



Reconnaissance dynamique de personnes dans les émissions audiovisuelles

Rémi Auguste

► To cite this version:

Rémi Auguste. Reconnaissance dynamique de personnes dans les émissions audiovisuelles. Vision par ordinateur et reconnaissance de formes [cs.CV]. Université de Lille 1, 2014. Français. <tel-01114399>

HAL Id: tel-01114399

<https://tel.archives-ouvertes.fr/tel-01114399>

Submitted on 9 Feb 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



École Doctorale Sciences Pour l'Ingénieur Université Lille Nord-De-France

Doctorat de l'Université des Sciences et Technologies de Lille
(Spécialité informatique)

AUGUSTE Rémi

RECONNAISSANCE DYNAMIQUE DE PERSONNES DANS LES ÉMISSIONS AUDIOVISUELLES

Soutenue le 9 juillet 2014

Composition du jury :

Rapporteurs :

CHRISMENT Claude	Professeur des universités	Université Toulouse III
QUENOT Georges	Directeur de recherches	Laboratoire d'Informatique de Grenoble

Examinateur :

TISON Sophie	Professeur des universités	Université Lille 1
CARINCOTTE Cyril	Chef de département	Multitel

Directeur de thèse :

DJERABA Chaabane	Professeur des universités	Université Lille 1
------------------	----------------------------	--------------------

Co-encadrant de thèse :

MARTINET Jean	Maître de conférences	Université Lille 1
---------------	-----------------------	--------------------

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE
Laboratoire d'Informatique Fondamentale de Lille - UMR 802
U.F.R. d'I.E.E.A. - Bât M3 - 59655 VILLENEUVE D'ASCQ CEDEX
Tél : +33 (0)3 28 77 85 41 - email : direction@lifl.fr

Résumé

L’analyse automatique de contenu des vidéos en vue de leur annotation est un domaine de recherche en plein essor. Reconnaître les personnes apparaissant dans des émissions audiovisuelles permet une structuration automatique d’une quantité grandissante d’archives audiovisuelles. Nous présentons une approche dynamique originale de reconnaissance de personnes dans les flux vidéo. Cette approche est dynamique car elle tire avantage de la richesse des informations contenues dans la vidéo, contrairement aux approches statiques basées uniquement sur les images. L’approche proposée comprend deux volets. Le premier volet consiste à isoler toutes les occurrences de personnes d’une émission, et à les regrouper en clusters en se basant sur un descripteur original : les histogrammes spatio-temporels, ainsi que sur une mesure de similarité dédiée. L’originalité vient de l’intégration d’informations temporelles dans le descripteur, qui permet une estimation plus fiable de la similarité entre les occurrences de personnes. Le second volet propose la mise en œuvre d’une méthode de reconnaissance faciale. Différentes stratégies sont envisagées, d’une part pour identifier les occurrences de personnes selon les trames qui composent la séquence, et d’autre part pour propager les identités au sein des groupes selon leurs membres. Ces deux aspects de notre contribution ont été évalués à l’aide de corpus de données réelles contenant des émissions issues des chaînes BFMTV et LCP. Les résultats des expérimentations menées indiquent que l’approche proposée permet d’améliorer notablement la précision de reconnaissance en prenant en compte la dimension temporelle.

Mots clefs : Vision par ordinateur, analyse vidéo, histogrammes de couleurs, reconnaissance de personnes, ré-identification de personnes.

Abstract

The annotation of video streams by automatic content analysis is a growing field of research. The possibility of recognising persons appearing in TV shows allows to automatically structure ever-growing video archives. We present an original and dynamic approach to person recognition from video streams. This approach is dynamic as it benefits from the motion information contained in videos, whereas the static approaches are solely based on still images. The proposed approach is composed of two parts. In the first one, we extract persontracks from the shows and cluster them using a new descriptor and its associated similarity measure : space-time histograms. The originality of our approach is the integration of temporal data into the descriptor. Experiments show that it provides a better estimation of the similarity between persontracks. In the second part of our approach, we propose to use a facial recognition method which aims at "naming" the clusters. Various strategies are considered to assign an identity to a persontrack using its frames and to propagate this identity to members of the same cluster. These two aspects of our contribution have been evaluated using a corpus of real life TV shows broadcasted on BFMTV and LCP TV channels. The results of our experiments show that our approach significantly improves the precision of the recognition process thanks to the use of the temporal dimension.

Keywords : Computer vision, video analysis, color histograms, person recognition, person re-identification.

Remerciements

Je remercie tout d'abord mon directeur de thèse, Professeur Chaabane Djeraba, pour m'avoir soutenu et fait confiance depuis de nombreuses années et pour m'avoir aidé à m'épanouir dans la recherche.

Je tiens à remercier mon co-encadrant, Jean Martinet, pour sa patience et son suivi attentif, ainsi que pour les nombreuses remarques toujours constructives et les nombreux débats que nous avons eus ensemble tout au long de cette thèse.

Je remercie le Professeur Claude Chrisment ainsi que le Directeur de recherches Georges Quenot d'avoir accepté de rapporter mon mémoire de thèse et le Professeur Sophie Tison et Cyril Carincotte d'être examinateurs. Je les remercie de m'avoir fait l'honneur de participer à mon jury de soutenance de thèse.

Merci, bien évidemment, à toute l'équipe FOX pour l'ambiance chaleureuse et néanmoins studieuse qui règne dans l'équipe : Amel Aissaoui, José Mennesson, Pierre Tirilly, Affa Dahmane, Taner Danisman et Marius Bilasco.

Je remercie Nadia Meknini et Rebeca Murillo dont les stages de qualité, liés à ces travaux, ont permis d'explorer efficacement certaines pistes.

Je salue tous les membres du consortium PERCOL pour la superbe ambiance qui régnait dans notre groupe. Le travail réalisé est impressionnant, nous avons bien mérité la victoire. Je garde un souvenir inoubliable de nos workshops.

Je remercie José Mennesson, Pierre Tirilly, Robin Pochet et Audrey Debarbieux pour leur aide précieuse dans la relecture de ce manuscrit de thèse.

Je remercie Audrey Debarbieux pour son soutien permanent qui m'a particulièrement bien aidé dans les moments difficiles.

Enfin, je remercie mes amis et ma famille pour leur patience, leurs encouragements et leur compréhension concernant mon absence, surtout vers la fin de la thèse.

Contexte de la thèse

Cette thèse a été en partie financée par l'agence nationale pour la recherche (ANR) via le défi REPÈRE¹ (reconnaissance de personnes dans des contenus audiovisuels) : <http://www.defi-repere.fr/>. Il s'inscrit dans les objectifs du programme Contenus et Interactions (CONTINT) de l'ANR, en partenariat avec la direction générale de l'armement (DGA).

Trois consortiums ont été financés, pour une durée de 36 mois. L'objectif est de réaliser un système de reconnaissance de personnes dans des émissions audiovisuelles. Les différentes sources d'information utilisées sont :

- l'image dans laquelle les personnes sont visibles,
- les textes en incrustation dans lesquels le nom des personnes apparaît,
- la bande son dans laquelle la voix des locuteurs est reconnaissable,
- le contenu du signal de parole dans lequel le nom des personnes est prononcé.

La performance de ces systèmes a été évaluée chaque année au moyen d'une campagne d'évaluations, qui porte sur des émissions audiovisuelles, journaux, débats, divertissements, en langue française.

Les travaux de thèse présentés dans ce manuscrit ont été mis à profit dans le consortium PERCOL (*person recognition in audiovisual content*) qui est composé de Aix-Marseille Université, l'Université Lille 1, l'Université d'Avignon et d'Orange Labs. Les corpus utilisés pour évaluer les propositions que nous avons faites dans ce travail de thèse sont issus des données mises à disposition par le défi REPÈRE. Il s'agit de vidéos annotées d'émissions télévisuelles diffusées par les chaînes BFMTV et LCP. Le consortium PERCOL a obtenu le meilleur classement lors de la clôture du défi REPÈRE.



1. ANR 2010-CORD-102-01

Table des matières

I Introduction - État de l'art	1
1 Flux vidéo et reconnaissance de personnes	3
1.1 Progrès de l'acquisition vidéo	3
1.2 Dimension sociétale de la vidéo	4
1.3 Conséquences	5
1.4 Applications de la reconnaissance de personnes	5
1.4.1 Difficultés de la reconnaissance de personnes	6
1.4.2 Cas particulier des émissions télévisées	7
1.5 Propositions	8
2 La reconnaissance de personnes	11
2.1 Statique vs. dynamique	11
2.1.1 Approches statiques	12
2.1.2 Approches dynamiques	18
2.2 Regroupement des occurrences de personnes	19
2.2.1 Les histogrammes de couleurs et leurs extensions	21
2.2.2 Mesures de distance entre histogrammes	23
2.2.3 Distance entre séquences	25
2.2.4 Espace de représentation des couleurs	28
2.2.5 Clustering d'histogrammes	37
2.3 Étiquetage d'ensembles	38
2.4 Synthèse de l'état de l'art	39
3 Mise en perspective de nos contributions	41
3.1 Approche classique de reconnaissance	41
3.2 Proposition générale	43
II Regroupement des occurrences vidéo de personnes	47
4 Regroupement des OVP	49
4.1 Histogrammes spatio-temporels	50
4.2 Interprétations	50
4.2.1 Illustration	50
4.2.2 Information de mouvement	51
4.2.3 Interprétation de la dimension temporelle des HST	52
4.3 Stratégies de construction des HST	52
4.3.1 Accumulation des pixels	52

4.3.2	Fenêtre glissante	53
4.3.3	Fenêtre sautante	54
4.3.4	Séparation des canaux colorimétriques	55
4.4	Mesure de similarité	56
4.5	Complexité	57
4.5.1	Complexité de la construction	57
4.5.2	Coût mémoire des descripteurs	58
4.5.3	Complexité des mesures de similarités	59
4.6	Matrices de similarités	59
4.7	Choix du nombre de partitions des histogrammes spatio-temporels	59
4.8	Clustering d'histogrammes spatio-temporels	61
4.9	Résumé des propositions	62
5	Validation du regroupement d'OVP	63
5.1	Présentation des expérimentations	63
5.2	Présentation des données de test	63
5.3	Prétraitements des données	64
5.4	Calcul des matrices de similarités	65
5.5	Paramétrage des HST	65
5.5.1	Variation du nombre de partitions	65
5.5.2	Variation de l'espace de couleurs	67
5.5.3	Comparaison avec un descripteur de textures	68
5.5.4	Comparaison des mesures de similarités	69
5.5.5	Stratégie de construction	70
5.6	Significativité de la mesure de similarité	72
5.6.1	Test de Student	72
5.6.2	Séries de données testées	73
5.6.3	Significativité de la similarité	73
5.6.4	Significativité de l'augmentation du nombre de partitions	73
5.6.5	Résumé de la significativité de nos résultats	73
5.7	Qualité du regroupement	74
5.7.1	Regroupement hiérarchique ascendant	74
5.7.2	Mesure de pureté	74
5.7.3	Mesure de fragmentation	75
5.7.4	Vrais/faux positifs/négatifs	77
5.7.5	Indice de Rand	78
5.7.6	Rappel/Précision	78
5.7.7	F-mesure	79
5.7.8	Indice Fowlkes–Mallows	79
5.7.9	Résumé de la qualité du regroupement	81
5.8	Précision et clustering	82
5.8.1	Précision et pureté	82
5.8.2	Précision et fragmentation	83
5.8.3	Précision et Rand	83
5.8.4	Résumé de la mesure de précision pour l'évaluation du clustering	83
5.9	Comparaison avec des méthodes existantes	84
5.9.1	Précision	84
5.9.2	Efficience	86

5.9.3	Précision à coût mémoire constant	87
5.10	Résumé des résultats des expérimentations	90
III	Nommage des personnes	91
6	Nommage de groupes	93
6.1	Introduction	93
6.2	Nommage d'une occurrence à partir de ses trames	93
6.2.1	Utilisabilité d'une trame	94
6.2.2	Reconnaissance basée sur une trame unique	94
6.2.3	Reconnaissance basée sur plusieurs trames	96
6.2.4	Synthèse du nommage d'une occurrence à partir de ses trames . .	98
6.3	Nommage d'un groupe à partir de ses occurrences	98
6.3.1	Sélection d'une occurrence unique	99
6.3.2	Choix du nombre d'occurrences	99
6.4	Résumé sur le nommage des groupes d'occurrences	101
7	Validation des approches de nommage	103
7.1	Expérimentations	103
7.1.1	Présentation des données	103
7.1.2	Utilisation des données	104
7.2	Taux de reconnaissance de référence	106
7.2.1	Calcul de la précision	106
7.2.2	Résultat de référence	106
7.3	Identification des OVP	107
7.4	Variation de la proportion de visages considérés	108
7.4.1	Selon l'ordre de la séquence	109
7.4.2	Du milieu vers les extrémités	111
7.4.3	De façon aléatoire	111
7.4.4	Discussion sur la proportion de visages à considérer	112
7.5	Propagation d'identités à partir d'OVP nommées	113
7.5.1	Identités issues d'un vote à la majorité relative	114
7.5.2	Identités issues d'un vote à la majorité absolue	114
7.5.3	Discussion sur la détermination des identités initiales	115
7.6	Variation de la proportion d'occurrences utilisées	115
7.6.1	Propagation par sélection aléatoire	115
7.6.2	Propagation par ordre de similarité	116
7.6.3	Propagation par score de confiance	117
7.6.4	Discussion sur les stratégies de propagation	118
7.7	Conclusion	119
IV	Conclusion et perspectives	121
8	Synthèse de nos contributions et perspectives	123
8.1	Synthèse de nos contributions	123
8.2	Mise en œuvre de nos contributions	124

8.3 Perspectives de travail	125
---------------------------------------	-----

Première partie

Introduction - État de l'art

Chapitre 1

Les flux vidéo et la reconnaissance de personnes

Un promeneur dans la ville de Londres est filmé en moyenne par 300 caméras de surveillance. Le lecteur de cette thèse aura probablement une caméra vidéo proche de lui, dans sa poche ou peut-être sur son bureau. Si celui-ci lit ce manuscrit de thèse sur un ordinateur portable ou une tablette, il est probablement observé par une caméra située au-dessus de l'écran. En effet, la vidéo est omniprésente aujourd'hui.

1.1 Progrès de l'acquisition vidéo

Cette omniprésence s'explique premièrement par les progrès réalisés concernant les dispositifs d'acquisition vidéo. Les premiers capteurs vidéo capables de convertir une image optique en signal électrique datent des années 1930 avec les tubes caméras [3]. Ces tubes étaient trop encombrants pour permettre leur portabilité avant le milieu des années 1970 (cf. Figure 1.1). Ils ont été remplacés à partir de 1999 par les capteurs CCD et CMOS. Ceux-ci sont composés d'une matrice de capteurs. Chacun est responsable d'un point de l'image (pixel). Les dispositifs d'acquisition ont pu être grandement miniaturisés. Les différentes améliorations successives ont permis la fabrication à grande échelle de dispositifs toujours plus complexes. Ceux-ci se sont ouverts au marché grand public. Ce matériel est ainsi progressivement devenu accessible au plus grand nombre. Le prix à la consommation pour les équipements photo et vidéo ne représente aujourd'hui que le dixième de leur prix de 1998¹. Cela explique en partie la démocratisation des équipements vidéo.

L'omniprésence de la vidéo ne s'explique pas uniquement par les progrès techniques qui entourent l'acquisition de la vidéo, mais aussi par les progrès concernant son stockage et sa diffusion. L'évolution des supports d'enregistrement vidéo est directement liée aux avancées en matière d'acquisition vidéo et de stockage informatique. De 1956 à 2000, le principal médium d'enregistrement est la cassette vidéo. L'information est encodée sur une bande magnétique souple. La vidéo est principalement stockée analogiquement sur ce support. Le début de l'enregistrement numérique marque la fin des cassettes vidéo. De 2000 à aujourd'hui, la cassette vidéo a été progressivement remplacée par le DVD. Les

1. Données INSEE BDM : Indice des prix à la consommation (mensuel, ensemble des ménages, métropole, base 1998) - Nomenclature COICOP : 09.1.2.1 - Équipement photo et cinéma, instruments d'optique.



FIGURE 1.1 – Une des premières caméras portables. Il s'agit du modèle SL-F1 Betamax de Sony. L'enregistrement sur cassette se fait dans le boîtier relié à la caméra, qui peut être transporté par une seconde personne.

supports optiques (CD, DVD et Blu-Ray) stockent l'information vidéo en marquant un disque en rotation avec un faisceau laser. Le support Blu-Ray ne s'est pas encore imposé auprès du public pour la sauvegarde de vidéo. De nos jours, les supports physiques tendent à disparaître de l'environnement de l'utilisateur au profit de la dématérialisation et du stockage en ligne (*cloud*). Ce dernier est possible grâce à la démocratisation de l'accès à Internet. En 2013, 79,6% des Français ont accès à Internet². De plus, l'accès haut débit³ se généralise (70% des internautes français). Il est ainsi possible, pour une part de plus en plus importante de la population, de transférer une vidéo en haute définition en moins de temps qu'il n'en faut pour la visionner. Les différentes avancées techniques qui entourent la vidéo, de son acquisition à son partage, ont permis la démocratisation de son usage. L'INSEE estime qu'en 2010, quasiment tous les foyers de France étaient équipés de télévision, de magnétoscope ou lecteur DVD. De plus, tous les ordinateurs portables et smartphones vendus aujourd'hui sont équipés d'une webcam et le taux d'équipement des foyers en téléphones portables et en connexion Internet est en constante augmentation. En 2012, 46% des Français sont équipés de smartphones⁴. Ainsi, une part très importante de la population française est capable de réaliser l'acquisition, l'affichage et la diffusion par Internet de vidéos.

1.2 Dimension sociétale de la vidéo

Aujourd'hui, la démocratisation de la vidéo est telle qu'elle est devenue un phénomène de société, l'augmentation du nombre de chaînes télévisées en témoigne. La première chaîne télévisée nationale a été créée en 1935. En 1986, on comptait 6 chaînes nationales.

2. Données INSEE : Tableaux de l'Économie Française - Édition 2014 - avril 2014.

3. Un accès Internet haut débit est un accès à Internet offrant un débit d'au moins 500 kbit/s.

4. Données de l'institut Médiamétrie 2012.

Aujourd’hui, après le passage à la télévision numérique terrestre (TNT), on en compte plus de 80 en France. Cela représente donc 80 heures de contenu vidéo diffusé pour chaque heure qui s’écoule. En 2008, les Français ont regardé la télévision en moyenne 3h24⁵ par jour. De plus, avec l’augmentation de la pénétration d’Internet dans les foyers et l’augmentation de la vitesse des connexions, de nombreux sites de partage de vidéos sont apparus. Parmi les plus connus, on peut notamment citer YouTube et Dailymotion. Les différents sites de réseaux sociaux permettent aussi le partage de vidéos ; c’est le cas de Facebook, VKontakte et Google+. La fusion du réseau social Google+ avec la plate-forme de partage de vidéos Youtube illustre parfaitement l’importance sociale qu’acquiert la vidéo avec le temps. 100 heures de vidéo sont mises en ligne chaque minute sur la plateforme de partage de vidéos YouTube. Plus d’un milliard d’utilisateurs uniques consultent YouTube chaque mois. Tous les mois, les internautes regardent plus de six milliards d’heures de vidéo sur YouTube, soit presque une heure par personne dans le monde.

1.3 Conséquences

L’omniprésence de la vidéo fait qu’aujourd’hui, il devient difficile de traiter la quantité de vidéos disponibles pour en tirer des informations pertinentes. En ce qui concerne les organismes d’archivage vidéo, prenons comme exemple l’Institut National de l’Audiovisuel (INA) : ses archives couvrent presque 70 ans d’histoire de la télévision, avec notamment le premier journal télévisé français datant du 26 juin 1949. On estime qu’il faudrait 300 ans pour voir et écouter de façon ininterrompue toutes les archives de l’INA.

La question de la recherche de contenus dans cette masse colossale de vidéos se pose naturellement. On doit ainsi s’intéresser à ce que cherchent les utilisateurs dans les vidéos. Dans la page *Trends* du moteur de recherche Google⁶, pour les années 2011, 2012 et 2013, 5 requêtes parmi les 10 requêtes les plus populaires dans le monde concernent des personnes. Pour ces trois années, la requête mondiale la plus populaire sur Internet concerne une personnalité. Les 6 vidéos YouTube les plus vues sur Internet⁷ ont toutes le nom d’une personne dans leur titre. Enfin, si on consulte le site de l’INA, on remarque qu’une partie importante du site est dédiée à la recherche de vidéos de personnalités⁸. Ainsi, les personnes contenues dans les vidéos sont importantes pour les utilisateurs. Pour faciliter la recherche de vidéos contenant des personnes, il est utile de pouvoir annoter de telles vidéos pour pouvoir les indexer et effectuer des recherches. Le volume de données et la complexité de la tâche sont trop importants pour être réalisée par des personnes. Il est donc nécessaire d’automatiser cette tâche.

1.4 Applications de la reconnaissance de personnes

La problématique de la reconnaissance de personnes dans les vidéos est à la croisée de nombreux axes de recherche : l’indexation multimédia, la fouille de données, la vision par ordinateur, l’intelligence artificielle, la biométrie, etc. Les applications de la reconnaissance de personnes à partir de la vidéo sont multiples. On retrouve la reconnaissance de

5. Données de l’institut Médiamétrie 2008.

6. URL : <http://www.google.fr/trends>.

7. URL : <http://youtube-trends.blogspot.fr>.

8. URL : <http://www.ina.fr/pages-carrefours/toutes-les-personnalites>.

personnes dans la sécurité, par exemple aux postes frontières de certains pays, pour vérifier que l'identité réelle de la personne et celle indiquée dans son passeport correspondent. De même, la reconnaissance de personnes est utilisée pour déverrouiller automatiquement certains smartphones quand son propriétaire l'utilise. La reconnaissance de personnes à partir de vidéos se retrouve aussi dans le domaine de l'indexation vidéo. L'objectif est d'identifier les personnes présentes dans une vidéo pour ensuite effectuer des recherches ou des recoupements à partir de ces informations. Cette application intéresse notamment les réseaux sociaux, afin d'identifier les utilisateurs et de faciliter le partage. Les organismes d'archivage s'y intéressent pour sélectionner, organiser et documenter les vidéos afin de les éditorialiser sous forme de collections thématiques.

1.4.1 Difficultés de la reconnaissance de personnes

D'une façon générale, les problèmes que l'on rencontre lors de la reconnaissance de personnes concernent deux aspects : les variations d'apparence de la personne que l'on souhaite reconnaître d'une part, et les conditions de prise de vue de l'autre. La personne peut se montrer non-coopérative en prenant des postures particulières, allant du simple fait de baisser la tête jusqu'à l'occultation partielle ou complète de celle-ci (cf. Figure 1.2). Porter des lunettes, un couvre-chef, un foulard, du maquillage, présenter une pilosité particulière, etc. peut rendre les mécanismes de détection et de reconnaissance inefficaces. La plupart des approches de reconnaissance supposent la coopération, au moins passive, du sujet [116].



FIGURE 1.2 – Exemple dans lequel une personne se dissimule à l'aide de sa capuche, ses cheveux, ainsi que ses lunettes de soleil.

La seconde difficulté vient des conditions de prise de vue. Elle concerne le positionnement de la caméra par rapport aux personnes, ou les conditions d'éclairage de la scène. Le dispositif d'acquisition de l'image conditionne souvent le type d'approche pouvant être utilisé. En effet, celui-ci peut être de basse résolution, présentant beaucoup de bruit⁹,

9. Le bruit peut prendre la forme d'artefacts graphiques, de crénelage des silhouettes ou de nombreux

ou n'être capable que d'acquérir des images en niveaux de gris. C'est le cas notamment de la plupart des caméras de surveillance. Les conditions peuvent être défavorables si la lumière est trop faible ou orientée de façon à n'éclairer qu'une petite partie du visage.



FIGURE 1.3 – Exemple d'un éclairage du visage non propice à la reconnaissance de la personne (avec des artefacts de compression dûs au changement brusque de luminosité).

1.4.2 Cas particulier des émissions télévisées

Le contexte qui nous intéresse dans cette thèse est celui des émissions télévisées (journaux télévisés, débats, émissions et chroniques culturelles). Elles présentent de nombreuses caractéristiques intéressantes pour reconnaître les personnes. Les conditions d'éclairage sont maîtrisées, notamment pour les séquences vidéo filmées en studio. Les conditions de prise de vue sont idéales et la personne filmée est souvent face à la caméra. Enfin, les séquences vidéo constituant l'émission sont filmées dans un intervalle de temps réduit ce qui nous permet de faire une hypothèse de constance de leur apparence visuelle au cours de l'émission. Malgré les nombreux atouts que semblent présenter les émissions télévisées pour la reconnaissance de personnes, des difficultés subsistent.

