souvent sur des distances assez courtes qui dépassent rarement le mètre.

Modulation sur-échantillonnée
Le principe de la modulation sur-échantillonnée est de transmettre les données en sur-échantillonnant le signal lumineux. Deux méthodologies sont régulièrement mises en place dans la littérature.

La première consiste à utiliser deux lumières polarisées orthogonalement représentant les bits 1 et 0. Comme ni l'œil humain ni les caméras commerciales ne peuvent capter le changement de polarisation de la lumière, aucun scintillement n'est observé. Cependant, lorsqu'un polariseur avec un certain angle d'orientation est placé devant la caméra, les variations de polarisation de la lumière sont converties en variations d'intensité lumineuse et sont sur-échantillonnées par la caméra [143]. Cette méthodologie permet de transmettre des messages à une distance de l'ordre d'une dizaine de mètres. Cependant, il est nécessaire que l'orientation de l'émetteur et du récepteur soit contrôlée afin de limiter les erreurs de transmission.

Le second type de modulation exploite le fonctionnement des caméras et notamment le mécanisme du Rolling Shutter. Lors de l'enregistrement d'une image, le capteur réalise l'acquisition ligne par ligne. Les impulsions lumineuses, transmises à des fréquences inférieures à la fréquence de balayage et supérieures aux fréquences de clignotement, peuvent donc être captées comme des bandes de luminosité différentes comme illustrées figure 3.10. L'analyse des changements de luminosité permet de pouvoir démoduler le message.

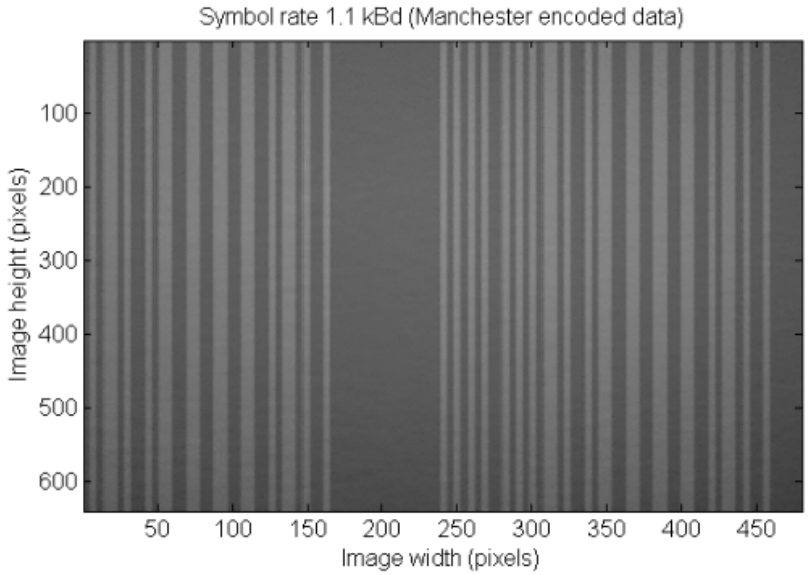


FIGURE 3.10 – Capture d'un message à l'aide du Rolling Shutter" [31]

Dans [31], les auteurs exploitent ce principe pour transmettre un message d'une LED vers

un téléphone portable. La LED, située à une distance de 35 cm d'un mur, est utilisée comme émetteur. Les messages sont encodés en utilisant le codage Manchester et sont envoyés par une modulation OOK. Le téléphone, situé à 9 cm du mur, récupère les images via son capteur CMOS, et démodule le signal après analyse des images. L'exploitation du Rolling Shutter permet d'atteindre des taux de transmission supérieurs à la fréquence d'acquisition de la caméra. Cependant, ces méthodes nécessitent que la caméra soit proche de la source lumineuse.

Modulation sous-échantillonnée

Afin d'être invisible à l'œil nu, la fréquence d'oscillation de la source lumineuse doit être suffisamment élevée. Cependant, la fréquence d'acquisition de la plupart des caméras est inférieure à 60 fps, ce qui est inférieur à la limite de 100 Hz [61] nécessaire pour être imperceptible par l'humain. Le principe des méthodes sous-échantillonnée est alors d'échantillonner le signal émis par une LED avec la fréquence de la caméra. Différentes méthodes ont été mises en place dans la littérature pour réaliser des transmissions sous-échantillonnées, parmi lesquelles les modulations de fréquence (UFSOOK) et de phase (UPSOOK) sont les plus souvent utilisées.

Le principe de la modulation Under-sampled Frequency Shift 0n-Off Keying (UFSOOK) [118] est d'utiliser deux signaux carrés de fréquences différentes, pour représenter les bits "0" et "1". Ces fréquences, f_{s1} et f_{s2}, peuvent s'exprimer sous la forme :

$$s(t) = \begin{cases} [cos(2\pi(m+0.5)f_{camera}t)] & si\ bit = 1 \\ [cos(2\pi m f_{camera}t)] & si\ bit = 0 \end{cases} \quad avec\ 0 < t < T_c \tag{3.2}$$

Où [ ] est la fonction signal carré de fréquence $f_{S1} = (m+0.5)f_{camera}$ et $f_{S0} = m.f_{camera}$ avec $m \in Z$. Comme $f_{S1} = (m+0.5)f_{camera}$, la LED a deux états différents dans deux images successives (lumière OFF et lumière ON ou vice versa). Par ailleurs, lorsqu'un zéro est envoyé, l'oscillation de la LED est synchronisée avec la fréquence d'image de la caméra et la même image est vue dans deux prises de vue consécutives. Ainsi, comme illustré figure 3.11 deux acquisitions successives permettent de décoder un 0 si on obtient deux images identiques (sombres ou lumineuses) et un 1 si on obtient deux images ayant des luminosités opposées.

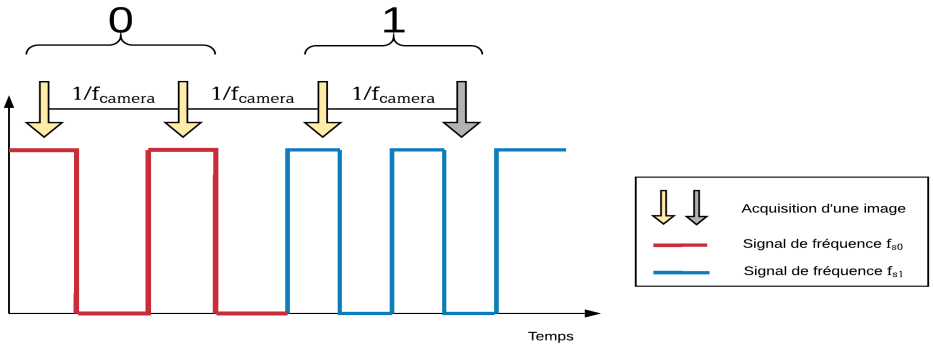


FIGURE 3.11 – Modulation UFSOOK

Comme pour UFSOOK, la modulation Undersampled Phase Shift 0n-Off Keying (UPSOOK) utilise deux signaux rectangulaires pour envoyer les bits 0 et 1. Ces signaux ont la même fréquence et la même amplitude, mais ont des phases opposées, comme illustré figure 3.12. Le débit

de symboles du signal est réglé sur la fréquence d'acquisition pour garantir que l'émetteur soit synchronisé avec la caméra. Par conséquent, chaque symbole de données est échantillonné une seule fois par le récepteur. Ainsi, un bit est décodé à chaque image. Il est cependant possible, en cas de mauvaise synchronisation que le message obtenu soit inversé et il est nécessaire de mettre en place une stratégie de rectification [86].
Les modulations sous-échantillonnées offrent le meilleur potentiel pour les scénarios à courte et longue portée.

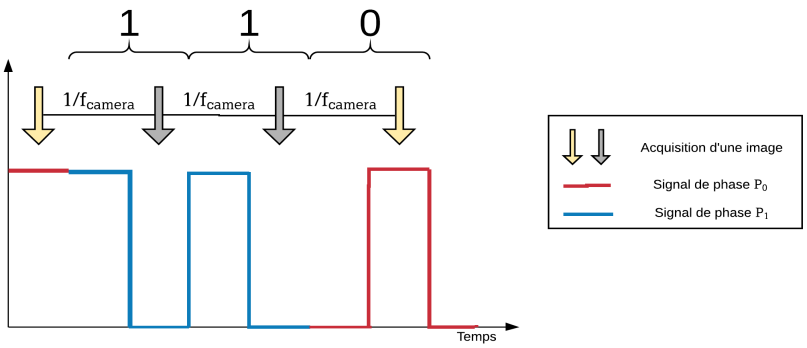


FIGURE 3.12 – Modulation UPSOOK

Dans notre cas d'application, les caméras utilisées pour la réception des messages sont les mêmes que celles utilisées pour la captation de l'évènement. Il est alors nécessaire que les vidéos diffusées ne soient pas altérées par le système de communication choisi. De ce fait, les méthodes basées sur une modulation sur-échantillonée ou par écran ne sont pas adaptées à notre cas d'application. De plus, la communication doit pouvoir être effectuée à une distance comprise entre 3 et 10 mètres. Les méthodes basées sur une modulation sous-échantillonnées sont donc celles les plus adaptées à notre contexte.

Présentation de la méthode proposée

Nous proposons donc une méthode d'identification de locuteur basée sur l'utilisation de la technologie Visible Light Communication. La figure 3.13 présente la vue d'ensemble de la méthode proposée.

Nous associons à chaque élu un identifiant numérique permettant de le caractériser de manière unique. Quand le microphone est actif (le locuteur parle.), la lumière LED est allumée et un code unique prédéfini pour chaque locuteur est transmis à l'aide de la modulation UFSOOK.

Côté récepteur, les lumières LED sont d'abord localisées et séparées des flux vidéo acquis, en utilisant la méthode proposée dans la partie 3.3. Une fois la lumière LED localisée, la région d'intérêt est étudiée pour retrouver l'identification de l'orateur. Pour ce faire, nous calculons la moyenne en luminance de la région d'intérêt comprenant la LED. Cette valeur de luminance est ensuite comparée à un seuil afin de déterminer l'état allumé ou éteint de la lampe. Ce seuil est adapté automatiquement en fonction de la valeur moyenne des 11 dernières fenêtres, permettant ainsi d'adapter le seuil aux changements rapides de luminosité de la scène.

Les états sont analysés deux à deux afin de déterminer la valeur du bit, puis les successions de bits sont étudiées afin de récupérer les messages.

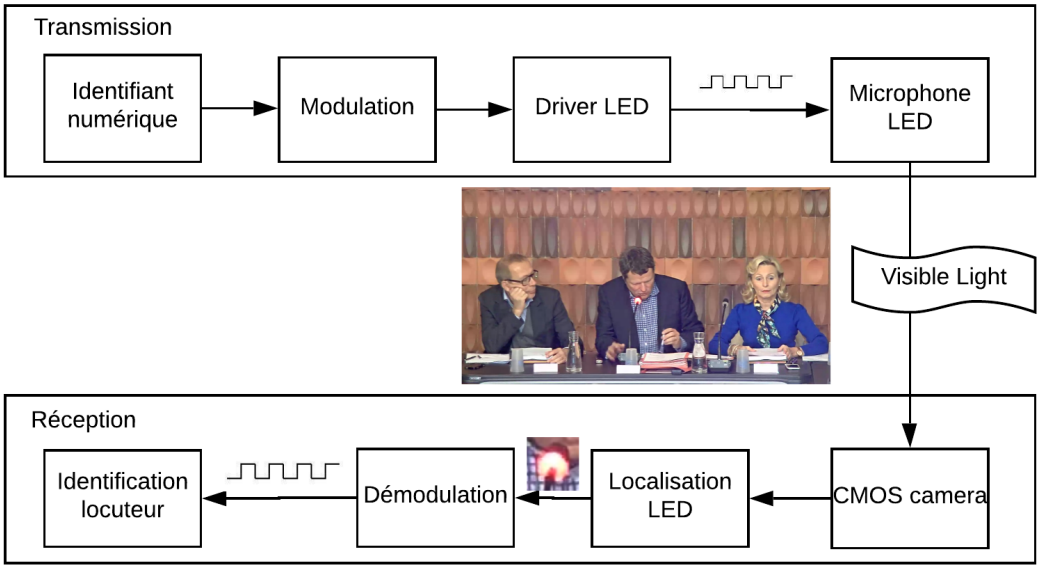


FIGURE 3.13 – Vue d'ensemble de la méthode d'identification de locuteurs utilisant la communication par lumière visible.

Composition des messages

Afin d'identifier les locuteurs, il est nécessaire d'envoyer un message contenant une information sur l'identité de l'élu. Nous proposons d'associer à chaque microphone, et donc à chaque élu, un identifiant numérique, permettant de faire le lien entre le microphone et l'ensemble des informations du locuteur. Du fait du nombre limité de participant , un maximum de 8 bits permet d'assigner un identifiant à chaque microphone.

Les messages étant envoyés en continu une fois la LED allumée, il est nécessaire de trouver le début du message. Pour ce faire, les trames envoyées sont composées d'un en-tête suivi du message.

Dès qu'un en-tête est retrouvé, et connaissant la taille du message, il est alors possible de récupérer l'information transmise. Cependant, la suite de bits composant l'en-tête peut également faire partie de la composition du message lui-même, comme illustré figure 3.14. Il en résulterait alors une augmentation des messages détectés, dont certains incorrects.

Afin de palier ce problème, nous proposons d'utiliser un en-tête dynamique, c'est-à-dire un en-tête dont la valeur change entre deux messages. Ainsi, une négation logique est appliquée sur l'en-tête à chaque envoi de message. Par exemple, si l'entête utilisée pour le premier message est '10' alors celle utilisée pour le message suivant sera '01', puis la suivante sera de nouveau '10' et ainsi de suite.
Ce protocole permet de s'assurer que l'en-tête n'est pas un corps de message en comparant l'en-tête avec les bits situés une longueur de trame plus loin, comme illustré figure 3.15. Si ces bits forment l'en-tête inversé, le message est effectivement compris entre ces deux en-têtes. Si au

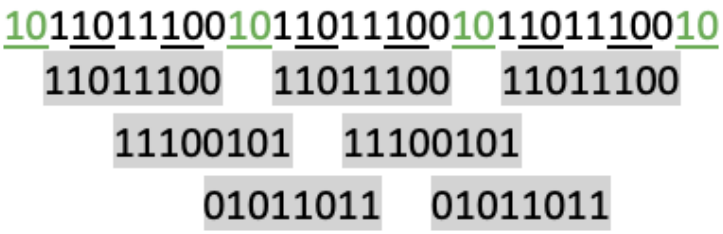


FIGURE 3.14 – Messages de 8 bits détectés avec une entête fixe '10'

contraire, ces bits ne sont pas une permutation de l'en-tête présumé, c'est que ce dernier n'est pas un en-tête, mais une partie du message .

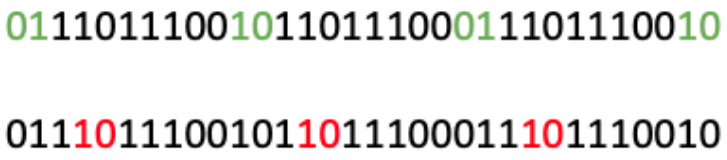


FIGURE 3.15 – Utilisation d'un en-tête dynamique : l'inversion des en-têtes permet de trouver uniquement les messages envoyés.

Les trames envoyées ayant une longueur faible, il est possible de transmettre l'identification du locuteur en un laps de temps faible. Dans le cas où une trame de 10 bits est envoyée (2 bits d'en-tête et 8 bits de messages), 20 images sont nécessaires pour transmettre une trame. Il est alors possible d'obtenir l'identification du locuteur 0.4 secondes après l'allumage du microphone dans le cas d'une caméra fonctionnant à 50 fps.

3.4.3 Expérimentation

L'évaluation de cette méthode a été réalisée à partir de vidéos contenant deux LEDs. Quatre messages différents , définis dans la table 3.4, ont été transmis.

Id. message	En-tête	Données transmises
1	1/0	1 1100 1011
2	10/01	1010 1111
3	10/01	1100 1011

TABLE 3.4 – Caractéristiques des messages envoyés

Afin de générer les signaux correspondant et ainsi contrôler les LEDs, deux Générateurs de Fonctions Arbitraire (AWG) AGILENT 33120A ont été utilisés.

Pour évaluer les performances de notre méthode, différentes caméras ont été utilisées. La majorité des caméras actuelles proposent des fréquences d'acquisition réglables. Les fréquences les plus courantes sont 25 fps, 30 fps et 50 fps. Cependant, ces fréquences d'acquisition ne sont pas toujours respectées. Nous avons remarqué que de nombreuses caméras acquéraient les images avec

des fréquences oscillantes autour des fréquences données. Cette variation induit de nombreuses erreurs de transmission, comme illustré figure 3.16. En effet, au lieu d'avoir lieu durant les états allumés ou éteints, les acquisitions peuvent avoir lieu lors d'un changement d'état de la LED.

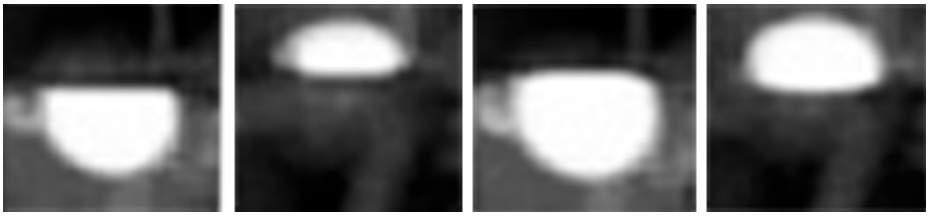


FIGURE 3.16 – Erreurs dues à une mauvaise fréquence d'acquisition : la LED change d'état durant l'acquisition de l'image

Il est alors nécessaire d'utiliser une caméra ayant une fréquence d'acquisition stable. Ainsi, les expérimentations menées dans cette partie, ont été réalisées sur des vidéos capturées avec un reflex numérique Nikon D5600. Les paramètres utilisés lors des expérimentations sont résumés dans la table 3.5.

Caractéristiques	**Valeurs**
Frame Rate	50 fps
Shutter-speed	1/4000 s
ISO	2000
Ouverture	3.5
Focale	18

TABLE 3.5 – Paramètres d'acquisition de la caméra

Afin de garantir une transmission pour diverses installations de caméras, différentes distances et angles de vue ont été étudiés. La table 3.6 résume les caractéristiques des vidéos utilisées. Les vidéos 1, 2, 3 et 4, ont été réalisées en plaçant la caméra en face des LEDs. Les vidéos 1 et 2 ont été faites à une distance de 2,5 mètres. Les distances pour les vidéos 3 et 4 sont respectivement de 5 mètres et 7,5 mètres. Enfin, la vidéo 5 a été réalisée à une distance de 2,5 mètres, avec un angle d'environ 45 ° par rapport à l'axe des LEDs.
Pour chaque vidéo, le nombre maximal de messages est indiqué. Il s'agit du nombre de messages qui pourrait être obtenu dans le cas où l'enregistrement d'une vidéo commence en même temps que l'émission d'un en-tête.

Id Vidéo	Distance (m)	Caractéristiques	Longueur (Nb image)	Nb max de messages
1	2.5	Vue de face	1650	82
2	2.5	Vue de face	2350	117
3	5	Vue de face	2450	122
4	7.5	Vue de face	5300	265
5	2.5	Vue de biais	2350	117

TABLE 3.6 – Caractéristiques des vidéos capturées

Les images et les trames sont analysées grâce à un programme développé en Python, en utilisant les fonctions de la bibliothèque OpenCV.

3.4.4 Résultats

Les tables 3.7 et 3.8 présentent les résultats obtenus en utilisant notre méthode dans différentes vidéos, dont les caractéristiques sont résumées dans le tableau 3.6. Pour chaque vidéo, le nombre de messages trouvés et le pourcentage de messages corrects sont indiqués ainsi que le taux d'erreur bit (BER) qui est calculé sur la séquence vidéo.
Les résultats sont présentés en fonction de trois critères : la taille des régions d'intérêt, la taille de l'en-tête et la distance entre la LED et la caméra.

Influence de la taille de fenêtre

Comme notre méthode se base sur la moyenne de la luminance de la zone d'intérêt, la taille de cette région influe sur le résultat. Nous proposons donc dans cette partie d'analyser l'influence de la taille dans la réception des messages. Pour ce faire, nous proposons d'analyser les valeurs de moyenne, ainsi que les taux d'erreur (BER) pour les tailles utilisées dans la section 3.3.2. En effet, nous souhaitons vérifier si l'étude de la fenêtre produite par notre méthode (de taille 19x19 pixels) permet de récupérer efficacement l'information ou si il est nécessaire de choisir une fenêtre plus faible (3x3 ou 9x9 pixels).

La figure 3.17 présente un extrait des résultats obtenus en appliquant notre méthode sur la vidéo 3. Nous pouvons remarquer que plus la taille de la région d'intérêt est grande, plus les valeurs moyennes des niveaux haut et bas sont proches. Ceci s'explique facilement car de plus en plus de pixels de l'arrière-plan sont pris en compte dans le calcul de la moyenne de la fenêtre. Cependant, pour les trois cas d'application, les états sont toujours identifiés de la même manière comme état haut ou état bas. Cela implique que les résultats en terme de messages correctement identifiés et de BER sont similaires.
Dans notre cas d'application l'arrière-plan est uniforme et fixe permettant d'obtenir des résultats similaires avec différentes tailles de fenêtre. Il sera cependant conseillé de sélectionner une taille de fenêtre faible, notamment dans le cas d'arrière-plan dynamique afin de prendre en compte uniquement les valeurs des LEDs. Il sera alors nécessaire de sélectionner, dans la fenêtre de taille 19x19 obtenue par notre méthode présentée dans la section 3.3, une sous fenêtre centrée sur la LED.

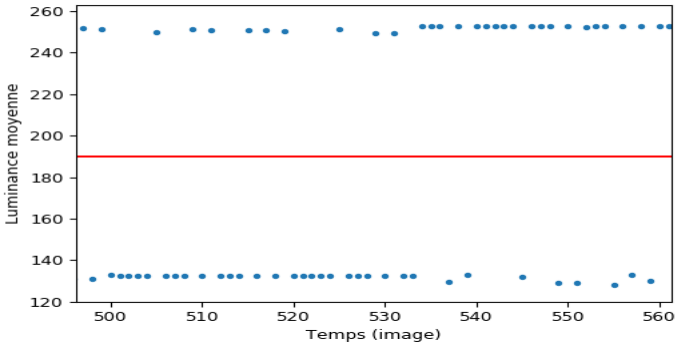
Influence de la taille d'en-tête

Le tableau 3.7 présente les résultats obtenus avec un en-tête d'un bit de longueur '1' ou '0' (message 1) et de deux bits de longueurs '10' ou '10' (message 2).

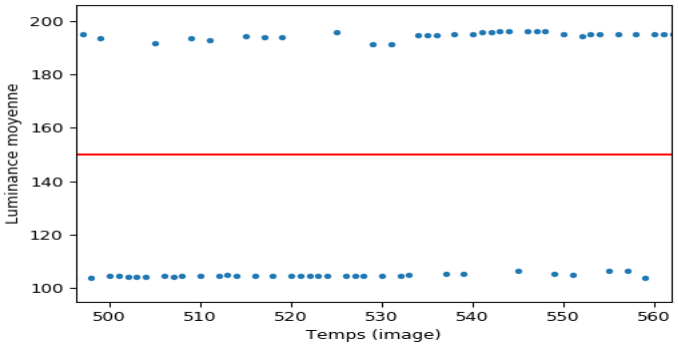
Vidéo	Entête	Nb messages trouvés	Nb messages corrects	BER
1	'1'/'0'	148	60	32%
	'10'/'01'	98	53	23%
2	'1'/'0'	302	31	41%
	'10'/'01'	131	74	22%

Table 3.7 – Influence de la taille de l'entête.

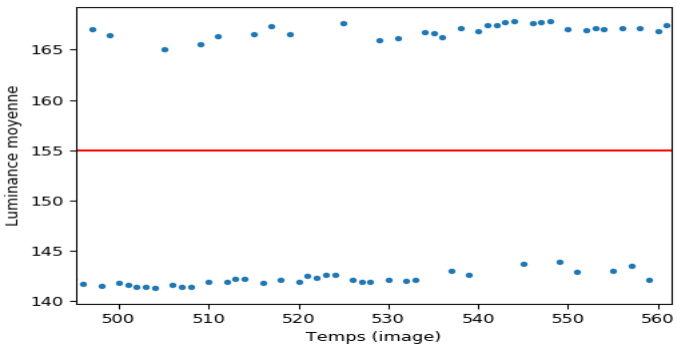
Nous pouvons remarquer que le nombre de messages trouvés est près de deux fois supérieur au nombre de messages émis. Cela vient du fait qu'une erreur dans la transmission d'un état de



(a) Luminance moyenne pour une fenêtre de taille de 3x3 pixels



(b) Luminance moyenne pour une fenêtre de taille de 9x9 pixels



(c) Luminance moyenne pour une fenêtre de taille de 19x19 pixels

FIGURE 3.17 – Influence de la taille des régions d'intérêt sur l'extraction des états

LED se produit, alors le bit reçu est inversé par rapport à celui devant être transmis. Si ce bit lors d'une trame précédente ou suivante est transmis correctement, alors ces bits sont considérés comme en-tête et le message est alors extrait. Afin de diminuer le nombre de ces messages incorrects, il est alors nécessaire d'utiliser un en-tête d'une longueur supérieure ou égale à 2 bits.