En ce qui concerne les conditions de prise de vue, la caméra est fréquemment focalisée sur une personne en particulier, qui peut être en train de parler. Le nombre et la configuration des personnes présentes à l'écran sont variables. Des personnes peuvent être vues de dos ou de profil, c'est le cas notamment dans les émissions de débats où les personnes se font face. Les vêtements et accessoires des personnes peuvent rendre difficile leur reconnaissance. Certaines personnalités portent des lunettes de soleil, des couvre-chefs, des bijoux, du maquillage ou une pilosité, ce qui modifie leur apparence visuelle. En ce qui concerne la posture des personnes, elle est également changeante, ce qui constitue une autre source de variabilité d'apparence. Des personnes occultent leur visage avec leur main en parlant, ou encore des microphones (ou autres éléments du décor) peuvent cacher en partie le visage d'une personne (cf. Figure 1.4) rendant ainsi difficile leur reconnaissance. Enfin, il existe des problèmes intrinsèques à la vidéo. On peut citer

pixels incohérents dispersés.



FIGURE 1.4 – Exemple où une partie des personnes présentes à l'image ne sont pas filmées de face.



FIGURE 1.5 – Exemples d'occultation du visage des personnes.

la compression de la vidéo qui peut introduire des artefacts visuels notamment lors de mouvements importants au sein des vidéos (par exemple, la personne tourne la tête, un présentateur baisse la tête pour lire un texte, etc.). Par exemple, quand une personne se déplace dans le champ de vision de la caméra, l'apparence de la personne peut devenir très bruitée (cf. Figures 1.3 et 1.6).

1.5 Propositions

Les émissions télévisées contiennent un très grand nombre d'images et la reconnaissance de personnes dans une image a un coût calculatoire non-négligeable. De plus, comme nous venons de le voir, les émissions audiovisuelles ne sont pas toujours propices à la reconnaissance des personnes. Ainsi, l'utilisation exhaustive de toutes les trames pour reconnaître les personnes d'une émission audiovisuelle ne semble pas être une bonne approche. Les émissions audiovisuelles sont souvent enregistrées dans un intervalle de temps relativement court, par exemple de quelques heures dans le cas d'un débat. L'apparence des personnes (vêtements, coiffure, maquillage, bijoux, pilosité, etc.) ne varie donc pas au cours d'une émission donnée. Dans le cas où l'émission contient des reportages enre-



FIGURE 1.6 – Exemple d’artefacts de compression, suite au déplacement de la caméra et des personnes présentes à l’image.

gistrés à différentes périodes, l’apparence des personnes au sein de ceux-ci ne varie pas. Nous proposons de grouper les différentes occurrences d’une personne d’une émission en nous basant sur leur apparence, sous l’hypothèse que cette apparence ne varie pas au cours de l’émission. Ensuite, pour chaque groupe représentant une personne, nous déterminons l’identité de la personne en utilisant un algorithme de reconnaissance sur un sous-ensemble d’occurrences choisies dans le groupe. L’identité déterminée est alors propagée à l’ensemble du groupe.

Notre approche présente deux principaux avantages : le premier est qu’elle permet l’identification d’occurrence vidéo de personne par le biais de la propagation, en particulier quand l’identification directe échoue. Le second est que notre approche limite le recours à des algorithmes de reconnaissance, coûteux en temps de calcul, pour identifier les personnes. Les stratégies que nous proposons minimisent le nombre d’identifications nécessaires. Pour cela, nous prenons en compte à la fois l’aspect spatial et l’aspect temporel des personnes dans la vidéo. Comme nous le verrons par la suite, de nombreuses approches ne considèrent que l’aspect spatial des personnes. Nous pensons que l’aspect temporel des vidéos peut, en combinaison avec l’aspect spatial, augmenter la robustesse de la ré-identification et de la reconnaissance des personnes dans les vidéos.

La présentation détaillée de nos travaux est organisée de la manière suivante :

- Le **Chapitre 2** donne une présentation de l’état de l’art concernant la reconnaissance de personnes. Il distingue les méthodes statiques des méthodes dynamiques, et présente des méthodes de regroupement d’occurrences (clustering) basées sur des descripteurs couramment utilisés pour représenter l’apparence des personnes.
- Le **Chapitre 3** donne une vue d’ensemble de nos contributions. Il introduit également quelques définitions servant de base à notre travail.
- Le **Chapitre 4** propose un descripteur pour représenter chacune des occurrences vidéo de personnes, afin de les mettre en correspondance. Ce descripteur, appelé histogramme spatio-temporel, fournit une représentation discriminante des personnes présentes dans les occurrences vidéo. Les signatures servent de base à un

processus de regroupement dont l'objectif est de séparer les identités dans des groupes d'occurrences.

- Le **Chapitre 5** apporte une validation expérimentale des histogrammes spatio-temporels comme descripteurs discriminants pour les occurrences vidéo de personnes. Ce chapitre étudie l'évolution de la précision de notre système en fonction du paramétrage. Nous identifions l'espace de couleur le plus approprié, le nombre de partitions optimal, ainsi que la stratégie de construction la plus adaptée. Une fois ces paramètres déterminés, Nous comparons les résultats obtenus, pour une tâche de recherche, dans les différents cas avec ceux obtenus à l'aide de notre approche. Enfin, nous évaluons le regroupement d'occurrences vidéo de personnes construit à partir de la matrice des similarités entre histogrammes spatio-temporels.
- Le **Chapitre 6** détaille les différentes stratégies que nous envisageons, d'une part pour assigner une identité aux occurrences de personnes selon les trames qui composent la séquence, et d'autre part pour propager les identités au sein des groupes selon leurs membres. L'objectif est de limiter le nombre d'identification d'occurrences, et de propager les identités aux groupes. Cela permet d'identifier plus d'occurrences de personnes qu'un système dépourvu de propagation, et améliore sensiblement la précision tout en nécessitant moins de calculs.
- Le **Chapitre 7** valide expérimentalement nos stratégies de nommage des personnes. Dans un premier temps, nous présentons les expérimentations qui servent à déterminer un taux de reconnaissance de référence. Celui-ci sert à évaluer les performances des approches pour déterminer l'identité des occurrences vidéo à partir de leurs trames. Après avoir assigné une identité à certaines occurrences choisies, nous propageons cette identité à l'ensemble du groupe. Le taux de reconnaissance de référence (*baseline*) permet d'évaluer les performances de la propagation selon le nombre d'occurrences vidéo de personnes considérées.
- Le **Chapitre 8** conclue ce manuscrit en résumant les points principaux de nos contributions, et propose quelques perspectives que nous envisageons d'explorer suite à ce travail de thèse.

Chapitre 2

La reconnaissance de personnes

Pour reconnaître une personne visuellement, la partie du corps la plus discriminante pour les êtres humains est le visage [20]. Les visages sont des objets tridimensionnels, et les informations utiles pour la reconnaissance peuvent être trouvées dans la géométrie et la texture du visage, ainsi que dans ses mouvements [71]. Nous nous intéressons à l'aspect temporel présent dans les vidéos pour mieux discriminer les personnes. Pour cela, nous classons les approches de l'état de l'art selon deux catégories : les approches statiques et les approches dynamiques. Les *approches statiques* (historiquement les premières en raison de l'évolution de la capacité de traitements des ordinateurs) se basent sur une ou plusieurs images de test pour reconnaître l'identité d'une personne. Dans le cas de plusieurs images, il s'agit d'un ensemble d'images représentant la même personne, sans relation temporelle ou de séquence entre elles. Ainsi, ces approches ne prennent pas en compte l'aspect temporel. Les *approches dynamiques* se basent sur des séquences d'images en considérant la relation temporelle qui lie ces images entre elles. Parmi les approches dynamiques, certaines réalisent une combinaison des résultats, obtenus par une approche statique appliquée à plusieurs images d'une séquence vidéo.

Les différentes approches de l'état de l'art de la reconnaissance de personnes présentent des limitations ne permettant pas leur emploi systématique sur toutes les trames d'une émission. Nous nous intéressons ainsi aux approches en *ré-identification* qui permettent de regrouper les occurrences de personnes selon leur similarité visuelle. Les approches en ré-identification de personnes basées sur les histogrammes présentent des caractéristiques intéressantes pour notre contexte. Nous allons ainsi étudier différentes variantes d'histogrammes et les métriques associées à ceux-ci. Les histogrammes sont pour la plupart basés sur les couleurs pour représenter une image ou une vidéo. Nous présentons ainsi différents espaces de couleurs. Nous terminons ce chapitre en présentant des approches de clustering et l'étiquetage d'ensembles.

2.1 Statique vs. dynamique

Dans cette section, nous discutons des approches statiques de reconnaissance en présentant quelques travaux représentatifs. Ensuite, nous présentons une étude globale sur les méthodes dynamiques (basées sur la vidéo). Nous poursuivons l'état de l'art de la reconnaissance de personnes sur une discussion à propos des limitations de celle-ci.

2.1.1 Approches statiques

Dans un premier temps, nous nous intéressons aux approches statiques de reconnaissance de personnes en distinguant les approches globales des approches locales. Ces deux catégories d'approches ne présentent pas les mêmes caractéristiques et ne nécessitent pas le même niveau de détails des visages. Ainsi, les approches globales peuvent se contenter de visages de petite taille [8, 72, 115], parfois de très petite taille (jusqu'à 12×11 pixels pour [115]). Les approches locales nécessitent que les points caractéristiques du visage soient visibles, facilement identifiables et précisément localisables [116]. De nombreuses méthodes locales appliquent des traitements dédiés aux approches globales sur des points d'intérêt [81, 63, 78, 24, 94]. Il est donc naturel de présenter les approches globales avant les approches locales. Nous présentons dans cette section les méthodes globales et locales de reconnaissance statique, puis nous donnons un tableau récapitulatif de comparaison des approches statiques. Nous présentons enfin une catégorie d'approches qui utilisent un ensemble d'images de test (au lieu d'une unique image) pour la reconnaissance d'une identité.

Méthodes globales

Les méthodes globales de reconnaissance statique de personnes utilisent souvent une approche statistique. Une image de visage peut être vue comme une matrice de pixels. Il est possible de transformer cette matrice en vecteur en la linéarisant, c'est-à-dire en mettant bout à bout les lignes qui la composent. Sinon, il est possible de construire un vecteur de descripteur sur l'image. Les méthodes globales basées sur une approche statistique consistent à créer un espace vectoriel de représentation des visages à partir des vecteurs basés sur les visages de la base d'apprentissage. Un visage test (ou requête) est projeté dans cet espace en un vecteur. Le visage peut être ensuite localisé dans une région correspondant à une classe (identité) particulière. Dans ce cas, le système renvoie l'identité correspondante.

Pour créer un tel espace de représentation des visages, une des approches les plus fréquentes est de réaliser une analyse en composantes principales (ACP) [85]. Cette analyse est usuellement utilisée pour réduire un espace vectoriel en ne conservant que ses composantes principales. La première composante principale correspond à un axe, issu d'une combinaison linéaire des variables d'origine, autour duquel la variance des éléments présents dans l'échantillon est maximale. La deuxième composante, orthogonale à la première, correspond au deuxième axe où la variance des éléments est la plus importante. Chaque composante correspond à une dimension de cet espace, et l'ACP permet de réduire le nombre de dimensions de l'espace d'origine considérées. Dans notre cas, les images de visages forment des vecteurs dans un espace contenant autant de dimensions qu'il y a de pixels. Ainsi, à partir de l'ensemble des visages de la base d'apprentissage, l'ACP détermine les composantes principales à partir de la variance constatée pour ces données d'apprentissage. Les *Eigenfaces* (cf. Figure 2.1), proposées par Sirovich et al. [100], sont un exemple d'approche basée sur l'ACP. Les Eigenfaces sont les composantes principales qui décomposent le visage en vecteurs caractéristiques (*Eigenfaces vectors*). Ces vecteurs sont les vecteurs propres de la matrice de la matrice de variance-covariance des visages humains. Cette méthode [100] est considérée comme un des premiers exemples réussis de reconnaissance globale de visages [116]. Cependant, la pertinence d'utiliser les composantes présentant la variance maximale entre les classes est discutable, car ces



FIGURE 2.1 – Exemple de représentation d'un visage à partir d'Eigenfaces, pour un nombre de vecteurs propres allant de 10 à 300 par pas de 15. Exemple tiré de la documentation en ligne de la librarie OpenCV (<http://docs.opencv.org>).

composantes ne sont pas nécessairement les plus pertinentes pour réaliser la classification [101] – elles n'ont d'ailleurs pas été créées dans ce but. Un autre espace de projection de visage peut être construit en utilisant l'analyse discriminante linéaire (ADL)¹ [36]. L'ADL consiste en une réduction de la dimensionnalité en prenant en compte la classe (dans notre cas l'identité) des données. L'objectif est que les éléments d'une même classe soient proches dans cet espace et que la distance entre les éléments appartenant à deux classes différentes soit grande. Cette approche a été utilisée dans les travaux de Belhumeur et al. [16], de Swets et al. [101] et de Zhao et al. [117] pour modéliser des visages, sous forme de *Fisherfaces* (cf. Figure 2.2), dans le but de les reconnaître. Les Fisherfaces sont réputées comme donnant des meilleurs taux de reconnaissance que les Eigenfaces. En revanche, elles sont plus sensibles aux conditions d'éclairage. Il est possible d'utiliser les vecteurs des visages linéarisés pour essayer de réaliser la classification directement sur ces vecteurs. C'est ce que propose Phillips [87], où une machine à vecteurs de support (*support vector machine*, SVM) [25] est entraînée sur ces vecteurs, annotés selon leur classe (identité). L'objectif des SVM est de trouver les frontières de séparation entre les différentes classes qui maximisent la séparation entre elles. Une approche consiste à modéliser le visage par une transformée en cosinus discrète [1] (*discrete cosine transform*, DCT). Hafed et al. [47] proposent d'appliquer la DCT sur des visages normalisés. La normalisation consiste ici à redresser l'image pour que l'axe entre les yeux soit horizontal. Les images sont, de plus, recadrées pour que les visages soient centrés. Bien que cette approche soit globale dans sa description du visage, elle nécessite de localiser les yeux. Cette limitation se retrouve couramment dans les méthodes locales.

1. ADL est plus connue sous l'appellation anglaise de *linear discriminant analysis* (LDA).



FIGURE 2.2 – Exemple de représentation d'un visage à partir de Fisherfaces, pour un nombre de caractéristiques allant de 1 à 14. Exemple tiré de la documentation en ligne de la librairie OpenCV (<http://docs.opencv.org>).

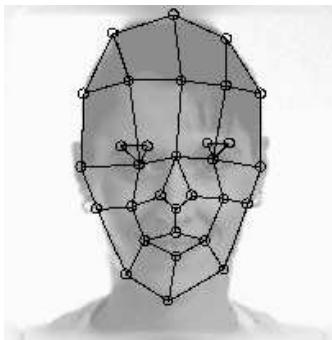


FIGURE 2.3 – Exemple de masque correspondant aux points d'intérêt du visage positionné par la méthode de Wiskott et al. [110].

Méthodes locales

Les approches les plus intuitives de la reconnaissance de personnes consistent à s'intéresser aux caractéristiques géométriques du visage – il s'agit des approches dites locales [116]. Elles consistent à détecter les points caractéristiques du visage (yeux, nez, bouche, oreilles, etc.) et de mesurer la position de chacun de ces points dans l'espace du visage [60, 61]. Dans les travaux de Wiskott et al. [110], les auteurs créent un masque qu'ils appliquent aux visages. Chaque noeud du masque correspond à un point d'intérêt (voir la Figure 2.3). Chaque point d'intérêt correspond à ce que les auteurs nomment un *jet*. Celui-ci représente localement l'image par des ondelettes de Gabor [34]. L'approche de Wiskott et al. combine ainsi la distance entre les points d'intérêt (comme dans l'approche de Kanade [60]), en ajoutant pour chaque point d'intérêt une coordonnée correspondante à l'index du *jet*.

Nefian et al. [81] ont une approche très différente de celles décrites précédemment. Le visage est modélisé par une fenêtre glissante sur un axe vertical (avec un peu de recouvrement). Chaque position est modélisée par une transformée 2D en cosinus discrète

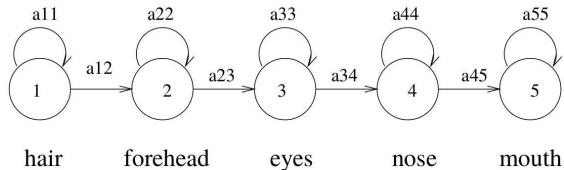


FIGURE 2.4 – Illustration de la structure des états du visage modélisé par un HMM, ainsi que des probabilités de transition. (Illustration tirée de [81]).

(*2D-DCT*). Ces différents modèles sont utilisés pour créer une représentation en modèle de Markov caché (*hidden Markov model*, HMM). Les états du HMM (voir la Figure 2.4) correspondent à des régions concrètes du visage (cheveux, front, yeux, nez et bouche). Cette approche est moins sensible aux conditions d'éclairage que les approches précédentes. En revanche, elle est plus sensible à la pose de la tête et aux expressions du visage [81].

Kirby et Sirovich [63] proposent une méthode basée sur les Eigenfaces (méthode détaillée précédemment dans les méthodes globales) pour reconnaître des personnes. Pour compléter le principe des Eigenfaces, une approche appelée *Eigenfeatures* a été développée par la suite. Elle combine une métrique géométrique faciale, mesurant la distance entre des points caractéristiques du visage comme les yeux ou le nez avec l'approche classique des Eigenfaces. Le visage est découpé en régions sémantiques comme dans l'approche précédente de Nefian et al. [81] (cheveux, front, yeux, etc.). Chaque région est modélisée comme dans l'approche Sirovich et Kirby [100]. Moghaddam et Pentland [78] se basent sur le principe des Eigenfeatures appliquées aux yeux, au nez, à la bouche et aux joues. Ils combinent aux Eigenfeatures une approche permettant de déterminer le point de vue par rapport au visage de chaque personne. Cela permet d'ajouter une certaine robustesse face aux variations de la posture des sujets. Chang et al. [24] ont proposé d'appliquer les Eigenfaces à l'oreille, dans l'approche *Eigen-ears*. Ce travail a été étendu par la suite dans les travaux de doctorat de Saleh [94]. L'idée de cette méthode est de faire une combinaison de la reconnaissance des oreilles d'une part et de la reconnaissance du visage d'autre part. Les oreilles sont souvent visibles quand la personne n'a pas une pose exactement frontale. Le choix d'incorporer les oreilles en complément du visage semble donc pertinent. Les auteurs montrent que l'efficacité de la reconnaissance des oreilles est similaire à celle du visage dans les mêmes conditions expérimentales. De plus, les auteurs montrent que les oreilles de vrais jumeaux sont différenciables visuellement. La limitation principale de cette approche concerne les occultations du visage et des oreilles. Ils peuvent être cachés par les cheveux, et la présence des boucles d'oreille peut perturber la reconnaissance.

Le problème principal des approches locales est qu'elles nécessitent une détection très précise des points d'intérêt [116]. Encore aujourd'hui, la précision de la détection ne permet pas d'exploiter correctement de telles approches. En revanche, ces dernières ne sont pas sensibles aux variations de pose du sujet ou aux mauvaises conditions d'éclairage, dans la mesure où il est possible de localiser les points d'intérêt. Ahonen et al. [2] proposent de décrire localement l'intégralité du visage. Pour cela, chaque point du visage est décrit par un motif binaire local (*local binary pattern*, LBP) [84]. Le code LBP d'un pixel consiste en la séquence de ses 8 pixels voisins², après une binarisation³ en utilisant

2. Il est possible de considérer pour un LBP plus de pixels que les 8 voisins immédiats.

3. Une binarisation est un seuillage où les valeurs en dessous d'un seuil sont mises à 0 et les valeurs

la valeur du pixel comme seuil. Le visage est ensuite découpé selon une grille, et un histogramme est construit sur chaque case de la grille en comptabilisant les différents codes LBP présents. Tous les histogrammes sont ensuite concaténés pour former un descripteur de l'ensemble du visage appelé *LPH* (LBP histogram).

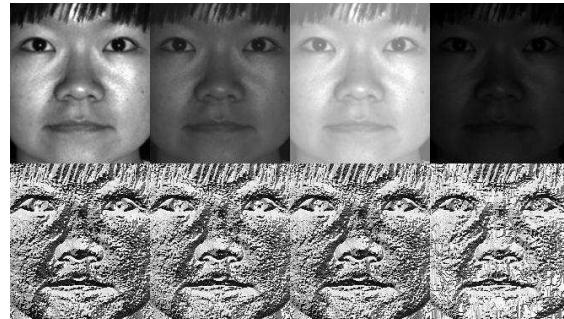


FIGURE 2.5 – Exemple de représentation d'un visage à partir de LBP, en variant la luminosité de l'image source de façon uniforme. Exemple tiré de la documentation en ligne de la librarie OpenCV (<http://docs.opencv.org>).

Comparaison des approches statiques

Afin de comparer les différentes approches de reconnaissance de personnes, le *Defense Advanced Research Projects Agency* (DARPA) et le *National Institute of Standards and Technology* (NIST) ont établi en 1996 une base d'images annotées FERET (Face Recognition Technology) [88]. Elle contient des images en niveaux de gris, et également des images en couleurs dans la seconde version (publiée en 2003). Cette base de données contient 14.051 photos de visages représentant 1.199 individus. La base d'images a été découpée en plusieurs ensembles : les visages de face pour l'apprentissage (fa) et pour le test (fb), des visages ayant subi un changement d'illumination (fc), les visages aux trois-quarts gauches (hl) et droits (hr), les visages de profils gauches (pl) et droits (pr). Nous avons compilé dans le Tableau 2.1 les résultats publiés pour les différentes approches sur la base FERET. Nous remarquons que les approches globales et locales atteignent des taux de reconnaissance élevés pour les visages de faces de la base FERET. On remarque de plus que la précision de ces approches diminue quand les conditions d'éclairage changent. Cette diminution est plus importante lorsque que la pose des personnes varie. Cela indique que les conditions d'éclairage ont moins d'impact que la pose sur la précision de la reconnaissance.

Nous avons présenté un ensemble représentatif des approches de reconnaissance statique de personnes et discuté des limites de chaque approche. Nous allons maintenant nous intéresser aux approches considérant plusieurs trames de la vidéo.

Approches statiques basées sur les trames

L'ouvrage de Li et al. [71] propose une répartition des approches existantes en reconnaissance de personnes basée sur la vidéo en trois catégories :

au-dessus sont mises à 1.

Références	Type d'approche	Données d'apprentissage	Données de test	Précision (en %)
Moghaddam et al. [78] (Eigenfeatures) (source : [110])	local	150 fa 150 hl 150 pl	150 fb 150 hr 150 pr	99 38 32
Wiskott et al. [110]	local	250 fa 250 hr 250 pr	250 fb 181 hl 250 pl	98 57 84
Ahonen et al. [2] (LBPH)	local	NC fa NC fa	NC fb NC fc	97 79
Zhao et al. [117] (LDA)	global	1.316 fa 1.316 fa	298 fb 298 fc	83 32
Zhao et al. [117] (LDA+PCA)	global	1.316 fa 1.316 fa	298 fb 298 fc	95 59
Chang et al. [24] (oreille) (visage) (oreille + visage)	local	197 fa 197 fa 197 fa	197 fb 197 fb 197 fb	72,7 69,3 90,9
Chang et al. [24] (oreille) (visage) (oreille + visage)	local	197 fa 197 fa 197 fa	197 fc 197 fc 197 fc	66,7 64,9 86,5
Belhumeur et al. [16] (Fisherfaces) (source : [113])	global	600 aléa.	800 aléa.	77,87

TABLE 2.1 – Comparaison des résultats obtenus sur la base FERET à l'aide de différentes approches de reconnaissance statique de personnes.

1. les approches basées sur les trames considérées individuellement,
2. les approches de mise en correspondance d'ensembles,
3. les approches basées sur un sous-espace mutuel.

Dans un premier temps, nous allons voir les approches de cette première catégorie qui considèrent la séquence vidéo comme un ensemble non-ordonné d'images. Ces approches ne sont donc pas dynamiques, puisqu'elles ignorent l'ordre des trames ainsi que tout autre aspect temporel qui pourrait les lier. Cependant, ces approches sont donc intéressantes dans notre étude. Ces approches considèrent chaque trame de façon indépendante, et fusionnent les résultats de reconnaissance obtenus pour chacune afin de déterminer l'identité finale [71]. Plusieurs techniques de fusion de décisions peuvent être appliquées pour fournir l'identité finale. Selon les auteurs de [71], les stratégies de fusion les plus fréquemment utilisées sont celles qui ont été proposées par Liu et al. [98] et par Shakhnarovich et al. [73]. Si on exige une comparaison entre chaque visage de test et l'ensemble des visages extraits dans la phase de reconnaissance [96], alors la complexité de traitement est très élevée et les temps de calcul sont importants. Pour résoudre ce problème, une approche qui consiste à sélectionner uniquement les trames les plus représentatives des séquences, ou *trames clés* a été proposée par Gorodnichy [43]. Dans cette méthode, la reconnaissance nécessite l'apparition simultanée du nez et des yeux. Leurs emplacements sont utilisés pour décider si le visage est approprié ou non pour la reconnaissance. Si les deux yeux et le nez forment un triangle équilatéral, alors la suite des traitements est exécutée ; si-

non, la recherche dans la séquence continue jusqu'à ce qu'une trame contenant un visage approprié soit rencontrée.

2.1.2 Approches dynamiques

Contrairement aux approches précédentes, les approches dynamiques de reconnaissance de personnes dans des vidéos considèrent l'aspect temporel. Il s'agit d'exploiter une source d'informations continue dans la vidéo plus riche que les images dans le cas statique. Les approches dynamiques peuvent s'appliquer à plusieurs contextes avec des caractéristiques différentes comme par exemple dans le contexte de la vidéo surveillance.

Zhou et al. [118] proposent, dans le contexte où une seule caméra filme une personne se déplaçant face à la caméra sur un tapis roulant, de modéliser la séquence vidéo par un filtre particulier [29] mis à jour à chaque nouvelle trame de la vidéo. Ainsi, l'identité proposée par leur approche est affinée progressivement. L'avantage de leur approche est qu'elle permet de reconnaître les personnes en mouvement selon différentes démarches : lente, rapide, personne évoluant sur un plan incliné ou portant une charge. Le problème principal de cette approche est une grande sensibilité, autant aux variations de l'apparence (si celle-ci varie légèrement ou si les conditions de l'environnement perturbent le système) que de la pose de la personne [118]. Ce type d'approche considère la séquence vidéo comme un ensemble d'images ordonnées, contrairement aux approches qui considèrent les trames individuellement. Arandjelovic et al. [4] proposent de représenter chaque séquence d'images (extraites d'une séquence vidéo) par une distribution paramétrique et de calculer la similarité entre la distribution paramétrique de la séquence d'images à tester avec les distributions paramétriques de références. Ces dernières sont obtenues lors de la phase d'apprentissage. Cette approche produit de bons résultats sur une collection d'une centaine d'individus filmés dans des conditions d'éclairage similaires. La limitation principale de cette approche vient de la difficulté à localiser le visage selon la pose du visage des personnes. Lee et al. [68] proposent de créer des sous-espaces, à partir des espaces vectoriels des visages, représentants les différentes poses du visage pour chaque identité. Cette approche détermine l'identité en cherchant le sous-espace correspondant. Pour cela, les auteurs introduisent une distance entre une image et un sous-espace. Cette mesure prend de plus en compte les temps de transition de l'évolution d'un visage entre les sous-espaces correspondant aux poses pour vérifier la crédibilité du résultat. Cette approche nécessite un corpus d'apprentissage dans lequel toutes les variations de la pose du visage d'une personne sont présentes. Afin de modéliser des séquences d'images à l'aide de distributions, on peut considérer les approches basées sur un sous-espace mutuel. Yamaguchi et al. [112] ont proposé une méthode appelée *mutual subspace method* (MSM) qui permet de modéliser une séquence d'images dans un sous-espace linéaire. La similarité entre deux séquences d'images est définie par l'angle formé entre ces deux sous-espaces. Par la suite, pour rendre le sous-espace invariant aux changements de pose et aux changements d'éclairage, la MSM a été étendue par la *constrained mutual subspace method* (CMSM) [39]. Les contraintes supplémentaires permettent de réduire l'espace de recherche, améliorant ainsi les résultats obtenus à l'aide de cette approche par rapport à la précédente. D'autres méthodes basées sur le même principe que MSM ont été proposées, telles que la *kernel constrained mutual subspace method* (KCMSM), [39] et appliquées à la reconnaissance de personnes. Fukui et al. [38] ont réalisé une étude comparative entre les méthodes MSM, CMSM et KCMSM sur une tâche de reconnaissance d'objets 3D. Cette application, proche de la reconnaissance de personnes, montre

une amélioration du taux de reconnaissance sur une collection d'images 3D, en passant des MSM, aux CMSM, puis aux KCMSM.

De façon générale, les approches dynamiques que nous venons d'étudier nécessitent des conditions contrôlées pour obtenir de bons taux de reconnaissance. En effet, la plupart de ces approches dynamiques souffrent des mêmes limitations que les approches statiques, bien que la richesse des séquences d'images permette d'obtenir de meilleurs résultats de reconnaissance [116].

2.2 Regroupement des occurrences de personnes

Les différentes approches de l'état de l'art de la reconnaissance de personnes présentent des limitations ne permettant pas leur emploi systématique sur toutes les trames d'une émission. Nous nous intéressons ainsi aux approches de *ré-identification* qui permettent de regrouper les occurrences de personnes selon leur similarité visuelle. L'objectif est qu'à chaque groupe corresponde une identité. Elle consiste à regrouper toutes les occurrences des personnes selon leurs identités.

La plupart des approches existantes pour la ré-identification sont *globales* : elles décrivent l'apparence globale d'une personne (par exemple le buste, le corps complet, la silhouette, etc.) afin de générer sa signature pour la retrouver dans d'autres vidéos. Les approches *locales*, au contraire, utilisent uniquement des points d'intérêt pour générer cette signature. Les approches locales nécessitent une qualité d'image élevée, et sont sujettes à de nombreuses contraintes, notamment sur la pose de la personne et sur la luminosité. À cause de ces contraintes, elles sont rarement utilisées dans le contexte de la ré-identification de personnes [15].

De nombreuses approches globales de ré-identification de personnes existent. Parmi les plus connues, on peut citer les travaux de Nakajima et al. [80], qui présentent un système basé sur les informations de couleur et de forme pour créer la signature de l'aspect visuel d'une personne. Les silhouettes sont ensuite apprises et reconnues en utilisant un SVM. Cette méthode donne de bons résultats dans un environnement contraint, où l'arrière-plan est à la fois connu et statique. Bien que cette approche soit considérée par ses auteurs comme une approche de reconnaissance de personnes, elle nécessite de faire l'hypothèse comme c'est le cas dans notre approche, que l'aspect visuel (vêtements, coiffure, etc.) des personnes ne doit pas varier. En cas de variation de cet aspect, le classifieur n'est plus à même de reconnaître correctement les personnes. Ngo et al. [82] utilisent le nombre de points d'intérêt obtenus par l'algorithme de Shi et al. [99] sur des visages pour déterminer si deux visages correspondent à la même personne. Pour retrouver les points d'intérêt d'un visage dans un autre, les auteurs appliquent un algorithme de flot optique à partir des points d'intérêt du premier visage, et comptent le nombre de points d'intérêt retrouvés dans le second visage. Au-delà d'un certain seuil, les deux visages sont notés comme appartenant à la même personne. De façon similaire, Hamdoun et al. [49] sélectionnent les points d'intérêt obtenus par une méthode inspirée des points d'intérêt SURF [13]. Dans leurs travaux, les points d'intérêt d'une occurrence vidéo de personne forment un ensemble qui sert de signature. Pour ré-identifier une personne, les auteurs calculent la somme des différences absolues (*sum of absolute differences*, SAD) entre l'ensemble des points d'intérêt de l'occurrence de la personne et ceux des personnes à retrouver. Bird et al. [21] s'intéressent à la ré-identification dans un système multi-caméras. Les auteurs détectent les piétons en notant les personnes restant dans le champ de vision de la caméra

pendant une période relativement longue. Les caractéristiques utilisées pour corrélérer les régions correspondant aux personnes sont basées sur la couleur des vêtements. Une ADL est appliquée pour accentuer les différences entre les différents individus dans l'espace des caractéristiques, cela permet de retrouver plus facilement une personne passant d'une caméra à une autre.

Dans [62], Kettnaker et al. introduisent la formalisation bayésienne d'une tâche de surveillance utilisant plusieurs caméras. Ils exploitent à la fois les similarités dans les vues des différentes caméras, et les temps de transition d'une personne pour passer d'une caméra à l'autre pour la ré-identifier. Prosser et al. [89] proposent un système mono-caméra permettant la ré-identification des personnes en utilisant un score donné par un classifieur SVM. Baulm et al. [12] font la ré-identification de plusieurs personnes dans un réseau de plusieurs caméras. Au sein d'une même caméra, le suivi se fait de façon classique avec un outil de suivi de visages (ou *face tracker*). Pour retrouver une personne d'une caméra à une autre, un SVM est entraîné sur les visages. Hirzer et al. [53] testent différents descripteurs sur la base VIPeR [44], contenant plusieurs photos de personnes en extérieur. Ils comparent les résultats obtenus avec un descripteur de caractéristiques pseudo-Haar [106], des histogrammes d'orientation de gradients (*histogram of oriented gradients*, HOG) [26], des LBPs et la covariance des pixels dans l'espace rouge-vert-bleu (RGB). Ils obtiennent de meilleurs résultats avec la covariance et les descripteurs de caractéristiques pseudo-Haar. La combinaison des deux descripteurs permet d'obtenir les meilleurs résultats. Jungling et al. [58] ré-identifient les personnes dans un scénario de vidéo-surveillance multi-caméras. Le suivi des personnes est réalisé par un modèle de formes des personnes. La ré-identification entre les différentes occurrences est réalisée à partir des points d'intérêt SIFT [74] détectés sur les trames des occurrences vidéo des personnes. Le problème principal de cette approche est que le taux de bonnes ré-identifications chute quand le nombre de personnes à retrouver augmente. Il est compris entre 30% et 65% selon le choix des paramètres. Plus récemment, Gandhi et al. [40] ré-identifient les personnages du film "Rope" d'Alfred Hitchcock (1948) en utilisant un modèle d'apparence basé sur des ellipses de couleurs. Cette approche n'utilise pas de détecteur de personnes, mais fait glisser une fenêtre dans l'image et calcule le modèle d'apparence dans cette fenêtre pour chaque position. Si le modèle correspond à celui d'une des personnes, la position de la personne est alors conservée. Cette approche permet de retrouver les personnes dans de nombreux cas avec une certaine résistance à l'occultation. En revanche, l'approche génère de nombreuses fausses détections. Gheissari et al. [41] proposent d'utiliser une signature invariante à l'illumination et à la pose pour comparer les différentes parties du corps des personnes. Cette signature est générée en combinant à la fois des informations de couleur et de structure (position des membres du corps des personnes). Les informations de couleur, représentées dans l'espace de couleur HSV (*hue saturation value*), sont décrites par un histogramme de teintes et de saturations. Les informations structurelles sont obtenues en sur-segmentant le corps de la personne et en regroupant au fil des trames les bords saillants. Cette approche a tendance à regrouper les personnes ayant les mêmes poses plutôt que selon les identités [41]. Schwartz et al. [97], plus orientés vers la reconnaissance faciale, proposent une approche basée sur la forme, la texture et les informations de couleur pour ré-identifier les personnes. Les informations de forme sont extraites d'un HOG et les informations de texture sont extraites à l'aide de LBP. Enfin, les informations de couleur sont obtenues en faisant la moyenne sur des blocs de pixels. La ré-identification se base quant à elle sur la méthode des moindres carrés partiels afin de pondérer la décision

en fonction des trois caractéristiques. Les travaux de Truong Cong et al. [102, 103] proposent la ré-identification de passagers dans un wagon de train muni de deux caméras. Les connaissances a priori des caractéristiques du wagon permettent d'extraire les passagers malgré des conditions d'éclairage très variées. Pour chaque personne, un histogramme de couleurs, un spatiogramme et un chemin-couleur sont évalués. Les spatiogrammes ont été définis par Elmongui et al. [33], qui les proposent pour le classement et la recherche dans des bases de données numériques. Leurs performances sont comparées, dans l'approche de Truong Cong et al., pour la ré-identification avec les histogrammes de couleurs et les chemin-couleur. Dans ces conditions, les histogrammes de couleurs ne permettent pas de bien ré-identifier les personnes. Cela peut s'expliquer par les variations importantes de l'illumination entre les caméras du système. En revanche, les spatiogrammes permettent d'obtenir les meilleurs résultats, suivis de près par les chemin-couleur. De nombreuses approches se basent sur des histogrammes pour ré-identifier les personnes [102, 103, 97, 53]. Dans les travaux de Schwartz et al. [97] et Hirzer et al. [53], les histogrammes produisent de bons taux de ré-identification. Comme nous l'avons vu dans plusieurs approches présentées précédemment, les histogrammes peuvent prendre différentes formes selon le type de données qu'ils contiennent.

2.2.1 Les histogrammes de couleurs et leurs extensions

Les histogrammes trouvent de nombreuses utilisations dans plusieurs domaines. Essentiellement, ils servent à résumer des observations en les catégorisant dans différentes partitions (ou classes) : pour chaque partition, on conserve une information de comptage du nombre d'occurrences de cette partition constatées dans les observations.

Dans le domaine de la vision par ordinateur, il est courant de recourir aux histogrammes de couleurs. Leur popularité vient du faible coût calculatoire de construction, de leur faible coût mémoire, ainsi que de la description synthétique qu'ils donnent d'un phénomène. Ainsi, un histogramme h peut se définir de la manière suivante :

$$h(b) = \langle n_b \rangle, \quad b = 1, \dots, B \quad (2.1)$$

où B est le nombre de partitions de l'histogramme et n_b le nombre d'observations de la partition b . Par exemple, les histogrammes de couleurs classiques renseignent, pour chaque partition (qui correspond à un intervalle de couleurs), le nombre de pixels de la partition trouvés dans une image ou une région d'image. Les histogrammes de couleurs résument donc la distribution des couleurs d'une image ou d'une région d'image. Leur inconvénient principal réside dans le fait qu'un histogramme perd complètement toute information spatiale (agencement des pixels) présente dans une image.

Il est possible de réaliser des opérations sur les histogrammes. Une opération classique sur les histogrammes consiste à les normaliser, c'est-à-dire à rendre la somme des données de comptage égale à 1. Pour cela, le nombre d'observations de chaque partition est divisé par le nombre total d'observations. Ainsi, les valeurs de comptage pour chaque partition peuvent être interprétées comme des pourcentages, ce qui permet de comparer des histogrammes construits sur des échantillons de tailles différentes. Dans le cas d'images, cela permet de comparer des histogrammes construits sur des images de dimensions différentes. Pour ces comparaisons, diverses métriques permettent de mesurer la distance entre deux histogrammes. À partir de ce socle simple, de nombreuses variations des histogrammes ont été mises au point pour conserver plus d'informations relatives au phénomène étudié.

Les spatiogrammes

Les spatiogrammes de couleurs [33] ont la particularité de conserver, en plus des données de comptage, une information spatiale sur la position moyenne des pixels contenus dans chaque partition. En effet, la distribution des pixels dans les partitions des histogrammes se retrouve à l'identique dans les partitions des spatiogrammes. En plus de cette position moyenne, la covariance de la position spatiale (x, y) des pixels de chaque partition est aussi conservée sous la forme d'une matrice de covariances :

$$\begin{bmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{bmatrix} \quad (2.2)$$

Cette matrice permet de connaître la dispersion de chaque partition de couleur dans l'image. Elle est symétrique car $cov(x, y) = cov(y, x)$. À titre d'illustration, l'information spatiale conservée dans chaque partition du spatiogramme peut être représentée par une ellipse centrée sur la position moyenne, de taille $cov(x, x)$ pour le grand axe et $cov(y, y)$ pour le petit axe, et d'orientation $cov(x, y)$. Notons que $cov(x, x) = var(x)$ et $cov(y, y) = var(y)$. Des mesures de similarité dédiées permettent de comparer les spatiogrammes en tenant compte des informations spatiales des pixels.



FIGURE 2.6 – Exemple d'images produisant des histogramme de couleur identiques, mais des spatiogrammes différents. L'image de droite a été obtenue en mélangeant les pixels de l'image de gauche.

Par exemple dans la Figure 2.6, montrant deux images visuellement différentes mais composées exactement des mêmes pixels agencés différemment, les histogrammes de couleurs construits sur ces deux images seront identiques. En revanche, les spatiogrammes construits sur ces mêmes images seront différents. De plus, en comparant des spatiogrammes construits sur ces images, on obtiendra une faible similarité. Du fait que les spatiogrammes sont construits en ajoutant des informations aux histogrammes de couleurs, soulignons que les spatiogrammes *contiennent* les histogrammes de couleurs.

Les tempogrammes

Dans le cas de données susceptibles de varier dans le temps, les tempogrammes conservent des informations temporelles sur la localisation dans le temps des données comptées [45]. Historiquement, les tempogrammes ont été utilisés à l'origine dans le domaine de l'analyse musicale et du son [66]. En vision par ordinateur, les tempogrammes ont été très peu utilisés. Ceci est dû au fait que de nombreux algorithmes de vision se basent sur des images, sans information temporelle. En revanche, cette information

temporelle est utile dès lors que l'on s'intéresse à l'analyse de la vidéo, composée d'une séquence d'images ordonnées temporellement.

2.2.2 Mesures de distance entre histogrammes

Il existe de nombreuses mesures de distance définies pour les histogrammes et leurs extensions. Ces mesures ne présentent pas toutes les propriétés mathématiques des distances ; dans ce cas, nous parlons plutôt de dissimilarité. Une distance d doit satisfaire les conditions suivantes sur un ensemble noté \mathbb{E} :

- séparation : $\forall a, b \in \mathbb{E} : a = b \Leftrightarrow d(a, b) = 0$
- symétrie : $\forall a, b \in \mathbb{E} : d(a, b) = d(b, a)$
- inégalité triangulaire : $\forall a, b, c \in \mathbb{E} : d(a, b) + d(b, c) \geq d(a, c)$

L'une des distances les plus utilisées est la distance euclidienne. Dans un premier temps, nous nous intéressons aux métriques permettant de comparer deux histogrammes entre eux. Plus tard dans nos travaux, nous décrivons une vidéo à l'aide de plusieurs histogrammes rangés dans une séquence (éléments indexés par les entiers naturels). Nous nous intéressons donc ici à la comparaison d'ensembles d'histogrammes, nous étudions les approches d'alignement de séquences.

Distance euclidienne

En mathématiques, la distance euclidienne d_e est la distance usuelle entre deux points telle que l'on pourrait la mesurer avec une règle, et qui est donnée par le théorème de Pythagore. Dans la littérature plus ancienne, cette métrique est appelée la mesure de Pythagore.

Soient deux points P et P' définis dans un espace en dimensions n :

$$d_e(P, P') = \sqrt{\sum_{i=1}^n (P_i - P'_i)^2} \quad (2.3)$$

où P_i et P'_i sont les positions des points P et P' pour la $i^{\text{ème}}$ dimension.

La distance donnée par l'Équation 2.3 peut être utilisée sur des histogrammes ayant le même nombre de partitions. Dans ce cas, cela revient à faire la racine carrée de la somme des différences au carré des différentes valeurs qui composent les histogrammes. La distance euclidienne entre deux histogrammes h et h' se formule naturellement de la façon suivante :

$$d_e(h, h') = \sqrt{\sum_{b=1}^B (n_b - n'_b)^2} \quad (2.4)$$

où B est le nombre de partitions des histogrammes et n_b (respectivement n'_b) est la valeur associée à la partition b de h (respectivement h').

Coefficients de Bhattacharyya

Les coefficients de Bhattacharyya [19] sont une mesure approximative de la quantité de recouvrement entre deux échantillons statistiques. Les coefficients sont utilisés pour obtenir la mesure de Bhattacharyya d_b . Ces coefficients peuvent être utilisés pour déterminer la dissimilarité entre les deux échantillons via leur représentation sous forme

d'histogrammes. Le calcul des coefficients de Bhattacharyya utilise une forme rudimentaire d'intégration du recouvrement des deux échantillons [19].

La formule de la mesure de Bhattacharyya, entre deux histogrammes h et h' est ainsi :

$$d_b(h, h') = \sum_{b=1}^B \sqrt{n_b \times n'_b} \quad (2.5)$$

Le résultat de cette formule est grand quand chaque partition a des observations dans les deux histogrammes simultanément, et plus grand encore quand les partitions contiennent un large recouvrement.

La mesure de Bhattacharyya vaut 0 s'il n'y a aucun recouvrement. Cela est dû à la multiplication lors de la comparaison du nombre d'observations de chaque partition. Dans ce cas, cela signifie que les deux histogrammes sont parfaitement séparés.

Distance du χ^2

À la différence des coefficients de Bhattacharyya, la distance du χ^2 [67], d_{χ^2} permet de vérifier si deux échantillons de même taille sont issus d'une même loi de probabilité.

$$d_{\chi^2}(h, h') = \sum_{b=1}^B \frac{(n_b - n'_b)^2}{n_b + n'_b} \quad (2.6)$$

d_{χ^2} est nulle si les deux échantillons comparés sont identiques. Les différences en nombre d'observations dans une partition des deux histogrammes sont accentuées par le carré. Le dénominateur a pour rôle de pondérer cette différence par le nombre total d'observations considérées dans la partition. Cela permet de limiter l'influence des petites différences dans des partitions contenant un grand nombre d'observations [86].

Distance de Mahalanobis

La distance de Mahalanobis d_m est une statistique descriptive qui fournit une mesure relative de la distance entre des données et une référence. Elle a été introduite par P.C. Mahalanobis en 1936 [75]. La distance de Mahalanobis est utilisée pour identifier un échantillon inconnu en estimant sa dissimilarité avec un échantillon connu. Elle diffère de la distance euclidienne en prenant en compte la corrélation de l'ensemble des données, et elle est invariante à l'échelle (par nature).

$$d_m(h, h') = \sqrt{\sum_{b=1}^B (\mu_b - \mu'_b)^t \hat{\Sigma}_b^{-1} (\mu_b - \mu'_b)} \quad (2.7)$$

où B est le nombre de partitions des histogrammes h et h' , μ_b et μ'_b sont leurs moyennes respectives pour la $b^{\text{ième}}$ partition. $\hat{\Sigma}_b^{-1}$ est l'estimateur de leur covariance :

$$\hat{\Sigma}_b^{-1} = (\Sigma_b^{-1} + (\Sigma'_b)^{-1}) \quad (2.8)$$

où Σ_b et Σ'_b sont les covariances respectives des histogrammes pour leur $b^{\text{ième}}$ partition. La distance de Mahalanobis prend en compte la variance entre les partitions des deux histogrammes. Elle prend aussi en compte la différence de l'orientation de la variance entre deux histogrammes. De plus, la mesure de dissimilarité est faible si les données des deux histogrammes ne sont pas corrélées. En résumé, la distance de Mahalanobis permet de mesurer la corrélation entre les deux histogrammes.

Divergence Kullback-Leibler

Dans la théorie des probabilités et de l'information, la divergence de Kullback-Leibler (KL) [65, 64] est une mesure asymétrique de la différence entre deux distributions de probabilités P et Q . La divergence de KL est un cas particulier d'une classe plus large de divergences appelées les *f-divergences*. Elle a été introduite par Solomon Kullback et Richard Leibler en 1951 comme divergence directrice entre deux distributions. De façon formelle, la divergence de KL de Q par rapport à P , notée $D_{KL}(P||Q)$, mesure l'information perdue quand Q est utilisé pour estimer P . Pour deux distributions de probabilités discrètes P et Q la divergence de KL de Q par rapport à P est définie par :

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (2.9)$$

Une autre manière de voir cette divergence est que D_{KL} mesure le nombre de bits attendus pour coder un échantillon de P en utilisant un code basé sur Q au lieu d'utiliser un code basé sur P . De façon classique, P représente la "vraie" distribution des données ou d'observations et la mesure Q représente description ou une approximation de P .

Bien qu'elle soit souvent utilisée comme métrique ou distance, la divergence de KL n'est pas une distance au sens mathématique. En effet, elle n'est pas symétrique : la D_{KL} de P vers Q est généralement différente de la D_{KL} de Q vers P . De plus, la D_{KL} ne satisfait pas l'inégalité triangulaire.