Influence de la distance

Le tableau 3.8 présente les résultats obtenus pour différentes distances. Les deux messages sont transmis simultanément par deux LEDs, positionnées à égale distance de la caméra.

Vidéo	Distance (m)	Id. Message	Nb messages trouvés	Nb messages corrects	BER
3	5	2	132	78	20%
		3	126	83	17%
4	7.5	2	306	178	21%
		3	288	176	17%
5	2 (vue de biais)	2	149	82	21%
		3	125	75	17%

TABLE 3.8 – Influence de la distance sur des images comportant deux LEDs

Nous pouvons remarquer que les résultats obtenus à différentes distances et différentes orientations sont similaires. Les taux de messages corrects obtenus sont supérieurs à 50%, signifiant que plus d'un message sur deux est reçu sans erreur. Ainsi, il est nécessaire de recevoir plusieurs fois le même messages afin de garantir la réception du bon identifiant et ainsi obtenir les informations sur une personne.

Influence des générateurs

Les tests que nous avons conduits jusqu'à présent utilisaient des générateurs AGILENT 33120A dont le coût est incompatible avec le souhait de CitizenCam de proposer une solution de captation abordable. C'est pourquoi, nous proposons de générer les signaux grâce à des cartes électroniques Arduino Uno, permettant ainsi de réduire les coûts. Cinq nouvelles vidéos ont été enregistrées, avec les mêmes paramètres d'acquisitions que précédemment, pendant une durée de 40 secondes (100 messages émis au maximum). Les messages envoyés sont composés d'une en-tête "01/10" et d'un message de 8 bits "0010 0111". Les distances utilisées correspondent à celles retrouvées fréquemment dans les captations de CitizenCam.

Distance	Nb messages trouvés	Nb Messages corrects	BER
3	78	74	2%
5	74	70	2%
7.5	74	68	4%
8	87	79	7%
10	86	77	8%

TABLE 3.9 – Résultats obtenus avec l'utilisation d'un Arduino Uno

Les résultats, présentés dans la table 3.9, montrent que l'implémentation avec un Arduino Uno est une alternative plausible. En moyenne, 80% des messages envoyés sont trouvés et plus de 90% de ces messages sont corrects. Cela montre la possibilité d'utiliser ce système dans le cas de captation de conseils municipaux. Nous pouvons remarquer que les taux d'erreur (BER) obtenus sont plus faibles que ceux obtenus avec les générateurs AGILENT 33120A. En effet, la fréquence d'horloge est plus élevée que celle du générateur (16MHz/15MHz). L'échantillonnage du signal est alors plus précis permettant ainsi des transitions plus rapides entre les états induisant une diminution des erreurs de transmission.

Ces premiers résultats montrent, qu'il est possible de transmettre un message court à des distances correspondantes à celles utilisées dans la captation de conseils municipaux. De plus, la méthode proposée permet de transmettre plusieurs messages dans un même flux vidéo et d'utiliser les images pour la diffusion de l'évènement. La transmission d'un identifiant dans le cas d'une captation de conseil municipal par l'utilisation de la communication par lumière visible est alors envisageable. Les résultats montrent que près de 72% des messages envoyés sont correctement reçus. Il sera alors nécessaire d'attendre plusieurs messages identiques afin de confirmer le code reçu. Cependant, il est fréquent que les locuteurs allument leurs microphones avant de prendre la parole. De ce fait, il est possible de recevoir plusieurs messages avant que le locuteur prenne la parole, et ainsi de trouver son identité au moment où il s'exprime.

3.5 Discussions

Dans ce chapitre, une proposition de méthodologie pour le montage automatique dans le cadre des conseils municipaux a été proposée. La modélisation des connaissances, provenant du Code Général des Collectivités Territoriales et de connaissances d'experts, nous a permis d'identifier l'action d'intérêt principal : la prise de parole. Nous avons proposé une méthodologie de détection d'AOI en se basant sur la détection de l'allumage des LEDs des microphones. Notre méthode permet de détecter la prise de parole en temps réel dans une application multi-caméra et de générer un flux vidéo monté automatiquement présentant les interventions successives des orateurs lors d'un conseil municipal.

Afin d'améliorer la sélection de caméras, et de permettre la personnalisation des flux vidéo, une méthodologie d'identification des locuteurs a été proposée. Cette méthodologie originale, basée sur l'utilisation de la technologie Visible Light Communication s'est avérée pertinente. En plus de leurs utilisations pour la détection de prise de parole, les LEDs ainsi utilisées comme moyen de transmission, permettent de transmettre des informations sur la personne en train de parler. Comme il est nécessaire d'attendre plusieurs trames avant de confirmer le message, il n'est pas possible d'obtenir l'identité d'une personne en temps réel. Cependant, dans le cas de conseils municipaux, il n'est pas dérangeant d'avoir cette information quelques secondes après la prise de parole.

Nous avons présenté dans ce chapitre une méthode de détection de source lumineuse et de transmission d'informations par lumière visible dans le cadre des conseils municipaux. Ces méthodologies peuvent facilement être mises en œuvre pour n'importe quelle réunion ou conférences avec comme seul pré-requis la présence de microphone ayant une LED lors d'une prise de parole. De plus, la méthode de communication par la lumière que nous avons présentée pourrait être utilisée dans différents contextes. Un dispositif portable équipé d'une LED pourrait être utilisé pour identifier une personne en déplacement dans une scène filmée par des caméras. Cela implique néanmoins de localiser précisément cette personne.

Chapitre 4

Montage automatique pour la diffusion d'un match de basketball

Sommaire

4.1 Introduction

Chaque jour des évènements sportifs se produisent. Ces manifestations rassemblent un grand nombre de personnes, que ce soit en tant qu'acteur ou en tant que spectateur. Depuis le début du 20ème siècle, la télévision s'est intéressée à la retransmission sportive et en 1936, la première diffusion en direct d'un évènement a eu lieu lors des jeux olympiques à Berlin. Il est aujourd'hui fréquent de voir un évènement sportif diffusé en direct à la télévision. Les grands clubs sportifs ont les moyens et l'audience nécessaires pour se permettre les diffusions ou rediffusions. Cependant, les petits clubs, à l'échelle d'une ville ou d'un département ne peuvent pas s'offrir la visibilité des grands clubs sans une nécessaire réduction des coûts de captations. Aussi, l'automatisation de la sélection de la prise de vue est une des voies possibles pour y parvenir.

Des systèmes de sélections automatiques de caméras ont été proposés pour de nombreux sports : Basketball [30, 21, 32, 23], Football [4, 137, 46, 26], Hockey[20], ... La plus grande contrainte de ces systèmes vient du nombre important de joueurs, des déplacements rapides et pratiquement aléatoires de ces derniers rendant le contrôle des caméras et la sélection de la vue d'intérêt difficiles.

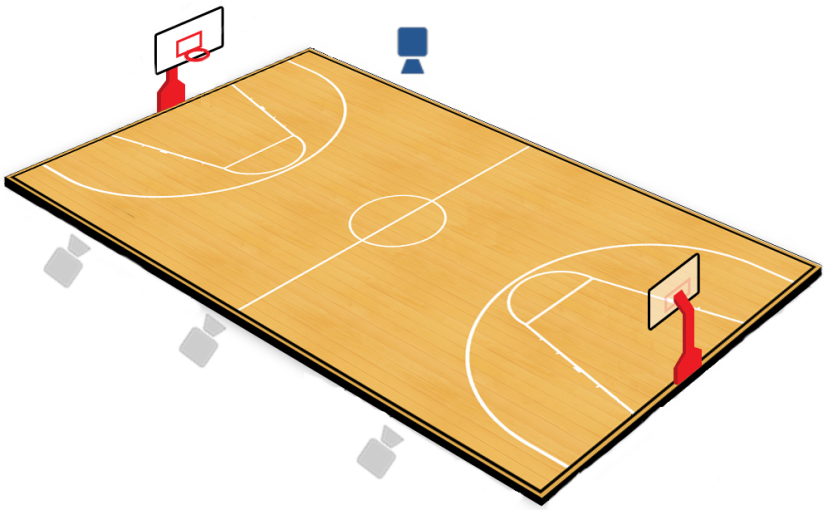


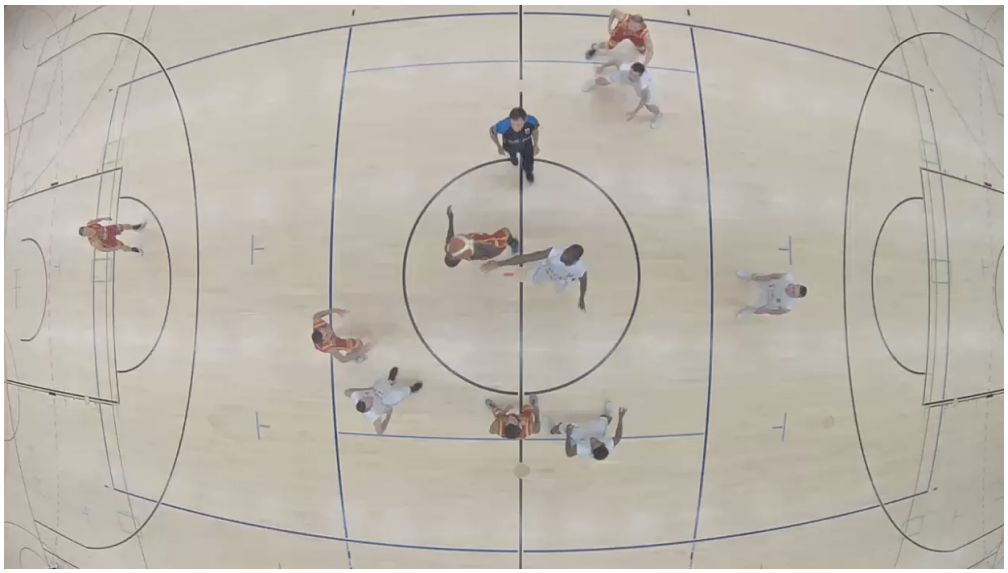
FIGURE 4.1 – Installation des caméras pour la captation d'un match de basketball

Dans ce chapitre, nous nous intéressons à la diffusion de matchs de basketball. Notre cas d'application est celui du centre sportif de Mondercange, au Luxembourg, accueillant l'équipe de BC Mess. Quatre caméras sont installées dans le gymnase afin de capter l'ensemble du terrain, comme illustré dans la figure 4.1. Une caméra grand-angle est située au-dessus du terrain (figure 4.2a) et 3 caméras sont installées sur le côté du terrain (figure 4.2b). Lors d'un match de basketball, seules les trois caméras latérales sont diffusées. La fonction première de la caméra azimutale est de permettre aux entraîneurs d'avoir des informations sur la position des joueurs.

Lorsqu'un spectateur veut regarder un match de basket, il est nécessaire qu'il sélectionne manuellement la caméra où se passe l'action. Ce changement de caméra est, contrairement aux

conseils municipaux, une action relativement fréquente. L'ensemble des joueurs se déplace d'un bout à l'autre du terrain tout au long du match. Il est alors nécessaire de souvent changer de caméra, ce qui rend la visualisation de la rencontre peu agréable pour le spectateur.

Pour résoudre ces problèmes, nous proposons de mettre en place un système de sélection automatique de caméras. Le système de captation que nous utilisons étant composé uniquement de caméras fixes, seules les étapes de planification et de sélection sont abordées. Nous allons dans un premier temps, selon la méthodologie proposée, instancier le modèle générique au contexte "basketball" pour en extraire les connaissances afin d'identifier les différents points d'intérêt pour la réalisation de la sélection. L'étude de l'état de l'art nous conduit à proposer de nouvelles méthodes permettant un suivi simple et efficace des joueurs dans le cadre d'un match afin de proposer une diffusion en direct des rencontres. Nous proposons également une nouvelle méthode permettant de détecter des actions d'intérêt afin de rendre possible la personnalisation des flux vidéos.
De plus, la réalisation d'un montage automatique nécessite d'extraire un grand nombre d'informations qui peuvent s'avérer utiles aux entraîneurs des équipes. En effet, les informations sur les distances parcourues, les vitesses des joueurs, ou encore les statistiques de tirs sont des statistiques utilisées par les entraîneurs afin d'améliorer les performances de leur équipe. Nous proposons donc d'extraire ces statistiques afin de proposer un système à moindre coût.



(a) Vue azimutale

(b) Caméras latérales

FIGURE 4.2 – Diffusion d'un match de basketball : vues disponibles

4.2 Modélisation du contexte d'un match de basketball

La première étape de notre méthodologie est d'identifier les différentes sources d'intérêts présentes lors d'un match de basket. Pour ce faire, nous pouvons nous appuyer sur le règlement de la fédération Internationale de Basketball (FIBA) [44].
Un match de basketball est défini de la manière suivante :

"Une rencontre de Basketball se dispute entre deux équipes de 5 joueurs chacune. L'objectif de chaque équipe est de marquer dans le panier de l'adversaire et d'empêcher l'autre équipe de marquer. Une rencontre est gagnée par l'équipe qui a marqué le plus grand nombre de points au score à l'expiration du temps de jeu."

Cette définition nous permet de mettre en évidence les personnes et actions d'intérêt principales. La POI est une "équipe" composée de "joueurs". Une action d'intérêt ressort : "la rencontre de Basketball" qui est composée des actions "marquer dans le panier" et "empêcher l'autre équipe de marquer". L'application de notre méthodologie conduit à rechercher les différents attributs des personnes et action d'intérêts afin de piloter les algorithmes d'extraction de leur caractéristiques. Ces attributs peuvent également servir pour proposer des flux vidéo personnalisés.

4.2.1 Définition des personnes d'intérêt

Lors d'une rencontre de basketball, plusieurs acteurs sont présents : les personnes sur le terrain (joueurs et arbitres), les personnes hors du terrain : les membres des équipes (remplaçants, entraîneurs, soigneurs) et les officiels (officiels de table de marque, commissaire). Enfin, il est fréquent que des spectateurs soient présents dans la salle où a lieu la rencontre. Dans le cadre d'une diffusion du match, l'intérêt des spectateurs se porte principalement sur ce qui se passe sur le terrain. Les joueurs sont donc logiquement au centre de l'attention. L'exploitation de notre modèle générique nous conduit à rechercher les attributs caractérisant les POI, à savoir des attributs de forme et d'espace. L'article 4 du règlement officiel de basketball donnent un certain nombre d'informations sur la formation d'une équipe et donc sur les joueurs.

Par exemple, concernant les joueurs pouvant être sur le terrain, il est noté qu'un "membre d'équipe est autorisé à jouer lorsque son nom est inscrit sur la feuille de marque avant le commencement de la rencontre". Enfin, si une équipe est composée d'un maximum de 12 membres autorisés à jouer, seuls cinq joueurs de chaque équipe seront présents sur le terrain. Les joueurs d'une équipe portent une tenue devant respecter les dispositions suivantes : "Les membres d'une même équipe porte une même tenue. Cette tenue est composée d'un maillot d'une même couleur dominante devant et derrière, d'un short de la même couleur dominante devant et derrière, identique à celle des maillots. Et enfin des chaussettes de la même couleur dominante pour tous les membres de l'équipe. L'équipe nommée en premier sur le programme (équipe locale) doit revêtir des maillots de couleur claire (de préférence blancs), la seconde équipe nommée sur le programme (équipe visiteuse) doit porter des maillots de couleur foncée." Les joueurs d'une même équipe portent donc la même tenue. Les maillots ayant une couleur dominante et différente pour chaque équipe, nous pouvons en déduire qu'une équipe à une couleur propre. De plus, "chaque membre d'équipe doit porter un maillot numéroté devant et derrière avec des chiffres pleins, d'une couleur contrastant avec celle du maillot. Des joueurs d'une même équipe ne peuvent pas porter le même numéro."

Un joueur est donc identifié par un nom, un numéro et l'équipe à laquelle il appartient. De plus, pendant le temps de jeu, un membre de l'équipe peut avoir plusieurs rôles. "Il peut être un

joueur lorsqu'il est sur le terrain, un remplaçant lorsqu'il n'est pas sur le terrain ou un joueur éliminé si il a commis cinq fautes." "Un joueur peut également être désigné par son entraîneur comme capitaine. Son rôle est de représenter son équipe sur le terrain de jeu."

Pour résumer, lors d'une rencontre de Basketball, deux équipes de 5 joueurs s'affrontent sur le terrain. Les membres d'une même équipe portent un maillot d'une couleur unie, permettant de différencier les deux équipes. Les joueurs d'une même équipe sont identifiés par leur nom et un numéro qui leur sont propres. Enfin, les joueurs peuvent avoir différents rôles lors de la rencontre, en fonction de leur présence, ou non, sur le terrain.

4.2.2 Définition des actions d'intérêt

Comme pour les POI, nous devons définir les caractéristiques des AOI à savoir des attributs temporels et spatiaux. Nous avons vu précédemment que la rencontre de basketball se dispute entre deux équipes. Une rencontre est définie temporellement de la manière suivante : "Une rencontre doit consister en 4 quart-temps de 10 minutes. Si le score est à égalité à l'expiration du quatrième quart-temps, le jeu doit continuer par autant de prolongations de 5 minutes que nécessaires pour casser l'égalité. Il doit y avoir un intervalle de 2 minutes entre le premier et le second quart-temps (première mi-temps), entre le troisième et le quatrième quart-temps (seconde mi-temps) et avant chaque prolongation. Il doit y avoir un intervalle de 15 minutes à la mi-temps." Un match de basketball est donc divisé en quatre quart-temps de 10 minutes. Les deux premiers quart-temps forment la première mi-temps, tandis que les deux suivants forment la seconde. Il est à noter que ces 10 minutes correspondent à 10 minutes de jeux et non à 10 minutes continues. Le chronomètre est arrêté lorsque, par exemple, une faute est commise ou lorsqu'un panier est marqué à moins de deux minutes de la fin du quatrième quart-temps. Il est redémarré lors d'une remise en jeu ou après un dernier lancer-franc.

Durant le temps de jeu, l'objectif de chaque équipe est de marquer dans le panier de l'adversaire et d'empêcher l'autre équipe de marquer. L'équipe ayant marqué le plus de paniers à la fin de la rencontre étant gagnante, il est vraisemblable que les joueurs aillent d'un panier à un autre fréquemment durant la rencontre. Un panier est marqué lorsque le ballon "pénètre dans le panier par le haut et reste dedans ou passe à travers entièrement."

L'action d'intérêt principale dans un match de basketball est donc les attaques et les défenses des équipes, c'est-à-dire jeu notable. Il est nécessaire pour le spectateur de pouvoir suivre le déplacement des joueurs afin de visualiser quand un panier est marqué ou quand une équipe empêche un panier d'être marqué.

Pendant une rencontre de basketball, plusieurs évènements peuvent se produire. Un joueur peut marquer un panier durant un tir au panier ou un lancer-franc. Un joueur peut également commettre une infraction au règlement. Deux types de fautes peuvent être réalisées. Tout d'abord les violations commises par un seul joueur (Règle cinq - Violations [44]) comme par exemple une sortie du joueur ou du ballon, une reprise de dribble, un marcher. Lorsqu'une violation est commise, le ballon est donné à l'équipe adverse pour une remise en jeu. Puis les fautes, c'est-à-dire : "une infraction aux règles impliquant un contact personnel illégal avec un adversaire et/ou un comportement antisportif". (Règle six - fautes [44]) : contact, faute technique, faute antisportives, ... Une faute est immédiatement sifflée par l'arbitre et un ou plusieurs lancers-francs sont accordés aux adversaires. Quelle que soit l'infraction constatée, une violation ou une faute est sifflée par un arbitre. Le chronomètre de jeu est alors arrêté. Il est redémarré lors de

la remise en jeu dans le cas d'une violation et après un dernier lancer-franc manqué dans le cas d'une faute.

Pour résumer, un grand nombre d'actions peuvent avoir lieu lors d'un match de basketball. Ces actions, que ce soit le jeu notable ou les fautes, ont une date de début et de fin. Ces actions comportent des noms et un certain nombre de caractéristiques les définissants.

4.2.3 Configuration du montage automatique

Les informations que nous avons extraites du règlement officiel de la fédération internationale de basketball nous ont permis d'instancier le modèle NIAM/ORM, comme le montre la figure 4.3.

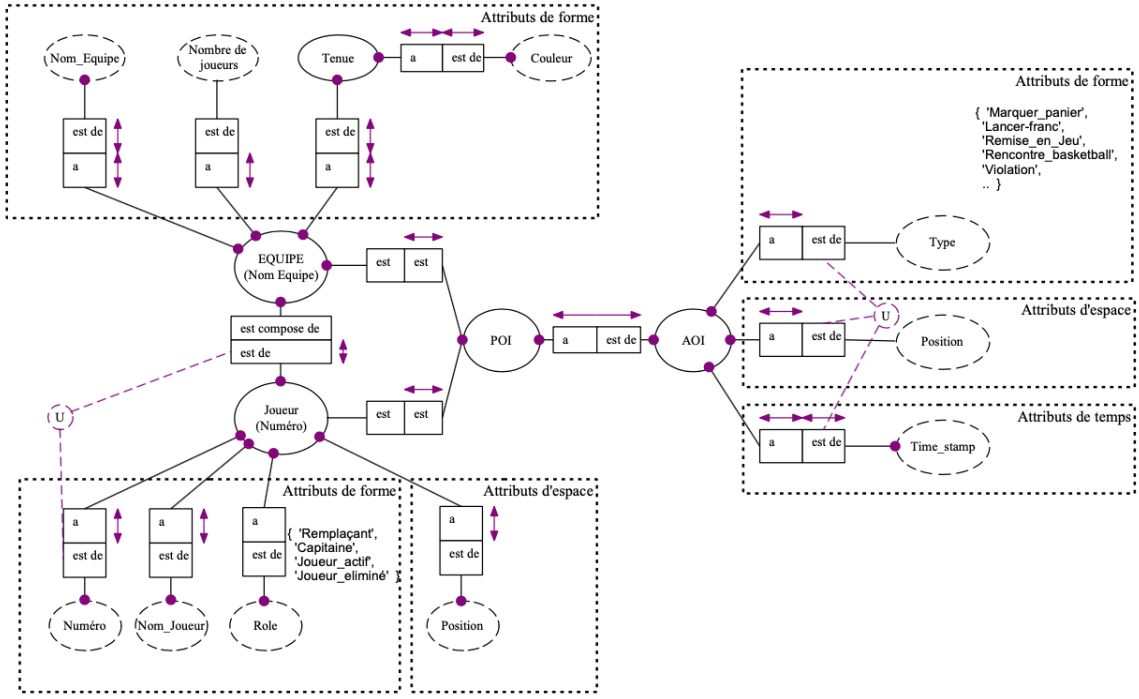


FIGURE 4.3 – Modélisation d'un match de basket

L'action d'intérêt principale que nous avons identifiée est appelée jeu notable, c'est-à-dire le déplacement des joueurs, qu'ils soient en attaque ou en défense, tout au long de la rencontre. En effet, dans le cas d'une diffusion en direct, un spectateur veut principalement voir les attaques et les défenses de son équipe préférée. Une attaque étant définie par le déplacement d'une équipe vers le panier adverse, les personnes d'intérêts sont les joueurs d'une équipe. Afin de proposer un flux vidéo monté en direct, il est ainsi nécessaire de localiser l'action sur le terrain. C'est-à-dire trouver le jeu notable tout au long de la rencontre.

Pour permettre la personnalisation des flux vidéos montés, il est donc nécessaire d'extraire des informations supplémentaires sur la scène. Les spectateurs peuvent vouloir s'intéresser à un joueur ou une équipe particulière. Il sera alors possible de mettre en place un système de différenciation des équipes se basant sur les noms d'équipe, ou bien sur la couleur qui est portée par tous les membres de l'équipe. Les membres de l'équipe sont quant à eux identifiables par un nom

et un numéro qui leur sont propres. Ils peuvent avoir différents rôles au cours de la rencontre et réaliser un certain nombre d'actions. Les différentes actions qui ont lieu lors d'un match sont quand à elles reconnaissables par plusieurs caractéristiques propres à chacune.