Nous avons présenté cinq mesures classiques pour comparer les histogrammes. En conclusion, si l'on souhaite prendre en considération la corrélation entre les partitions correspondantes dans les deux histogrammes, il faudra s'intéresser à la distance de Bhattacharyya. Si on souhaite avoir une distance qui dont la valeur ne dépende pas des tailles des partitions, la distance du χ^2 semble idéale. Afin de mettre en évidence que les histogrammes sont issus d'une même distribution statistique, il est possible d'utiliser la distance de Mahalanobis ou la divergence de Kullback-Leibler. La distance de Mahalanobis semble algorithmiquement plus simple que celle de Kullback-Leibler. Notre travail s'inspire de la distance du χ^2 et de la mesure de Mahalanobis pour définir une mesure de similarité entre les histogrammes spatio-temporels (cf. Section 4.4).

2.2.3 Distance entre séquences

Nous avons vu comment mettre en correspondance deux histogrammes. Une séquence vidéo pouvant être représentée par plus d'un histogramme, il est nécessaire de pouvoir mettre en correspondance deux ensembles d'histogrammes. Il existe plusieurs familles d'approches pour comparer des ensembles de symboles discrets (historiquement : des chaînes de caractères). Une première famille d'approches utilise les paires, prises dans les deux collections qu'on cherche à comparer, et propose une mesure de distance basée sur les distances entre les paires ; on parle de comparaison "deux à deux" (*pairwise comparison* en anglais). Dans ce cas, la notion de séquence est ignorée. Une seconde famille d'approches cherche le nombre minimum d'opérations d'insertions, de suppressions et de substitutions nécessaires afin de passer d'une séquence à l'autre [50, 69, 22, 11, 70, 27, 109]. Cette famille contient les différentes distances d'édition, la plus connue étant la distance d'édition de Levenshtein [69]. La distance d'édition trouve de nombreuses applications dans l'analyse de texte. Enfin, une dernière famille d'approches considère les collections

comme des séquences d'éléments et propose de trouver l'alignement optimal afin de proposer une mesure de distance.

La comparaison par paires ne suppose pas que les éléments de l'ensemble soient ordonnés d'une quelconque manière. Cela est particulièrement utile dans le cas où il n'y a aucune manière évidente de trier les éléments. Par exemple, si on compare deux ensembles d'images, il est difficile de justifier d'un ordonnancement particulier de celles-ci. Dans le cas de comparaisons par paires, toutes les paires d'éléments sont choisies et comparées pour établir la distance globale entre les deux collections. Cette famille ne nous intéresse pas dans le contexte de nos travaux, du fait que nous souhaitons tirer avantage de la notion de séquence.

La distance de Hamming [50] entre deux séquences de symboles discrets (e.g. chaînes de caractères) de même longueur est le nombre de positions où les symboles sont différents. En d'autres termes, la distance de Hamming mesure le nombre minimum de substitutions nécessaires pour transformer une chaîne en l'autre. Ou encore, elle mesure le nombre minimum d'erreurs qui auraient transformé une chaîne en l'autre. La distance de Damerau-Levenshtein [22] (du nom de Frederick J. Damerau et Vladimir I. Levenshtein) est une distance entre deux chaînes de caractères, donnant le nombre minimum d'opérations nécessaires pour transformer une chaîne en une autre. Ces opérations sont l'insertion, la suppression ou la substitution d'un caractère, ou la transposition de deux caractères adjacents. Dans l'article [27], Damerau distingue ces quatre opérations d'édition, et affirme qu'elles correspondent à plus de 80% des fautes d'orthographes commises dans les textes. L'article de Damerau considère les fautes d'orthographes qui pourraient être corrigées par au plus une opération d'édition. Le nom de distance de Damerau-Levenshtein est utilisé pour faire référence à la distance de Levenshtein prenant en compte la transposition.

Similarité de Jaro–Winkler

La similarité de Jaro–Winkler [109] est une mesure de similarité entre deux chaînes de caractères. Il s'agit d'une variante de la similarité de Jaro [55, 56], et est principalement utilisée pour détecter la duplication. Le score est normalisé pour que 0 dénote l'absence de similarité et 1 une correspondance exacte. Enfin, cette similarité est bien adaptée pour les chaînes relativement courtes telles que les noms de personnes.

La distance de Jaro S_J entre chaînes C_1 et C_2 est définie par :

$$S_J = \frac{1}{3} \left(\frac{m}{|C_1|} + \frac{m}{|C_2|} + \frac{m-t}{m} \right) \quad (2.10)$$

où $|C_1|$ et $|C_2|$ sont les longueurs des chaînes de caractères, m est le nombre de caractères correspondant et t est le nombre de transpositions. Deux caractères identiques de C_1 et de C_2 sont considérés comme correspondant si leur éloignement γ (i.e. la différence entre leurs positions dans leurs chaînes respectives) ne dépasse pas :

$$\gamma(C_1, C_2) = \left[\frac{\max(|C_1|, |C_2|)}{2} \right] - 1 \quad (2.11)$$

Le nombre de transpositions est obtenu en comparant le $i^{\text{ème}}$ caractère correspondant de C_1 avec le $i^{\text{ème}}$ caractère correspondant de C_2 . Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.

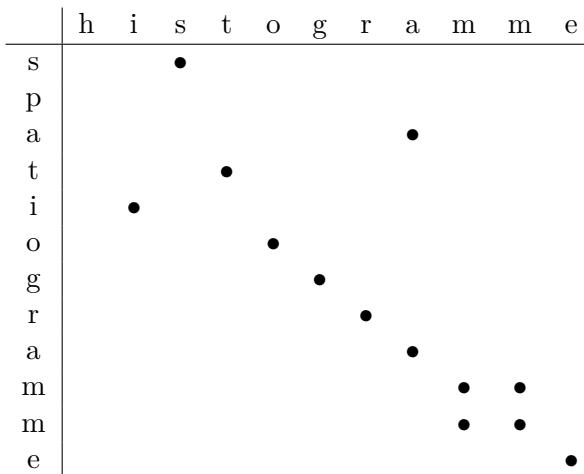


TABLE 2.2 – Exemple de dot-matrix comparant les séquences de texte "histogramme" et "spatiogramme".

La méthode introduite par Winkler utilise un coefficient de préfixe p qui favorise les chaînes commençant par un préfixe de longueur l (avec $l \leq 4$). En considérant deux chaînes C_1 et C_2 , leur similarité de Jaro-Winkler S_{JW} est :

$$S_{JW} = S_J + (l \times p(1 - S_J)) \quad (2.12)$$

où S_J est la similarité de Jaro entre C_1 et C_2 , l est la longueur du préfixe commun (maximum 4 caractères), et p est un coefficient qui permet de favoriser les chaînes avec un préfixe commun. Winkler propose pour valeur $p = 0,1$.

Nous avons présenté plusieurs approches de mise en correspondance de séquences. Nous nous intéressons maintenant à la dernière famille d'approches, basées sur l'alignement de séquences.

Approche dot-matrix

L'approche *dot-matrix* [42] peut être utilisée pour identifier de façon visuelle certaines propriétés dans des séquences. Par exemple, en l'absence de bruit, les insertions, les suppressions, les répétitions, les répétitions inversées sont facilement identifiables à l'aide *dot-matrix*. Pour construire une telle matrice, les deux séquences sont placées dans la première ligne et première colonne d'une matrice à deux dimensions. Un point (*dot*) est placé dans la matrice quand les caractères de ligne et de colonne correspondent. La dot-matrix de deux séquences très proches présente essentiellement des points formant une diagonale (cf. la sous-séquence commune "ogramme" dans l'exemple du Tableau 2.2). Les problèmes de ce type de représentation viennent du bruit, du manque de clarté, du manque d'intuitivité et de la difficulté d'en extraire un résumé statistique sur les positions correspondantes dans deux séquences. En effet cette représentation ne permet que de comparer deux séquences que visuellement.

Dynamic time warping (DTW)

La déformation temporelle dynamique, plus connue sous son nom anglais de *dynamic time warping* (DTW) [95], est un algorithme permettant de mesurer la similarité entre

deux séquences temporelles. La similarité entre les deux séries de données peut être établie même si le phénomène étudié se déroule à des **vitesses différentes** dans les deux séries d'échantillons. L'algorithme DTW peut être appliqué dans toute situation où les données peuvent être transformées en une représentation linéaire. Une application majeure concerne la **reconnaissance automatique de la parole** [105], où il est nécessaire de tenir compte de vitesses de locution très variables d'une personne à l'autre. Voici l'algorithme permettant de la calculer :

```

Data : séquences  $s_0$  et  $s_1$ 
Result :  $\frac{dtw(n-1, m-1)}{n+m-2}$ 
 $n \leftarrow |s_0|, m \leftarrow |s_1|;$ 
 $dtw \leftarrow Mat(n, m);$ 
for  $i = 0$  to  $n$  do
    |  $dtw(i, 0) \leftarrow 0;$ 
end
for  $j = 0$  to  $m$  do
    |  $dtw(0, j) \leftarrow 0;$ 
end
for  $i = 1$  to  $n$  do
    | for  $j = 1$  to  $m$  do
        | |  $d \leftarrow distance(s_0[i], s_1[j]);$ 
        | |  $dtw(i, j) \leftarrow d + \min[dtw(i - 1, j), dtw(i, j - 1), dtw(i - 1, j - 1)];$ 
    | end
end
```

Algorithme 1 : L'algorithme DTW réalisant l'alignement de deux séquences s_0 et s_1 .

De façon générale, DTW est une méthode qui recherche un appariement optimal entre deux séquences temporelles, sous certaines restrictions. Les séquences temporelles sont déformées par transformation non-linéaire de la variable temporelle, pour déterminer une mesure de leur similarité, indépendamment de certaines transformations non-linéaires du temps.

L'avantage de DTW est qu'elle se base sur une opération unitaire qui est l'évaluation de la distance entre deux éléments de la séquence, et permet ainsi la comparaison de séquences d'éléments numériques (non-nominaux), contrairement aux autres mesures de similarité entre chaînes de symboles nominaux.

En conclusion, dans le cas où l'on souhaite comparer deux séries d'histogrammes de même longueur, il est possible de comparer les histogrammes situés aux mêmes positions dans les deux séries en prenant la moyenne obtenue par une des métriques précédentes sur tous les histogrammes. Si les deux séries ont des tailles différentes mais que les histogrammes sont ordonnés dans le temps (i.e. des histogrammes construits sur chaque trame de la vidéo), il est possible d'utiliser la DTW associée à une des métriques entre deux histogrammes précédentes. Dans nos travaux, nous utilisons la DTW pour comparer des séquences d'histogrammes spatio-temporels (cf. Section 4.3.2).

2.2.4 Espace de représentation des couleurs

Nous avons vu qu'il existait de nombreuses formes d'histogramme et de nombreuses métriques pour les comparer. **Les histogrammes sont des outils généraux pour représenter la fréquence de phénomènes.**

Dans la vision par ordinateur, on peut distinguer les objets en s'intéressant principalement à leurs formes, à leurs textures ou à leurs couleurs. Nous allons nous intéresser à ces dernières. Il existe de nombreuses façons de représenter les couleurs, on parle couramment d'espace de représentation des couleurs.

L'œil humain

Avant de discuter de la représentation des couleurs en informatique, il est important de rappeler comment sont perçues les couleurs par l'œil humain car de nombreux espaces de couleurs ont été créés en s'inspirant de son fonctionnement.

L'homme perçoit une immense variété de couleurs différentes, il ne possède pourtant que trois types de récepteurs appelés cônes ayant chacun une sensibilité plus grande à certaines longueurs d'onde lumineuse [9] : les cônes bleus (B), les cônes verts (V) et les cônes rouges (R). Il est courant de qualifier les cônes bleus de S (pour short), les cônes verts de M (pour medium) et rouge de L (pour long) en référence à la longueur d'onde au maximum de sensibilité. Cette sensibilité est d'ailleurs différente d'un individu à l'autre [104].

Chaque type de cônes en lui-même ne peut détecter qu'une couleur particulière, dans la mesure où sa réponse ne fait que refléter le nombre de photons qu'il capte, indépendamment de leur longueur d'onde. Un photorécepteur n'est qu'un "compteur de photons" [32]. La perception des couleurs n'est possible qu'au niveau du cerveau par comparaison des signaux issus de deux classes de cônes. La réponse des cônes V et R étant très proche, ils servent principalement à détecter la structure spatiale des images.

Chez l'Homme, les cônes B sont les moins nombreux (4% à 5%), puis viennent les cônes V et les cônes R, avec des variations inter-individuelles importantes [92]. Les cônes forment une mosaïque avec chaque type disposé de manière aléatoire.

Représentation RVB

La représentation Rouge-Vert-Bleu (RVB, RGB en anglais) est la représentation la plus répandue des images. Les quantités de rouge, de vert et de bleu de chaque pixel sont exprimées indépendamment. Cependant, il a été montré qu'il existe une forte corrélation entre les différentes valeurs pour rouge, vert et bleu [35]. Ceci est mis en évidence par certaines méthodes permettant d'estimer la valeur d'une couleur en se basant sur la valeur des deux autres, par exemple pour résoudre un problème de saturation des couleurs [114].

La Figure 2.7, décomposant une trame d'une émission audiovisuelle selon les canaux rouge, vert et bleu, montre clairement que l'information portée par chaque canal est très proche de celle portée par les autres : le journaliste reste facilement reconnaissable dans les trois images.

Quand on calcule un histogramme de couleur sur les pixels RVB, deux approches différentes peuvent être utilisées. L'une consiste à découper le spectre de couleurs, donc à prendre la valeur combinée de rouge, de vert et de bleu. Une autre approche consiste à considérer ces trois valeurs indépendamment en les séparant dans trois canaux différents, et en construisant un histogramme monochrome pour chaque composante.

Représentation Y'UV

Y' est la composante de luminosité et la luminance est notée Y, le symbole prime (') correspond à la compression gamma. La luminance correspond à la luminosité perçue,

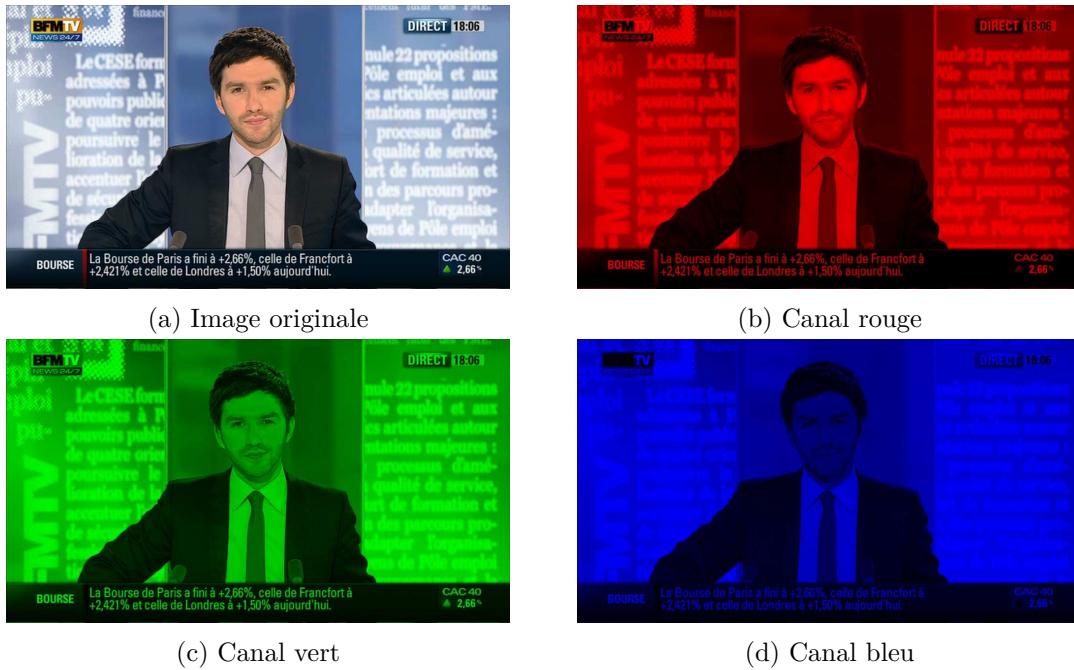


FIGURE 2.7 – Décomposition d’une image selon les canaux rouge-vert-bleu.

alors que la luminosité correspond à une grandeur en électronique, i.e. la tension appliquée sur l'affichage. Le modèle Y'UV définit un espace de couleurs composé d'un canal de luminosité (Y') et de deux canaux de chrominances (UV). Cet espace de représentation des couleurs a historiquement été introduit pour la télévision.

Les anciens systèmes noir et blanc utilisaient uniquement l'information de luminosité (Y'). Les informations de couleurs (U et V) (cf. Figure 2.8) ont été ajoutées séparément dans d'autres composantes pour assurer la rétrocompatibilité avec les affichages noir et blanc. L'espace de couleurs YUV encode la couleur d'une image en prenant en compte la perception humaine. La contribution de chaque canal à la représentation des couleurs est mis en évidence dans la Figure 2.9. L'espace de couleurs YUV permet de réduire la bande passante utilisée pour la composante de chrominance tout en permettant de masquer au maximum à l'œil humain les erreurs issues de transmission ou d'encodage. L'avantage principal de celui-ci est qu'il est interfaçable avec de l'équipement analogique ou numérique tel que des télévisions, des caméras ou des appareils photos qui se conforment au standard Y'UV.

La transformation de RGB vers YUV s'écrit :

$$Y = 0,299 \times R + 0,587 \times G + 0,114 \times B \quad (2.13)$$

$$U = -0,147 \times R - 0,289 \times G + 0,436 \times B \quad (2.14)$$

$$V = 0,615 \times R - 0,515 \times G - 0,100 \times B \quad (2.15)$$

Dans cette formule Y reste dans l'intervalle [0, 1], mais U et V peuvent prendre des valeurs positives ou négatives.

YCbCr est un espace de couleur similaire à Y'UV. La formule de transformation dans cet espace de couleur dépend de la recommandation suivie. En suivant la recommandation Rec 601-1, nous prenons la valeur 0,2989 pour rouge, la valeur 0,5866 pour vert et 0,1145 pour bleu :

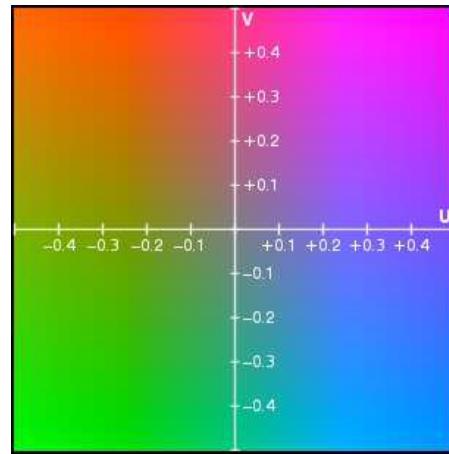


FIGURE 2.8 – Exemple d’une plage U-V, où $Y = 0,5$, représentée à l’intérieur de la gamme de couleurs RVB ; en noir et blanc, seule Y est utilisée, toutes ces couleurs rendent donc le même gris.

La transformation de RGB vers YCbCr (Recommandation 601-1) s’écrit :

$$Y = 0,2989 \times R + 0,5866 \times G + 0,1145 \times B \quad (2.16)$$

$$Cb = -0,1688 \times R - 0,3312 \times G + 0,5000 \times B \quad (2.17)$$

$$Cr = 0,5000 \times R - 0,4184 \times G - 0,0816 \times B \quad (2.18)$$

Représentation HSV

HSV (*hue saturation value*) et HSL (*hue saturation luminance*) sont les deux systèmes de représentation de couleurs du modèle RGB par coordonnées cylindriques les plus connus. Les deux représentations réarrangent la géométrie de RGB pour être plus intuitif et plus perceptuellement pertinent que la représentation cartésienne sous forme de cube. Développée dans les années 1970 pour les applications de dessins par ordinateur, HSL et HSV sont utilisés couramment aujourd’hui pour la sélection de couleurs sur une palette (cf. Figure 2.10), pour les logiciels d’édition d’image et un peu moins pour l’analyse d’image et la vision par ordinateur.

HSL est composé de la teinte (*hue*), de la saturation et de la lumière et est aussi souvent dénommé HLS ou TSL en français. HSV est composé de la teinte, de la saturation et de la valeur et aussi souvent écrit HSB avec le B désignant la luminosité (*brightness* en anglais) ou TSV en français. Un troisième modèle appelé HSI pour teinte, saturation et intensité existe aussi. Cependant, même si elles sont cohérentes entre elles, ces définitions ne sont pas standardisées et chaque abréviation peut être utilisée pour tous les modèles présentés précédemment ou pour tout modèle cylindrique de ce type.

Dans chaque cylindre, l’angle autour de l’axe vertical correspond à la teinte (cf. Figure 2.12), la distance à cet axe à la saturation et la distance le long de cet axe à la lumière, la valeur ou la luminosité. Soulignons que si les teintes dans HSL et HSV correspondent au même attribut, la définition de la saturation est très différente.

Dans chaque géométrie, les couleurs primaires et secondaires additives que sont le rouge, le jaune, le vert, le cyan, le bleu et le magenta, ainsi que des combinaisons linéaires

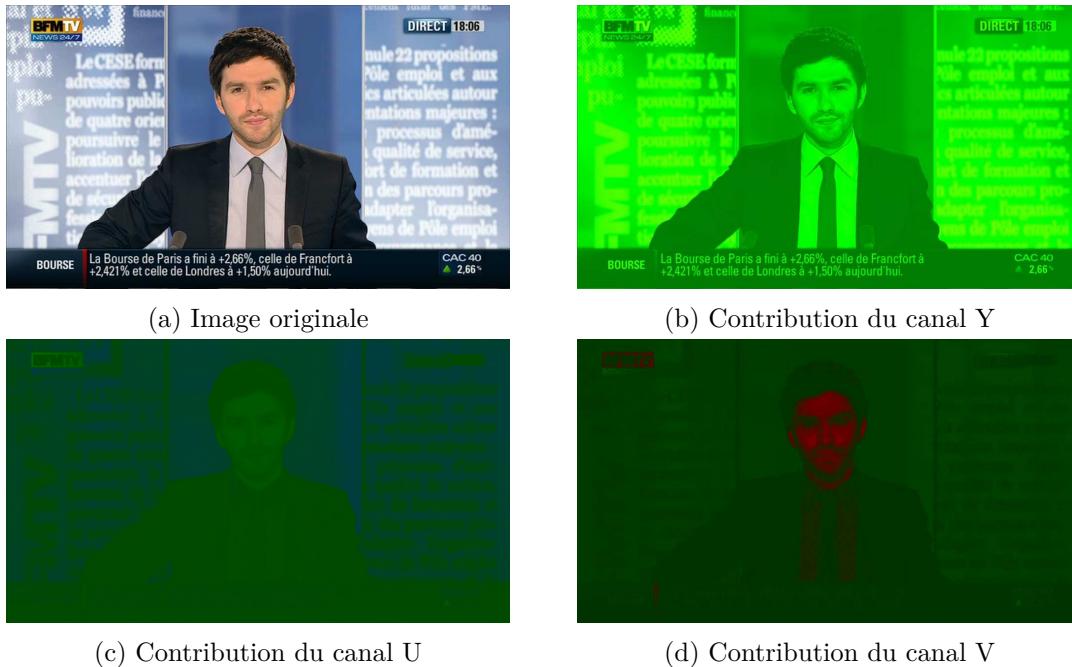


FIGURE 2.9 – Décomposition d'une image selon les canaux YUV avec mise en évidence de la contribution de chaque canal à la représentation des couleurs.

entre paires adjacentes de ceux-ci, appelées parfois "couleurs pures", sont situées sur le bord externe du cylindre pour une saturation de 1. Dans HSV, ces couleurs pures ont une valeur de 1 alors que dans HSL, elles ont une valeur de $\frac{1}{2}$. De plus, dans HSV, le mélange de ces couleurs pures avec du blanc, produisant des teintes, réduit la saturation. Alors que dans HSL, les teintes et les ombrages ont une saturation complète, seuls les mélanges avec du blanc et du noir, appelés tons, ont une saturation plus petite que 1.

Représentation OHTA

L'espace de représentation OHTA [83] a été créé afin d'obtenir une corrélation minimale entre les trois canaux de couleurs. Ces derniers contiennent donc des informations

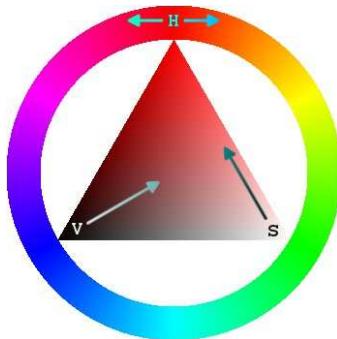


FIGURE 2.10 – Une roue de couleurs HSV permet à l'utilisateur de sélectionner une multitude de couleurs.