La modélisation et l'exploitation des connaissances dans le cadre du Basketball nous a permis d'identifier différentes sources d'intérêts. La source d'intérêt principale est l'action d'intérêt "jeu notable" qui peut être caractérisée par le déplacement des POI "joueurs". Nous présentons dans la section 4.3 une méthode de détection de cette AOI afin de permettre la diffusion en direct d'un match de Basketball.
De plus, afin d'autoriser une personnalisation des flux vidéo, il est nécessaire de détecter, suivre ou identifier les différentes POI, c'est-à-dire les joueurs et les équipes auxquelles ils appartiennent, et les AOI ayant lieu pendant le match. Nous présentons alors, dans la section 4.4.2, un méthode permettant la détection de l'AOI " lancer-franc" et dans la section 4.5 une méthode de suivi d'un joueur ou d'un groupe de joueurs.

4.3 Détection de l'AOI "jeu notable"

Afin de proposer une diffusion en direct d'une rencontre de basketball, il est nécessaire de localiser le jeu notable durant les quatre quarts-temps. Dans la littérature, cette action d'intérêt peut être détectée de trois façons différentes.

Daigo et Ozawa [30] s'intéressent à l'orientation des visages des personnes dans l'audience afin de contrôler l'orientation d'une caméra. La méthode présentée repose sur l'hypothèse que les spectateurs regardent le jeu notable lors d'un match. Cette méthode nécessite alors de disposer d'une caméra supplémentaire orientée vers l'audience. De plus, le temps de calcul nécessaire à l'obtention de l'orientation des visages rend la diffusion en direct difficile.

D'autres méthodes utilisent des informations sur la position de la balle [4]. En effet, dans les jeux de ballon, la balle est le centre d'attention des joueurs et est donc représentative de l'action. Grâce à ces caractéristiques, les auteurs proposent un système permettant de contrôler une caméra virtuelle en différé. Cependant, le ballon est difficile à détecter, car ses déplacements sont aléatoires en terme de direction et de vitesse. De plus, sa petite taille rend important le nombre d'occultations au cours du match.

Enfin d'autres méthodes s'intéressent à la position des joueurs sur le terrain. Carr et al. [17] utilisent le déplacement des joueurs afin d'orienter une caméra PTZ vers l'ensemble des joueurs. Deux caméras, fixées au plafond, filment l'ensemble du terrain. Ces caméras sont utilisées afin d'extraire la position des joueurs, en calculant une carte d'occupation [18], à une fréquence de 25 images par secondes. Le contrôle de la caméra PTZ est basé sur le déplacement du centre de gravité des joueurs. Les résultats obtenus montrent l'efficacité de la méthode pour le contrôle de la caméra. Ren et al. [117] utilisent les images de huit caméras, placées à des endroits appropriés autour du stade, afin de détecter et suivre les joueurs et le ballon. Pour chaque caméra, la position des joueurs est obtenue en soustrayant aux images un modèle de fond, obtenu par mélange de gaussienne [129]. L'utilisation d'un réseau de caméras réduit les occlusions et permet un suivi plus efficace du ballon et des joueurs. Cependant, du fait de l'utilisation de 8 caméras, cette méthode est cependant coûteuse en terme de matériel et de temps de calcul.

Afin de suivre l'action, nous avons décidé de ne pas utiliser la balle comme référence, mais le centre de gravité des joueurs. En effet, nous supposons que la position globale des joueurs est

représentative de la position de l'action. Du fait que cinq joueurs soient en attaque et que les cinq autres joueurs en défense, la moyenne des positions devrait être le centre de l'affrontement entre les deux équipes. Nous proposons donc de piloter la sélection des caméras latérales par l'analyse du déplacement du centre de gravité.

Le principe de notre méthode est illustré figure 4.4 où chaque processus est exécuté séquentiellement. L'objectif est d'extraire la position de chacun des joueurs afin de pouvoir calculer le centre de gravité moyen de la position de l'ensemble des équipes. Pour ce faire, nous utilisons la caméra azimutale qui permet d'obtenir la vision globale du terrain. La position du centre de gravité est ensuite utilisée pour sélectionner la caméra latérale filmant la zone localisée du centre de gravité. Afin d'utiliser cette méthode dans toutes les situations, une étape d'initialisation permet de définir les zones du terrain de basket que chaque caméra cible.

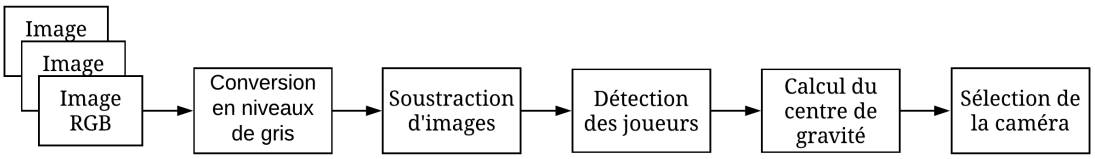


FIGURE 4.4 – Description de la méthode

4.3.1 Extraction de la position des joueurs

Afin de pouvoir calculer le centre de gravité, il est nécessaire d'extraire la position des différents joueurs présents sur le terrain. Pour effectuer cette extraction, deux catégories de méthodes peuvent être utilisées : les méthodes par soustraction de fond et les méthodes par soustraction d'images. La première catégorie s'intéresse à extraire la position des joueurs en calculant la différence entre une image et un modèle d'arrière plan. La seconde se base sur la soustraction d'images temporellement adjacentes afin de mettre en évidence le déplacement des joueurs à chaque instant.

La modélisation statistique la plus simple pour séparer l'arrière-plan est le modèle par simple gaussienne, où la fonction de densité de probabilité d'un pixel suit une loi gaussienne [140]. Cette modélisation par une gaussienne est une méthode pertinente dans le cas où l'arrière-plan n'a qu'une faible variabilité. Cette méthode se révèle cependant peu efficace dans le cas de fonds à fortes dynamiques et reste principalement dédiée aux scènes intérieures. Pour résoudre ce problème, le modèle par mélange de gaussiennes, au niveau du pixel est proposé par [129]. Le système est robuste aux faibles changements de luminosité en adaptant les valeurs des Gaussiennes. Dans le cas d'environnement dynamique, l'arrière-plan ne peut être modélisé efficacement avec peu de distributions Gaussiennes ce qui implique un temps de calcul important. Des méthodes floues ont également été utilisées pour modéliser l'arrière-plan [146, 124, 87]. Ces techniques proposent une modélisation de l'arrière-plan au niveau des pixels, basée sur l'utilisation d'histogrammes flous. Un des avantages du partitionnement flou réside dans le fait que le choix du nombre de partitions de l'histogramme influence peu la sensibilité comparé à une technique traditionnelle. Ces méthodes permettent une soustraction de fond moins bruitée, dans le cas de scène à forte dynamique, que les méthodes par mélange de gaussiennes. Cependant, les temps de calculs nécessaires à la modélisation de l'arrière plan rendent ces méthodes inutilisables pour une application

en temps réel.

Les méthodes par soustraction d'images se basent sur le fait que les différences entre deux images successivement sont liées aux déplacements des objets dans une scène. Pour ce faire, une soustraction élémentaire entre l'image à l'instant t et celle d'une image de référence. Cette différence D_t est calculée grâce à l'équation suivante : $D_t = |I_t - I_{t-k}|$, où I_t est l'image à l'instant t et I_{t-k} est l'image à l'instant t-k avec k supérieur ou égale à 1. Si la valeur d'un pixel de l'image D_t est supérieur à un certain seuil, alors ce pixel fait partie d'un objet mobile. Dans le cas contraire, le pixel est annoté comme arrière-plan.
Dans [37], les auteurs utilisent la différence entre trois images successives pour détecter les changements dans l'image. La différence entre les images aux instants t et t-1 et celle entre les images t et t+1 sont calculées. L'opérateur logique ET est utilisé, afin de réunir ces deux images des différences, et ainsi d'obtenir les changements dans l'image à l'instant t. De plus, un filtre médian est appliqué afin de réduire le bruit de type "poivre et sel". Makandar et al [90] cherchent à extraire la position de joueurs dans un match de Kabaddi. Leur méthode est constituée de deux étapes. La première consiste à sélectionner les images clefs, c'est à dire les images présentant le plus de différences avec les images temporellement adjacentes, dans toute la séquence vidéo. Une fois ces images sélectionnées, les positions des joueurs sont extraites en soustrayant à chaque image, l'image clef correspondante. Les résultats obtenus sont ainsi moins impactés par les changements d'illumination dans la scène. Cependant, la sélection des images clefs implique une utilisation en différée.

Les méthodes par soustraction d'images ont l'avantage d'être plus rapide que les méthodes basées sur la modélisation de l'arrière-plan, mais sont plus sensibles au bruit. Ces méthodes sont donc plus adaptées aux scènes avec un arrière-plan statique, avec peu de changement d'illumination. Le principal inconvénient des méthodes par soustraction d'images est qu'un objet ayant un déplacement trop faible entre deux trames risque de ne pas être détecté.

Dans notre cas, la scène que nous étudions est contrôlée : terrain intérieur et éclairé artificiellement. Ainsi, peu de variations se présentent tout au long d'une rencontre. De plus, souhaitant réaliser une diffusion en direct, il est nécessaire que l'extraction de la position des joueurs soit la plus rapide possible. C'est pourquoi, nous privilégions une méthode de soustraction d'images pour respecter ces contraintes temporelles.

Du fait que nous souhaitons extraire la position de l'ensemble des joueurs et non de joueurs spécifiques, seuls les informations d'espace sont nécessaires. De ce fait, l'information de couleur n'est pas utile. C'est pourquoi nous utilisons des images en niveaux de gris pour extraire les positions des joueurs. De plus, un des inconvénients des méthodes par soustraction d'images étant la détection d'objet avec un déplacement trop faible, nous choisissons de calculer la différence entre les trames aux instants t et t-4.

Pour valider notre choix, nous avons comparé les temps de calcul de la soustraction d'images avec ceux des méthodes de soustractions d'arrière-plan.
Comme le montre le tableau 4.1, la soustraction d'image est efficace, en terme de temps de calcul, par rapport à une méthode de soustraction d'arrière-plan.

La figure 4.5 montre le résultat d'une soustraction d'images dans le cadre d'un match de basketball. Puisque des artefacts peuvent apparaître pendant la soustraction, nous appliquons une analyse en composantes connexes [45] afin d'exécuter un seuillage dimensionnel. Les détections qui sont inférieures à la taille d'un joueur (comme les reflets ou la balle) sont ignorées. En d'autres termes, nous ne conservons que les objets ayant une surface rectangulaire supérieure ou

Méthodes	FPS
Soustraction d'images	**136**
Mélange de gaussienne [70]	43
Mélange de gaussienne [149]	58
Modélisation floue [146]	5

TABLE 4.1 – Comparaison des méthodes de soustraction d'images et de soustraction d'arrière-plan

égale à 400 pixels. Cette taille de zone est choisie empiriquement en fonction de nos conditions d'utilisation.

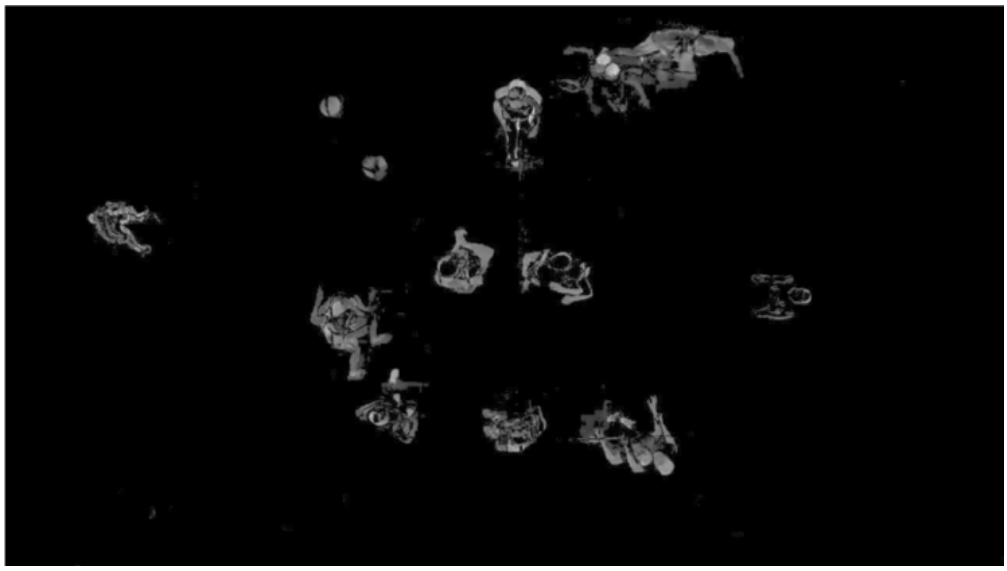
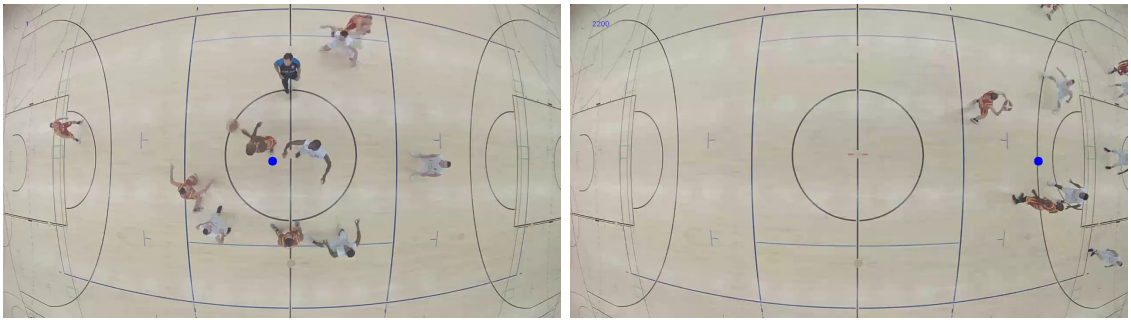


FIGURE 4.5 – Soustraction d'images entre l'image t et l'image t-4

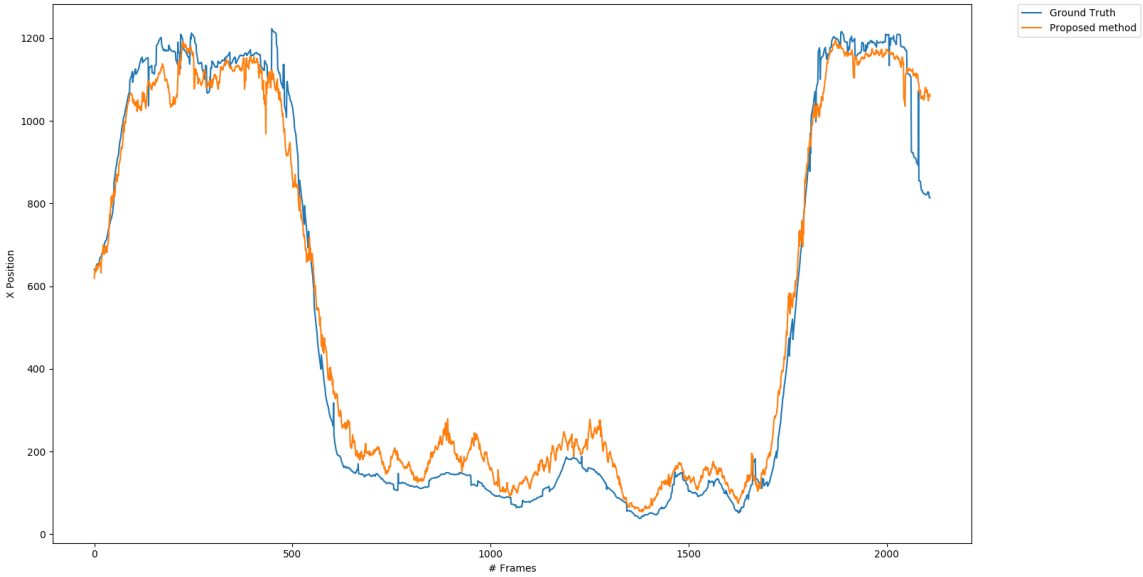
4.3.2 Extraction du centre de gravité

Pour valider nos hypothèses, nous avons comparé le déplacement du centre de gravité à une vérité terrain qui a été obtenue en calculant le centre de gravité de tous les joueurs, dont la position a été annotée manuellement. Après avoir analysé le déplacement du centre de gravité de la vérité terrain, nous obtenons la confirmation que son évolution dans le temps pourrait fournir des informations utiles pour la sélection de la caméra.
Dans le cas d'un match de basket, presque tous les joueurs bougent en même temps que le ballon (voir Fig. 4.6a). La position des joueurs est donc représentative de l'action. Il est possible que certains joueurs restent en arrière en attendant le retour de l'action. Si c'est le cas, leurs déplacements seront lents, et ne seront pas détectés par la soustraction d'images. Une autre catégorie de personne d'intérêt est également présente sur le terrain : les arbitres. Leurs rôles étant de contrôler le bon déroulement de la rencontre, ils se placent de manière générale à proximité du jeu notable. La détection de ces POI ne perturbe donc pas, voire renforce la localisation de l'action d'intérêt principale.

Le centre de gravité des joueurs correspond à la moyenne des positions de chaque joueur, pondérée par leur surface. En effet, lorsque plusieurs joueurs sont proches (par exemple au



(a) Centres de gravité (en bleu) pour l'image t=1 (à gauche) et t=1950 (à droite)



(b) Comparaison de l'évolution du centre de gravité avec la vérité terrain

FIGURE 4.6 – Évolution du centre de gravité

niveau du panier), il est possible que certains joueurs ne soient pas suffisamment éloignés pour être détectés séparément. Cependant la surface de la forme détectée augmente, symbolisant la présence de plusieurs joueurs. Une comparaison de l'évolution du centre de gravité calculé selon notre méthode avec la vérité terrain peut être vue dans la figure 4.6b. Cette vérité terrain a été obtenue en calculant le centre de gravité de tous les joueurs, dont la position a été annotée manuellement, dans une vidéo de 2110 images. On remarque que le centre de gravité obtenu par notre méthode est comparable à celui de la vérité de terrain.

4.3.3 Sélection de la caméra d'intérêt

Puisque nous avons 3 caméras, situées en bordure du terrain, nous avons défini trois zones dans l'image azimutale, comme nous pouvons le voir dans la figure 4.7. La proposition de base est que l'on change de caméra dès que les joueurs passent d'une zone à l'autre.

Cependant lorsque l'action se situe dans une zone inter caméras, de nombreux changements sont provoqués. Par exemple, lorsqu'une équipe tente de marquer un panier, il arrive souvent que les joueurs reculent pour espacer le jeu. Ce mouvement vers l'arrière peut entraîner un change-

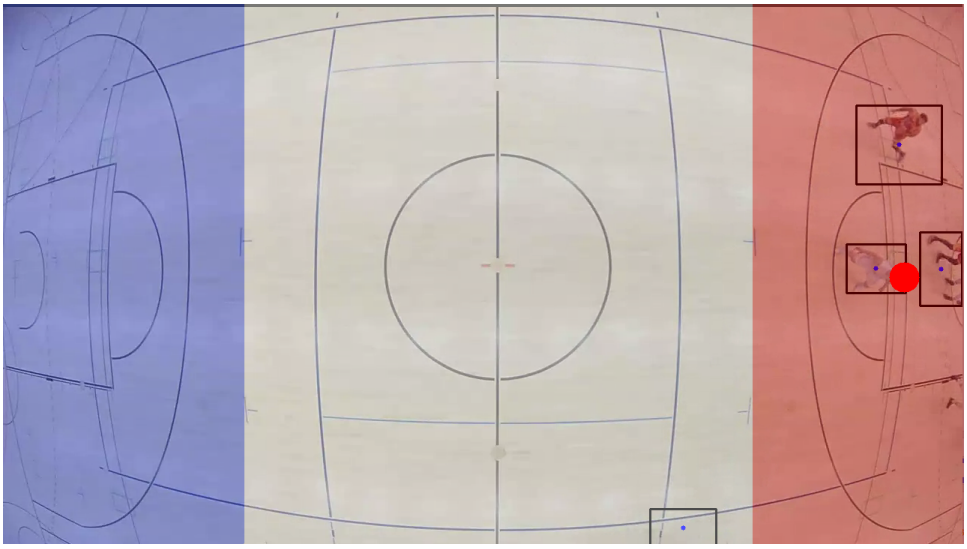


FIGURE 4.7 – Définissions des trois zones correspondant aux trois caméras

ment de caméra, bien que l'action se déroule sous le panier. Pour pallier ce problème, et assurer une plus grande stabilité dans la diffusion, nous proposons d'ajouter une fonction d'hystérésis. Cette fonction maintient la caméra près du panier jusqu'à ce qu'un grand nombre de joueurs soit revenu du côté de l'adversaire, comme nous pouvons le voir figure 4.8.

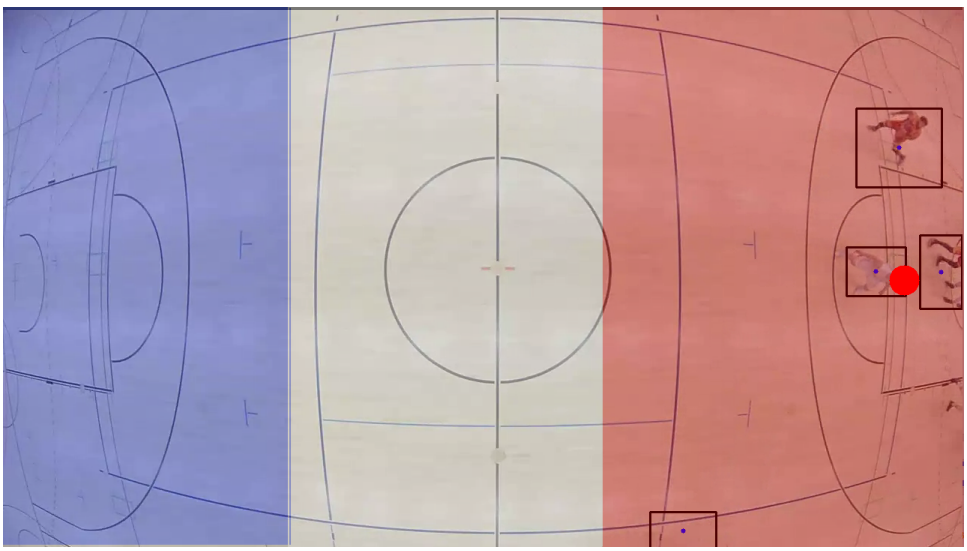


FIGURE 4.8 – Sélection de la vue grâce à une fonction hystérésis

Nous pouvons voir dans la figure 4.9 une comparaison entre trois montages différents pour une séquence de 100 secondes. La courbe bleue représente la sélection de caméras effectuée à partir de la position du centre de gravité de la vérité terrain. La courbe en vert correspond à l'exploitation de la méthode que nous proposons. Enfin la courbe orange correspond à la sélection effectuée par un opérateur.

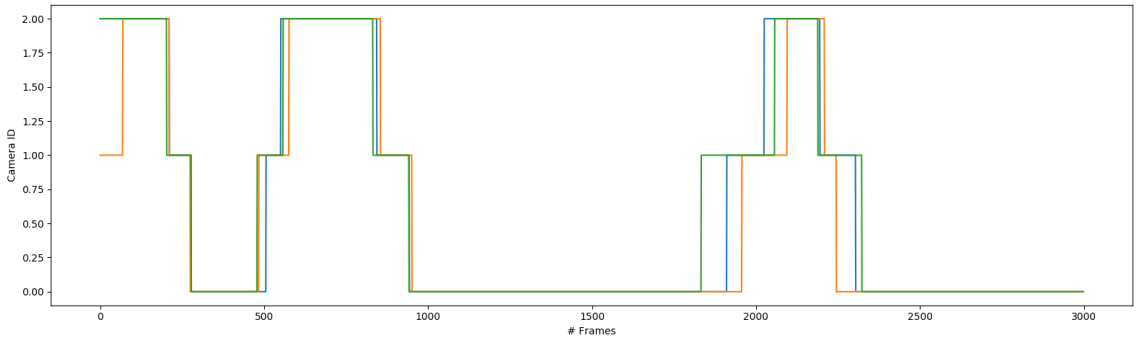


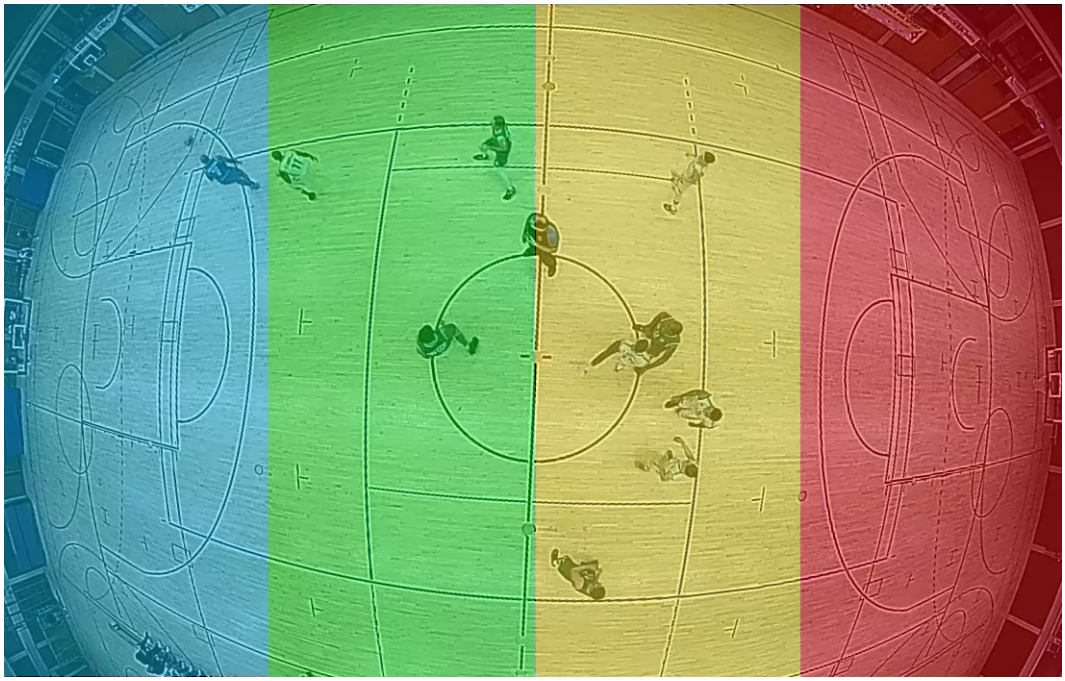
FIGURE 4.9 – Sélection de caméra en fonction du déplacement du centre de gravité (Bleu : vérité terrain, Vert : sélection automatique, Orange : sélection manuelle

Notre méthode permet la génération d'un flux monté proche d'un montage réalisé manuellement. Les flux vidéo générés par notre méthode et obtenus par l'opérateur sont similaires à 86,2%. Les différences proviennent principalement du fait que l'humain aura tendance à anticiper ou retarder le changement de caméra. Cependant, les différences en terme de temps sont plutôt faibles : en moyenne, la sélection d'une caméra par notre méthode est effectuée 23 images (soit 0.76s) avant celle effectuée manuellement. Globalement, les caméras sélectionnées sont les mêmes au cours du temps. Ainsi les spectateurs regarderont les mêmes actions que ce soit avec un montage manuel ou automatique. Enfin, le temps de traitement par image est d'environ 7-8 ms (130 fps), ce qui est compatible avec la diffusion en direct.

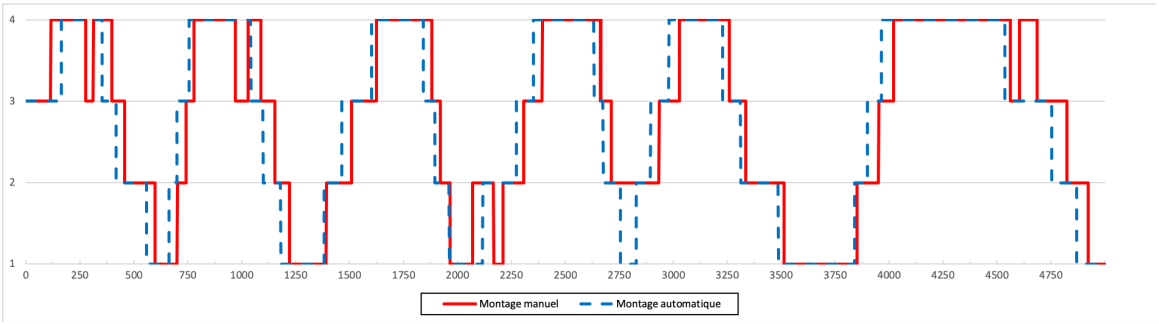
La figure 4.10 présente une extension de la méthode à un cas où 4 caméras sont disponibles. Afin de permettre la sélection du flux d'intérêt, le seul changement dans notre méthode est de redéfinir les zones correspondant aux caméras (figure 4.10a). Bien que la caméra supérieure ait un angle de vue plus important, aucun autre paramétrage supplémentaire n'est nécessaire.

La figure 4.10b présente la comparaison entre un montage manuel et le montage obtenu par notre méthode pour une séquence de jeu de 3 minutes. Nous pouvons remarquer que les caméras sélectionnées sont similaires. Le montage automatique a tendance à effectuer la transition en moyenne 27 images (0,9 secondes) avant le montage manuel. Cette différence peut s'expliquer par la volonté du monteur de conserver la vue actuelle plus longtemps. Nous pouvons également remarquer que le monteur effectue de brefs changements de caméra (à i=276, i=970 et i=2165). Ces changements s'expliquent par le fait que le monteur focalise son attention sur le porteur du ballon. Ces changements étant relativement courts (respectivement 35, 72 et 44 images), le changement n'est pas forcément agréable pour le spectateur. Le montage automatique délivre ainsi un flux vidéo qui apparaît plus stable.
Notre méthode montre des limites dans le cas de lancers-francs (images 4500 à 4700 de la figure 4.10b). En effet seule une partie des joueurs est impliquée par la réalisation d'un lancer-franc. Les joueurs restants ont tendance à reculer vers le centre du terrain. De ce fait, le centre de gravité change de zone, provoquant un changement de caméra. Nous proposons une solution palliant ce problème en détectant spécifiquement ces actions (voir section 4.4)

Nous avons donc proposé une méthode permettant de sélectionner la caméra où l'action d'intérêt se produit. Cette méthode s'appuie sur la détection de la position des joueurs à chaque



(a) Définition des quatre zones caméras



(b) Comparaison entre un montage manuel (rouge) et le montage automatique (bleu)

FIGURE 4.10 – Extension de la sélection à 4 caméras

instant t. Le centre de gravité de la position des joueurs des deux équipes est représentatif du jeu notable dans le match de basketball et nous permet de sélectionner la caméra présentant l'action. De plus, notre méthode est facilement adaptable en fonction du nombre de caméras : il est juste nécessaire de redéfinir le nombre et la position des zones correspondant aux caméras. Enfin, la complexité calculatoire étant faible, il est possible d'effectuer la sélection automatique de caméra en temps-réel, permettant la diffusion en direct de l'évènement.

4.4 Détection de l'AOI "lancer-franc"

Comme nous l'avons vu, le cas spécifique des périodes de lancers-francs peuvent influer sur la qualité du montage automatique du jeu notable. Ainsi nous proposons dans cette partie de détecter une AOI particulière dédiée aux lancers-francs. La détection de cette AOI entre également dans le cadre de la personnalisation du montage vidéo pour le spectateur. L'objectif est ainsi de proposer la diffusion en différé d'un flux monté ne montrant que les lancers-francs du

match, voire uniquement ceux de son équipe préférée.

Dans la littérature, la détection de lancers-francs est rarement abordée. Il n'existe, à notre connaissance, aucune méthode de détection de lancer-franc pour la diffusion en direct. Chen et al [22] réalisent la détection de lancer-franc en analysant le tableau des scores. Une équipe a 24 secondes pour marquer un panier lorsqu'elle récupère la balle. Ce temps est indiqué sur le chronomètre des tirs. Si ce chronomètre s'arrête, il peut s'agir d'une faute, suivi d'un lancer-franc. Si le score est incrémenté de 1 point, alors un lancer-franc est détecté. Cette méthode ne peut cependant pas détecter un lancer-franc manqué. Ramanathan et al. [116] s'intéressent à la détection d'évènements et à des joueurs clés dans un match de basketball. Ils utilisent un réseau de neurones récurrents (RNN) afin de classifier 11 évènements différents dont la "réussite d'un lancer-franc" et l'"échec d'un lancer-franc". Les résultats obtenus, en terme de précision, sont de 0.81 pour les lancers-francs réussis et de 0.47 pour les lancers-francs manqués. Cependant le coût calculatoire, ainsi que le faible taux de reconnaissance de cette méthode ne permettent pas de l'envisager dans notre scénario. La reconnaissance de lancer-franc est également réalisée pour de l'indexation de vidéo de match de basketball. Certaines méthodes s'appuient sur le son et le type de prise de vue [141], d'autres utilisent des informations textuelles [147]. Dans notre cas, nous ne pouvons obtenir de telles informations. Il est donc nécessaire de mettre en place une nouvelle méthode de détection de lancer-franc.

4.4.1 Intégration de connaissances

Selon les règles du basket [44], un lancer franc est un moment particulier d'un match de basket où un joueur à le droit d'effectuer "un tir au panier, sans opposition, à partir d'une position située derrière la ligne de lancer-franc et à l'intérieur du demi-cercle". Ces lancers sont accordés lorsqu'une faute personnelle, une faute antisportive ou une faute disqualifiante est sifflée. En fonction de la faute, un, deux ou trois lancers peuvent être accordés.

Un lancer franc s'accompagne de plusieurs caractéristiques précises sur la position que doivent avoir les joueurs pendant toute sa durée. La figure 4.11 résume les positions que les joueurs doivent respecter jusqu'à la fin du lancer franc :

- le tireur de lancer-franc *(A3)* doit se placer derrière la ligne de lancer franc, à l'intérieur du demi-cercle ;
- les joueurs occupant les places de rebond *(A1,A2, B1, B2, B3)* doivent prendre des positions alternées dans ces places qui sont considérées comme ayant une profondeur d'un mètre ;
- les joueurs qui ne sont pas dans les places de rebond *(A4, A5, B4, B5)* doivent rester derrière la ligne de lancer-franc prolongée et derrière la ligne de panier à 3 points jusqu'à ce que le lancer-franc ait pris fin.

Lors d'un lancer-franc, les joueurs doivent donc se trouver dans les zones blanches de la figure 4.11 pendant toute la durée du lancer franc. Par extension, aucun joueur ne peut se trouver dans la zone grise.

4.4.2 Méthodologie de détection de lancer-franc

En s'appuyant sur les connaissances que nous avons extraites du règlement, il est nécessaire d'extraire des attributs de position (Espace) et de temps. La détection d'un lancer franc peut se réaliser en analysant la position des joueurs. En effet, 6 joueurs sur 10 ont une position définie

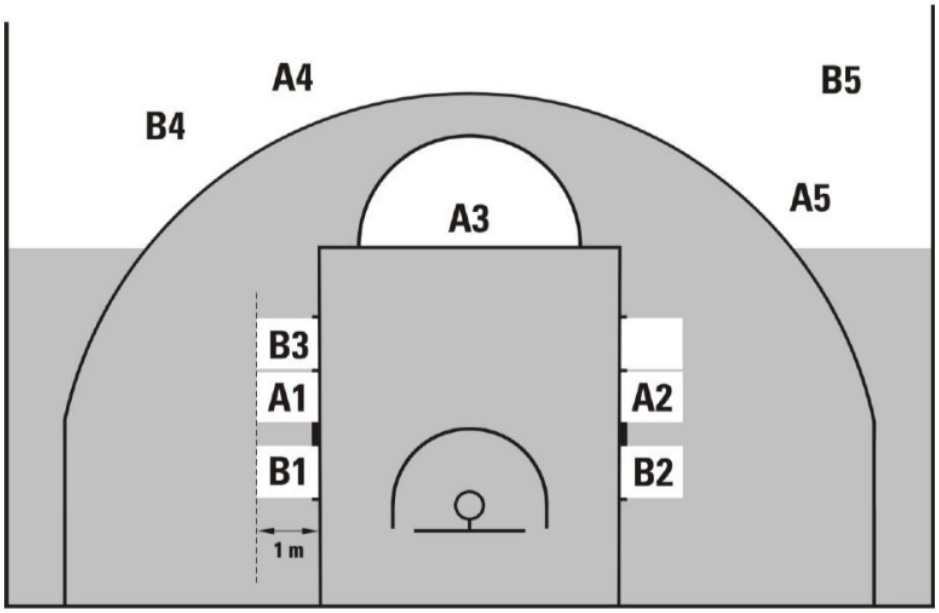


FIGURE 4.11 – Position des joueurs pendant les lancers-francs [44]

dans l'espace. Les 4 autres joueurs se trouvent dans une zone large, située derrière la ligne de lancer-franc. L'hypothèse que nous posons est que la détection de joueurs immobiles dans ces zones pendant un délai suffisant, définit une situation de lancer franc.

La méthode de localisation de joueurs que nous avons proposée dans la section 4.3.2 est basée sur l'hypothèse de joueur en mouvement. Elle ne peut donc pas s'appliquer ici, où les joueurs sont essentiellement statiques. Il est alors nécessaire de mettre en place une méthode permettant une extraction précise des joueurs. Pour ce faire, une méthode de soustraction d'arrière-plan peut être mise en place. La méthode de modélisation de d'arrière-plan par mélange de gaussiennes proposé par [149] est ainsi utilisée pour extraire la position des joueurs.

En reprenant les positions attendues lors d'un lancer-franc, nous pouvons définir les zones correspondantes dans notre cas (figure 4.12). Nous pouvons voir en bleu, les zones pour les joueurs se trouvant en position de rebond. En rouge, nous trouvons la zone de la raquette ou aucun joueur ne doit se trouver et enfin en vert, la zone correspondant à l'emplacement du joueur réalisant le lancer-franc. Les zones jaunes aux extrémités servent à vérifier l'absence de joueur devant la ligne de lancer-franc. Comme nous utilisons une caméra fish-eye, les zones définies correspondent à l'endroit où le joueur est détecté et non à la position des pieds du joueur.

Une fois ces zones définies, la comparaison des positions des joueurs et des zones permet de mettre en évidence les situations de lancers francs.

Ainsi, un lancer franc a lieu lorsque les joueurs sont à une "bonne" position pendant un certain laps de temps. Par bonne position, nous entendons :

— un joueur dans la zone de lancer-franc ;
— plus d'un joueur en zone de rebond ;
— aucun joueur dans la raquette ;
— aucun joueur dans les zones jaunes définies précédemment.

Les règles énoncées de la FIBA, stipulent que cinq joueurs se trouvent dans les zones de rebonds. Cependant, la déformation de l'image ne garantit pas que les 5 joueurs soient détectés,

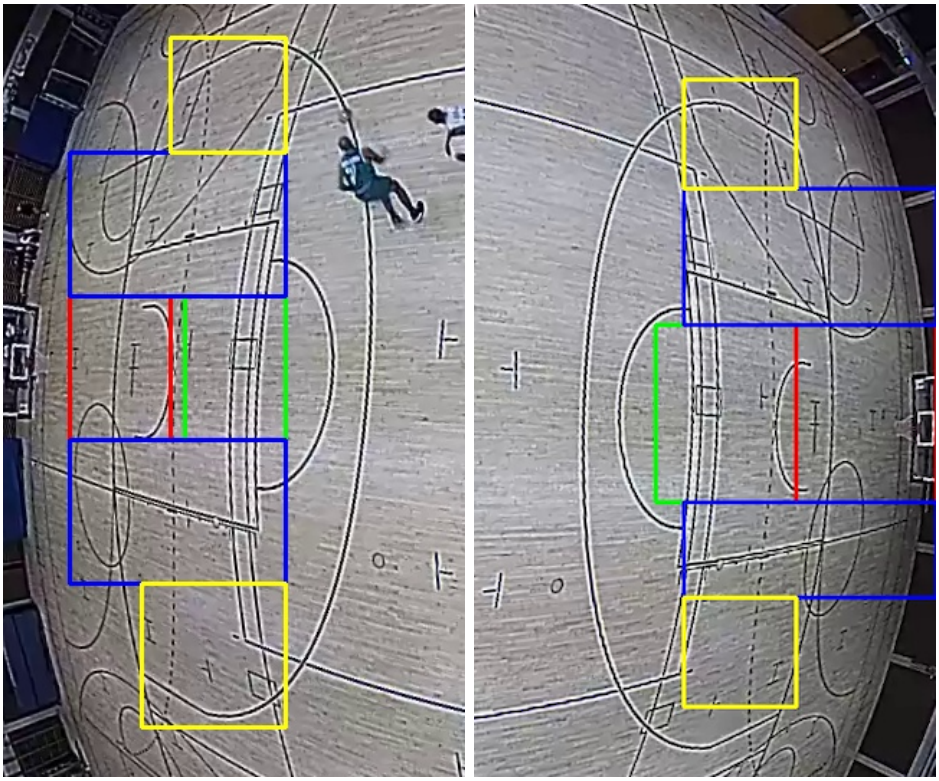


FIGURE 4.12 – Diffusion d'un match de basketball

certains joueurs peuvent se trouver en dehors de l'image. D'autres joueurs peuvent être masqués par les joueurs les plus proches de la zone de lancer-franc (les joueurs A2 et B3 dans la figure 4.11). La figure 4.13 présente un lancer franc correctement détecté. La couleur verte des zones montrent que les conditions sont respectées.

Une autre difficulté tient dans la position des arbitres. Ils se trouvent souvent au niveau de la raquette, afin de donner la balle au joueur effectuant le lancer-franc. Bien qu'ils sortent de la raquette avant le tir, les passages dans la raquette entre les tirs peuvent perturber la détection en continu d'une période de lancer-franc. De plus, il est nécessaire de sélectionner un laps de temps adéquat. Si le temps sélectionné est trop court, nous risquons de détecter des lancers-francs lorsque les joueurs essaient de marquer un panier et se retrouvent dans la configuration attendue pour un lancer-franc. Dans le cas contraire, il sera possible que de nombreux lancers-francs ne soient pas détectés.

4.4.3 Expérimentation

Afin de détecter un lancer franc, il est nécessaire que les joueurs se trouvent dans les positions définies pendant le temps du lancer. De ce fait, nous considérons la présence d'un lancer franc lorsque toutes les images sont reconnues comme lancer franc pendant un certain délai. Ce délai induit permet de ne pas considérer les passages rapides dans les zones définies comme faisant parti d'un lancer franc. Cependant, cette temporisation implique qu'un lancer franc sera détecté plusieurs images après son commencement réel. Il est alors nécessaire de choisir un délai permet-

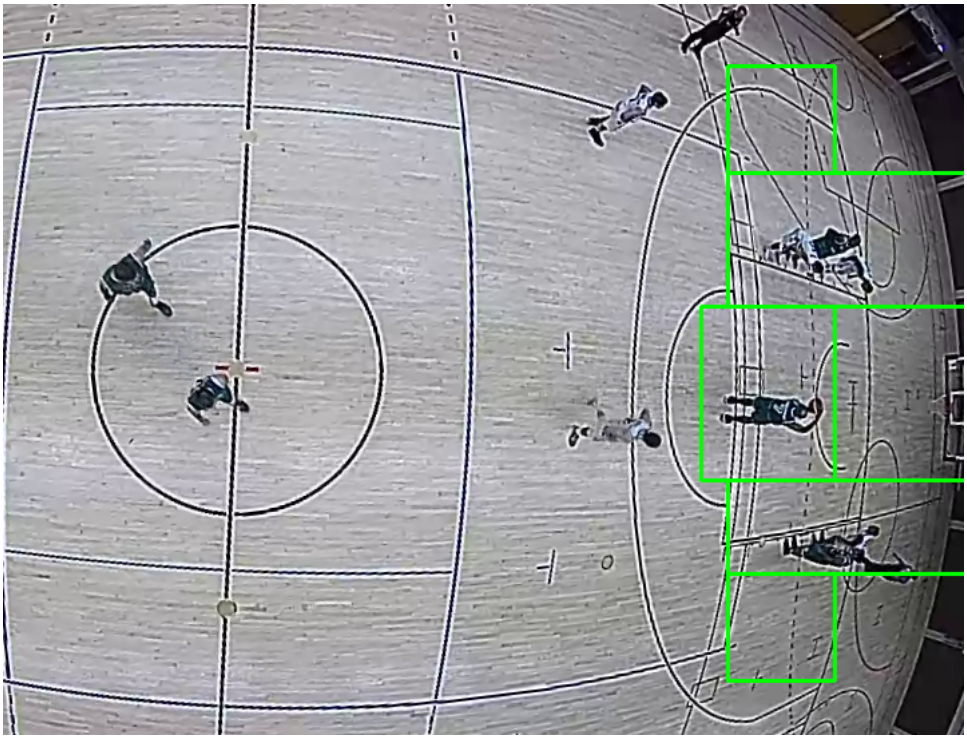
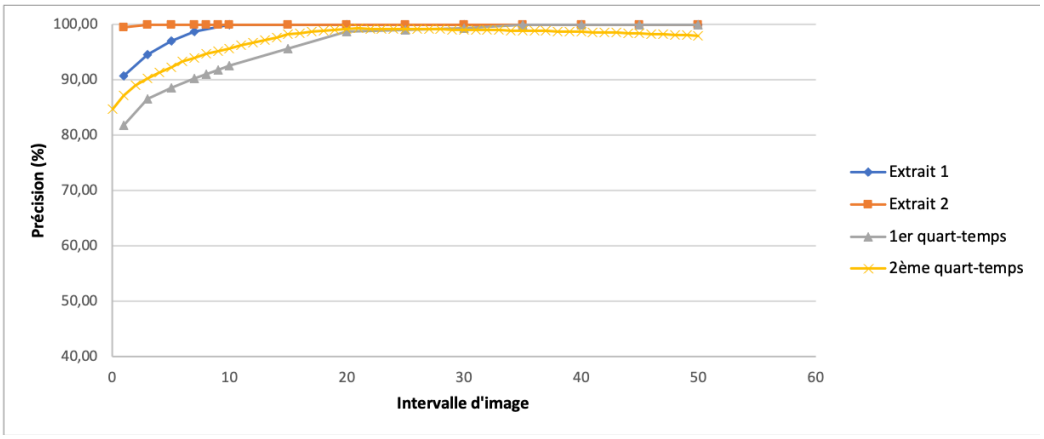


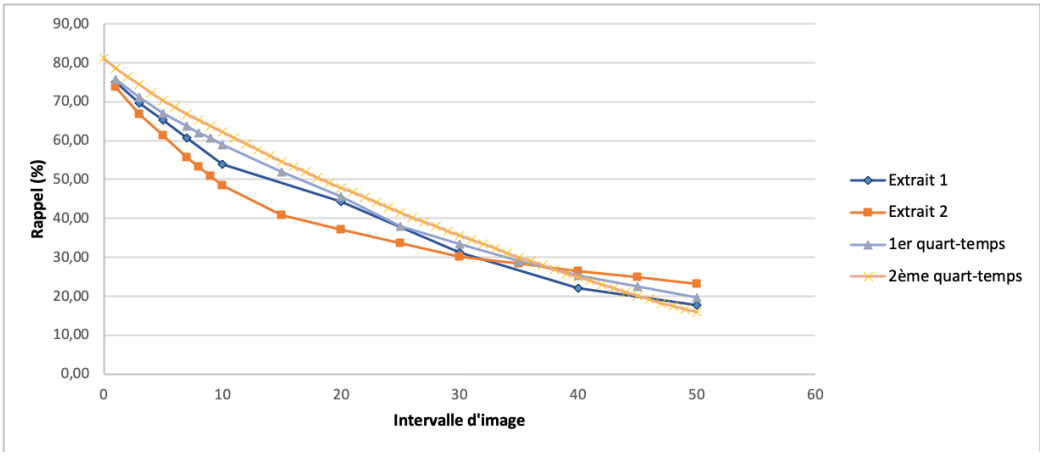
FIGURE 4.13 – Détection d'un lancer-franc

tant la diminution des faux positifs, tout en maximisant la détection des lancers francs.
La figure 4.14 présente les résultats obtenus, en terme de précision, de rappel et de F-mesure pour différents délais. Les extraits 1 et 2 sont des vidéos comprenant 1 ou 2 lancers-francs et ont une durée de 1 minute 30 (2250 images) pour le premier et 6 minutes (9000 images) pour le second. Les deux autres vidéos sont respectivement le premier quart-temps (15m18 - 22950 images) et le second quart-temps (16m42 - 25050 images) d'un match de basketball. Pour ces quatre extraits vidéos, les images contenant un lancer franc et celles n'en contenant pas ont été annotées manuellement. Les deux premiers extraits nous ont permis de mettre en évidence l'efficacité de la méthode pour la détection des lancers-francs. Les deux seconds nous permettent de mettre en évidence la robustesse de notre méthode par rapport aux faux positifs.