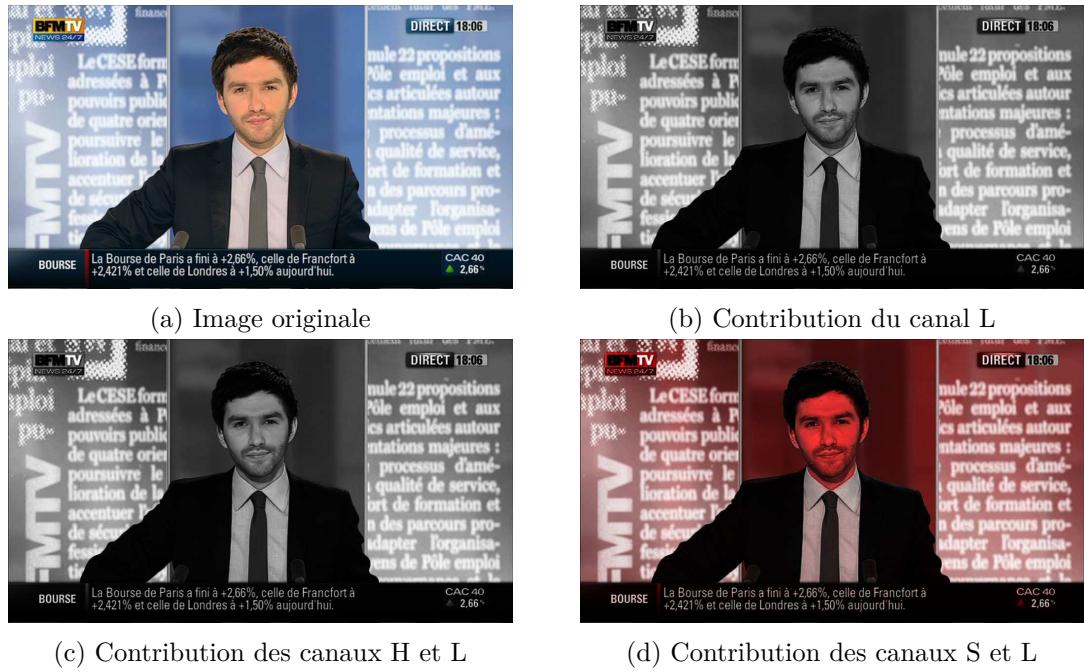


FIGURE 2.11 – Décomposition d’une image selon les canaux HSL avec mise en évidence de la contribution de chaque canal à la représentation des couleurs.



FIGURE 2.12 – Teintes du cercle chromatique.

différentes. Il est difficile de décrire ce que représente chaque canal car la représentation OHTA est obtenue de façon artificielle, la Figure 2.13 donne un aperçu de la contribution de chaque canal pour représenter une couleur. En effet, une analyse en composantes principales (ACP) a été réalisée sur l'espace de couleur RVB à partir d'une collection de 8 images montrant 8 scènes différentes supposées représentatives des images naturelles possibles [83]. Ces 8 scènes (voir Figure 2.14) montrent : un cylindre, un bâtiment public, un bord de mer, une fille (la Figure 2.15 présente l'image originale utilisée), une chambre, une maison, une voiture et un visage.

De cette ACP, les trois composantes principales I_1 , I_2 et I_3 ont été conservées. Ces dernières permettent d'approximer toutes les couleurs. L'espace de couleur ainsi généré est une transformation linéaire simple de l'espace RVB.

Cette transformation se fait grâce aux équations suivantes :

$$I_1 = \frac{1}{3}(R + G + B) \quad (2.19)$$

$$I_2 = \frac{1}{2}(R - B) \quad (2.20)$$

$$I_3 = \frac{1}{4}(2G - R - B) \quad (2.21)$$

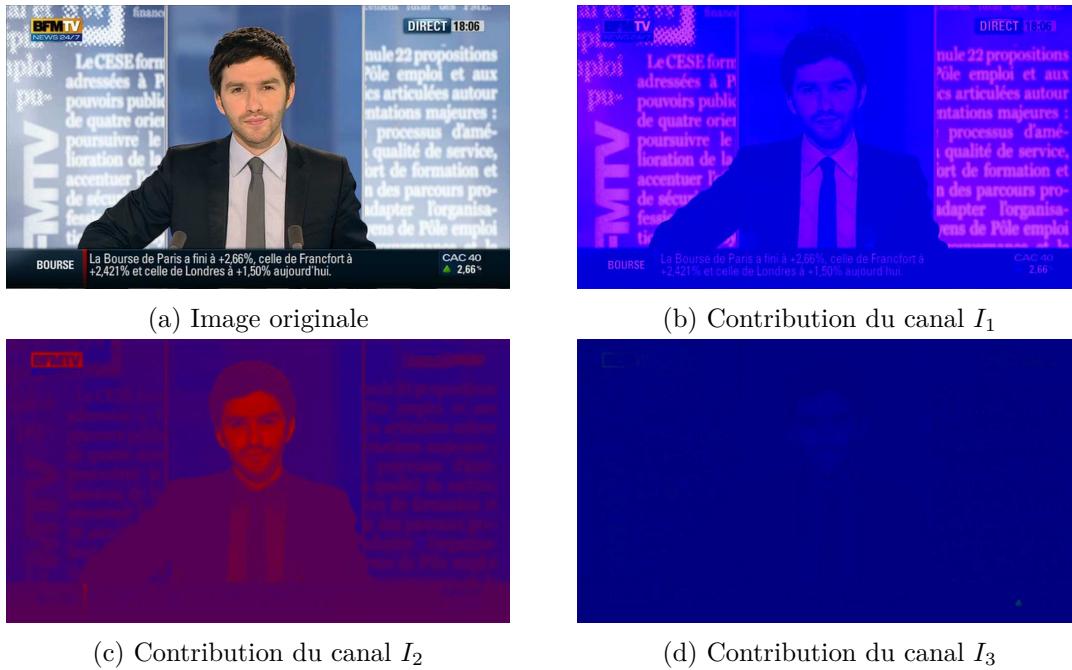


FIGURE 2.13 – Décomposition d'une image selon les canaux de l'espace de représentation OHTA avec mise en évidence de la contribution de chaque canal à la représentation des couleurs.

Espaces physiologiques

Parmi les autres espaces de couleurs, certains essaient de reproduire le fonctionnement de l'œil humain. C'est le cas des espaces de couleurs créés par la Commission Internationale de l'Eclairage (CIE). Les plus connus sont CIE XYZ et CIELAB.

CIE XYZ, dont le nom complet est "CIE 1931 XYZ color space", a été créé à la fin des années 1920 par Willan David Wright [111] et John Guild [46]. Leurs résultats expérimentaux ont été combinés dans la spécification de l'espace de couleur CIE RGB duquel CIE XYZ a été dérivé. La contribution de chaque canal à la représentation des couleurs est mis en évidence par la Figure 2.16. L'espace de couleur XYZ simule le phénomène physique de perception de couleur par l'œil humain alors que l'espace de couleur L*a*b* simule la perception des couleurs par le cerveau.

Quand l'œil humain juge de la luminosité relative de différentes couleurs dans un environnement bien éclairé, il a tendance à percevoir la lumière dans la partie verte du spectre comme étant plus lumineuse que celle dans le rouge et le bleu à intensité égale. La fonction qui décrit la luminosité perçue pour les différentes longueurs d'onde est proche de la fréquence de réponse des cônes M (cf. Section 2.2.4).

Le modèle CIE tire parti de ce fait en définissant Y comme la luminosité. Z est quasiment égal à la stimulation du bleu, ou la réponse du cône S, et X est un mélange linéaire des courbes de réponse des cônes choisis pour être non négatif. La valeur du tristimulus XYZ est donc proche, à la réponse des cônes LMS de l'œil humain. La définition de Y en tant que luminosité a l'avantage que pour chaque valeur de Y, le plan formé par XZ contient toutes les chromacées⁴ de celle-ci.

4. Une chromacie caractérise la couleur indépendamment de son intensité.

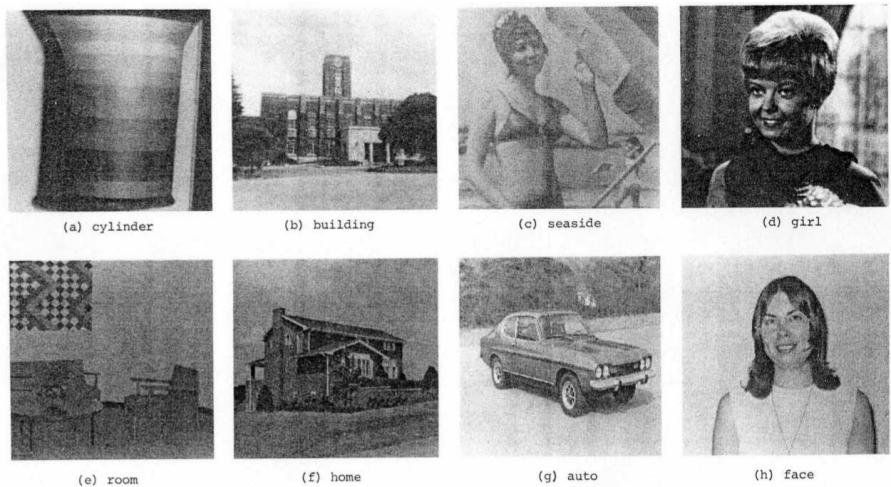


FIGURE 2.14 – Collection d’images ayant servi pour la construction de l’espace de représentation OHTA. Ces images sont issues de l’article [83]. Les images utilisées étaient en couleur, mais seules les versions en niveaux de gris sont facilement disponibles aujourd’hui.

La représentation de couleurs CIELAB, aussi appelée $L^*a^*b^*$, a été créée afin de modéliser la façon dont le cerveau perçoit les couleurs. Ainsi, les couleurs sont modélisées à partir de trois canaux. Le canal L^* , contient la valeur de luminosité de 0 à 100, du plus sombre au plus lumineux. Le canal a^* contient une valeur pour l’axe rouge-vert allant de -299 pour une couleur verte à +300 pour une couleur rouge, en passant par 0 pour le gris. Enfin, le canal b^* représente l’axe bleu-jaune, de la même façon que précédemment (cf. Figure 2.17).

Les composantes a^* et b^* sont plus souvent notées par une valeur de +127 à -128 comportant ainsi 256 niveaux permettant d’être codées sur 8 bits en base hexadécimale pour être corrélées avec le système RGB.



FIGURE 2.15 – Image originale couleur utilisée pour la construction de l’espace de représentation OHTA.

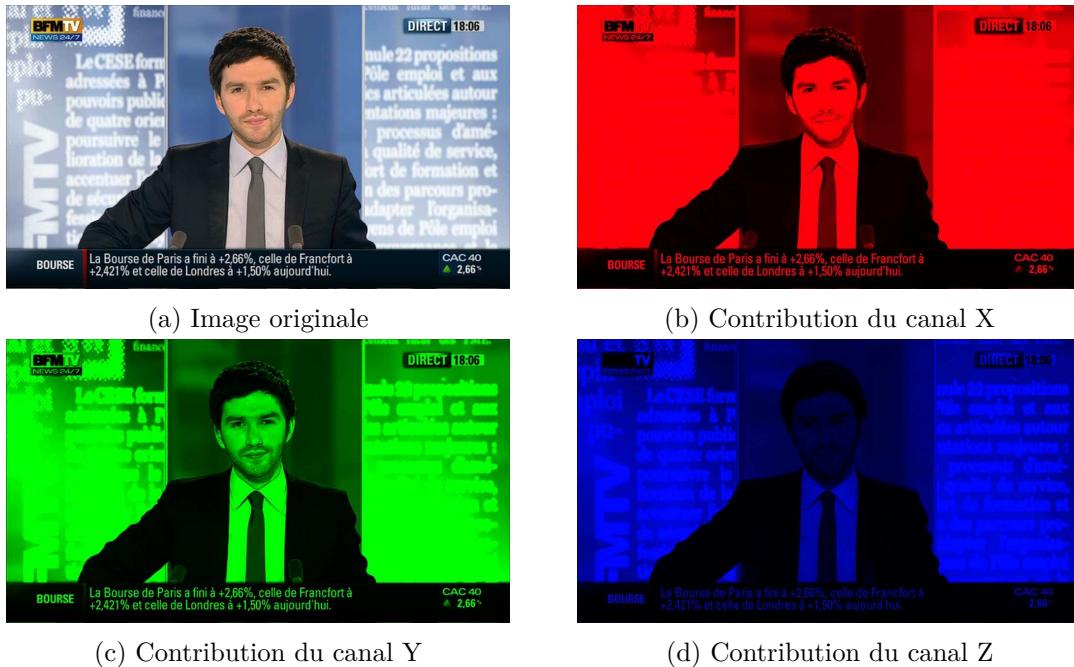


FIGURE 2.16 – Décomposition d'une image selon les canaux de l'espace de représentation XYZ avec mise en évidence de la contribution de chaque canal à la représentation des couleurs.

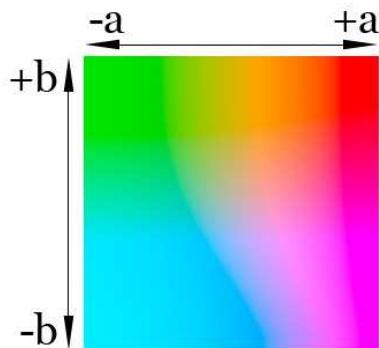


FIGURE 2.17 – Représentation des valeurs des canaux a^* et b^* pour une luminance de 75% dans l'espace de représentation des couleurs $L^*a^*b^*$.

Choix de l'espace de couleurs

Il existe des espaces variés pour représenter les couleurs qui composent une image. Ces différents espaces de représentation des couleurs ont des propriétés différentes. Il convient donc de choisir un espace de couleurs adapté à l'application envisagée. Certains espaces de couleurs permettent de décomposer l'image selon des canaux portant des informations sémantiquement différentes. Ce n'est pas le cas de l'espace de couleur RGB, il convient donc de traiter cet espace de couleur comme une combinaison linéaire plutôt que comme trois canaux distincts. Les images et les trames d'une vidéo sont couramment encodées dans l'espace de représentation RGB. Pour les étudier dans un autre espace, il

est nécessaire d'effectuer une conversion. Cette conversion peut avoir un coût important lorsque l'on considère une vidéo, car il est nécessaire de décoder chaque trame en RGB, puis de la convertir dans l'autre espace de représentation. Le coût de cette conversion dépend du nombre de trames, de leur taille, ainsi que des calculs propres à la conversion.

2.2.5 Clustering d'histogrammes

Afin de ré-identifier les personnes dans une vidéo, il est nécessaire de regrouper les différentes apparitions de celles-ci. Nous avons étudié les différents types d'histogrammes ainsi que les mesures de dissimilarités qui permettent de les comparer. Nous allons maintenant étudier différentes approches pour le clustering d'histogrammes.

Le regroupement (*clustering*) se retrouve aussi dans la littérature sous le nom de *partitionnement de données*. Il est particulièrement utilisé pour la fouille et l'analyse de données. Le regroupement vise à diviser l'ensemble des données en groupes homogènes, pour que les données de chaque groupe (ou sous-ensemble) partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité). La méthode à appliquer pour réaliser ce regroupement est conditionnée par les propriétés des données. Ainsi, de nombreuses méthodes de regroupement existent pour répondre aux différents cas.

Précédemment, nous avons présenté une certaine forme de regroupement donnée par l'ACP et l'ADL (cf. Section 2.1). Parmi les approches les plus représentatives du regroupement, nous allons présenter le regroupement hiérarchique et l'algorithme *k*-moyennes (*k-means*).

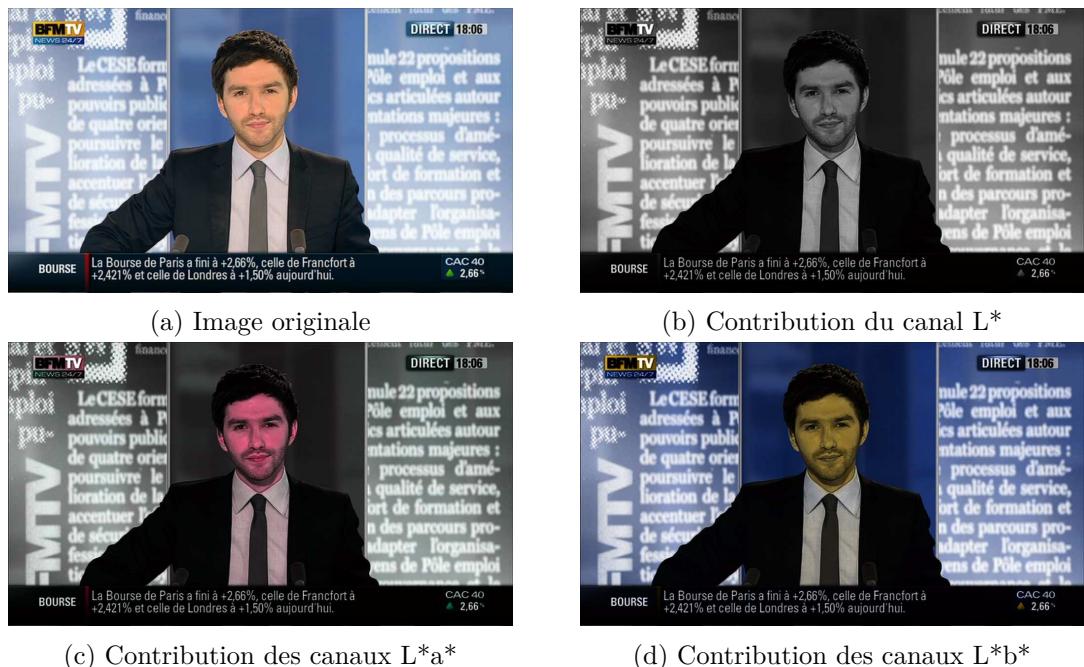


FIGURE 2.18 – Décomposition d'une image selon les canaux de l'espace de représentation $L^*a^*b^*$ avec mise en évidence de la contribution de chaque canal à la représentation des couleurs.

Le clustering hiérarchique

Il existe deux variantes du clustering hiérarchique [54] :

- à partir de la situation initiale où chaque élément est dans un cluster différent, l'algorithme fusionne itérativement les clusters (approche ascendante)
- à partir de la situation initiale où tous les éléments sont dans un même cluster, l'algorithme divise les clusters à chaque itération (approche descendante).

Pour décider de la fusion ou de la division d'un cluster (selon l'approche choisie), une mesure de similarité entre les éléments est nécessaire. Cette mesure ne doit pas nécessairement être une distance, au sens mathématique. Ainsi, les mesures de similarités que nous avons présentées à la Section 2.2.2 conviennent, après transformation en similarité, pour réaliser ce regroupement. Cette méthode continue de fusionner ou de diviser les clusters jusqu'à ce qu'un critère d'arrêt soit atteint. Par exemple, dans un regroupement hiérarchique ascendant, les groupes sont fusionnés deux par deux en minimisant l'inertie interclasse jusqu'à atteindre un équilibre [77].

L'algorithme k -moyennes

Dans l'algorithme k -moyennes [51, 52], un cluster est représenté par un centroïde qui est une moyenne des éléments affectés au cluster. Ce centroïde ne fait pas forcément partie des éléments de l'ensemble des données sur lequel l'algorithme est appliqué. Le nombre de clusters k que l'on souhaite obtenir est un paramètre donné à l'algorithme. Celui-ci va ensuite résoudre le problème d'optimisation qui est de trouver k clusters, où chaque cluster contient les éléments tels que la distance entre ceux-ci et le centre du cluster soit minimale. Ce problème étant NP-difficile [30], une approximation est faite en initialisant aléatoirement les k centroïdes et en cherchant un minimum local. Ainsi, il est courant d'exécuter l'algorithme de nombreuses fois avec des initialisations différentes jusqu'à ce que le résultat converge vers une valeur minimale.

Clustering d'histogrammes

L'algorithme k -moyennes a pour inconvénient que le nombre total de clusters doit être déterminé a priori. De plus, il nécessite de devoir calculer un centroïde en prenant la moyenne des éléments qui le constituent. Dans notre cas, le calcul d'un histogramme spatio-temporel moyen n'est pas défini. Ce clustering n'est donc pas adapté à notre problème. Comme nous avons défini des mesures de dissimilarité entre histogrammes et que nous proposons dans notre approche une mesure de similarité entre histogrammes spatio-temporels, nous allons nous intéresser au clustering hiérarchique pour effectuer le regroupement des personnes qu'ils décrivent. L'hypothèse est que chaque groupe contient une unique identité.

2.3 Étiquetage d'ensembles

Nous avons vu comment regrouper des histogrammes en se basant sur une mesure de similarité. Nous allons maintenant voir comment la littérature aborde le problème du nommage des groupes à partir des éléments qui les constituent. Dans notre cas il s'agit d'occurrences vidéo de personnes. Comme nous l'avons vu précédemment (cf. Sections 2.1 et 2.2), certaines occurrences vidéo de personnes peuvent être associées à une identité. Il

s'agit donc d'utiliser ces identités pour déterminer l'identité globale d'un groupe. Le cas idéal est quand toutes occurrences vidéo de personnes d'un groupe ont la même identité. Les approches de l'état de l'art traitant ce problème [93] sont pour la plupart basées sur le vote. Celui-ci trouve ses fondements dans la théorie des jeux et dans le domaine de la politique, plus précisément dans le cadre de la démocratie. Une majorité est définie comme le sous-ensemble contenant le plus d'éléments d'un ensemble [10]. On distingue principalement deux grandes méthodes de scrutin : l'élection à la majorité relative et l'élection à la majorité absolue [10]. Dans ces deux méthodes, l'élection est remportée par le choix le plus fréquent. Dans une élection à la majorité relative, le choix le plus fréquent remporte automatiquement l'élection. Dans celle à la majorité absolue, le choix le plus fréquent doit représenter au moins la moitié des votes, sans cela l'élection n'est pas validée. Ainsi, une élection à la majorité absolue peut échouer et ne pas donner d'issue. Le résultat d'une élection associe la majorité avec un score. Celui-ci est calculé en comptant le nombre de voix reçues par la majorité divisé par le nombre total de votants [10].

2.4 Synthèse de l'état de l'art

Nous avons vu qu'il est possible d'aborder le problème de la reconnaissance de personnes de façon locale ou de façon globale. Les approches globales permettent de mieux reconnaître les personnes car elles ont accès à des caractéristiques plus précises des personnes. Elles nécessitent néanmoins d'avoir une prise de vue de bonne qualité pour que ces caractéristiques soient exploitable.

Parmi les approches locales, nous distinguons les approches dynamiques basées sur la vidéo des approches statiques qui considèrent les images ou les trames de la vidéo indépendamment. Les approches dynamiques sont relativement peu nombreuses. Elles bénéficient pourtant de l'accès à l'information temporelle que l'on peut supposer utile pour reconnaître une personne. Ces méthodes se basent pour la plupart sur les approches statiques pour effectuer la reconnaissance de la personne et souffrent donc en partie des limitations de celles-ci. Rappelons que les limitations principales des approches de reconnaissance concerne le sujet (posture, pilosité, bijoux, etc.) et les conditions de prise de vue (éclairage, bruit, angle par rapport au sujet, etc.). De ce fait, toutes les occurrences d'une personne ne peuvent être reconnues.

Nous proposons alors de tirer profit de l'aspect temporel pour reconnaître les personnes dans une vidéo. Pour cela, nous proposons de regrouper les occurrences vidéo d'une même personne. Nous avons montré que pour effectuer une telle ré-identification les approches globales basées sur des histogrammes donnent de bons résultats. Nous nous sommes ainsi inspirés de ces approches pour définir un nouveau type d'histogramme qui tire profit à la fois de données spatiales et de données temporelles. Ils sont construits sur les trames de la vidéo, représentées dans l'espace de couleur RGB, considérées de façon linéaire. Cela permet d'éviter les coûts de conversion vers d'autres espaces de représentation. Le problème des conditions d'éclairage ne se pose pas vraiment dans les émissions audiovisuelles, cet aspect étant bien maîtrisé par les équipes de tournage, dans l'environnement contrôlé qu'est le studio de télévision.

Chapitre 3

Mise en perspective de nos contributions

Nous avons présenté l'état de l'art permettant de justifier notre approche. Dans un premier temps, nous allons formaliser les différentes notions essentielles à notre approche. Nous allons mettre en perspective nos propositions à travers une approche de l'état de l'art de reconnaissance de personnes à partir de la vidéo, que nous proposons d'étendre.

Avant toute chose, il est nécessaire d'introduire la notion de vidéo dont nous nous servons dans notre travail :

Définition 1. *Une vidéo V_i , appartenant à l'ensemble des vidéos du corpus, est elle-même un ensemble ordonné de plans s_j^i :*

$$(V_i, \leq) = \{s_0^i, s_1^i, \dots, s_{|V_i|-1}^i\} \quad (3.1)$$

Nous supposons que nous disposons d'un ensemble de vidéos de travail constituant un corpus. Nous donnons la définition suivante d'un plan s_j (s pour *shot*) de la vidéo :

Définition 2. *Un plan s_j d'une vidéo V_i est défini comme un ensemble ordonné de trames contiguës acquises en continu par une même caméra :*

$$(s_j^i, \leq) = \{f_t, f_{t+1}, \dots, f_{t+|s_j^i|-1}\} \quad (3.2)$$

où t est l'indice temporel correspondant au numéro de la première trame (ou *frame*, en anglais) du plan d'une vidéo. Cet indice induit de fait une relation d'ordre (\leq) entre les trames d'un plan. Un plan se termine dès que la vidéo présente des trames acquises par une autre caméra ou que la vidéo présente une coupure.

On retrouve ces deux définitions dans la plupart des systèmes de reconnaissances de personnes dans les vidéos [116].

3.1 Approche classique de reconnaissance de personnes dans les vidéos

Cette approche se compose de plusieurs modules (cf. Figure 3.1) [116]. Le premier module consiste en un découpage en plans. Cela permet de découper la vidéo en segments homogènes en termes de conditions de prise de vue et de personnes présentes. La deuxième

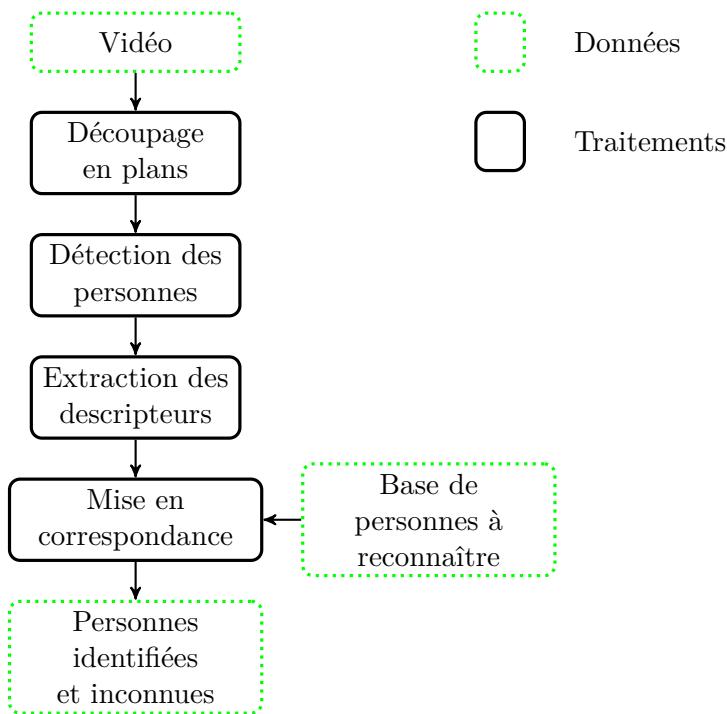


FIGURE 3.1 – Schéma présentant une approche classique de reconnaissance de personnes dans les flux vidéo.

étape consiste à détecter les personnes dans tout le corpus. Pour cela, les approches de l'état de l'art utilisent le plus souvent un détecteur de visages, le plus courant étant le détecteur basé sur des pseudos caractéristiques de Haar proposé par Viola et Jones [106]. La détection du visage permet d'extraire les personnes dans l'image. L'ensemble des pixels qui composent la personne forment un blob :

Définition 3. *Un blob est une région composée de pixels contigus qui partagent une propriété commune.*

La séquence des blobs successifs d'une personne dans un plan forment un *persontrack* [31]. Ce terme anglais de *persontrack* ne semble pas avoir d'équivalent en français ; nous proposons de le nommer ***occurrence vidéo de personne*** (OVP).

Définition 4. *Une occurrence vidéo de personne est une séquence de blobs issus de trames contiguës tirées d'un même plan, représentant une unique personne.*

Pour isoler une occurrence vidéo de personne d'un plan de la vidéo, les détections consécutives sont fusionnées à l'aide d'un algorithme de suivi. Quand plusieurs personnes sont présentes dans un plan donné, alors le plan contient plusieurs occurrences vidéo. Nous notons \mathbb{O}_i l'ensemble des occurrences vidéo des personnes appartenant à la vidéo V_i :

$$\mathbb{O}_i = \{o_0, o_1, \dots, o_{|\mathbb{O}_i|-1}\} \quad (3.3)$$

Les visages d'une occurrence vidéo de personne sont ensuite utilisés pour reconnaître le sujet en utilisant un algorithme dédié à partir d'une base d'apprentissage contenant les

différentes identités à reconnaître, ainsi que les descripteurs permettant d'associer une identité à une occurrence vidéo de personne. Cette base est construite au préalable à partir d'occurrences vidéo de personnes ou d'images de personnes annotées. L'ensemble des identités considérées est noté :

$$\mathbb{I} = \{\iota_0, \iota_1, \dots, \iota_{|\mathbb{I}|-1}\} \quad (3.4)$$

Cet ensemble (non ordonné) est défini de manière commune à l'ensemble du corpus vidéo. À chaque occurrence vidéo de personne est associée une identité ι . Nous définissons la fonction *id* :

$$\begin{aligned} id : & \mathbb{O} \rightarrow \mathbb{I} \\ & o \rightarrow \iota \end{aligned} \quad (3.5)$$

comme la fonction qui indique l'identité ι associée à l'occurrence vidéo de personne o , dans un corpus annoté. Nous avons formalisé les notions permettant de présenter l'approche classique de reconnaissance de personnes dans les vidéos et de comprendre nos contributions que nous allons maintenant présenter.

3.2 Proposition générale

Notre proposition générale est inspirée de l'approche classique de reconnaissance présentée à la Figure 3.2. Un premier volet (cf. Figure 3.2) s'intéresse au **regroupement des occurrences de personnes** (aussi appelé *ré-identification*). Elle est détaillée dans la Partie II de nos travaux. Nous faisons l'hypothèse qu'au sein d'une vidéo, l'apparence (visage, cheveux, costume, etc.) d'une personne ne varie pas. Nous proposons, dans le Chapitre 4, un descripteur pour représenter chacune des occurrences vidéo de personnes, afin de les mettre en correspondance. Ce descripteur, appelé **histogramme spatio-temporel**, fournit une représentation de l'aspect visuel (couleurs), spatial (positions dans l'image), ainsi que temporel (temps d'apparition) des personnes présentes dans les occurrences vidéo. L'objectif est de créer une signature propre à chacune, qui soit la plus discriminante possible. Les signatures servent de base à un processus de *regroupement (clustering)* dont l'objectif est de séparer les identités dans des groupes d'occurrences. Autrement dit, l'objectif de cette étape est de ranger dans un même groupe $\Omega_{\iota,i}^*$ toutes les occurrences de personnes d'une vidéo V_i ayant la même identité ι :

$$\Omega_{\iota,i}^* = \{o \in \mathbb{O}_i | id(o) = \iota\} \quad (3.6)$$

Notre approche est dynamique en cela qu'elle exploite l'aspect temporel des vidéos. Les histogrammes spatio-temporels permettent dans une certaine mesure de localiser les apparitions des couleurs dans une occurrence vidéo de personne et de donner des indications quant au mouvement des couleurs dans le temps et dans l'espace. La plupart des approches de l'état de l'art ne considèrent pas le temps dans la description d'une occurrence d'une personne [15], elles ne donnent qu'une représentation de l'aspect visuel des personnes.

Le second volet (cf. Figure 3.2) est consacré au **nommage des personnes**, en se basant sur les groupes définis dans la partie précédente. Elle est détaillée dans la partie III de nos travaux. Nous utilisons dans le Chapitre 6 une approche de reconnaissance de l'état de l'art pour identifier un sous-ensemble d'occurrences choisies d'un groupe, ces identités sont propagées, avec une stratégie adaptée, à toutes les occurrences du groupe.

La reconnaissance de personne nécessitant d'importants calculs, l'objectif est limiter le nombre d'occurrences à considérer et de propager les identités dans les groupes. Cela nous permet de nommer plus d'occurrences de personnes qu'une approche dépourvue de propagation, améliore sensiblement la précision et nécessite moins de calculs. De cette façon, la plupart des limitations de la reconnaissance sont contournées pour identifier les personnes au sein d'une même vidéo. La fonction inverse id^{-1} permet, à partir d'une identité ι , de retrouver l'ensemble O_ι des occurrences vidéo correspondant à cette identité. Cette fonction s'apparente à une fonction de recherche des occurrences d'une personne dans la vidéo. L'objectif de ce travail de thèse est de proposer une approche originale pour définir des fonctions \hat{id} et \hat{id}^{-1} comme des approximations des fonctions id et id^{-1} . Dans les Parties II et III, nous évaluons nos propositions afin de les valider expérimentalement (respectivement dans les Chapitres 5 et 7). Dans le Chapitre 8, nous concluons en résumant les points principaux de nos contributions, et nous proposons quelques perspectives que nous allons explorer suite à ce travail.

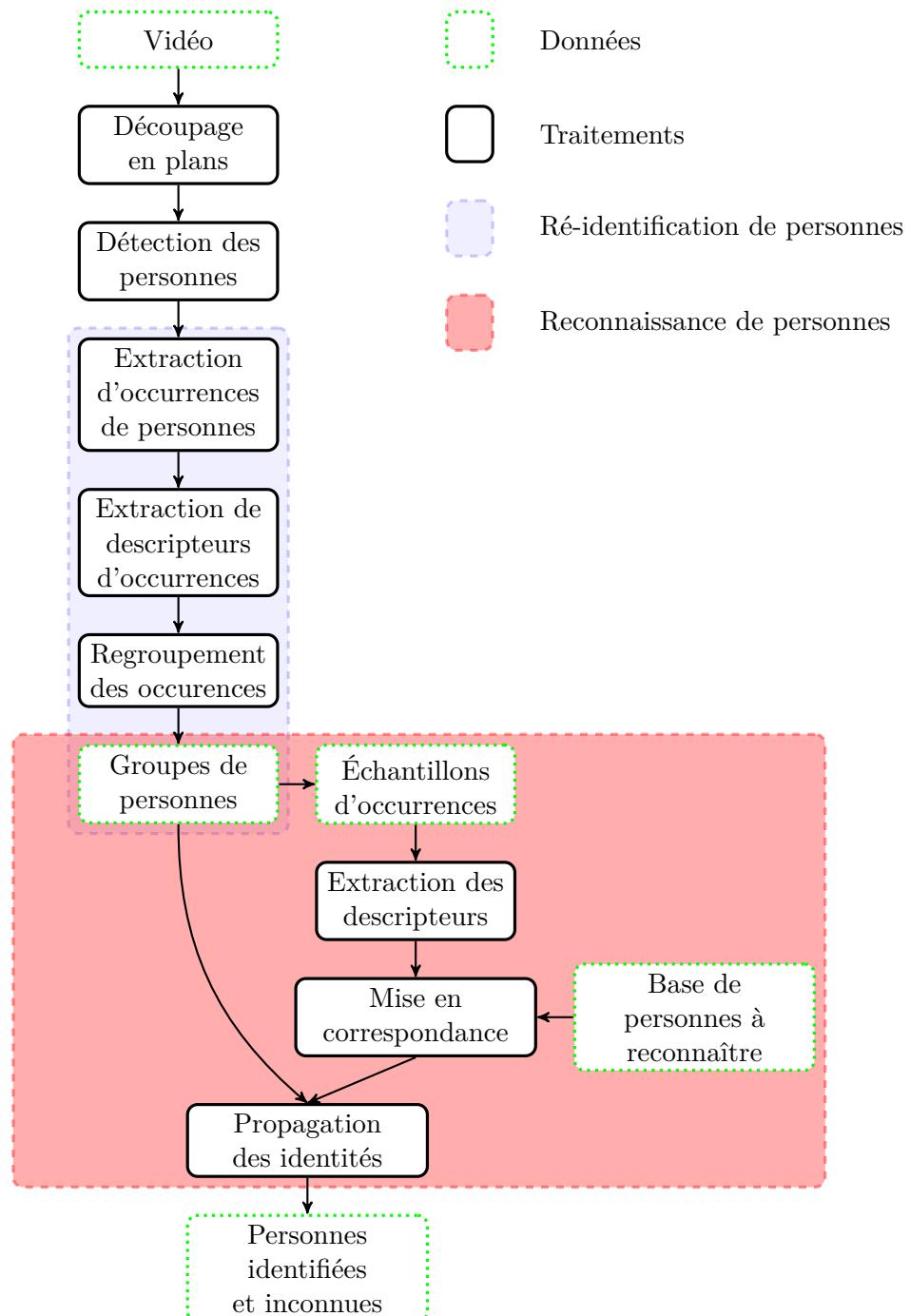


FIGURE 3.2 – Schéma présentant notre approche globale pour la reconnaissance dynamique de personnes dans les flux vidéo.

Deuxième partie

Regroupement des occurrences vidéo de personnes

Chapitre 4

Approche proposée pour le regroupement des OVP

Dans ce chapitre, nous proposons une méthode de regroupement des occurrences vidéo de personnes basée sur leur apparence globale. Pour cela, nous nous plaçons au niveau d'une occurrence vidéo de personne dont les dimensions spatiales x et y sont normalisées. La dimension temporelle t est également normalisée de telle sorte que les coordonnées tridimensionnelles (x, y, t) d'un pixel dans le volume que représente l'occurrence vidéo sont comprises entre $(0, 0, 0)$ et $(1, 1, 1)$ (cf. Figure 4.1).

Nous détaillons dans ce chapitre un descripteur original, l'histogramme spatio-temporel, dédié à la représentations des personnes. Nous détaillons plusieurs manières de construire **un histogramme spatio-temporel** à partir d'une occurrence de personne. Une mesure de similarité des histogrammes spatio-temporels est définie, permettant de comparer les occurrences vidéo selon leur ressemblance visuelle. Ce descripteur associé à la mesure de

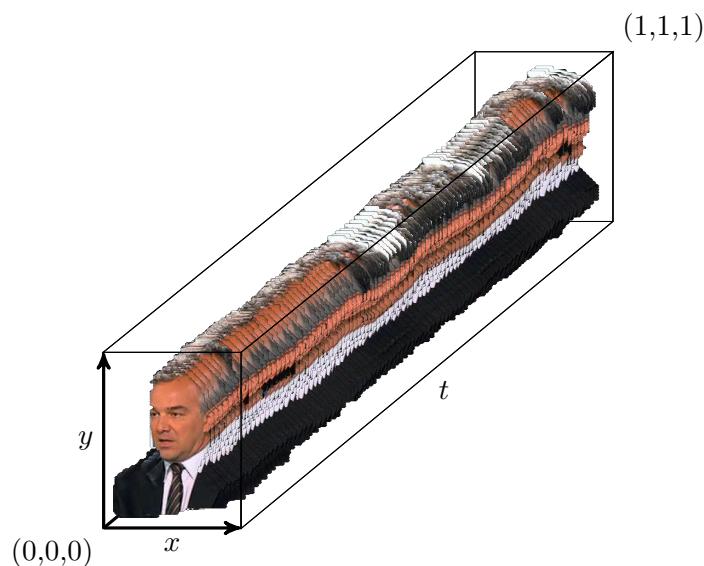


FIGURE 4.1 – Illustration de la notion d'occurrence vidéo de personne (OVP) (*person-track*).

similarité tient une place centrale dans le processus de regroupement d'occurrence de personne en vue de la création des groupes $\Omega_{t,i}$ définis dans l'Équation 3.6. La question de leur complexité et discutée dans ce chapitre. Ils servent à la construction de matrices de similarités d'occurrence de personne, qui sont à la base du clustering des histogrammes spatio-temporels.

4.1 Histogrammes spatio-temporels

Nous définissons ici de façon formelle l'histogramme spatio-temporel qui est un descripteur qui nous permet de créer une signature d'occurrence vidéo de personne. Ce descripteur est une des contributions principales de notre travail. La définition que nous proposons pour les histogrammes spatio-temporels est une extension des spatiogrammes proposés dans les travaux de Truong Cong [102], eux-mêmes étant une extension des histogrammes de couleurs classiques.

La structure de données de l'histogramme spatio-temporel hst_o , construite à partir de l'occurrence vidéo o , est définie ainsi :

$$hst_o(b) = \langle n_b, \mu_b, \Sigma_b \rangle, \quad b = 1, \dots, B \quad (4.1)$$

où n_b est le nombre de pixels de la partition b (*bin* en anglais) et B le nombre total de partitions. La position moyenne dans l'espace et dans le temps, μ_b , est définie par :

$$\mu_b = (\bar{x}_b, \bar{y}_b, \bar{t}_b) \quad (4.2)$$

En notant x_b et y_b la position normalisée des pixels de la partition dans l'occurrence vidéo, et t_b leur indice temporel normalisé, leurs valeurs moyennes sont dénotées par \bar{x}_b , \bar{y}_b et \bar{t}_b . Σ_b est la matrice de covariance de ces positions spatio-temporelles :

$$\Sigma_b = \begin{pmatrix} cov(x_b, x_b) & cov(x_b, y_b) & cov(x_b, t_b) \\ cov(y_b, x_b) & cov(y_b, y_b) & cov(y_b, t_b) \\ cov(t_b, x_b) & cov(t_b, y_b) & cov(t_b, t_b) \end{pmatrix} \quad (4.3)$$

Cette matrice de covariance est symétrique car $cov(a, b) = cov(b, a)$. Rappelons que les histogrammes spatio-temporels, tels que définis dans l'Équation 4.1, contiennent des spatiogrammes, et de manière identique, les spatiogrammes contiennent des histogrammes de couleurs (cf. Section 2.2.1).

4.2 Interprétations

Nous présentons ici des interprétations qu'il est possible de donner aux informations contenues dans un histogramme spatio-temporel. Comme nous l'avons vu dans la section précédente, un histogramme spatio-temporel conserve, pour chaque partition, le nombre de pixels appartenant à cette partition, leur position moyenne dans l'espace et dans le temps, ainsi qu'une matrice de covariance de ces positions.

4.2.1 Illustration

Si on considère le volume tridimensionnel que représente une occurrence vidéo de personne, chaque partition d'un histogramme spatio-temporel peut être décrite par un

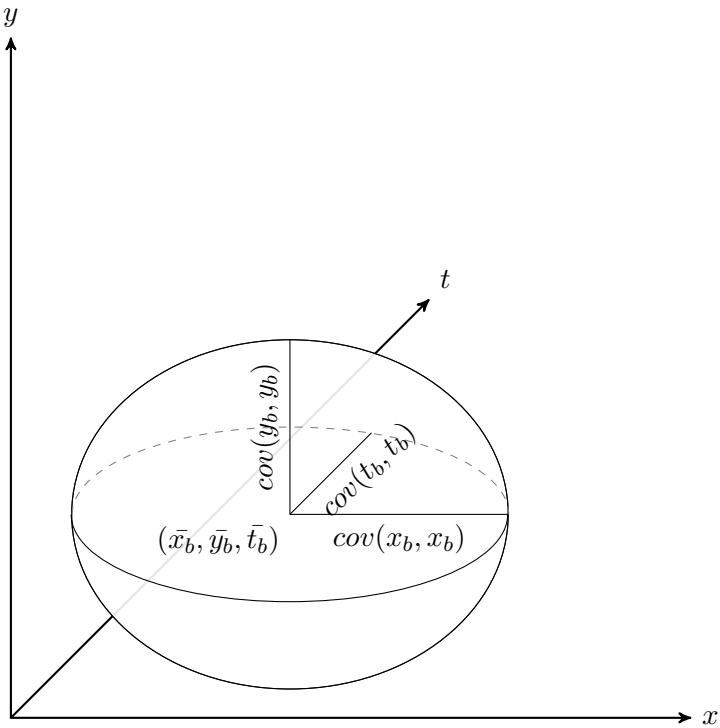


FIGURE 4.2 – Illustration de la représentation d'une partition b dans le volume tridimensionnel de l'occurrence vidéo à l'aide d'un ellipsoïde.

ellipsoïde dans ce volume. Les positions moyennes \bar{x}_b , \bar{y}_b et \bar{t}_b sont le centre de la forme et les covariances $cov(x_b, x_b)$, $cov(y_b, y_b)$ et $cov(t_b, t_b)$ donnent les paramètres (longueurs des demi-axes) de l'ellipsoïde. Les covariances entre les dimensions informent de l'orientation dans l'espace de l'ellipsoïde. Cette illustration permet d'appréhender les informations contenues dans un histogramme spatio-temporel. Cependant, cette représentation n'est qu'une approximation, le volume de l'ellipsoïde ne correspondant pas au nombre de pixels de chaque partition.

4.2.2 Information de mouvement

Les covariances entre les coordonnées x_b , y_b et t_b mettent en évidence des informations de mouvement. Les covariances $cov(x_b, t_b)$ et $cov(y_b, t_b)$ indiquent le sens et l'amplitude du mouvement global de la partition. L'indice du temps étant croissant le long de la vidéo, le signe de la covariance témoigne de la direction du mouvement. Ainsi, $cov(x_b, t_b) > 0$ (respectivement $cov(x_b, t_b) < 0$) dénote un moment de translation vers la droite (respectivement vers la gauche) des pixels de la partition. De même, le signe de $cov(y_b, t_b)$ témoigne d'une translation vers le bas (valeur positive), ou vers le haut (valeur négative). Une covariance nulle témoigne de l'absence de mouvement pour cette dimension. La covariance entre x_b et y_b indique si x_b et y_b varient de la même façon dans le temps. Ainsi, si x_b et y_b sont tous deux croissants ou décroissants (i.e. déplacement dans le temps des pixels de la partition selon une diagonale), la covariance sera positive. Si l'un est croissant alors que l'autre est décroissant, la covariance sera négative.

4.2.3 Interprétation de la dimension temporelle des HST

La distribution temporelle des pixels d'une partition peut être uniforme dans le cas de couleurs présentes durant toute la durée de l'occurrence vidéo ; ou variable dans le cas d'une présence ponctuelle ou discontinue de ces couleurs. Ainsi, les informations temporelles (position moyenne \bar{t}_b et $cov(t_b, t_b)$, correspondant à la variance de t_b) contenues dans les histogrammes spatio-temporels permettent de distinguer les schémas d'apparition des pixels pour chaque partition (apparition ponctuelle, discontinue, persistante, etc.). La variance $var(t_b)$ mesure l'étalement dans le temps des pixels de la partition, centrés sur la moyenne \bar{t}_b . Par exemple, une présence continue se traduira par une position temporelle moyenne \bar{t}_b proche de la moitié du temps total de la séquence, et une variance $var(t_b)$ égale à moitié de la durée de la séquence vidéo. Une apparition ponctuelle impliquera une moyenne \bar{t}_b centrée sur les pixels de la partition.

4.3 Stratégies de construction

L'objectif de ce travail est de proposer un descripteur pour les occurrences vidéo de personnes permettant de discriminer les identités. Nous avons envisagé différentes représentations d'occurrences vidéo de personnes à base d'histogrammes spatio-temporels :

- représentation par un unique histogramme spatio-temporel construit par accumulation des pixels de toutes les trames de l'occurrence vidéo,
- représentation par un ensemble d'histogrammes spatio-temporels construit en échantillonnant les trames à l'aide d'une fenêtre glissante/sautante,
- représentation par un ensemble d'histogrammes spatio-temporels construits pour chaque canal de couleur.

4.3.1 Accumulation des pixels

La construction par cumul des trames consiste à accumuler l'information issue de chaque trame de la vidéo (cf. Figure 4.3).

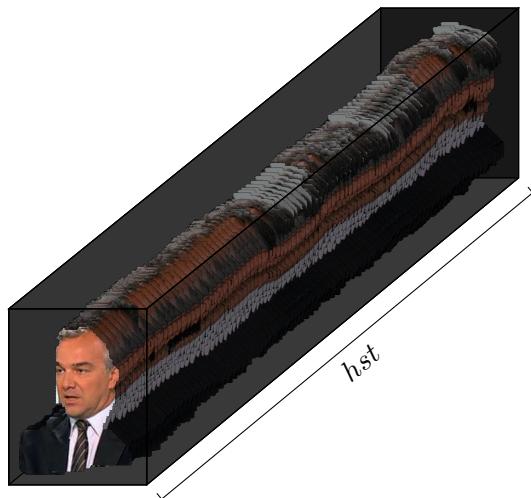


FIGURE 4.3 – Schéma illustrant la construction d'un histogramme spatio-temporel par accumulation des pixels de toutes les trames de la séquence vidéo.