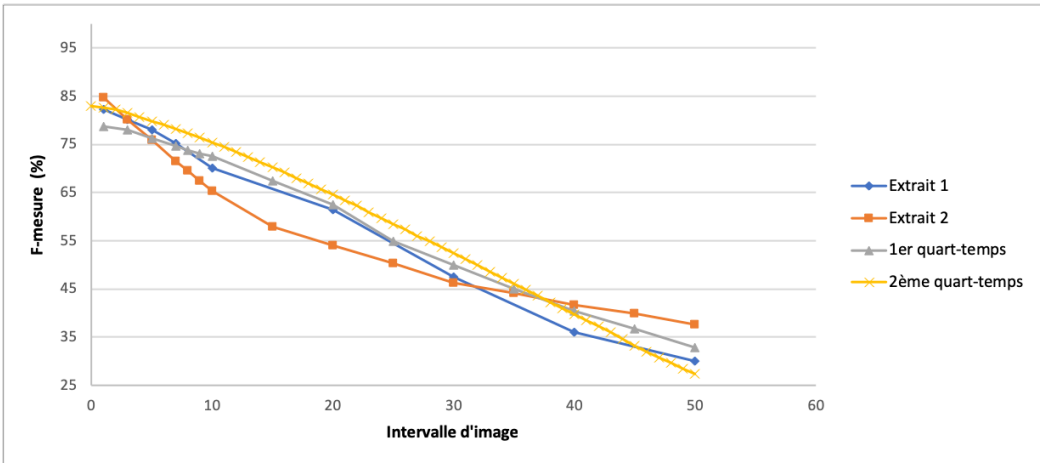
Nous pouvons remarquer que plus le délai est important, plus la précision augmente (figure 4.14a). Cependant, plus l'intervalle est grand, plus le rappel diminue (figure 4.14b). Le calcul de la F-mesure sur notre lot de données (Figure 4.14c) montre l'intérêt de privilégier un délai court pour annoter le flux vidéo comme contenant un lancer franc. Cette diminution s'explique du fait que les images nécessaires pour valider la présence de lancer-franc ne seront pas annotées "lancer-franc". Dans le cas où un délai de 50 images (2 secondes) est nécessaire pour valider un lancer-franc, les 50 premières images seront considérées comme des faux négatifs. De plus, si au cours du lancer-franc, une image n'est pas correctement détectée, 50 nouvelles images sont nécessaires pour annoter le lancer-franc.
Dans le cas d'une diffusion en différé, un lissage temporel peut être mis en place afin de ne pas prendre en compte les interruptions dans les lancers francs. De plus, il est possible de prendre en compte l'intervalle, et d'annoter la vidéo à partir de la première image détectée en tant que lancer-franc.



(a) Précision



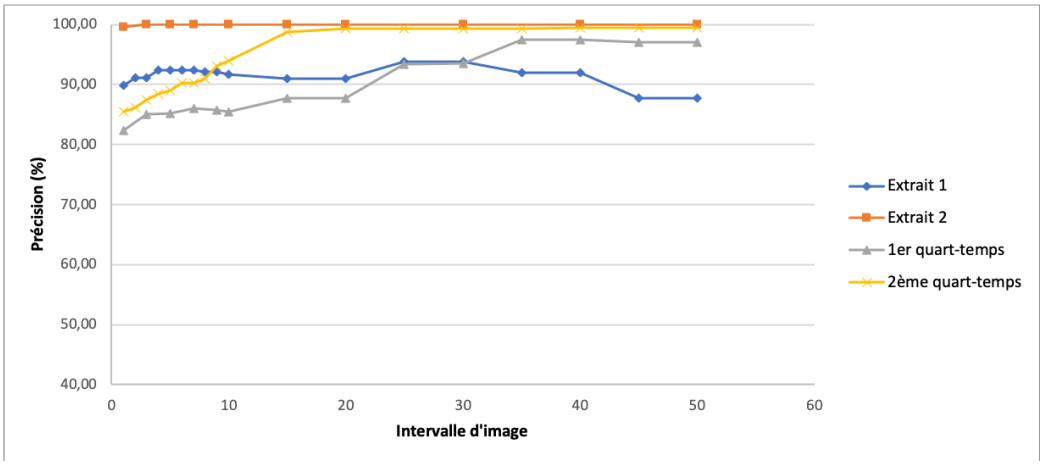
(b) Rappel



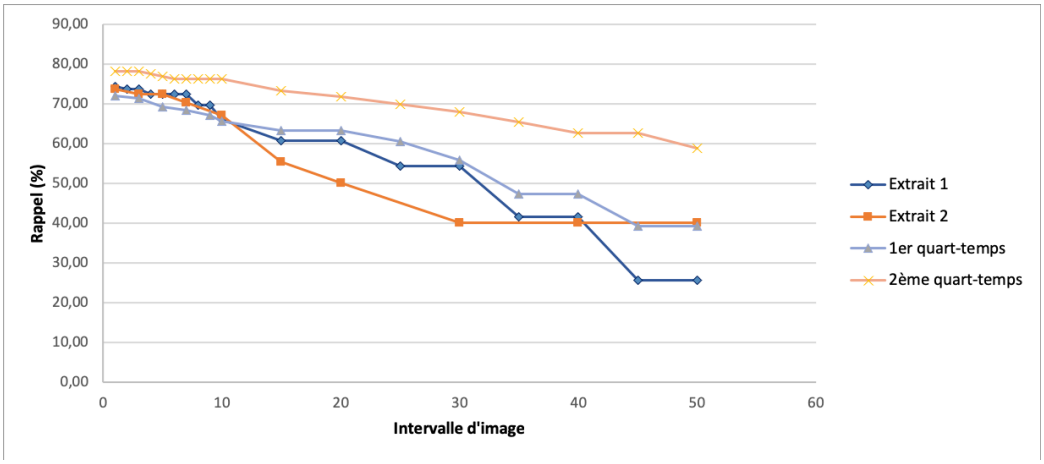
(c) F-mesure

FIGURE 4.14 – Précision, rappel et F-mesure pour différents valeurs de temporisation

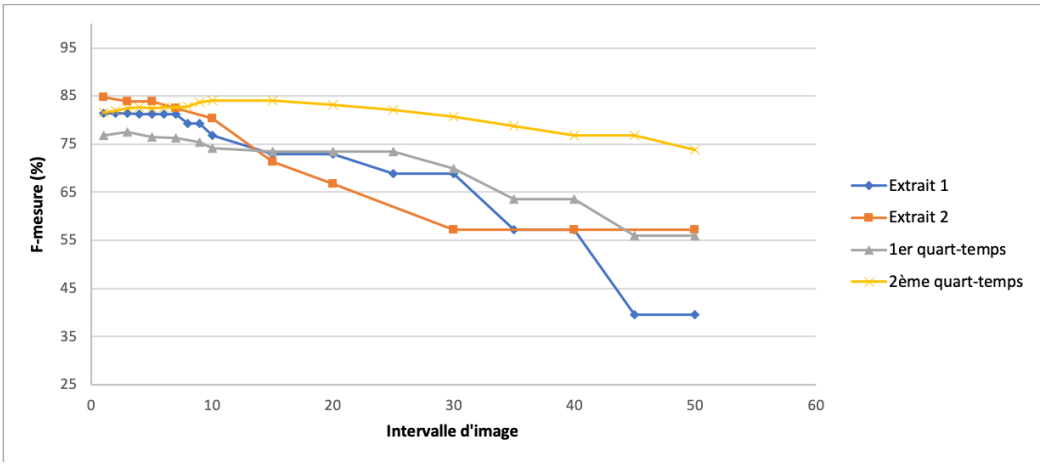
La figure 4.15 présente les résultats obtenus dans le cas d'une diffusion différée. Nous pouvons remarquer que les résultats sont supérieurs à ceux obtenus lors d'une diffusion en direct. Dans les



(a) Précision



(b) Rappel

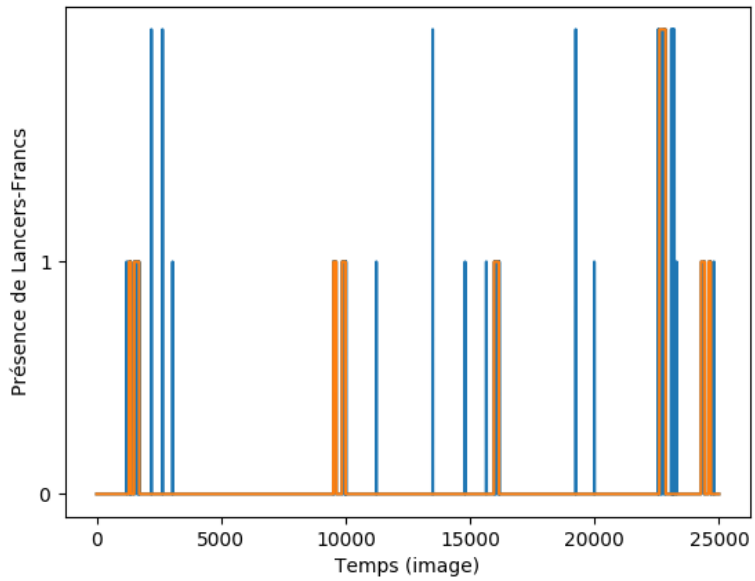


(c) F-mesure

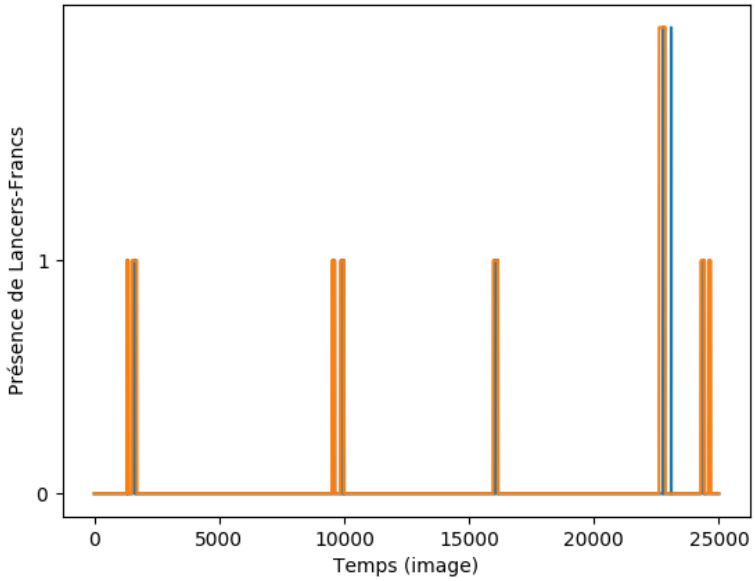
FIGURE 4.15 – Précision et rappel pour différentes valeurs de temporisation dans le cas d'une diffusion différée

deux cas, un intervalle de trames réduit présente de meilleurs résultats en termes de précision,

de rappel et de F-mesure. Afin d'éviter la présence de faux positifs, nous avons décidé d'utiliser un délai de 10 images afin de détecter correctement les périodes de lancers-francs.



(a) Détection des lancers-francs pour une intervalle de temps de 1



(b) Détection des lancers-francs pour une intervalle de temps de 10

FIGURE 4.16 – Détections (en bleu) des lancers-francs pour différents intervalles de temps comparés à la vérité terrain (en orange).

La figure 4.16 compare les détections pour un délai égal à 1 image (figure 4.16a) avec celles d'un intervalle égal à 10 images (figure 4.16b) pour le second quart-temps du match de basket-ball. Les lancers-francs du coté gauche du terrain sont annotés '1', ceux du côté droit '2'. Nous pouvons remarquer que le nombre de faux positifs diminue fortement avec une temporisation de 10 images. Concernant les faux négatifs, la figure 4.17 représente les lancers-francs détectés

à partir de l'image 9500. Nous pouvons remarquer quelques absences de détection sur certaines images (entourées en rouge) qui font augmenter le nombre de faux négatifs. Nous proposons de lisser cette réponse en ne considérant pas les absences de détections de moins de 3 images. Un lancer franc est ainsi considéré dans son ensemble.

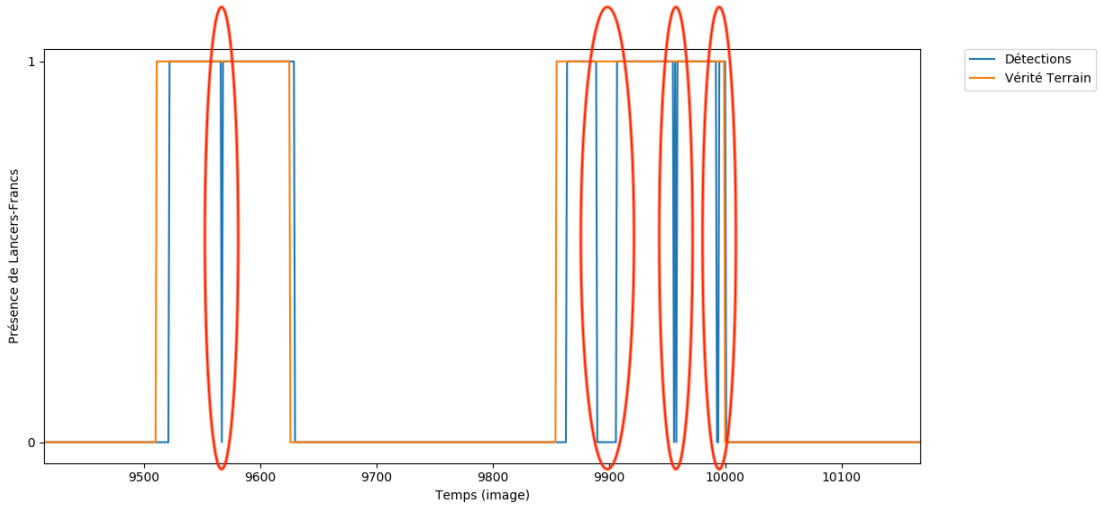


FIGURE 4.17 – Absences de détection de lancer-franc (faux négatifs) pour un intervalle de 10 images

La méthode proposée permet donc de détecter efficacement la présence d'un lancer-franc dans un flux vidéo de basket. Elle nous permet ainsi d'offrir aux spectateurs le montage automatique d'une diffusion personnalisée en visualisant, par exemple, un résumé des lancers francs. La configuration permet également de ne choisir que les lancers francs d'une équipe particulière, en utilisant les connaissances sur les couleurs des maillots. De plus, le temps de calcul moyen par image est de 20 ms par image, ce qui permet une annotation des lancers francs en temps réel.

4.5 Suivi des POI "joueurs"

La détection de lancers-francs correspond à proposer une première possibilité pour les spectateurs de pouvoir personnaliser les flux vidéos montés. Afin d'offrir d'autres choix de personnalisation, nous proposons dans cette partie de nous intéresser aux POI "Joueurs". En effet, les spectateurs peuvent vouloir regarder le match de basketball en s'intéressant à un joueur en particulier. De plus, les informations sur les joueurs présentent un intérêt pour des spectateurs experts : les entraîneurs. Afin de pouvoir améliorer les entraînements des joueurs, les coaches des équipes ont besoin d'obtenir un certain nombre de statistiques, comme par exemple la distance parcourue, les déplacements lors de certaines actions ou encore le nombre de paniers réussis. Proposer ces statistiques aux entraîneurs, en plus de la captation des matchs, permet de fournir une solution complète aux clubs sportifs.
Afin de sélectionner la caméra contenant le joueur d'intérêt, et d'extraire ses statistiques, il est alors nécessaire de pouvoir identifier le joueur à chaque instant. Pour ce faire, différentes méthodes peuvent être mises en place. Certaines de ces méthodes sont similaires à celles présentées dans le chapitre 3, comme par exemple la reconnaissance faciale. D'autres méthodes sont spécifiques au contexte des évènements sportifs comme la reconnaissance des maillots [36], ou l'utilisation

de dispositifs transmettant les informations par radio-fréquences [53, 46]. Cependant, du fait de l'aspect dynamique des scènes sportives, l'identification des joueurs en continu, à l'aide de méthode se basant sur les caméras, est une tâche difficile, dont les difficultés principales sont les suivantes.

— *Les occultations* : Pendant le suivi, les joueurs peuvent être cachés pendant un certain temps et il peut être difficile d'associer les formes détectées dans une image aux bons joueurs. Deux sous-cas peuvent être identifiés : lorsque deux joueurs se croisent et lorsqu'un joueur en cache un autre pendant une phase statique.
— *Les rotations* : L'apparence des joueurs change en fonction de la façon dont ils tournent et de l'endroit où ils se trouvent. L'apparence d'un joueur est différente en fonction de l'angle de vue
— *Les accélérations* : Le déplacement rapide des joueurs cause des changements importants entre des images successives. Les changements de vitesses fréquents dans les matchs de basket rendent le suivi des joueurs difficile.
— *Les groupes* : La plupart des situations au basketball n'impliquent que deux joueurs, à l'exception de certaines phases comme par exemple le début de la rencontre, les blessures, les célébrations d'un panier ou encore la fin du match. Dans ces situations, beaucoup de joueurs similaires se tiennent les uns à côté des autres, rendant difficile de différencier les joueurs.

Ces différentes difficultés montrent les problèmes que les méthodes de suivi peuvent rencontrer dans le cadre de rencontre de basketball. En raison de ces contraintes majeures, il est alors nécessaire de développer une solution spécifique au lieu d'utiliser des algorithmes génériques afin de rendre le suivi plus précis et plus rapide. Dans cette partie, nous proposons une méthode originale de suivi de joueurs [84]. Nous la comparons ensuite aux méthodes de référence pour ce domaine.

4.5.1 Méthodes de suivi de personne

De nombreuses méthodes de suivi dans le domaine du sport ont été proposées dans la littérature [91]. Nous nous intéressons dans cette étude aux algorithmes de suivi implémentés dans la librairie OpenCV. Ces algorithmes sont Boosting, KCF, MedianFlow, Multiple Instance Learning (MIL) et Tracking-learning-detection. Il s'agit de méthodes régulièrement utilisées dans les problématiques de suivi de personnes. Puisque les trackers MedianFlow [71] et MIL tracker [6] ne sont pas adaptés à notre application (déplacement aléatoire, rotation rapide), notre étude se focalisera sur les autres méthodes. Une comparaison plus exhaustive est proposée par Janku et al. [68].

— La méthode de Boosting est basée sur l'algorithme AdaBoost, qui utilise le fond environnant comme exemples négatifs afin de trouver les caractéristiques les plus discriminantes de l'objet suivi. Comme elle est basée sur l'apparence, les changements des joueurs tels que les rotations ou les changements légers sont de manière générale correctement pris en compte [50].
— L'objectif principal des filtres de corrélation kernelisés (KCF) est de distinguer la cible de l'environnement. Cet algorithme traduit et met à l'échelle différents patchs afin de trouver le meilleur. Cette méthode présente l'avantage d'avoir des temps de calcul relativement faibles grâce à certaines améliorations calculatoires [58, 59].

— La méthode *Tracking-Tracking-learning-detection (TLD)* repose sur la détection d'un objet et de son voisinage proche, puis de la détection de l'objet dans les cadres suivants. En fonction de la présence ou non de l'objet, le cadre est mis à jour [71]. Cet algorithme est censé être capable de traiter les mouvements rapides, ainsi que les occlusions partielles.

KCF, Boosting et TLD sont les méthodes les plus appropriées pour notre étude, c'est pourquoi nous proposons de comparer les résultats de notre solution à ces algorithmes.

4.5.2 Présentation de la méthode

Notre solution, illustrée figure 4.18, est implémentée en Python en utilisant la librairie OpenCV. Cette méthode est basée sur l'utilisation de caractéristiques propres aux joueurs à savoir : leur position dans l'image précédente, la vitesse et l'orientation de leur déplacement, ainsi que la couleur de maillot qu'ils portent.

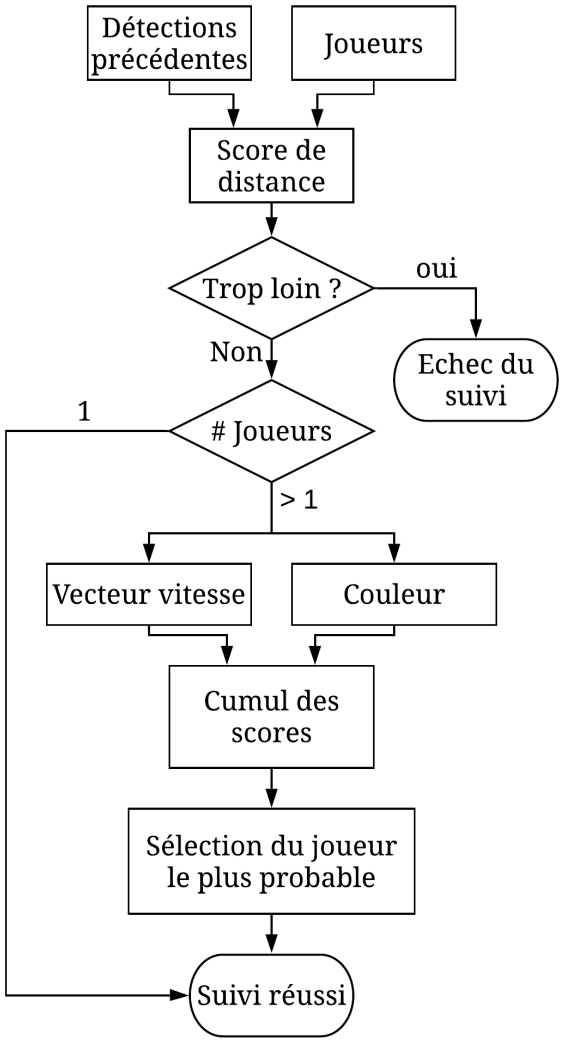


FIGURE 4.18 – Principe de fonctionnement de la méthode proposée pour le suivi des joueurs

Afin de pouvoir calculer ces caractéristiques, il est nécessaire de pouvoir extraire la position

de chaque joueur à chaque image. La caméra azimutale que nous utilisons pour la détection des joueurs possédant un objectif panamorphique, une première étape corrige la distorsion induite.

La seconde étape consiste à extraire les objets en mouvement en utilisant une modélisation de l'arrière plan. La soustraction de ce modèle à l'image courante permet de mettre en avant une zone où un déplacement de joueur s'est produit.

Afin d'améliorer l'image obtenue par la soustraction, des transformations morphologiques de fermeture et d'ouverture sont réalisées.

Il est alors possible d'extraire les positions des joueurs de l'image nettoyée à l'aide d'un algorithme d'analyse en composantes connexes [131].

Les différents objets sont ensuite reliés aux détections précédentes en fonction de critères tels que la surface, la distance, la direction du vecteur vitesse et la couleur principale.

Estimation de la distorsion

L'utilisation des images brutes fournies par la caméra n'est pas directement possible puisqu'elle utilise un objectif panamorphique. Comme la distorsion n'est pas radiale, il n'existe pas de moyen facile de la corriger sans avoir accès aux caractéristiques principales de la caméra. Cela rend l'exploitation des images difficile.

Une des possibilités est d'estimer le modèle de distorsion tout en posant l'hypothèse qu'il est radial. Cela permet d'obtenir de bons résultats pour le centre de l'image, mais en déforme les bords, comme nous pouvons le voir figure 4.19.

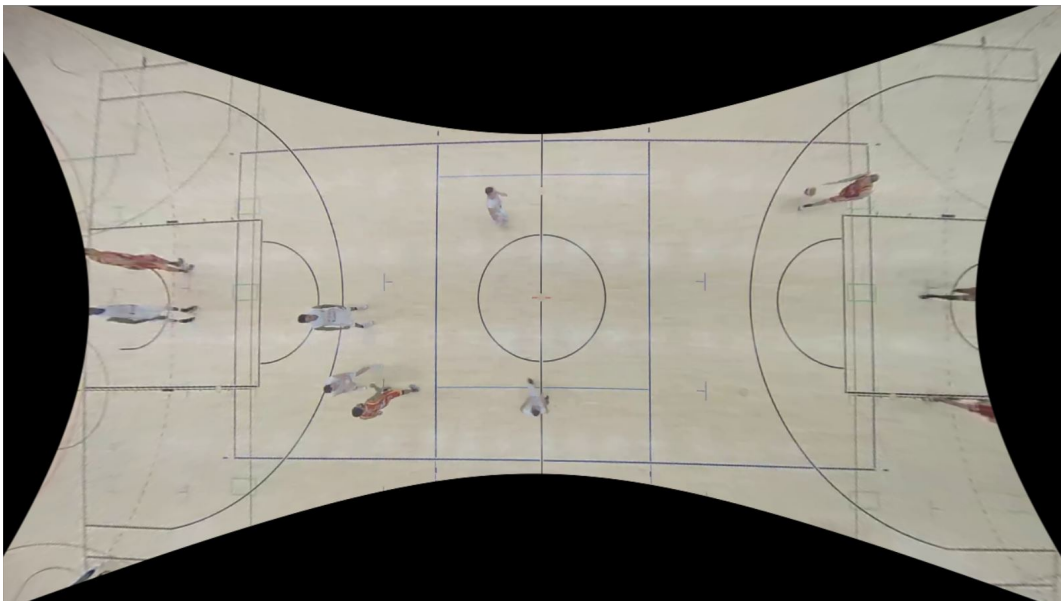


FIGURE 4.19 – Correction radiale

Pour pallier les mauvais résultats de l'approche et réduire le temps de traitement nécessaire pour corriger l'image, nous proposons d'utiliser un rapport pixel/mm créé manuellement, variant en fonction de la position dans l'image comme illustré figure 4.20. Un des principaux avantages de cette correction est sa simplicité et sa relative précision. De plus, il devient aussi possible d'exprimer les paramètres en centimètres et non en pixels.

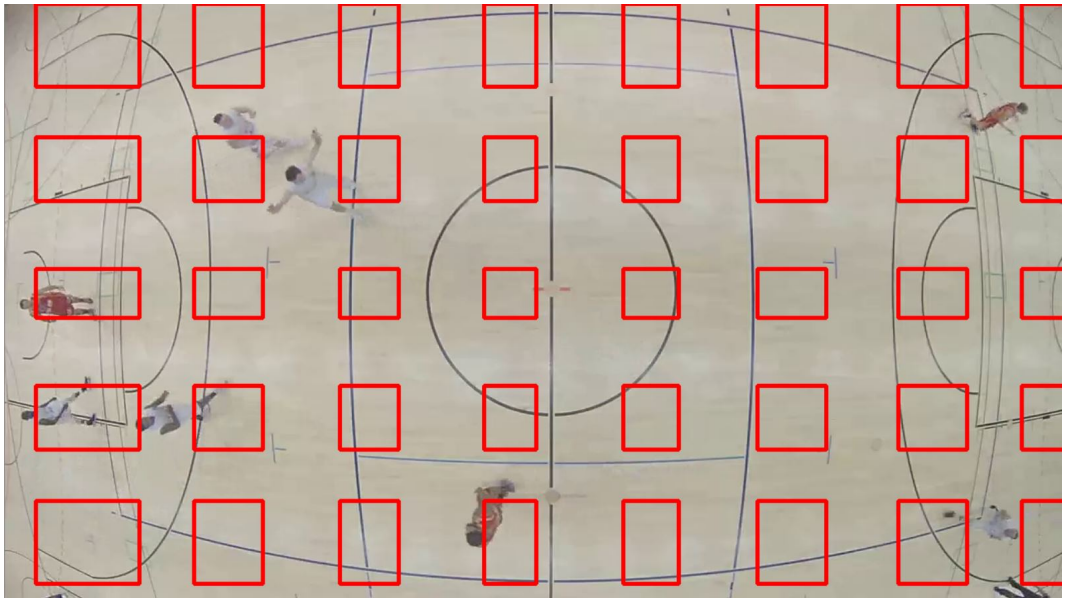


FIGURE 4.20 – Estimation de la distorsion

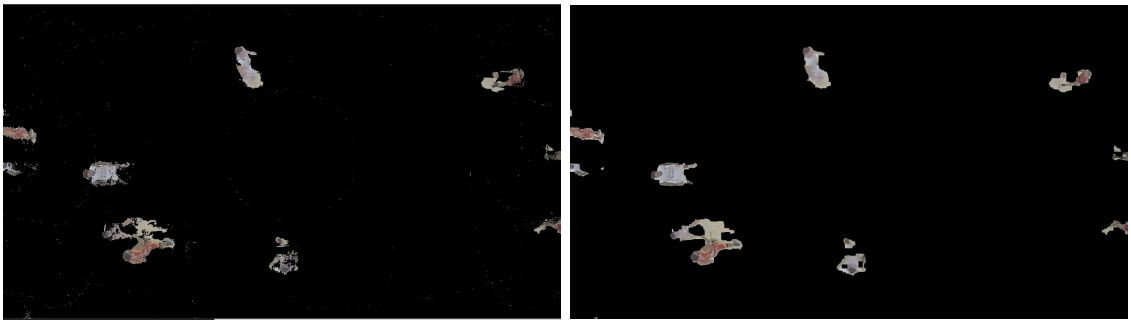
Modélisation de fond et extraction des joueurs

Afin de pouvoir détecter les joueurs, une soustraction d'arrière-plan [70] est utilisée. Comme la caméra est fixe et que le fond est statique, la taille de l'historique utilisé pour calculer cette soustraction peut être conséquent, pour rendre le modèle plus robuste. L'application de cette modélisation induit une détérioration de l'image résultante (voir figure 4.21a), en particulier au niveau des lignes du terrain qui sont plus floues. De plus, certains pixels appartenant aux joueurs peuvent être considérés comme faisant partis du fond, impliquant que l'image de certains joueurs est divisée en morceaux.Il est alors nécessaire d'améliorer l'image obtenue suite à la soustraction du fond afin de permettre une extraction correcte des joueurs. Trois transformations morphologiques sont appliquées à l'image. D'abord, une ouverture, avec un élément structurant de taille $\lambda = 1$, définit en 8-voisinages (figure 4.21c), permet de supprimer le flou. Puis une fermeture, avec un élément structurant de taille $\lambda = 4$, défini en 8-voisinages (figure 4.21d), permet de connecter les pixels appartenant à la même forme. Enfin, une ouverture finale est appliquée afin de lier les contours des objets en supprimant les aspérités (élément structurant de taille $\lambda = 1$, définit en 8-voisinages, voir figure 4.21e). Le résultat final est montré dans la figure 4.21b.

La dernière étape du traitement consiste à extraire les formes correspondant aux joueurs. En utilisant l'image obtenue, un algorithme d'analyse en composantes connexes, basé sur un suivi de contour est utilisé [131]. Cet algorithme classique a pour principal avantage d'être robuste et éprouvé. Ensuite, les rectangles encadrants calculés sur les formes détectées sont filtrés afin d'éviter de considérer plusieurs fois les même joueurs. Enfin, les joueurs trop petits sont supprimés en se basant sur un critère de surface.

Évaluation de la distance

La distance est une caractéristique essentielle car il n'est pas possible qu'un joueur puisse se déplacer plus vite qu'une certaine limite. Sur la base des informations disponibles, nous définissons pour chaque joueur un score de distance aux détections défini dans la formule 4.1, où *distance* est la distance euclidienne entre un joueur (position de l'objet suivi à l'instant t-1) et un objet



(a) Soustraction de l'arrière plan (b) Image finale

(c) Première ouverture (d) fermeture (e) Dernière ouverture

FIGURE 4.21 – Étapes successives de la soustraction d'arrière plan

candidat (position d'un objet à l'instant t) , et *max* est la valeur maximale pour relier un joueur avec un objet. Cette valeur a été fixée empiriquement à 3 mètres pour les expérimentations et nous ne gardons que les joueurs à 1,2 mètre ($score > 0,6$) pour l'association des joueurs aux cibles.

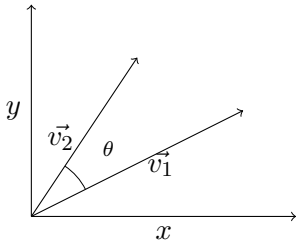
$$score = \begin{cases} 1 & if\ distance < min, \\ 0 & if\ distance > max, \\ \frac{max - distance}{max - min} & else \end{cases} \tag{4.1}$$

Si un seul objet candidat se trouve à proximité d'un joueur précédemment détecté, alors nous pouvons associer cet objet au joueur. Dans le cas où plusieurs objets se trouvent dans le voisinage d'un joueur, il est alors nécessaire d'utiliser des critères supplémentaires afin de sélectionner l'objet candidat correspondant réellement au joueur.

Direction du vecteur vitesse

Lorsque deux joueurs se croisent, il peut être difficile d'associer correctement les nouveaux objets aux positions des joueurs. L'utilisation des vecteurs vitesse peut permettre de lever les ambiguïtés dans le cas où les deux vecteurs ne sont pas colinéaires et qu'au moins un des modules des vecteurs vitesse n'est pas trop bas. En effet, lorsque la vitesse d'un joueur est basse, il est possible que ce joueur change rapidement de direction, rendant ainsi le suivi défaillant. Le score, défini dans la formule 4.2 est calculé en fonction de l'angle entre les deux vecteurs (voir fig. 4.22).

$$score = \begin{cases} 1 & if\ \theta_{degres} < min, \\ 0 & if\ \theta_{degres} > max, \\ \frac{max - \theta_{degres}}{max - min} & else \end{cases} \tag{4.2}$$



$$\theta_{degres} = \left|\arctan\frac{\vec{v_1}\cdot\vec{x}}{\vec{v_1}\cdot\vec{y}} - \arctan\frac{\vec{v_2}\cdot\vec{x}}{\vec{v_2}\cdot\vec{y}}\right| * \frac{180}{\pi}$$

FIGURE 4.22 – Orientation du vecteur vitesse

Utilisation de la couleur pour la séparation de groupe d'objets candidats
Lorsque deux joueurs sont trop proches, il est difficile d'associer avec certitude les objets candidats aux joueurs précédemment détectés. Pour augmenter la précision, nous utilisons le lien qui existe entre le joueur et l'équipe et plus particulièrement la couleur des maillots d'une équipe. Cette couleur équipe est identifiée à l'initialisation du système et sert de référence. Ainsi, les informations colorimétriques peuvent aider à séparer deux objets candidats proches. L'utilisation de l'espace colorimétrique LAB (L pour la luminosité, a pour l'échelle vert-rouge et b pour le bleu-jaune [66]) permet de séparer efficacement les couleurs, tout en prenant en compte les changements de luminosité pouvant intervenir. Nous proposons de calculer un rapport entre le nombre de pixel de la couleur équipe de référence sur le nombre total de pixels de la forme candidate. Pour faire l'affectation après la séparation d'un groupe d'objets candidats, nous nous basons sur la comparaison des ratios avant le rassemblement des joueurs avec les ratios calculés après la séparation.

La figure 4.23 présente le résultat obtenu par notre méthode.

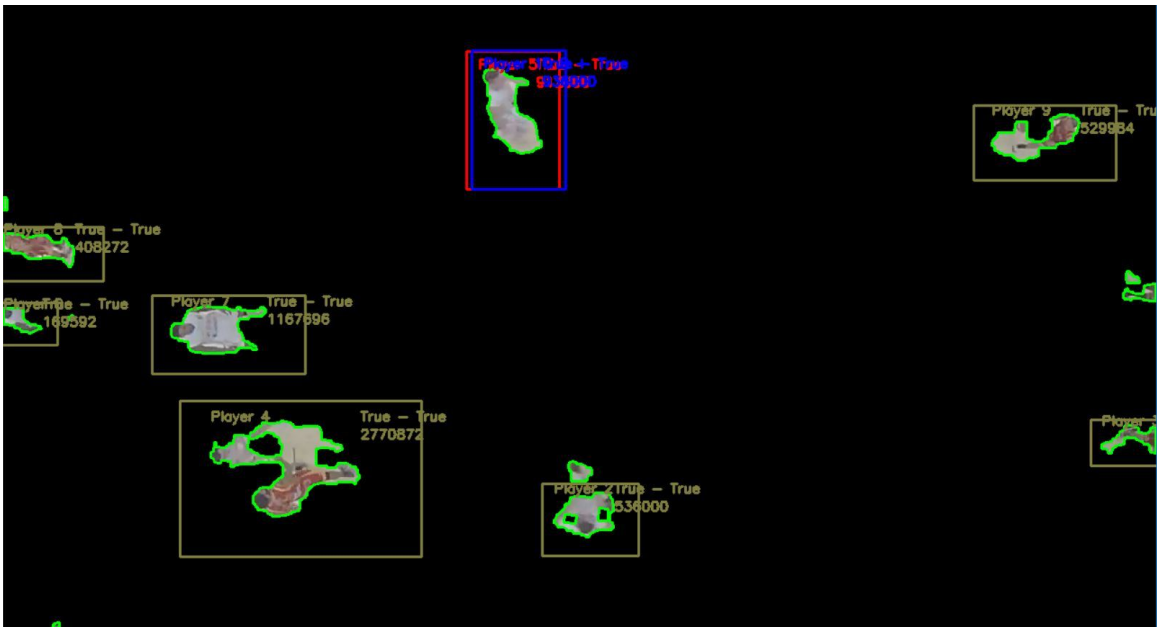
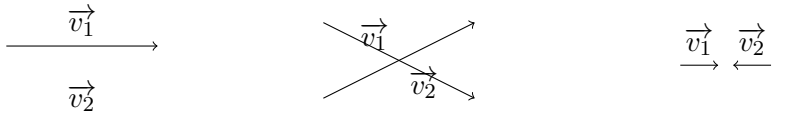


FIGURE 4.23 – Résultats de la méthode de suivi proposée

4.5.3 Comparaison des méthodes

Afin d'évaluer l'efficacité de notre méthode et de la comparer aux méthodes présentées dans la section 4.5.1, des extraits de match de basketball ont été sélectionnés. Ces extraits comprennent des situations problématiques dans le cadre du suivi de personnes. Nous avons annoté manuellement la position de chaque joueur dans différentes séquences correspondant à des situations spécifiques (entre 5 et 10 extraits pour chaque cas, voir fig. 4.24).



(a) Parallèle (b) Croisement (c) Statique

FIGURE 4.24 – Situations problématiques

Ces extraits nous servent de vérité de terrain, et permettent de calculer un score de précision [34] basé sur la quantité de situations résolues. Les résultats sont présentés dans la table 4.2. Le temps de traitement par image (T.T.I) exprime le temps moyen (en millisecondes) nécessaire pour détecter et suivre un joueur dans une image.

Situation	Mouvement	KCF	Boosting	TLD*	Méthode proposée
Joueur unique	Course	++	++	-	++
	Accélération	++	+	+	++
	Faible vitesse	++	++	+	++
Joueurs différents	Statique	++	-	-	++
	Croisement	++	+	+	++
	Parallèle	+	-	- -	+
Même équipe (rouge)	Statique	-	- -	-	-
	Croisement	+	-	-	++
	Parallèle	+	- -	- -	-
Même équipe (blanc)	Statique	-	- -	+	- -
	Croisement	+	-	-	++
	Parallèle	-	- -	-	-
T.T.I.		0.017	0.049	0.157	0.012

$++ : acc \geq 75\%$, $+ : acc \geq 50$, $- : acc \leq 50\%$, $-- : acc \leq 25\%$
* Des erreurs peuvent impliquer une interversion des joueurs durant quelques trames.

TABLE 4.2 – Précision des différents algorithmes

Pour résumer, nous pouvons voir que deux algorithmes sont performants dans notre étude : ceux basés sur la méthode KCF et celui proposé. En effet, ces deux méthodes sont capables de suivre les joueurs dans la plupart des situations problématiques, exceptées dans les endroits de l'image où la distorsion est trop importante. Les mauvais résultats de la méthode par Boosting s'expliquent par le fait que l'algorithme est souvent bloqué sur les lignes. En ce qui concerne l'algorithme TLD, le principal problème que nous observons est que le joueur est parfois perdu

pendant quelques images et qu'un autre joueur est choisi au hasard dans l'image. Cependant, après quelques images, le joueur initial est retrouvé. Lisser la trajectoire permettrait de supprimer ces sauts et ainsi d'améliorer le suivi. Néanmoins, ces deux algorithmes sont beaucoup plus lents que KCF.
Les résultats sont intéressants en terme de scalabilité. Le principal problème avec la méthode KCF est que nous devons définir un objet de suivi pour chaque joueur que nous voulons suivre, ralentissant la vitesse du processus. Au contraire, la majeure partie du temps de traitement de notre solution est prise en compte par la soustraction de l'arrière-plan et l'extraction de la position des joueurs. Cela signifie que le suivi des nouveaux joueurs demande simplement de calculer les scores d'association, ce qui est rapide. Notre solution est donc plus adaptée au suivi de plusieurs joueurs.

Nous pouvons remarquer que la plupart des difficultés identifiées pour les méthodes de suivi trompent les algorithmes génériques mais peuvent être résolues en utilisant des connaissances sur le contexte comme la couleur, la position ou la vitesse des joueurs. L'implémentation d'une solution dédiée rend l'algorithme bien plus performant en terme de précision mais surtout en terme de temps de calcul. Ces résultats prometteurs nous permettent d'envisager l'utilisation de cette méthode pour extraire des statistiques de la scène, de proposer un montage centré sur un joueur particulier ou encore de pouvoir détecter de nouvelles actions spécifiques comme par exemple les contre-attaques.

Même si les résultats sont prometteurs, l'algorithme implémenté, comme les autres algorithmes, est cependant confronté à certains problèmes. La plupart d'entre eux sont liés à la similitude entre la couleur de fond et la tenue d'une équipe. Par exemple, si deux joueurs de la même couleur se croisent à basse vitesse, le joueur peut être perdu. Pour résoudre ces problèmes, quelques améliorations peuvent être apportées.

FIGURE 4.25 – Rencontres problématiques de joueurs : (a) Croisement de joueurs ayant des vitesses de déplacements différentes. - (b) Déplacement de joueurs suivant une trajectoire elliptique

Tout d'abord, le suivi se concentre sur un seul joueur, signifiant qu'il n'est pas possible d'utiliser les informations de suivi des autres joueurs. Par exemple, on peut voir dans la fig. 4.25a les trajectoires de 2 joueurs. Si nous essayons de suivre le joueur 2, et que la couleur ne permet pas la séparation des joueurs, l'algorithme rencontre des problèmes pour trouver le joueur pertinent quand ils se croisent car le module de vitesse $|\overrightarrow{v_2}|$ est trop bas pour donner des informations (le joueur peut reculer). Cependant, si le joueur 1 est également suivi, du fait que $|\overrightarrow{v_1}|$ est significatif, la combinaison de ces informations par un système de règles pourrait aider à résoudre ce problème.

D'autre part, l'algorithme peut être mis en échec si les joueurs sont dans la configuration de

figure 4.25b. La vitesse est assez grande pour tenir compte de l'angle du vecteur vitesse mais le suivi risque probablement d'échouer. Une réponse à cela pourrait être d'améliorer l'étape de soustraction d'arrière-plan comme indiqué dans [146], afin d'obtenir une extraction des joueurs plus précise.

4.6 Discussions

Dans ce chapitre, l'instanciation de notre méthodologie pour le montage automatique dans le cadre du basketball a été faite : la modélisation des connaissances, provenant du règlement officiel de la FIBA et de connaissances d'experts, nous a permis d'identifier le point d'intérêt principal : le jeu notable représenté par le déplacement global des joueurs des deux équipes. Nous avons proposé une méthodologie de détection de l'action principale en se basant sur le déplacement du centre de gravité des joueurs. Notre méthode permet de suivre l'action avec un temps de calcul nous permettant l'exploitation en temps-réel et ainsi de proposer un montage automatique diffusant le flux vidéo de la caméra montrant l'action.
Afin d'améliorer la sélection de caméra, une méthodologie de détection de lancer-franc a été proposée. Une étude du règlement nous a permis d'extraire des connaissances sur les lancers-francs et ainsi de définir une méthodologie de détection de lancer-franc, basée sur la position des joueurs. Cette méthodologie a été mise en place pour des cas concrets. La méthode proposée fonctionne en temps-réel, permettant ainsi la sélection automatique de caméra dans le cas d'une diffusion en direct.
La détection de lancer-franc nous permet de plus de pouvoir mettre en place un premier levier de personnalisation pour le spectateur. Il lui est ainsi possible de visualiser tous les lancers-francs ayant eu lieu lors du match, ou lors d'un quart-temps spécifique. De nombreuses autres actions peuvent être détectées afin de pouvoir répondre aux exigences de nombreux spectateurs. La détection des actions que nous avons identifiées lors de la modélisation du match de basketball (paniers, remises en jeu, fautes, contres-attaques, ...) permet d'augmenter les possibilités de la personnalisation du contenu.

La proposition d'une nouvelle méthode de suivi de joueur permet d'envisager leur identification tout au long de la rencontre. Cette identification autorise une personnalisation plus complète des flux vidéos en permettant d'une part aux spectateurs de pouvoir visualiser les actions spécifiques d'un joueur particulier et d'autre part, de permettre aux entraîneurs d'obtenir des informations statistiques sur leurs joueurs.

Nous avons présenté dans ce chapitre des méthodes de détection de personnes et d'actions dans le cadre de match de basketball. Cependant, ces méthodes sont facilement adaptables à d'autres scénarios avec comme seul prérequis la présence d'une caméra azimutale. Ainsi, il est facile de réaliser une sélection automatique de caméra pour les sports collectifs tels que le handball, le hockey ou le football. En effet, comme pour le basket, les joueurs vont chercher à marquer dans le but des adversaires. L'étude du centre de gravité permet alors de localiser l'action d'intérêt.
Dans le cas des sports où les joueurs sont séparés par un filet, une étude des déplacements de chaque côté du terrain doit permettre de mettre en place une méthodologie de sélection de caméras.
Enfin, on peut penser que les méthodes proposées puissent être utilisées dans tous les scénarii intérieurs tels que la surveillance ou la robotique.

Conclusions et perspectives

Les principales contributions de ce travail se décomposent en quatre points. Le premier est un état de l'art sur les systèmes de montages automatiques. Le second est la proposition d'une méthodologie générique de montage automatique basée sur la modélisation des connaissances issues de l'état de l'art. Le troisième concerne l'intégration d'ontologies pour la proposition de méthodes d'extraction d'informations pertinentes de la scène. Enfin, le quatrième est la proposition de nouvelles méthodes d'extraction d'informations pertinentes dans différents contextes.

Le premier chapitre de cette thèse nous sert à présenter les différents systèmes de montage automatique issus de la littérature. Ainsi, les différents verrous et limites de ces systèmes ont pu être mis en avant. L'un de ces verrous est lié à une forte dépendance au contexte. Les systèmes de montage automatique sont développés pour un cas d'application particulier, ce qui rend leur adaptation difficile à des contextes variés. De plus, les systèmes proposés s'intéressent à la diffusion en direct ou en différé. Ces deux types de diffusions présentant un intérêt pour le spectateur, il nous est alors apparu intéressant de proposer un système capable de prendre en compte ces deux versions. De même, la prise en compte des desiderata des spectateurs est rarement abordée dans la réalisation du montage et constitue une originalité notable.

Nous proposons, dans le second chapitre, une méthodologie permettant la mise en place d'un système de montage automatique. L'analyse fonctionnelle des systèmes existants, par la méthode SADT, aboutit à la proposition d'un cadre générique pour la conception de système de montage automatique. L'architecture proposée prend en compte les préférences utilisateur et le type de diffusion. L'extraction d'informations de la scène étant à la base de la réalisation d'un montage automatique, l'identification des sources d'intérêt pour un contexte donné est primordiale. La capitalisation et la structuration des connaissances nécessaires, par la méthode NIAM/ORM, nous conduit à proposer une modélisation générique d'un évènement, et à l'identification d'une ontologie liant les personnes d'intérêt (POI) et les actions d'intérêt (AOI).

Les différentes étapes de la méthodologie sont appliquées lors de la mise en place de deux systèmes de montage automatique présentés aux chapitres 3 et 4. Pour permettre une diffusion en direct, ainsi que pour proposer des premiers éléments pour la personnalisation des flux vidéo, les étapes suivantes de notre méthodologie sont appliquées :

— acquisition de connaissance sur le contexte ;
— instanciation du modèle de connaissance générique ;
— identification des sources d'intérêts et de leurs attributs ;
— extraction des attributs des sources d'intérêts sélectionnées.

Nous nous intéressons, dans le troisième chapitre, à la mise en place d'un système de montage automatique dans le cas de conseils municipaux. L'intégration des connaissances nous permet de déterminer la "prise de parole" comme AOI principale. Les limites des méthodes existantes nous amènent à proposer une nouvelle méthode de détection de locuteurs, se basant sur la lumière émise au niveau de la collerette des microphones. Notre méthode permet la détection des prises de parole en temps réel, avec une justesse supérieure à 98%. Une étude de la communication par lumière visible, nous montre qu'il est également possible d'utiliser la lumière des microphones comme vecteur de transmission de l'identité des locuteurs. La principale originalité de cette approche est l'utilisation de la même image (ou flux vidéo) pour, à la fois, visualiser la scène d'intérêt et décoder l'information transmise. Cela permet ainsi la personnalisation des flux vidéos montés en fonction des POI choisies par le spectateur.

La diffusion des matchs de basketball est abordée dans le chapitre 4 avec la proposition d'une

méthode de sélection automatique de caméras pour la diffusion de l'AOI "jeu notable". Une méthode de détection de l'action est proposée. Elle se base sur la position des POI que constituent les joueurs et l'arbitre. Une caméra azimutale est utilisée afin d'obtenir la position du centre de gravité des joueurs dont le déplacement permet de sélectionner, en temps-réel, la caméra où se passe l'action. Les évaluations numériques montrent que les flux vidéo obtenus par notre méthode sont équivalent à ceux générés par un monteur. La détection de l'AOI "Lancer franc" et le suivi des POI "Joueurs" sont également abordés afin de proposer aux spectateurs la personnalisation des flux vidéo.

Pour résumer, ces travaux de thèse ont permis de proposer une approche méthodologique générique de sélection automatique de caméras. L'intégration des connaissances sur le contexte aide à la mise en place d'un système de montage automatique en guidant l'identification des sources d'intérêt dans la scène. La création et l'intégration des ontologies liées au contexte permettent de proposer des méthodes de traitement d'images assurant la détection, le suivi et l'identification des AOI et POI dans des situations non contraintes à partir de vues hétérogènes. La méthodologie proposée a été validée par son application à deux types d'évènement.

Ces travaux ouvrent des perspectives à plusieurs niveaux.