Cela revient à ranger les pixels de chaque trame dans le même histogramme. L'approche par cumul a l'avantage d'être relativement robuste face aux trames bruitées (trames contenant un flash, problèmes d'encodage, etc.), selon la proportion de celles-ci. En effet, le bruit devrait théoriquement se retrouver dans des partitions spécifiques. L'avantage du cumul est que les partitions correspondant aux bruits seront marginalisées dans la description de l'occurrence. Cependant, l'inconvénient de cette approche est que lorsque deux histogrammes spatio-temporels sont construits à partir de séquences vidéo de longueurs différentes, l'étape de normalisation modifie les échelles temporelles, et biaise l'évaluation de similarité. Ce problème peut être contourné par l'utilisation d'une fenêtre glissante.

4.3.2 Fenêtre glissante

La construction par fenêtre glissante consiste à construire un ensemble d'histogrammes spatio-temporels pour décrire l'occurrence vidéo, en utilisant une fenêtre temporelle, de taille fixe (n trames), qui "glisse" le long de la vidéo selon un décalage k donné. De cette manière, un histogramme spatio-temporel est construit pour les sous-séquences correspondant à chaque position de la fenêtre. Cette approche est plus coûteuse en termes de calcul que la construction par cumul, car la plupart des trames sont utilisées plusieurs fois pour la construction. Il y a un chevauchement important dans les positions successives de la fenêtre. La taille du chevauchement entre deux positions successives de la fenêtre est de $n - k$. Néanmoins, cette approche permet de représenter des sous-segments de même

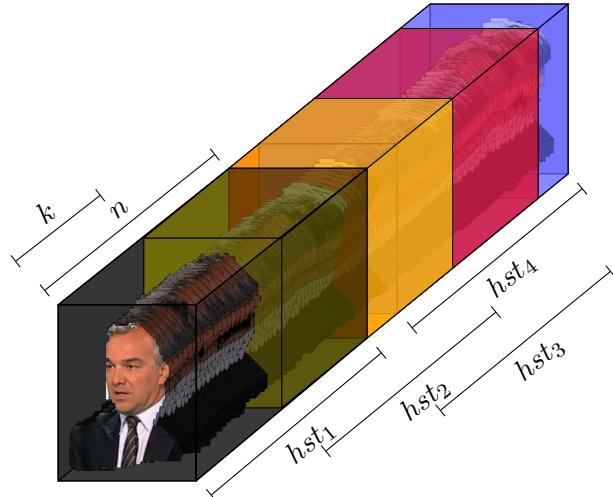


FIGURE 4.4 – Schéma illustrant la construction de plusieurs histogrammes spatio-temporels sur une fenêtre de n trames glissant le long de l'occurrence vidéo avec un décalage de k trames.

longueur, contrairement à la construction par cumul.

L'avantage de cette approche est que la taille des segments est ainsi la même pour chaque histogramme spatio-temporel, et l'évaluation de la similarité ne souffre pas du biais de la construction par cumul. Sans cette échelle de temps commune, le même phénomène étudié dans deux séquences vidéo se déroulerait à des vitesses différentes si les séquences ont des longueurs différentes.

Ainsi, utiliser une fenêtre de taille constante permet de modéliser les vidéos avec une même échelle temporelle commune. Le choix de la longueur de la fenêtre est un des deux paramètres de cette stratégie. Fixer la taille maximale de la fenêtre à la longueur de la séquence vidéo la plus courte du corpus offre l'avantage qu'aucune vidéo ne sera plus courte que la longueur de la fenêtre. Cependant, cette approche est discutable car les phénomènes temporels étudiés ont une forte probabilité d'être répartis dans plusieurs fenêtres de façon arbitraire.

4.3.3 Fenêtre sautante

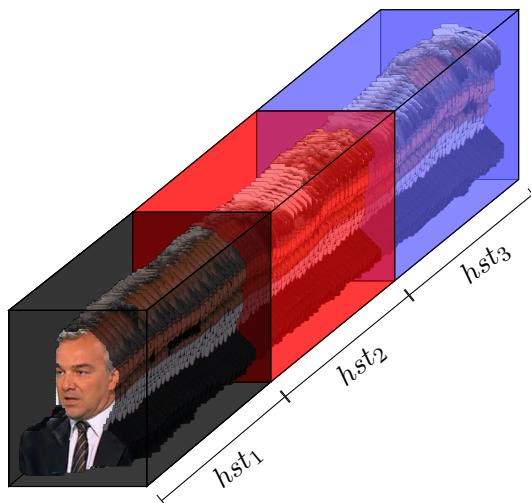


FIGURE 4.5 – Schéma illustrant la construction de plusieurs histogrammes spatio-temporels sur une fenêtre sautante, de taille n trames, le long de l'occurrence vidéo.

Nous envisageons la construction d'histogrammes spatio-temporels par fenêtre sautante comme alternative à la construction par fenêtre glissante. Il s'agit du cas particulier de la fenêtre glissante où $n = k$. Dans le cas d'une fenêtre sautante, il n'y a ainsi aucun chevauchement dans les positions successives de la fenêtre, sauf éventuellement pour une position si l'intervalle entre l'avant-dernière position de la fenêtre et la fin de l'occurrence vidéo de personne est inférieur à la taille d'une fenêtre. Dans ce cas, la dernière position de la fenêtre sera calculée pour que la fin de la dernière fenêtre coïncide avec la fin de la séquence vidéo. Ce cas de figure est illustré dans la Figure 4.5 pour hst_3 qui chevauche en partie la fenêtre utilisée pour la construction de hst_2 . Cette approche ne résout pas le fait que les phénomènes temporels risquent d'être répartis dans plusieurs fenêtres de façon arbitraire.

Enfin, le choix de la stratégie de comparaison se pose de nouveau. Nous avons pour chaque occurrence vidéo de personne un ensemble, de taille variable (selon la durée de l'occurrence vidéo de personne considérée) d'histogrammes à comparer. Pour répondre à ce problème, nous avons présenté les mesures de similarités permettant de comparer une séquence quelconque de taille n avec une autre de taille m (cf. Section 2.2.3). Par exemple, l'algorithme de Dynamique Time Warping peut convenir pour répondre à ce problème.

4.3.4 Séparation des canaux colorimétriques

La construction séparée à partir des canaux est une stratégie orthogonale à celles présentées précédemment. Il s'agit de traiter séparément les différentes dimensions de l'espace de représentation des couleurs. Par exemple, dans le cas de l'espace de couleur RVB, les trois canaux de couleurs R, V et B peuvent être utilisés séparément. Cette méthode se retrouve, comme nous l'avons vu précédemment, souvent dans la littérature [108]. Cependant, les avantages de cette approche sont encore sujets à débat. Séparer les trois canaux pour les traiter indépendamment plutôt qu'en spectre de couleur revient à supposer que les différents canaux sont porteurs d'informations différentes. Or, l'espace de représentation RVB est très fortement corrélé, comme indiqué à la Section 2.2.4, les canaux sont donc porteurs d'informations liées. Il est possible d'utiliser une des stratégies précédentes et de l'appliquer sur chaque canal séparément. Soulignons que cela multiplie

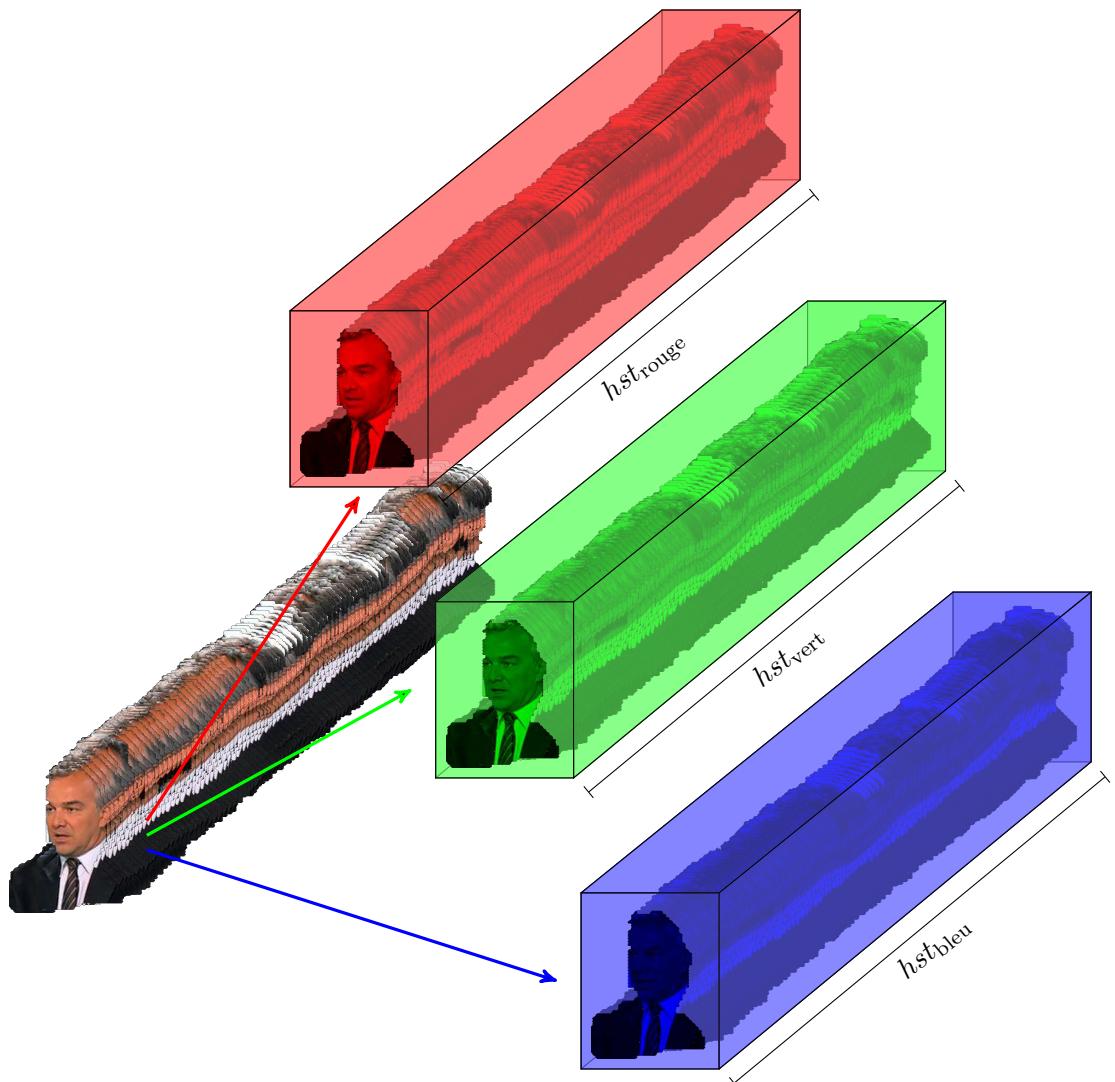


FIGURE 4.6 – Schéma illustrant la construction d'histogrammes spatio-temporels pour chaque canal et par accumulation des pixels sur les trames de la vidéo

le coût calculatoire par le nombre de canaux considérés.

4.4 Mesure de similarité

Nous définissons maintenant une mesure de similarité destinée à comparer les histogrammes spatio-temporels. Cette mesure permet de comparer les occurrences vidéo de personnes d'une même vidéo V_i . Nous nous inspirons pour cela de la métrique utilisée par les spatiogrammes, en y incluant la dimension temporelle [103]. Plusieurs mesures de similarité entre les histogrammes spatio-temporels ont été envisagées et testées. Comme nous l'avons vu dans la Section 4.1, les histogrammes spatio-temporels combinent des données de comptage de l'observation des couleurs avec des mesures sur la distribution spatio-temporelle de ces observations. Ainsi, les mesures de similarité que nous proposons intègrent les similarités entre les données de comptage et entre les caractéristiques des distributions.

Afin de vérifier si deux histogrammes sont issus d'une même distribution statistique, il est possible d'utiliser la distance de Mahalanobis (cf. Section 2.2.2). Pour mémoire, cette distance mesure la probabilité que la distribution dans l'espace et dans le temps de deux histogrammes spatio-temporels suivent la même distribution gaussienne. Dans notre cas, il s'agit d'une distance ψ_b inspirée de la distance de Mahalanobis, mesurant la similarité de la partition d'indice b dans les deux histogrammes spatio-temporels.

$$\psi_b = 1 - \sqrt{(\mu_b - \mu'_b)^t \hat{\Sigma}_b^{-1} (\mu_b - \mu'_b)} \quad (4.4)$$

où la matrice de covariance est estimée de la manière suivante :

$$\hat{\Sigma}_b^{-1} = (\Sigma_b^{-1} + (\Sigma'_b)^{-1}) \quad (4.5)$$

Soulignons que la distance de Mahalanobis sert ici à évaluer une similarité. Pour cette raison, nous utilisons son complément à 1.

La distance du χ^2 a la propriété de fournir une mesure de la dissimilarité entre deux partitions qui est proportionnelle à la taille de celles-ci. Cette propriété est intéressante en cela qu'une partition de taille importante, avec une différence relativement faible d'un histogramme spatio-temporel à l'autre, influera peu sur la mesure. Nous définissons la distance χ_b^2 entre deux partitions (notez l'absence de somme), portant le même indice b , selon la distance χ^2 :

$$\chi_b^2(n_b, n'_b) = 1 - \frac{(n_b - n'_b)^2}{n_b + n'_b} \quad (4.6)$$

La distance du χ^2 est aussi utilisée sous forme de similarité. Les distances du χ^2 et de Mahalanobis permettent de tenir compte d'aspects différents des histogrammes. Nous proposons ainsi une combinaison multiplicatrice de la mesure donnée par la similarité du χ^2 et la similarité de Mahalanobis. Une similarité est utile pour réaliser le clustering hiérarchique d'histogrammes spatio-temporels. Ainsi, la mesure de similarité entre deux histogrammes spatio-temporels hst_o et $hst_{o'}$ de même taille B est définie de la manière suivante :

$$s(hst_o, hst_{o'}) = \sum_{b=1}^B \psi_b \times \chi_b^2(n_b, n'_b) \quad (4.7)$$

Pour mémoire, ces distances, mesurées entre deux histogrammes normalisés, donnent des valeurs dans l'intervalle $[0, 1]$. Leur produit est donc aussi dans l'intervalle $[0, 1]$.

Cette métrique est utilisée dans notre approche pour estimer la similarité de deux histogrammes spatio-temporels représentant des occurrences vidéo de personnes, et ainsi établir une correspondance entre ces occurrences pour la ré-identification. Notre hypothèse est que la similarité $s(hst_o, hst_{o'})$ est grande (proche de 1) quand $id(o) = id(o')$, c'est-à-dire quand les deux occurrences correspondent à une même personne. Au contraire, cette similarité devrait être faible (proche de 0) quand $id(o) \neq id(o')$. Nous rappelons la nécessité de comparer des occurrences vidéo de personnes provenant d'une même vidéo V_i pour que l'apparence globale des personnes ne varie pas.

4.5 Complexité

Après avoir présenté les différentes stratégies possibles pour la construction des histogrammes spatio-temporels, nous allons maintenant nous intéresser à la complexité de chacune. La complexité des histogrammes spatio-temporels est comparable aux autres approches qui considèrent la composante temporelle des vidéos comme le cumul de spatiogrammes et le cumul d'histogrammes. Nous nous intéressons aux coûts des plusieurs aspects : la construction et la comparaison. En ce qui concerne la construction, le coût calculatoire ainsi que le coût en espace mémoire est étudié. Pour la comparaison, seul le coût calculatoire est considéré car son coût mémoire peut être déduit de celui de la construction.

4.5.1 Complexité de la construction

Pour construire un histogramme de couleurs sur une séquence vidéo, il est nécessaire de parcourir tous les pixels de chaque trame, qui composent cette séquence vidéo. Chaque pixel est comptabilisé dans la partition de l'histogramme qui correspond à sa couleur. Une fois que tous les pixels de la vidéo ont été comptés, il est nécessaire de parcourir toutes les partitions b pour normaliser les données de comptage par le nombre total de pixels.

Les spatiogrammes prennent en compte la position des pixels dans chaque trame en plus des données de comptage. Pour chaque pixel, la position moyenne \bar{x}_b et \bar{y}_b de la partition est mise à jour pour prendre en compte la position x, y de ce nouveau pixel. Le calcul des covariances doit aussi être mis à jour avec les coordonnées de ce pixel. Cela introduit de nouveaux calculs lors de l'étape de finalisation pour terminer le calcul des covariances.

Les histogrammes spatio-temporels prennent en compte, en plus des positions spatiales x, y la position temporelle t de chaque pixel (cf. Section 4.1). Il convient donc de mettre à jour, dans chaque partition, la moyenne \bar{t}_b ainsi que les 3 covariances supplémentaires associées à cette dimension. Le coût de construction par cumul des histogrammes spatio-temporels et des spatiogrammes est ainsi similaire.

Le coût calculatoire $T_{cumul}(p, f, B)$ de la construction des histogrammes spatio-temporels, des spatiogrammes et des histogrammes de couleurs peut être estimé en utilisant la formule suivante :

$$T_{cumul}(p, f, B) = O(f \times p + B) \quad (4.8)$$

où p est le nombre de pixels par trame, f le nombre de trames et B le nombre de partitions de l'histogramme considéré.

Les trames, d'une séquence vidéo peuvent être représentées indépendamment les unes des autres par plusieurs spatiogrammes ou plusieurs histogrammes de couleurs. Dans ce cas, le coût calculatoire de la construction est plus élevé car à chaque trame est associé à un descripteur. Il est donc nécessaire de réaliser l'étape de finalisation des calculs (terminer le calcul des covariances et normaliser les partitions). La formule estimant le coût calculatoire $T_{ind}(p, f, B)$ de cette construction est :

$$T_{ind}(p, f, B) = O(f \times (p + B)) \quad (4.9)$$

Ce coût calculatoire ne tient pas compte du temps nécessaire à l'allocation d'un descripteur de B partitions. Il faut pourtant instancier autant de descripteurs qu'il y a de trames dans la vidéo. Dès lors que l'on considère des séquences vidéo de plusieurs centaines de secondes, le coût de l'instanciation devient non-négligeable par rapport au coût total de la construction.

4.5.2 Coût mémoire des descripteurs

Après avoir estimé l'ordre du coût calculatoire de la construction des différents histogrammes pour les deux principales stratégies de construction, nous nous intéressons au coût mémoire.

Soit d le nombre de dimensions de la position des pixels pris en compte par un histogramme. Les histogrammes de couleurs ne prennent en compte aucune position, $d = 0$ dans ce cas. Les spatiogrammes prennent uniquement en compte les données spatiales x, y , ainsi $d = 2$ dans ce cas. Enfin, les histogrammes spatio-temporels prennent en compte les positions x, y, t en considération, $d = 3$ dans ce cas.

Cela permet d'exprimer le coût mémoire $M_{cumul}(B, d)$ pour les histogrammes spatio-temporels, le cumul de spatiogrammes et le cumul d'histogrammes selon la formule suivante :

$$M_{cumul}(B, d) = O(c \times B), c = 1 + d + \frac{d(d+1)}{2} \quad (4.10)$$

$$= O(d^2 \times B) \quad (4.11)$$

où B est le nombre de partitions et c la quantité de données conservée par le modèle. Par exemple, dans le cas des histogrammes spatio-temporels, c est égal à 10 car, sont conservés : le compte des pixels à un coût de 1, la position moyenne $\bar{x}_b, \bar{y}_b, \bar{t}_b$ à un coût de 3 et la matrice de covariance à un coût de 6. La matrice de covariance est de dimension 3×3 mais du fait de sa symétrie, seul 6 éléments ont besoin d'être mémorisés. La complexité de c et de d^2 sont de même ordre ainsi $c = d^2$ en ordre de complexité.

Dans le cas des descripteurs considérant indépendamment chaque trame, le coût mémoire $M_{ind}(B, d, f)$ est de nouveau bien plus élevé que celui basé sur le cumul de trames :

$$M_{ind}(B, d, f) = O(f \times c \times B), c = 1 + d + \frac{d(d+1)}{2} \quad (4.12)$$

Dans cette équation, f est de nouveau le nombre de trames de la vidéo. t est de façon générale bien supérieur à c et à B . Toutefois, ceci n'est pas vrai lorsqu'on considère une construction sur une fenêtre temporelle. Dans ce cas, f peut être inférieur à B et éventuellement à c .

4.5.3 Complexité des mesures de similarités

En ce qui concerne la comparaison entre deux histogrammes spatio-temporels construits par cumul, le coût calculatoire $T_{cumul}(B, d)$ est :

$$T_{cumul}(B, d) = O(B \times d^3) \quad (4.13)$$

Ce coût est similaire pour les spatiogrammes et les histogrammes.

Les descripteurs considérant chaque trame individuellement ou des fenêtres de trames ne peuvent pas utiliser d'algorithmes de comparaison similaires à ceux cumulant les pixels des trames sur l'ensemble de la vidéo du fait de la différence potentielle de la longueur des séquences. Le *dynamic time warping* permet de résoudre ce problème de comparaison *n-par-m* entre deux ensembles de modèles individuels (cf. Section 2.2.2). Cet algorithme a, en général, une complexité en $O(f^2)$, mais qui peut être améliorée de nombreuses façons [107]. Le coût total de la comparaison $T_{ind}(f, B, d)$ est ainsi :

$$T_{ind}(f, B, d) = O(f^2 \times B \times d^3) \quad (4.14)$$

En conclusion, les descripteurs conservant l'aspect temporel des vidéos ont un coût bien inférieur à ceux qui considèrent indépendamment chaque trame, autant en termes de calcul que de mémoire.

4.6 Matrices de similarités

Une fois que les histogrammes spatio-temporels ont été construits à partir des occurrences vidéo, ils peuvent être comparés pour mesurer leur similarité. Ainsi, une matrice de similarités de tous les histogrammes spatio-temporels peut être générée pour conserver ces résultats de similarité. Dans notre approche, elle est le point de départ pour réaliser le regroupement qui servira lors de la reconnaissance de personnes (voir la section suivante). De plus, cette matrice de similarités est centrale dans notre approche car elle est utilisée pour identifier, en complément de la vérité-terrain, les paramètres optimaux des histogrammes spatio-temporels.

La matrice de similarités M de taille $|O_v| \times |O_v|$ de la vidéo V_v est définie comme :

$$M = \begin{bmatrix} S_{11} & S_{12} & \dots \\ S_{21} & \ddots & \vdots \\ \vdots & \dots & \end{bmatrix} \quad (4.15)$$

où $S_{ij} = s(hst_{o_i}, hst_{o_j})$.

Lorsque la mesure de similarité est symétrique (cf. Section 4.4), la matrice M est également symétrique. Pour mémoire, toutes les mesures de similarité présentées dans ce chapitre ont pour propriété que la similarité entre un histogramme spatio-temporel et lui-même est de 1. La matrice de similarités contient ainsi la valeur 1 le long de sa diagonale principale.

4.7 Choix du nombre de partitions des histogrammes spatio-temporels

Comme nous l'avons évoqué dans la Section 4.1, les histogrammes spatio-temporels reposent sur des paramètres qui sont le nombre de partitions et l'espace de couleurs sur

lequel ils se basent. Ces paramètres ont un impact sur la qualité du regroupement qui détermine la précision du nommage. Il est donc nécessaire de choisir les meilleurs paramètres pour les histogrammes spatio-temporels. Le paramétrage est lié au genre de la vidéo. Par exemple, un paramétrage donné peut convenir pour des émissions audiovisuelles et donner de mauvais résultats pour des films. Ceci s'explique par le fait que les conditions de prises de vues sont très contrôlées pour les émissions audiovisuelles. Ainsi, l'éclairage et les prises de vues sont très encadrées. Pour un film, l'aspect artistique fait que les conditions de prise de vue sont beaucoup moins homogènes. Il est donc nécessaire de déterminer le paramétrage adapté à chaque contexte d'utilisation.

Pour déterminer le nombre de partitions optimal des histogrammes spatio-temporels pour un type de données vidéo (i.e. les émissions audiovisuelles), nous proposons une méthodologie se basant sur une vérité-terrain dédiée. Par exemple, dans le cadre d'émissions audiovisuelles, un corpus annoté nous permet de déterminer le meilleur nombre de partitions à utiliser pour ce type de données.