Tout d'abord, les méthodes proposées dans cette thèse se limitent à la sélection de caméras d'intérêt. Il serait nécessaire de poursuivre ces travaux en s'intéressant au contrôle des caméras. L'utilisation de caméras contrôlables (PTZ) permettrait, par exemple, de rendre le flux vidéo monté plus dynamique que celui présenté dans le chapitre 4, qui se base uniquement sur des caméras fixes. De plus, la prise en compte de règles cinématographiques permettrait d'améliorer la qualité visuelle des captations [47].

Par ailleurs, les flux vidéo montés n'ont été comparés qu'avec ceux réalisés par une seule personne. Bien que les montages obtenus par nos méthodes soient similaires à ceux proposés par le monteur, la réalisation de tests subjectifs permettrait d'analyser et d'améliorer la qualité des flux vidéo montés en fonction des retours des spectateurs [46].

Nous avons proposé dans le chapitre 3 d'utiliser la technologie Visible Light Communication dans le but d'identifier les locuteurs dans une scène. Il serait intéressant de poursuivre ces travaux afin de récupérer les informations d'une source lumineuse en mouvement. Ces recherches permettraient d'identifier des POI dans d'autres types d'événement comme par exemple l'identification de joueurs dans le contexte de match de Basketball. Il serait également possible d'appliquer cette technologie dans d'autres cas d'application. Il serait par exemple possible, dans un scénario de surveillance, de suivre et d'identifier des objets d'intérêts portant un dispositif lumineux (personnes avec un badge, robots équipés d'une LED,...) et de résoudre de nombreux problèmes liés au suivi de personnes comme l'occultation, le changement brutal de direction ou le croisement.

Une perspective à plus long terme concerne la recherche des méthodes optimales d'extraction des caractéristiques de la scène en fonction du contexte. En effet, le choix est réalisé par le concepteur du système, en fonction des exigences des spectateurs. L'approche systémique du modèle proposé dans le chapitre 2 permet de hiérarchiser les attributs et aide ainsi le concep-

teur du système à sélectionner les méthodes d'extraction de caractéristiques adéquates. Il serait intéressant d'améliorer cette sélection en mettant en place un système de sélection automatique de la méthode optimale d'extraction de caractéristiques basée sur la modélisation des connaissances du contexte. Cela permettrait de proposer une méthodologie semi-supervisée de montage automatique. Néanmoins, ce point pose le problème de l'évaluation de la qualité d'un traitement d'image, problème qui reste ouvert encore actuellement.

Enfin, des perspectives portent sur l'extension de notre méthodologie au montage automatique multi-événements. Les modélisations proposées dans cette thèse se concentrent sur la réalisation de montage automatique pour un seul événement. Par exemple, il serait intéressant de pouvoir visualiser tous les paniers marqués par un joueur lors d'une saison entière. Il est alors nécessaire d'intégrer aux systèmes de montage automatiques, la prise en compte des évènements précédents. Les systèmes d'indexation de flux vidéo présentent alors un intérêt. Une première étude des relations entre le montage automatique et l'indexation vidéo est proposée à l'annexe A. Cette étude montre que ces systèmes présentent des similitudes dans leurs processus respectifs. Le montage automatique cherche à produire un flux vidéo monté, à partir de flux vidéo provenant de différentes caméras. Pour ce faire, il est nécessaire d'extraire des connaissances sur les sources d'intérêts. Lors de l'indexation de vidéo, le flux vidéo monté est découpé en séquence puis analysé afin d'extraire les informations sémantiques de la séquence vidéo. Il semble alors intéressant que l'indexation des vidéos soit réalisée dès la création du flux vidéo monté. De plus, un des objectifs de l'indexation est la recherche de contenu vidéo à partir de données sémantiques. Les méthodes de recherche pourraient contribuer à la création de flux vidéo personnalisés en sélectionnant les séquences vidéo répondant aux desiderata des spectateurs. Ainsi, l'indexation vidéo se positionne comme un outil adéquat afin de proposer un montage automatique multi-évènements. La méthodologie que nous proposons s'appuie sur l'acquisition de connaissances afin d'identifier et d'extraire des concepts sémantiques dans la scène. Les systèmes de montage automatique pourraient servir pour réaliser l'indexation des flux vidéo dès leurs diffusion.

Annexe A

De l'indexation automatique de vidéos aux systèmes de montage automatique

L'explosion actuelle du nombre de vidéos disponibles sur internet rend obligatoire une classification permettant aux utilisateurs d'accéder aux contenus qu'ils souhaitent. C'est pourquoi les méthodologies d'indexation automatique se sont tournées ces dernières années sur les problèmes d'indexation de contenu vidéo.

L'indexation automatique de documents est le fait d'extraire, de représenter et d'organiser logiciellement des documents (textes, vidéos, images, musiques, ...) afin de pouvoir faciliter la recherche de ces documents. Cette problématique peut donc être divisée en deux parties : l'indexation de contenu et la recherche de documents. L'indexation consiste à enrichir un média avec des informations sémantiques ou visuelles sur le contenu du document permettant une recherche ultérieure en se basant sur le contenu.

Lors d'un montage automatique un certain nombre d'informations sont extraites de l'événement diffusé : le type, la durée, les actions et les personnes qui le composent, etc. Ces informations sont généralement utilisées pour un seul type de diffusion. Il serait donc intéressant d'étudier comment fonctionne l'indexation vidéo afin de pouvoir contribuer, dès la publication d'un événement à son indexation. De plus, la création d'un montage personnalisé d'un événement nécessite de rechercher dans un index les événements que le spectateur veut voir. Si un utilisateur veut voir tous les lancers francs lors d'un match de basketball, il est nécessaire de retrouver toutes les parties d'une vidéo annotées "lancer franc".

A.1 Indexation de vidéo

L'indexation automatique de vidéo se découpe généralement en trois étapes principales (voir fig. A.1) : la segmentation temporelle de la vidéo, la réduction de la taille des données à traiter et l'analyse du contenu [43].

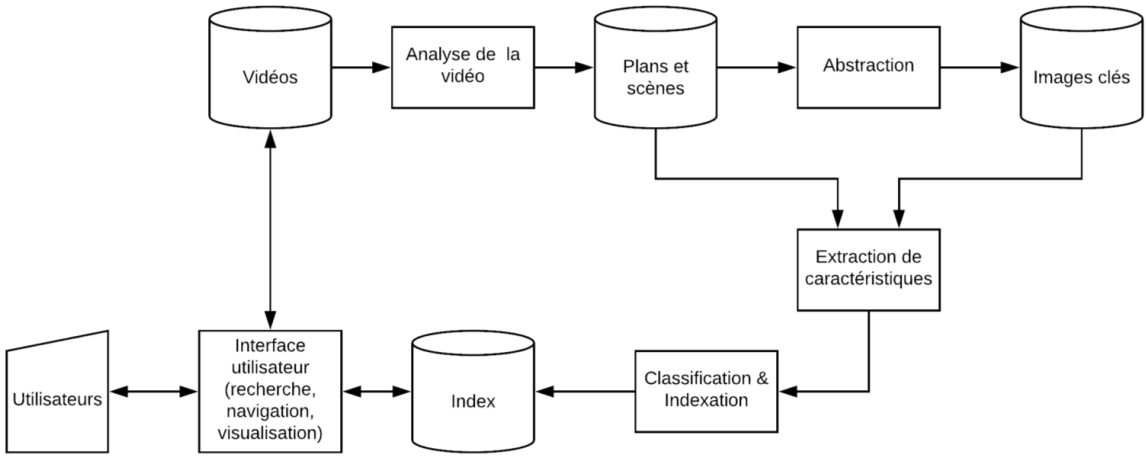


FIGURE A.1 – Système d'indexation vidéo basé sur le contenu [43]

A.1.1 Segmentation temporelle d'une vidéo

Indexer une vidéo entière ne présente que peu d'intérêt du fait du nombre important de concepts qui peuvent être exprimés. De plus, ces concepts interviennent à différents moments de

la vidéo et il est de découper la vidéo en différents segments pouvant être analysés et annotés séparément. La descriptions des différentes parties permettra une meilleure compréhension de la vidéo. Le découpage d'une vidéo peut se faire à différents niveaux. Les découpages les plus récurrents dans la littérature sont aux niveaux des scènes, des plans et des images. Il existe cependant d'autre décomposition dans la littérature comme la décomposition en histoires [56], ou au niveau de l'objet contenu dans une image.

Une scène est un groupement de plans temporellement adjacents et sémantiquement proches. Grâce à un assemblage logique de plans, la scène permet d'exprimer une idée, ce qui la rend particulièrement intéressante pour une indexation sémantique. Cependant le problème de segmentation d'une vidéo en scènes n'est pas complètement résolu et son approche repose généralement sur le découpage en plans.

Un plan est défini par une séquence d'images présentant une action continue qui est capturée par une seule caméra [109]. Deux approches se distinguent pour découper une vidéo brute en plans. La première compare les caractéristiques d'images successives. La différence calculée entre deux images permet de détecter les changements de plan. La seconde approche utilise des modèles mathématiques afin de modéliser les transitions entre plans [54, 12].

A.1.2 Réduction de la taille des données

La seconde étape de l'indexation consiste à réduire la taille des données à traiter. En effet, il n'est pas judicieux d'utiliser toutes les images d'un plan car il peut être composé d'un grand nombre d'images dont le contenu est redondant. Afin de réduire le nombre d'images à traiter, une image clé est extraite. Il s'agit d'une image représentant le mieux le contenu d'un plan. [134] réalisent cette extraction en comparant les niveaux de gris des pixels d'images successivement et sélectionnent les images présentant la plus grande différence. Plus récemment, [122] ont utilisé une classificateur à vaste marges (SVM) afin d'extraire les trames dominantes d'une vidéo.

A.1.3 Analyse du contenu

Enfin la dernière étape vise à générer les métadonnées pour chaque plan. Pour ce faire, les images clés sont analysées afin d'extraire des caractéristiques sur les plans. Le contenu d'une image clé peut être décrit à deux niveaux : numériquement (caractéristiques de couleurs, formes, textures, ...) ou sémantiquement (interprétation textuelle de la vidéo). Les méthodologies de recherche de vidéo basées sur le contenu (CBVR) utilisent les caractéristiques numériques pour trouver une image similaire / dissimilaire à celle d'entrée [99, 43, 109]. Cependant ces descripteurs ne correspondent pas à tous les besoins des utilisateurs. Il est difficile de trouver une vidéo précise à partir de caractéristiques. Il est plus simple pour un utilisateur de formuler une requête textuelle afin d'effectuer une recherche. Ainsi une indexation sémantique du contenu d'une vidéo permet une recherche plus simple. Les descripteurs sémantiques s'appuient généralement sur les caractéristiques numériques. Des classificateurs sont utilisés afin d'annoter chaque contenu avec un mot / terme / concept [126, 42]. Les relations entre ces concepts peuvent également être prises en compte afin de fournir une description haut niveau sur la scène.

Les connaissances extraites sont ensuite généralement stockées dans une base de données [109, 74] ou dans un fichier XML (Extensible Markup Language) [138] afin de permettre une recherche ultérieure

A.2 Recherche

Une fois que les vidéos ont été annotées, il est alors possible d'en effectuer la recherche qui est définie comme l'ensemble des opérations nécessaires pour répondre à la demande d'un utilisateur [128]. Une fois une requête formulée, une méthode de mesure de similarité est utilisée afin de chercher la vidéo la plus proche de la recherche. Différents types de requêtes peuvent être effectuée. Pour les recherches non sémantiques, les demandes peuvent être par un exemple (requête sous la forme d'une image / vidéo), par un dessin ou par un objet (localisation d'objet dans une image / vidéo). Pour les recherches sémantiques, les requêtes peuvent être effectuées par mot clefs ou en langage naturel.

Le calcul de la similarité de la vidéo avec la demande de l'utilisateur est une étape importante. Les méthodes pour mesurer les similarités entre les vidéos peuvent être basées sur les caractéristiques, sur le texte ou sur une combinaison [43].

Une mesure de similarité peut être la distance moyenne entre les caractéristiques de deux vidéos. Des mesures de bas niveau entre les images clefs, des caractéristiques sur les objets ou les déplacements peuvent être utilisés. Cependant, la similarité sémantique ne peut être représentée du fait de la différence entre les caractéristiques et les catégories sémantiques.

Les méthodes basées sur le texte calculent la similarité entre les termes de la recherche avec la description textuelle des concepts d'une vidéo. Il s'agit de la manière la plus simple de trouver une vidéo répondant à une requête. Cependant il est nécessaire d'être précis dans sa recherche afin d'obtenir le résultat le plus satisfaisant.

Les méthodes basées une combinaison de concept utilisent des ensemble de textes afin d'améliorer les résultats de recherche. Ces combinaisons sont apprises à partir de données d'exemple. L'avantage des méthodes par combinaison est le fait que les poids associés à chaque concept peuvent être automatiquement déterminés. Il est cependant difficile d'obtenir de nombreuses données d'apprentissages.

A.3 Liens indexation / montage automatique

Il existe de nombreux liens entre indexation et montage, comme illustré dans la figure A.2.

L'indexation commence par découper la vidéo à indexer en différentes parties : plans, images clés. Dans le cadre du montage automatique, l'opération inverse est réalisée. On sélectionne un nombre d'images temporellement adjacentes du flux vidéo d'une caméra : un plan. Ces plans sont ensuite assemblés afin de générer une vidéo. Afin de pouvoir sélectionner des plans, il est nécessaire d'acquérir des connaissances sur ces dernières. Cette analyse peut être effectuée au niveau de l'image (image clé) ou sur une séquence vidéo (ensemble de plan couvrant un seul évènement/situation).

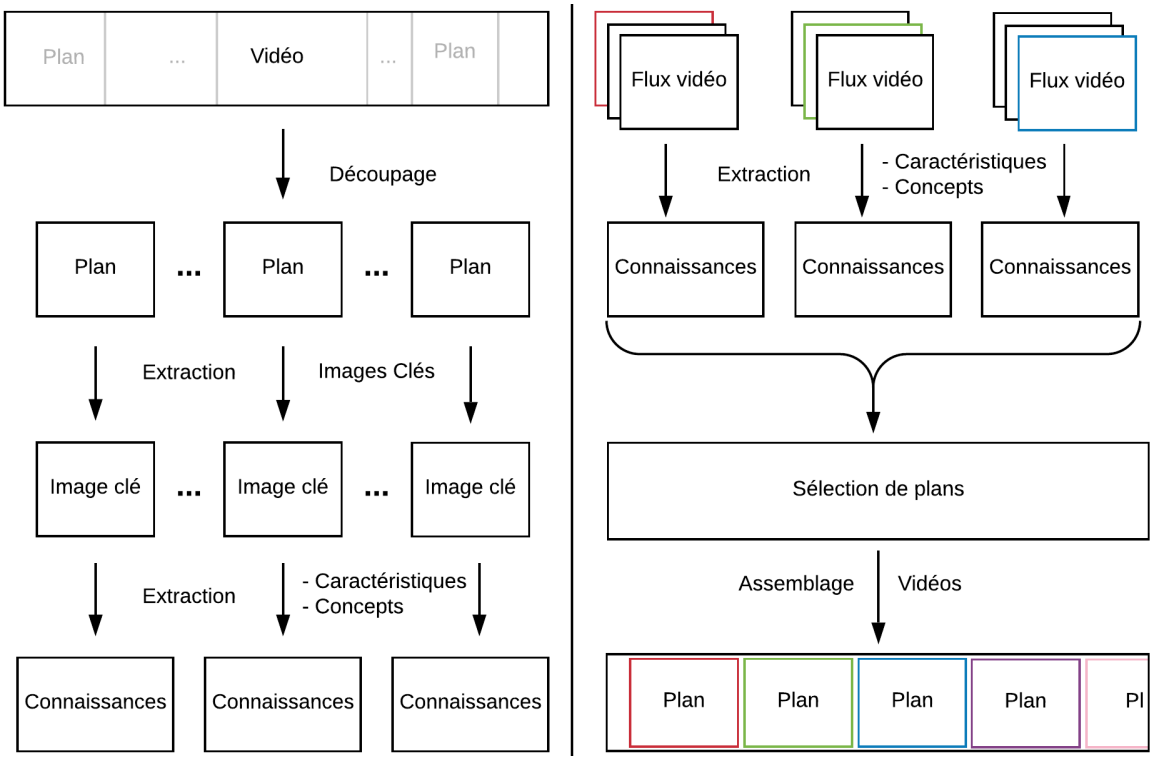


FIGURE A.2 – Comparaison entre indexation et montage automatique

L'extraction de connaissances est la première étape du montage automatique (Planifier), là où elle est l'étape finale dans les problématiques d'indexation vidéo. Il est donc opportun d'utiliser les connaissances extraites lors du montage pour les problématiques d'indexation.

Pour résumer, un certain nombre de constats montre l'apport de la réalisation de l'indexation lors de la génération d'un nouveau contenu vidéo par un système de montage automatique.

— Réaliser un montage automatique nécessite de connaître un certain nombre de concepts, dérivé de caractéristiques bas niveau, afin de sélectionner la caméra présentant une AOI ou POI. De ce fait, il est opportun d'annoter les vidéos dès que la connaissance est extraite.
— Notre méthodologie de montage automatique s'appuie sur l'acquisition de connaissances afin d'identifier les concepts sémantiques dans la scène. De ce fait, la définition des concepts pour l'indexation est déjà réalisé lors de l'étape de configuration du système.
— De plus, contrairement à un flux vidéo monté, nous disposons des flux vidéo de chaque caméra. Ainsi l'annotation de chaque flux vidéo permet la génération de montages différents, adaptés aux requêtes des utilisateurs.
— Enfin, l'architecture multi-événement de CitizenCam rend possible d'effectuer un montage, non pas sur une seule captation, mais sur plusieurs d'entre eux. Il est possible de réaliser un montage présentant par exemple les lancers francs d'un joueur sur toute une saison ou bien d'observer les participations d'un élu au cours de son mandat.

L'ensemble de ces constats nous montre qu'il est opportun d'intégrer l'indexation de contenu

dès son analyse pour le montage de flux vidéo. Parallèlement, les techniques de recherche des systèmes d'indexation peuvent permettre d'améliorer la personnalisation des contenus vidéos en répondant au mieux aux attentes des spectateurs. L'indexation permet de pouvoir générer des montages, non pas sur un seul évènement, mais sur un ensemble d'évènement. Il est alors possible de créer un montage comprenant les actions réalisées par une personne dans différents évènements.

Glossaire

AOI : Action of Interest
BER : Bit Error Rate
CBVR : Content-Based Video Retrieval
CGCT : Code Général des Collectivités Territoriales
CIFRE : Convention Industrielle de Formation par la Recherche
CMOS : Complementary Metal-Oxide-Semiconductor
FIBA : Fédération Internationale de Basket-ball Amateur
FPS : Frames Per Second
KCF : Kernelized Correlation Filter
LED : Lighting Emitting Diode
LiFi : Light Fidelity
MFCC : Mel-Frequency Cepstral Coefficient
NIAM : Nijssen Information Analysis Method
OCC : Optical Camera Communication
OCR : Optical Character Recognition
ORM : Object Role Modeling
OWC : Optical Wireless Communication
POI : Person of Interest
PTZ : Pan Tilt Zoom
ROI : Region of Interest
SADT : Structured Analysis and Design Technics
SOI : Source Of Interest
SVM : Support Vector Machine
TLD : Tracking Learning Detection
TTI : Temps de traitement par image
UFSOOK : Under Sampled Frequency Shift On Off Keying
VLC : Visible Light Communication
XML : eXtensible Markup Language

Bibliographie

[1] Sayf Albayati. An Overview of Visible Light Communication Systemps. *International Journal of Computer Science and Mobile Computing*, 8(6) :51–56, 2019.

[2] Benjamin Almecija, Vincent Bombardier, and Patrick Charpentier. Modeling Quality knowledge to design log sorting system by X rays tomography. *IFAC Proceedings Volumes*, 45(6) :1190–1195, May 2012.

[3] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic Editing of Footage from Multiple Social Cameras. *ACM Trans. Graph.*, 33(4) :1–11, July 2014.

[4] Yasuo Ariki, Shintaro Kubota, and Masahito Kumano. Automatic Production System of Soccer Sports Video by Digital Camera Work Based on Situation Recognition. In *Proceedings of the Eigth IEEE International Symposium on Multimedia*, pages 851–860, San Diego, CA, USA, December 2006. IEEE.

[5] Radhian Ferel Armansyah, Fadhli Dzil Ikram, Swizya Satira Nolika, and Trio Adiono. Efficient Sound-Source Localization system using low cost TDOA computation. In *2016 International Symposium on Electronics and Smart Devices (ISESD)*, pages 315–319, Bandung, Indonesia, November 2016. IEEE.

[6] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online Multiple Instance Learning. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*, pages 983–990, June 2009.

[7] Djamel-Eddine Benarab. *Automatic Swimmer Tracking Using Video Sequences : Application to Performance Analysis*. Theses, Université de Bretagne occidentale - Brest, December 2016.

[8] Michael Bianchi. AutoAuditorium : A fully automatic, multi-camera system to televise auditorium presentations. In *Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop*, 1998.

[9] Michael Bianchi. Automatic video production of lectures using an intelligent and aware environment. In *Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, pages 117–123, College Park, Maryland, USA, 2004. ACM Press.

[10] Stefano Bocconi, Frank Nack, and Lynda Hardman. Automatic generation of matter-of-opinion video documentaries. *Web Semantics : Science, Services and Agents on the World Wide Web*, 6(2) :139–150, April 2008.

[11] Vincent Bombardier and Olivier Nartz. Robot – Vision au Pôle AIP-Priméca Lorraine. In *Robot – Vision Au Pôle AIP-Priméca Lorraine*, Besançon, France, 2007.

[12] John Boreczky and Lynn Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3741–3744, 1998.

[13] RRayana Boubezari, Hoa Le Minh, Zabih Ghassemlooy, and Ahmed Bouridane. Smartphone Camera Based Visible Light Communication. *Journal of Lightwave Technology*, 34(17) :4121–4127, September 2016.

[14] René Bouillot and Gérard Galès. *Cours de Vidéo*. Dunod, Paris, 2008.

[15] Carlos Busso, Sergi Hernanz, Chi-Wei Chu, Soon-il Kwon, Sung Lee, Panayotis G. Georgiou, Isaac Cohen, and Shrikanth Narayanan. Smart room : Participant and speaker localization and identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages 1117–1120, March 2005.

[16] Timothy Callemein, Wiebe Van Ranst, and Toon Goedeme. The autonomous hidden camera crew. *(:unav)*, May 2017.

[17] Peter Carr, Michael Mistry, and Iain Matthews. Hybrid Robotic/Virtual Pan-tilt-zom Cameras for Autonomous Event Recording. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 193–202, New York, NY, USA, 2013. ACM.

[18] Peter Carr, Yaser Sheikh, and Iain Matthews. Monocular Object Detection Using 3D Geometric Primitives. In *Computer Vision – ECCV 2012*, volume 7572, pages 864–878. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[19] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint Video of Human Actors. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 569–577, New York, NY, USA, 2003. ACM.

[20] Christine Chen, Oliver Wang, Simon Heinzle, Peter Carr, Aljoscha Smolic, and Markus Gross. Computational sports broadcasting : Automated director assistance for live sports. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2013.

[21] Fan Chen and Christophe De Vleeschouwer. Personalized production of basketball videos from multi-sensored data under limited display resolution. *Computer Vision and Image Understanding*, 114(6) :667–680, June 2010.

[22] Fan Chen, Damien Delannay, and Christophe De Vleeschouwer. An Autonomous Framework to Produce and Distribute Personalized Team-Sport Video Summaries : A Basketball Case Study. *IEEE Transactions on Multimedia*, 13(6) :1381–1394, December 2011.

[23] Jianhui Chen. *Towards Automatic Broadcast of Team Sports*. PhD thesis, University of British Columbia, 2018.

[24] Jianhui Chen and Peter Carr. Autonomous Camera Systems : A Survey. In *Workshop on Intelligent Cinematography and Editing*, pages 18–22, 2014.

[25] Jianhui Chen and Peter Carr. Mimicking Human Camera Operators. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 215–222, January 2015.

[26] Jianhui Chen and James J. Little. Where should cameras look at soccer games : Improving smoothness using the overlapped hidden Markov model. *Computer Vision and Image Understanding*, 159 :59–73, June 2017.

[27] Yun-Chung Chung, Jung-Ming Wang, and Sei-Wang Chen. A Vision-Based Traffic Light Detection System at Intersections. *Journal of Taiwan Normal University : Mathematics, Science & Technology*, 47(1) :67–86, October 2017.

[28] Tarcisio Coianiz and Lorenzo Torresani. Analysis and encoding of lip movements. In *Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 51–60. Springer, Berlin, Heidelberg, March 1997.

[29] Ross Cutler and Larry Davis. Look who's talking : Speaker detection using video and audio correlation. In *2000 IEEE International Conference on Multimedia and Expo.*, volume 3, pages 1589–1592 vol.3, 2000.

[30] Shinji Daigo and Shinji Ozawa. Automatic pan control system for broadcasting ball games based on audience's face direction. In *21st ACM International Conference on Multimedia*, page 444. ACM Press, 2004.

[31] Christos Danakis, Mostafa Afgani, Gordon Povey, Ian Underwood, and Harald Haas. Using a CMOS camera sensor for visible light communication. In *2012 IEEE Globecom Workshops*, pages 1244–1248, Anaheim, CA, USA, December 2012. IEEE.

[32] Fahad Daniyal, Murtaza Taj, and Andrea Cavallaro. Content and task-based view selection from multiple video streams. *Multimedia Tools and Applications*, 46(2-3) :235–258, January 2010.

[33] Eleonora D'Arca, Neil M. Robertson, and James R. Hopgood. Look who's talking : Detecting the dominant speaker in a cluttered scenario. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1532–1536, May 2014.

[34] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning, ACM*, volume 06, June 2006.

[35] Anthony Dearden, Yiannis Demiris, and Oliver Grau. Learning models of camera control for imitation in football matches. In *4th International Symposium on Imitation in Animals and Artifacts*, pages 227–231, January 2007.

[36] Damin Delannay, Nicolas Danhier, and Christophe De Vleeschouwer. Detection and recognition of sports(wo)men from multiple views. In *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–7, August 2009.

[37] Shahbe Desa and Qussay Salih. Image substraction for real time moving object extraction. In *Proceedings. International Conference on Computer Graphics, Imaging and Visualization, 2004. CGIV 2004.*, pages 41–45, Penang, Malaysia, 2004. IEEE.

[38] Joseph H. DiBiase, Harvey F. Silverman, and Michael S. Brandstein. Robust Localization in Reverberant Rooms. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays : Signal Processing Techniques and Applications*, Digital Signal Processing, pages 157–180. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[39] George Doddington. Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE*, 73(11) :1651–1664, 1985.

[40] Petr Doubek, Indra Geys, Tomáš Svoboda, and Luc Van Gool. Cinematographic Rules Applied to a Camera Network. In *The Fifth Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, pages 17–29, 2004.

[41] Khaoula Elagouni, Christophe Garcia, Franck Mamalet, and Pascale Sebillot. Combining Multi-scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 120–124, Gold Coast, Queenslands, TBD, Australia, March 2012. IEEE.

[42] Jianping Fan, A. K. Elmagarmid, Xingquan Zhu, W. G. Aref, and Lide Wu. ClassView : Hierarchical video shot classification, indexing, and accessing. *IEEE Transactions on Multimedia*, 6(1) :70–86, February 2004.

[43] Mr Amit Fegade and Vipul Dalal. A Survey on Content Based Video Retrieval. *international journal of Modern Trends in Engineering and Research*, 3(7) :9, 2014.

[44] FIBA. REGLEMENT OFFICIEL DE BASKETBALL, September 2018.

[45] Christophe Fiorio and Jens Gustedt. Two linear time Union-Find strategies for image processing. *Theoretical Computer Science*, 154(2) :165–181, February 1996.

[46] Vamsidhar Reddy Gaddam, Ragnhild Eg, Ragnar Langseth, Carsten Griwodz, and Pål Halvorsen. The Cameraman Operating My Virtual Camera is Artificial : Can the Machine Be as Good as a Human ? *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(4) :1–20, June 2015.

[47] Vineet Gandhi. *Automatic Rush Generation with Application to Theatre Performances.* PhD thesis, Université de Grenoble, 2014.

[48] Daniel Gatica-Perez, Jean-Marc Odobez, Kevin Smith, and Guillaume Lathoud. Tracking People In Meetings With Particles. In *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services*, page 10, 2005.

[49] Michael Gleicher. Towards Virtual Videography. In *Proceedings of the Eighth ACM International Conference on Multimedia*, pages 375–378, 2000.

[50] H. Grabner, M. Grabner, and H. Bischof. Real-Time Tracking via On-line Boosting. In *Procedings of the British Machine Vision Conference 2006*, pages 6.1–6.10, Edinburgh, 2006. British Machine Vision Association.

[51] Terry Halpin. ORM/NIAM Object-Role Modeling. In Peter Bernus, Kai Mertins, and Günter Schmidt, editors, *Handbook on Architectures of Information Systems*, International Handbooks on Information Systems, pages 81–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[52] Terry A. Halpin. UML Data Models From An ORM Perspective. *Journal of Conceptual Modeling*, 1(1-10) :12, 1998.

[53] Pal Halvorsen, Simen S\a egrov, Asgeir Mortensen, David K. C. Kristensen, Alexander Eichhorn, Magnus Stenhaug, Stian Dahl, Hakon Kvale Stensland, Vamsidhar Reddy Gaddam, Carsten Griwodz, and Dag Johansen. Bagadus : An Integrated System for Arena Sports Analytics : A Soccer Case Study. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 48–59, New York, NY, USA, 2013. ACM.

[54] Arun Hampapur, Ramesh Jain, and Terry E. Weymouth. Production model based digital video segmentation. *Multimedia tools and applications*, 1(1) :9–46, 1995.

[55] John H.L. Hansen and Taufiq Hasan. Speaker Recognition by Machines and Humans : A tutorial review. *IEEE Signal Processing Magazine*, 32(6) :74–99, November 2015.

[56] A. G. Hauptmann and M. J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings*, pages 168–179, April 1998.

[57] Bernd Heisele, Thomas Serre, and T. Poggio. A Component-based Framework for Face Detection and Identification. *International Journal of Computer Vision*, 74(2) :167–181, August 2007.

[58] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In *Computer Vision – ECCV 2012*, volume 7575, pages 702–715. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[59] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3) :583–596, March 2015.

[60] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4) :1738–1752, April 1990.

[61] Christoph S. Herrmann. Human EEG responses to 1-100 Hz flicker : Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain Research*, 137(3-4) :346–353, April 2001.

[62] Rami Hodrob and Mustafa Jarrar. ORM to OWL 2 DL Mapping. In *Proceedings of the International Conference on Intelligent Semantic Web – Applications and Services*, pages 131–137, 2010.

[63] Thanarat Horprasert, David Harwood, and Larry S. Davis. A robust background subtraction and shadow detection. In *In Proceedings of the Asian Conference on Computer Vision*, 2000.

[64] Dries Hulens, Toon Goedemé, and Tom Rumes. Autonomous Lecture Recording with a PTZ Camera While Complying with Cinematographic Rules. In *4 Canadian Conference on Computer and Robot Vision*, pages 371–377, May 2014.

[65] Hayley Hung and Sileye O Ba. Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 830–833, Dallas, TX, USA, 2010. IEEE.

[66] ISO 11664-4. ISO 11664-4 : 1976 L* a* b* Colour Space. *Joint ISO/CIE Standard, ISO*, pages 11664–4, 2008.

[67] Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. Gaze-Driven Video Re-Editing. *ACM Trans. Graph.*, 34(2) :21 :1–21 :12, March 2015.

[68] Peter Janku, Karel Koplik, Tomáš Dulík, and Istvan Szabo. Comparison of tracking algorithms implemented in OpenCV. *MATEC Web of Conferences*, 76 :04031, January 2016.

[69] George H. Joblove and Donald Greenberg. Color Spaces for Computer Graphics. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '78, pages 20–25, New York, NY, USA, 1978. ACM.

[70] Pakorn Kaewtrakulpong and Richard Bowden. An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. In *Video-Based Surveillance Systems*, pages 135–144. Springer, Boston, MA, 2002.

[71] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-Backward Error : Automatic Detection of Tracking Failures. In *2010 20th International Conference on Pattern Recognition*, pages 2756–2759, Istanbul, Turkey, August 2010. IEEE.

[72] Yoshinari Kameda, Satoshi Nishiguchi, and Michihiko Minoh. CARMUL : Concurrent automatic recording for multimedia lecture. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings*, pages 677–680, Baltimore, MD, USA, July 2003. IEEE.

[73] Manisha M. Kasar, Debnath Bhattacharyya, and Tai-hoon Kim. Face Recognition Using Neural Network : A Review. *International Journal of Security and Its Applications*, 10(3) :81–100, March 2016.

[74] Laxmikant. S. Kate, Monica M. Waghmare, and Amrit Priyadarshi. An approach for automated video indexing and video search in large lecture video archives. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–5, January 2015.

[75] D. Kato, Y Yamada, K. Abe, A. Ishikawa, K. Ishiyama, and M. Obata. Analysis of the Camerawork of Broadcasting Cameramen. *SMPTE Journal*, 106(2) :108–116, February 1997.

[76] Kenji Kira and Larry A. Rendell. The Feature Selection Problem : Traditional Methods and a New Algorithm. In *Proceedings Tenth National Conference on Artificial Intelligence*, pages 129–134, January 1992.

[77] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, 2 :1137–1143, 1995.

[78] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2) :273–324, December 1997.

[79] Dimitrios Kosmopoulos, Anastasios Doulamis, Alexandros Makris, Nikolaos Doulamis, Sotirios Chatzis, and Stuart E. Middleton. Vision-based production of personalized video. *Signal Processing : Image Communication*, 24(3) :158–176, March 2009.

[80] Radek Kubicek, Pavel Zak, Pavel Zemcik, and Adam Herout. Automatic Video Editing for Multimodal Meetings. In *Computer Vision and Graphics*, Lecture Notes in Computer Science, pages 260–269. Springer, November 2008.

[81] Henry O. Lancaster. The Chi-squared Distribution. *iometrical Journal*, 13(5) :363–364, 1971.

[82] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. AV16.3 : An Audio-Visual Corpus for Speaker Localization and Tracking. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 182–195. Springer, Berlin, Heidelberg, June 2004.

[83] Jean Louis Le Moigne. *La Modélisation Des Systèmes Complexes.* Dunod, Paris, 1999. OCLC : 552032089.

[84] Colin Le Nost, Florent Lefevre, Vincent Bombardier, Patrick Charpentier, Nicolas Krommenacker, and Bertrand Petat. Automatic video editing : Original tracking method applied to basketball players in video sequences. In *8th International Conference on Image and Signal Processing, ICISP 2018*, Cherbourg, France, July 2018.

[85] Qiong Liu, Yong Rui, Anoop Gupta, and Jonathan J. Cadiz. Automating camera management for lecture room environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 442–449, Seattle, January 2001.

[86] Pengfei Luo, Min Zhang, Zabih Ghassemlooy, Stanislav Zvanovec, Shulan Feng, and Philipp Zhang. Undersampled-Based Modulation Schemes for Optical Camera Communications. *IEEE Communications Magazine*, 56(2) :204–212, February 2018.

[87] Lucia Maddalena and Alfredo Petrosino. Self Organizing and Fuzzy Modelling for Parked Vehicles Detection. In *Advanced Concepts for Intelligent Vision Systems*, Lecture Notes in Computer Science, pages 422–433. Springer, Berlin, Heidelberg, September 2009.

[88] N. Madhu and R. Martin. A Versatile Framework for Speaker Separation Using a Model-Based Speaker Localization Approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7) :1900–1912, September 2011.

[89] Nilesh Madhu and Rainer Martin. A scalable framework for multiple speaker localization and tracking. In *In Proceedings of the International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC 2008*, 2008.

[90] Aziz Makandar and Daneshwari Mulimani. Key frame extraction and Object Detection in the Sports Video. In *International Conference on Soft Computing Techniques in Engineering and Technology*, page 4, 2016.

[91] M. Manafifard, H. Ebadi, and H. Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159(Supplement C) :19–46, June 2017.

[92] David A. Marca and Clement L. McGowan. *SADT : Structured Analysis and Design Technique*. McGraw-Hill, Inc., New York, NY, USA, 1987.

[93] Thomas Marill and D Green. On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1) :11–17, January 1963.

[94] Sujeet Mate. *Automatic Mobile Video Remixing and Collaborative Watching Systems*. PhD thesis, Tempere University of Technology, Tempere, February 2017.

[95] Aditya Mavlankar, Piyush Agrawal, Derek Pang, Sherif Halawa, Ngai-Man Cheung, and Bernd Girod. An interactive region-of-interest video streaming system for online lecture viewing. In *18th International Packet Video Workshop*, pages 64–71, December 2010.

[96] Muhammad Owais Mehmood. *People Detection Methods for Intelligent Multi-Camera Surveillance Systems*. PhD thesis, Ecole Centrale de Lille, September 2015.

[97] Alberto Mendez-Villanueva, Martin Buchheit, Ben Simpson, and Pitre Bourdon. Match Play Intensity Distribution in Youth Soccer. *International journal of sports medicine*, 34, September 2012.

[98] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116 :374–388, 1976.

[99] Shivappa M Metagar, Anil S Naijk, and V D Chavan. A survey on Content Based Video Retrieval and Analysis using Image Processing. *International Journal of Modern Trends in Engineering and Research*, 03(02) :48–52, 2016.

[100] Petr Motlicek, Stefan Duffner, Danil Korchagin, Hervé Bourlard, Carl Scheffler, Jean-Marc Odobez, Giovanni Del Galdo, Markus Kallinger, and Oliver Thiergart. Real-Time Audio-Visual Analysis for Multiperson Videoconferencing. *Advances in Multimedia*, 2013 :175745 :1–175745 :21, August 2013.

[101] Sugata Mukhopadhyay and Brian Smith. Passive Capture and Structuring of Lectures. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 477–487, New York, NY, USA, 1999. ACM.

[102] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *Proceedings 2002 IEEE International Conference on Robotics and Automation*, volume 1, pages 1043–1049 vol.1, 2002.

[103] Rayat Neerja and Walia Ekta. Face recognition using improved fast PCA algorithm. In *2008 Congress on Image and Signal Processing*, pages 554–558, Sanya, Hainan, China, January 2008. IEEE.

[104] Viet Nguyen, Yaqin Tang, Ashwin Ashok, Marco Gruteser, Kristin Dana, Wenjun Hu, Eric Wengrowski, and Narayan Mandayam. High-rate flicker-free screen-camera communication with spatially adaptive embedding. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, San Francisco, CA, USA, April 2016. IEEE.

[105] Gerardus M. Nijssen and Terry A. Halpin. *Conceptual Schema and Relational Database Design*. Prentice-Hall, Sidney, 1989.

[106] Jim Owens. *Television Sports Production*. Elsevier, fourth edition edition, 2007.

[107] Derek Pang, Sameer Madan, Serene Kosaraju, and Tarun Vir Singh. Automatic Virtual Camera View Generation for Lecture Videos. Technical report, Stanford, 2010.

[108] Pascaline Parisot and Christophe De Vleeschouwer. Scene-specific classifier for effective and efficient team sport players detection from a single calibrated camera. *Computer Vision and Image Understanding*, 159 :74–88, June 2017.

[109] B. V. Patel and B. B. Meshram. Content based video retrieval. *The International journal of Multimedia & Its Applications*, 4(5) :77–98, October 2012.

[110] Vincent Pinel and Christophe Pinel. *Dictionnaire technique du cinéma - 3e éd.* Armand Colin, June 2016.

[111] Claudio Pinhanez and Aaron F. Bobick. Intelligent Studios : Using Computer Vision to Control TV Cameras. In *IJCAI'95 Workshop on Entertainment and AI/Alife*, 1995.

[112] Claudio Pinhanez and Aaron F Bobick. Approximate World Models : Incorporating Qualitative and Linguistic Information into Vision Systems. In *AAAI Conference*, page 8, 1996.

[113] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11) :1119–1125, November 1994.

[114] Francis Quek, David McNeill, ashid Ansari, Xin-Feng Ma, Robert Bryll, Susan Duncan, and Karl-Erik McCullough. Gesture cues for conversational interaction in monocular video. In *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99*, pages 119–126, September 1999.

[115] J. Ross Quinlan. Introduction of decision trees. *Machine Learning*, 1(1) :81–106, March 1986.

[116] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[117] Jinchang Ren, Ming Xu, James Orwell, and Graeme A. Jones. Multi-camera video surveillance for real-time analysis and reconstruction of soccer games. *Machine Vision and Applications*, 21(6) :855–863, October 2010.

[118] Richard D. Roberts. Undersampled frequency shift ON-OFF keying (UFSOOK) for camera communications (CamCom). In *2013 22nd Wireless and Optical Communication Conference*, pages 645–648, May 2013.

[119] Yong Rui, Anoop Gupta, Jonathan Grudin, and Liwei He. Automating lecture capture and broadcast : Technology and videography. *Multimedia Systems*, 10(1) :3–15, June 2004.

[120] Nirzhar Saha, Shareef Ifthekhar, Nam-Tuan Le, and yeong Min Jang. Survey on optical camera communications : Challenges and opportunities. *IET Optoelectronics*, 9(5) :172–183, 2015.

[121] Catarina B. Santiago, Armando Sousa, and Luis Paulo Reis. Vision system for tracking handball players using fuzzy color processing. *Machine Vision and Applications*, 24(5) :1055–1074, July 2013.

[122] Ngowda Shruthi and Sachan Priyamvada. Dominant frame extraction for video indexing. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 1799–1803, May 2017.

[123] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas. Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1) :133–137, January 2009.

[124] Mohamad Hoseyn Sigari, Naser Mozayani, and Hamid Reza Pourreza. Fuzzy Running Average and Fuzzy Background Subtraction : Concepts and Application. *IJCSNS International Journal of Computer Science and Network Sercurity*, 8(2) :138–143, 2008.

[125] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, April 1994.

[126] Cees Snoek, Marcel Worring, Jan-Mark Geusebroek, Dennis Koelma, Frank Seinstra, and Arnold Smeulders. Semantic Video Indexing. In *Multimedia Retrieval*, pages 225–249. Springer, August 2007.

[127] David Sodoyer, Bertrand Rivet, Laurent Girin, Jean-Luc Schwartz, and Christian Jutten. An Analysis of Visual Speech Information Applied to Voice Activity Detection. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–601–I–604, Toulouse, France, 2006. IEEE.

[128] Fabrice Souvannavong. *Indexation et recherche de plans vidéo par le contenu sémantique*. PhD thesis, Télécom ParisTech, 2005.

[129] C. Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 :246–252, 1999.

[130] Sébastien Stillittano, Vincent Girondel, and Alice Caplier. Lip contour segmentation and tracking compliant with lip-reading application constraints. *Machine Vision and Applications*, 24(1) :1–18, January 2013.

[131] Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1) :32–46, April 1985.

[132] Hwang Tae-Hyun, Joo In-Hak, and Cho Seong-Ik. Detection of Traffic Lights for Vision-Based Car Navigation System. In *Advances in Image and Video Technology*, pages 682–691. Springer, Berlin, Heidelberg, December 2006.

[133] Yoshinao Takemae, Kazuhiro Otsuka, and Junji Yamato. Automatic Video Editing System Using Stereo-based Head Tracking for Multiparty Conversation. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 1817–1820, New York, NY, USA, 2005. ACM.

[134] Sudeep D. Thepade and Ashvini A. Tonge. An optimized key frame extraction for detection of near duplicates in content based video retrieval. In *2014 International Conference on Communication and Signal Processing*, pages 1087–1091, April 2014.

[135] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio Segmentation and Speaker Localization in Meeting Videos. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 1150–1153, 2006.

[136] Neha Vishwakarma. Face Recognition System Using Principal Component Analysis (PCA). *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)*, 5(6) :1759–1766, 2016.

[137] Jinjun Wang, Changsheng Xu, Engsiong Chng, Hanqing Lu, and Qi Tian. Automatic composition of broadcast sports video. *Multimedia Systems*, 14(4) :179–193, September 2008.

[138] Xiaoli Wei, Weiming Shen, and Shuangshuang Jiang. A Novel Algorithm for Video Retrieval Using Video Metadata Information. In *2009 First International Workshop on Education Technology and Computer Science*, volume 2, pages 1059–1062, March 2009.

[139] A. W. Whitney. A Direct Method of Nonparametric Measurement Selection. *IEEE Transactions on Computers*, C-20(9) :1100–1103, September 1971.

[140] Christopher R. Wren, Ali Azarbayejani, Trevor Darrell, and Alex P. Pentland. Pfinder : Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :780–785, July 1997.

[141] Min Xu, Ling-Yu Duan, Changsheng Xu, M. Kankanhalli, and Qi Tian. Event detection in basketball video using multiple modalities. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, volume 3, pages 1526–1530 vol.3, December 2003.

[142] Takeshi Yamada, Satoshi Nakamura, and Kiyohiro Shikano. Robust speech recognition with speaker localization by a microphone array. In *Procedings of the Fourth International Conference on Spoken Language*, volume 3, pages 1317–1320 vol.3, October 1996.

[143] Zhice Yang, Zeyu Wang, Jiansong Zhang, Chenyu Huang, and Qian Zhang. Wearables Can Afford : Light-weight Indoor Positioning with Visible Light. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '15*, pages 317–330, Florence, Italy, 2015. ACM Press.

[144] Takao Yokoi and Hironobu Fujiyoshi. Virtual camerawork for generating lecture video from high resolution images. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4 pp.–, July 2005.

[145] Roberto Yus, Eduardo Mena, Sergio Ilarri, Arantza Illarramendi, and Jorge Bernad. MultiCAMBA : A system for selecting camera views in live broadcasting of sport events using a dynamic 3D model. *Multimedia Tools and Applications*, 74(11) :4059–4090, June 2015.

[146] Zhi Zeng, Jianyuan Jia, Dalin Yu, Yilong Chen, and Zhaofei Zhu. Pixel Modeling Using Histograms Based on Fuzzy Partitions for Dynamic Background Subtraction. *IEEE Transactions on Fuzzy Systems*, 25(3) :584–593, June 2017.

[147] Yifan Zhang, Changsen Xu, Yong Rui, Jinqiao Wang, and Hanqing Lu. Semantic Event Extraction from Basketball Games using Multi-Modal Analysis. In *2007 IEEE International Conference on Multimedia and Expo*, pages 2190–2193, July 2007.

[148] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. Face recognition : A literature survey. *ACM Computing Surveys*, 35(4) :399–458, December 2003.

[149] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27, 2006.

Résumé

Les travaux présentés dans cette thèse CIFRE liant le CRAN et la société CitizenCam visent la captation et la diffusion d'évènements publics à moindre coût. Ainsi, l'entreprise souhaite proposer un système de montage automatique, adaptable à chaque contexte d'application et prenant en compte les desiderata des spectateurs. L'étude bibliographique sur le montage automatique de séquences vidéo, présentée dans le premier chapitre montre que les méthodes existantes sont très spécifiques au contexte applicatif et de ce fait très peu généralisables. L'objectif du deuxième chapitre est donc de proposer une approche méthodologique du montage automatique, basée sur une structure générique pouvant être adaptée au contexte, tout en prenant en compte des préférences utilisateurs. Cette approche, basée sur la modélisation des connaissances du contexte applicatif par la méthode NIAM-ORM, nous permet d'identifier les personnes (POI) et actions (AOI) d'intérêts. La connaissance modélisée facilite également le choix et le paramétrage des algorithmes d'extraction des caractéristiques des POI et AOI nécessaires au montage.
Le chapitre 3 s'intéresse à la mise en place d'un système de montage automatique dans le cas de conseils municipaux avec la proposition d'une méthode originale de détection de locuteur et son identification basée sur le concept VLC. La diffusion des matchs de basketball est abordée dans le chapitre 4 avec la proposition d'une méthode de sélection automatique de caméras pour la diffusion de l'AOI "jeu notable" avec deux personnalisations que sont la détection de lancer francs et le suivi de joueurs.
La méthodologie proposée est ainsi validée par son application à ces deux types d'évènements.

Mots-clés: Modélisation de connaissances, Montage automatique, Détection et Identification de personnes, Visible Light Communication

Abstract

This thesis, resulting from a collaboration between CRAN and CitizenCam, aims to capture and broadcast public events at a lower cost. Thus, the company wishes to offer an automatic editing system, adaptable to each application context and taking into account the spectators' requirements. A bibliographical study on the automatic editing of video sequences is presented in the first chapter. This study shows that the existing methods are very specific to the application context and thus not very generalizable. The objective of the second chapter is therefore to propose a methodological approach to automatic editing, based on a generic framework adaptable according to the context, while taking into account user preferences. This approach, based on the knowledge modelling of the application context using the NIAM-ORM method, allows us to identify people (POI) and actions (AOI) of interest. The modelled knowledge also facilitate the choice and configuration of algorithms for extracting the POI and AOI features required for editing.
Chapter 3 focuses on implementation of an automatic editing system for municipal councils with the proposal of an original speaker detection method and its identification based on the VLC concept. The broadcasting of basketball matches is covered in Chapter 4 with the proposal of an automatic camera selection method for broadcasting of the AOI "relevant game" with two customizations that are free throw detection and player tracking.
Thus, the proposed methodology is validated by its application to this two types of events.

Keywords: Knowledge modeling, Automatic editing, Detection and identification of people, Visible Light Communication