Pour évaluer le nombre de partitions des histogrammes spatio-temporels, nous proposons de nous baser sur la précision dans une tâche de recherche. Celle-ci consiste à trier les occurrences vidéo de personnes dans l'ordre décroissant de leur similarité pour chaque ligne de la matrice de similarité. La précision du résultat nous permet d'évaluer les performances de la mise en correspondance des histogrammes spatio-temporels et ainsi de juger de la qualité du nombre de partitions.

Pour réaliser cette tâche, nous avons besoin d'un jeu de données annotées contenant plusieurs occurrences de plusieurs personnes. Le nombre d'occurrences pour chaque personne peut être différent. Avoir plusieurs personnes présentes dans une seule occurrence des données permet d'introduire un certain bruit permettant de vérifier la robustesse du système.

Pour commencer, il faut construire un histogramme spatio-temporel à partir de chaque occurrence pour ensuite calculer la matrice de similarités décrite précédemment dans l'Équation 4.15. Celle-ci peut être vue comme un ensemble de lignes :

$$M = [M_1, \dots, M_{|M|}]^T \quad (4.16)$$

où chaque ligne M_i de la matrice M donne les mesures de similarité de l'histogramme spatio-temporel hst_{o_i} vers tous les histogrammes spatio-temporels. En se basant sur les lignes de la matrice de similarités M , nous souhaitons trier, pour chaque occurrence, l'ensemble des occurrences par ordre de similarité décroissante. Pour cela, nous définissons la matrice R contenant, dans chaque ligne, les occurrences triées selon leur valeur de similarité constatée dans M :

$$R = \{r_{ij} = o_k \mid \text{rang}(o_k, M_i) = j\} \quad (4.17)$$

où $\text{rang}(o_k, M_i)$ est le rang de la valeur de similarité S_{ik} de l'occurrence o_k dans la ligne M_i , quand on les considère selon une similarité décroissante : rang 1 pour la plus similaire, 2 pour la suivante, etc. La première valeur est la mesure de similarité entre l'histogramme spatio-temporel représentant l'occurrence et lui-même (i.e. valeur 1). Pour chaque ligne M_i , on souhaite mesurer la précision de la mise en correspondance de o_i avec les autres occurrences de la vidéo portant la même identité que o_i . Le problème de la précision classique est que le nombre d'occurrences correspondant à chaque identité varient selon celle-ci. Pour résoudre ce problème, il faut s'intéresser à la précision à n (*precision at n*)

n , parfois abrégée $P@N$) [90], qui correspond à la précision atteinte en considérant les n premiers résultats où n est le nombre de réponses attendues. Dans notre cas, pour l'occurrence o_i , n_i est le nombre d'occurrences vidéo de V_v portant la même identité que o_i :

$$n_i = |id^{-1}(id(o_i))|, \forall o_i \in O_v \quad (4.18)$$

La précision à n_i pour o_i est donné par P_i :

$$P_i = \frac{|\{r_{ij} \in R | j \leq n_i\} \cap id^{-1}(id(o_i))|}{n_i} \quad (4.19)$$

Ensuite, nous calculons la moyenne pondérée des précisions P_i :

$$\bar{P} = \frac{\sum_{i=1}^{|M|} P_i n_i}{\sum_{i=1}^{|M|} n_i} \quad (4.20)$$

Pondérer la moyenne par le nombre d'occurrences de chaque identité permet d'éviter d'introduire un biais, dans la mesure où des identités ayant un petit nombre d'occurrences contribuerait de façon trop importante à la moyenne.

Cette précision moyenne sert de mesure d'efficacité pour des histogrammes spatio-temporels. Pour cela on souhaite obtenir une précision élevée et un nombre de partitions réduit pour réduire les temps de calculs. Afin d'évaluer le paramétrage des histogrammes spatio-temporels, le rapport entre la précision moyenne et le nombre de partitions est calculé. Plus le nombre de partitions augmente, plus ce rapport devrait diminuer rapidement. Cela est dû au fait que la précision moyenne augmente de plus en plus faiblement et surtout moins rapidement que le nombre de partitions

4.8 Clustering d'histogrammes spatio-temporels

Après la construction des histogrammes spatio-temporels à partir des occurrences vidéo de personnes et leur mise en correspondance dans une matrice de similarité, nous réalisons un clustering d'histogrammes spatio-temporels basé sur cette matrice. Ce clustering sert à regrouper les occurrences vidéo par identité, dans le but d'assigner une identité à chaque groupe, sous l'hypothèse que toutes les occurrences d'un groupe partagent la même identité. Comme nous l'avons expliqué dans la Section 2.1, les algorithmes de reconnaissance de personnes ne sont capables de nommer une personne que si l'image utilisée respecte de nombreuses contraintes (visage face à la caméra, pas d'occultation du visage, expression neutre, etc.). Relativement peu d'occurrences sont composées d'images réunissant toutes ces contraintes, et peu d'occurrences peuvent ainsi être nommées par les algorithmes de reconnaissance. En regroupant les occurrences, il est possible de propager l'identité trouvée d'une ou plusieurs occurrences à l'ensemble du groupe, c'est-à-dire donner l'identité à toutes les occurrences vidéo de personne qui composent le groupe et d'étiqueter le groupe avec cette identité. La propagation est effectuée sous la même hypothèse énoncée plus haut que les occurrences du groupe partagent la même identité. Cette propagation présente deux avantages :

- il devient possible d'attribuer une identité aux occurrences que l'algorithme de reconnaissance n'a pas su reconnaître.
- en propageant l'identité la plus fréquente dans le groupe, une correction peut être apportée aux occurrences éventuellement mal nommées.

Le taux de bonne reconnaissance s'en trouve ainsi amélioré.

4.9 Résumé des propositions

Dans ce chapitre, nous avons présenté un descripteur original : les histogrammes spatio-temporels, intégrant les aspects visuel, spatial, ainsi que temporel des vidéos pour construire une signature d'une occurrence vidéo d'une personne. Le fonctionnement des histogrammes spatio-temporels a été décrit. Cette présentation souligne la contribution de la composante temporelle par rapport aux composantes visuelles et spatiales dans la description d'une vidéo. Les coûts calculatoires et mémoires des histogrammes spatio-temporels ont été étudiés et comparés à ceux des histogrammes de couleurs et des spatio-grammes. Nous avons aussi proposé différentes stratégies de construction d'histogrammes spatio-temporels adaptés à plusieurs contextes, en fonction des objectifs recherchés (complexité, précision, etc.).

Nous avons proposé une mesure de similarité pour mettre en correspondance des histogrammes spatio-temporels ; nous pouvons ainsi mesurer la similarité visuelle entre plusieurs occurrences de personnes. De plus, nous avons présenté une manière d'utiliser la matrice de similarités générée à partir d'un ensemble d'histogrammes spatio-temporels pour paramétrier correctement ces derniers. Enfin, nous avons présenté la réalisation d'un regroupement d'occurrences vidéo de personnes basé sur la matrice de similarité. Nous supposons que le regroupement permet de ranger les différentes occurrences d'une même identité dans un même groupe. Les différents groupes seront utilisés dans la Partie III concernant la reconnaissance afin de les identifier et propager cette identité au groupe en utilisant des stratégies.

Maintenant que nous avons présenté de façon théorique notre proposition de ré-identification de personnes, le chapitre suivant présente une validation expérimentale des différentes parties qui la constituent.

Chapitre 5

Validation du regroupement d'occurrences vidéo de personnes

5.1 Présentation des expérimentations

Après avoir présenté de façon théorique nos propositions pour le regroupement de personnes, nous allons les valider de façon expérimentale. Pour cela, nous allons utiliser le corpus de vidéos issu du défi ANR REPERE (présenté dans le contexte de la thèse), proposant de plus de 100 vidéos d'émissions audiovisuelles, dans lesquelles les personnes ont été annotées de façon manuelle.

Dans un premier temps, nous allons vérifier que les résultats de mise en correspondance entre histogrammes spatio-temporels que nous obtenons ont du sens. Un test statistique confirme qu'il y a une différence significative dans la similarité entre des histogrammes spatio-temporels d'occurrences vidéo de même personne et d'histogrammes spatio-temporels de personnes différentes. Cela montre que les résultats ne sont pas obtenus de façon aléatoire et que notre approche permet effectivement de discriminer les occurrences vidéo de personnes.

Nous allons ensuite regarder l'évolution de la précision de notre système en fonction du paramétrage en cherchant à identifier l'espace de couleur le plus approprié, le nombre de partitions optimal, ainsi que la stratégie de construction la plus adaptée.

Une fois ces paramètres déterminés, nous identifions ces mêmes paramètres pour différentes approches de l'état de l'art comme les histogrammes de couleurs, les spatiogrammes et les histogrammes de LBP. Nous comparerons les résultats obtenus, pour une tâche de recherche, dans les différents cas avec ceux obtenus avec notre approche.

Les mesures de similarités données par les différentes approches seront ensuite utilisées pour effectuer le regroupement d'occurrences vidéo de personnes. Les différents groupes obtenus seront évalués selon de nombreux critères afin de déterminer quelle approche convient le mieux pour selon l'application considérée.

5.2 Présentation des données de test

Le corpus de données fourni pour le défi ANR REPERE consiste en plusieurs heures d'émissions télévisées annotées partiellement. Ces données viennent de deux chaînes télévisées françaises : LCP et BFMTV. Plusieurs émissions de ces chaînes sont présentes

dans le corpus, elles ont des longueurs variables et la façon dont chaque émission est filmée varie aussi. Certaines contiennent des plans filmés en extérieur.

Les données sont encodées au format vidéo MPEG avec une taille, à l'affichage (*Display Aspect Ratio*), de 720x576 pixels. En revanche, dans le cas de la chaîne LCP, les vidéos sont encodées avec une taille de 544x576 pixels (*Storage Aspect Ratio*) qui doit être redimensionnée en 720x576 pixels pour obtenir le ratio original de l'image.

Les annotations sont fournies dans des fichiers XML en utilisant le schéma de données du logiciel VIPER (*VIdeo Performance Evaluation Resource*)¹. Les annotations ne concernent pas les vidéos entières, mais uniquement un certain nombre de segments. Un segment annoté pour une personne débute sur l'apparition à l'image d'une personne et termine lors de sa disparition. Pour chaque de segment, une trame clef a été sélectionnée par l'annotateur. Cette trame est choisie aléatoirement avec pour contrainte d'éviter les trames situées à la limite de deux plans. Si cette trame clef contient le visage d'une personne annotée pour ce segment, il est détourné par un polygone, dessiné manuellement par l'annotateur. La quantité d'annotation de personnes dans chaque vidéo varie entre 30% et 90% de la longueur totale de l'émission.

En utilisant les vidéos d'origine et les annotations, nous avons extrait des occurrences vidéo de personnes dont l'identité est connue. En effet, toutes les personnes des vidéos du corpus ne sont pas annotées, c'est le cas notamment des personnes au sein d'une foule ou du public. La plupart des personnes sont présentes dans de nombreuses occurrences vidéo réparties le long de la vidéo. Ceci permet d'établir une collections de tests conséquente qui nous servira de vérité terrain lors de nos expérimentations.

Au total, le corpus est composé de 303 personnes différentes, dont l'identité est donnée par les annotations. Chaque personne apparaît en moyenne dans 15 émissions différentes. Les présentateurs apparaissent naturellement plus fréquemment que les autres personnes : ils peuvent apparaître dans plus de 50 occurrences vidéo par émission alors que certaines personnes peuvent n'apparaître qu'une seule fois.

5.3 Prétraitements des données

Les occurrences vidéo de personnes sont extraites de 141 émissions différentes. Les annotations ont été utilisées pour vérifier que chaque occurrence vidéo de personne contienne au plus une personne. Soulignons que tous les visages présents dans un segment annoté ne sont pas annotés dans le corpus REPERE, selon des critères de tailles et de sémantique. C'est le cas des scènes avec un public, notamment dans les scènes en extérieur. Nous avons filtré manuellement les occurrences vidéo pour nous assurer de la qualité du corpus. Ceci nous permet d'éviter toute confusion entre ces personnes lors de l'évaluation. Car bien que les visages soient annotés en position sur les trames clefs, les segments annotés ne tiennent pas compte du changement de plan. Il n'y a donc aucune garantie de la correspondance des visages en dehors des trames clefs. De plus, tous les visages ne sont pas annotés, même sur les trames clefs.

Ensuite, un algorithme combinant de la détection de visages et les annotations a été utilisé pour retirer toutes les occurrences qui pourraient contenir des personnes non-annotées. Le détecteur nous permet de mettre en évidence toutes les séquences vidéo dont deux visages ou plus ont été détectés dans pour une même trame.

1. Le logiciel VIPER est disponible à l'url <http://viper-toolkit.sourceforge.net/>.

Ainsi, à la fin du processus de sélection, nous obtenons 5279 occurrences vidéo de 303 personnes différentes. Chacune est présente en moyenne dans 5 émissions. Les journalistes sont plus représentés que les invités.

5.4 Calcul des matrices de similarités

Chaque occurrence vidéo de personne de notre corpus a été utilisée pour construire des histogrammes spatio-temporels, spatiogrammes et histogrammes de couleur afin de comparer ces trois descripteurs. Ces descripteurs ont été construits en utilisant différentes combinaisons de paramètres :

- nombre de partitions différents (10, 50, 100, 150, 200, 250, 300, 350, 400, 500, 800, 1.000, 1.500, 2.000, 2.500, 5.000, 10.000 et 100.000),
- des espaces de représentation des couleurs différents (RGB, OHTA, HSV),
- des stratégies de constructions différentes (accumulation, fenêtres glissantes, fenêtres sautantes et séparation des canaux).

Les matrices de similarités ont été générées en utilisant des mesures de similarités différentes (χ^2 , Bhattacharyya, Bhattacharyya combinée à Mahalanobis et χ^2 combiné à Mahalanobis).

5.5 Paramétrage des HST

Afin de paramétrier de façon optimale les histogrammes spatio-temporels, nous avons exploré de nombreuses combinaison de paramètres. Pour les histogrammes spatio-temporels deux paramètres entre en jeu lors de la construction : le nombre de partitions de l'histogramme et l'espace de représentation des couleurs. Ces deux paramètres auront un impact sur la qualité du regroupement final. Il est important de noter que le coût calculatoire de la mesure de similarité dépend du nombre de partitions.

5.5.1 Variation du nombre de partitions

Dans cette expérimentation, nous faisons varier graduellement le nombre de partitions B des histogrammes spatio-temporels, construits sur l'espace de couleur RGB et nous observons l'impact sur la précision mesurée (cf. Équation 4.20 de la Section 4.7).

Pour mémoire, notre mesure de similarité entre histogrammes spatio-temporels consiste en une similarité de Mahalanobis pondérée par la similarité du χ^2 (cf. Équation 4.7).

Nous avons testé différentes valeurs prises dans l'intervalle [10; 100.000]. Cela nous permet d'avoir une bonne résolution du comportement de la précision lors de ses plus grandes variations. L'augmentation du nombre de partitions a été arrêtée après que la précision mesurée a commencé à diminuer. Nous avons choisi de prendre la dernière mesure loin de ce point d'inflexion afin de confirmer le comportement de la courbe.

Dans les résultats de notre expérimentation, présentés dans la Figure 5.1, nous remarquons que la précision augmente rapidement jusqu'à environ 500 partitions. Au-delà de ce seuil, la précision continue d'augmenter mais de moins en moins rapidement, jusqu'à atteindre un point d'arrêt autour de 2.000 partitions pour entamer une diminution de la précision. soulignons que la courbe ne donne la précision que dans l'intervalle des partitions [10; 10.000]², nous avons calculé la précision jusqu'à 100.000 partitions. Cette

2. La courbe devenait peu lisible en allant jusqu'à 100.000 partitions.

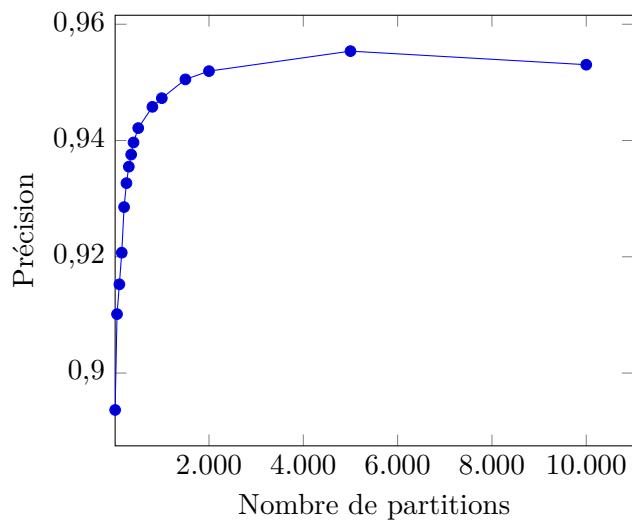


FIGURE 5.1 – Évolution de la précision en fonction du nombre de partitions des HST entre 10 et 10.000 partitions.

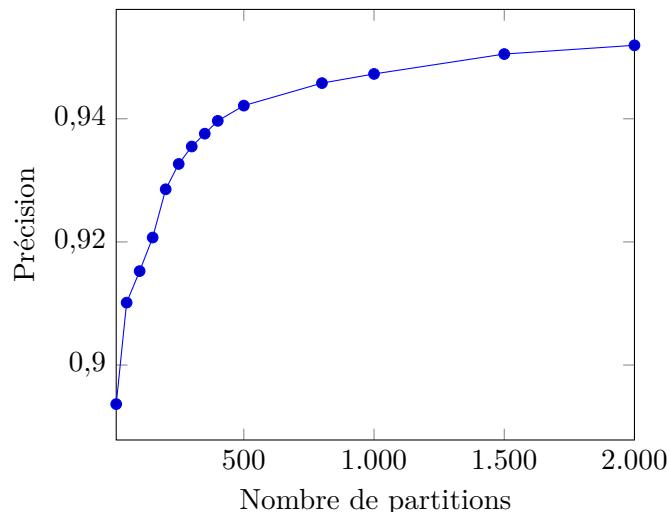


FIGURE 5.2 – Évolution de la précision en fonction du nombre de partitions des HST entre 10 et 2.000 partitions.

diminution se confirme effectivement au-delà de 10.000 partitions.

Ce seuil est particulièrement visible dans la Figure 5.2, qui montre l'évolution de la précision entre 10 et 2.000 partitions.

Une différence de 6 points de base (de 0,89 à 0,95) sur la précision peut sembler faible. Quand on rapporte cela à notre application de recherche d'occurrences vidéo de personnes, ce sont presque 320 occurrences supplémentaires qui sont correctement renvoyées par le système. Cette différence n'est pas anodine et mérite l'effort supplémentaire, en complexité, à fournir.

Ainsi, l'augmentation du nombre de partitions permet de mieux discriminer les oc-

currences vidéo de personnes jusqu'à une certaine limite. Une fois cette limite atteinte, on suppose que l'information est diluée dans plusieurs partitions et perd en cohérence. Dès lors, augmenter le nombre de partitions ne contribue qu'à faire diminuer la précision tout en augmentant le coût calculatoire des histogrammes spatio-temporels.

5.5.2 Variation de l'espace de couleurs

Nous allons maintenant étudier l'évolution de cette même précision quand l'espace de représentation des couleurs varie (le nombre de partitions restant constant). Cela va nous permettre d'étudier le comportement des histogrammes spatio-temporels sur différents espaces de couleurs. Les espaces de couleurs HSV, OHTA et RGB sont comparés.

Le nombre de partitions a été fixé à 350. Ce nombre de partitions, permet de comparer la précision sur les différents espaces de représentation avant le seuil de 2.000 partitions, au-delà duquel la précision commence à diminuer. En nous plaçant suffisamment loin de ce seuil, nous savons que la précision est plus volatile, l'impact du choix de l'espace de représentation sera ainsi mieux mis en évidence.

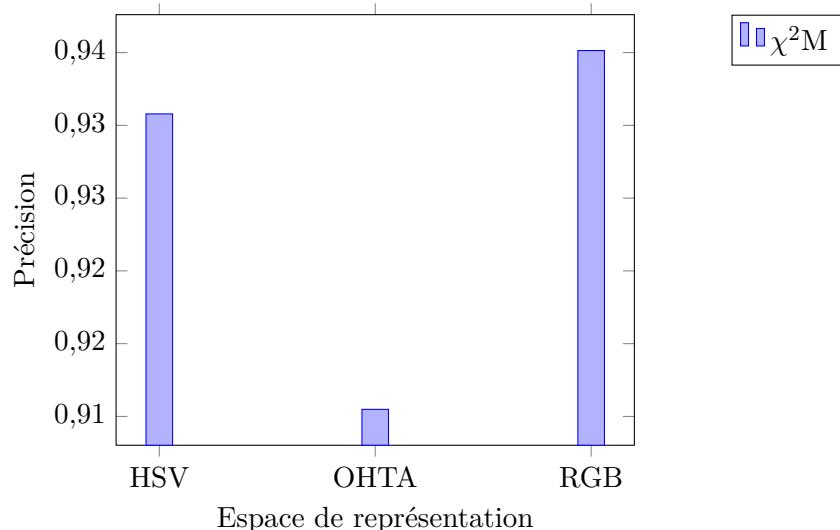


FIGURE 5.3 – Précision pour des espaces de représentation des couleurs différents et un nombre de partitions fixé à 350.

Dans la Figure 5.3, on remarque que l'espace de représentation OHTA donne les moins bons résultats. Les espaces de couleur HSV et RGB donnent des précisions proches, bien supérieures à celles obtenues OHTA. RGB obtient la meilleure précision.

Le descripteur de couleur OHTA produit des résultats inférieurs à ceux obtenus sur les espaces HSV ou RGB. Cela est probablement dû à la construction même de cet espace de couleur, ayant pour objectif de réduire au maximum la corrélation entre les différents canaux. Les canaux n'étant pas corrélés, le découpage linéaire de l'espace de couleur formé par les trois canaux est moins pertinent. Il serait plus intéressant d'exploiter les différents canaux de l'espace de couleur OHTA et de les décrire séparément. Néanmoins, cela aurait pour conséquence de tripler le coût de la construction et de la comparaison.

De façon générale, le descripteur couleur RGB donne les meilleurs résultats en termes de précision et ne présente aucun surcoût, les images étant par convention matérielle

exprimées dans cet espace de couleur. Dans le cas d'autres espaces, une conversion depuis RGB est nécessaire. Comme nous l'avons présenté dans l'état de l'art sur les espaces de couleurs (Section 2.2.4), cette conversion peut être très coûteuse à calculer car elle dépend du nombre de pixels ainsi que du nombre de trames sur lesquelles sont construits les histogrammes spatio-temporels.

5.5.3 Comparaison avec un descripteur de textures

La comparaison entre des histogrammes spatio-temporels construits sur la représentation de couleur RGB et la représentation de texture LBP a été réalisée. Comme nous l'avons présenté dans l'état de l'art sur la ré-identification de personnes, plusieurs approches exploitent des histogrammes de LBP [53, 97, 15, 84].

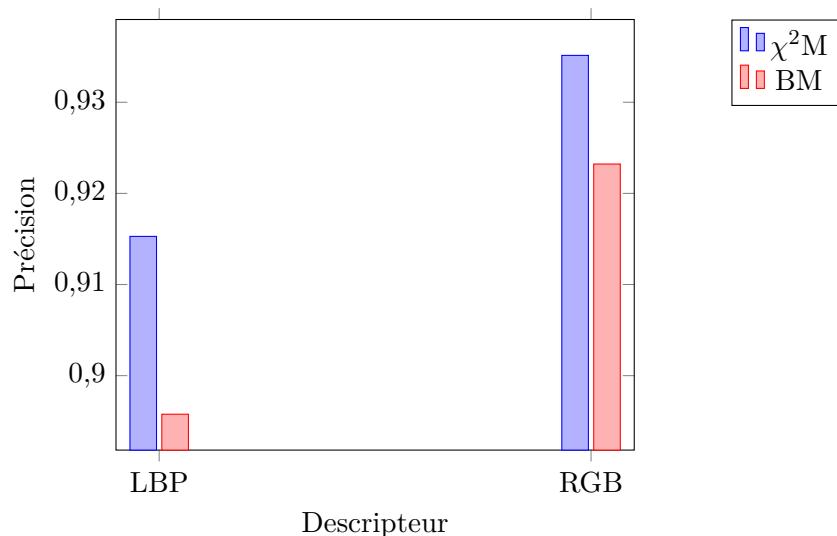


FIGURE 5.4 – Précision entre des histogrammes spatio-temporels construits sur l'espace de couleur RGB et le descripteur de texture LBP pour un nombre de partitions constant, fixé à 350, selon la mesure de similarité considérée.

La précision obtenue, présentée dans la Figure 5.4, avec l'approche exploitant la représentation de la texture par le descripteur LBP offre des résultats inférieurs à ceux obtenus par les espaces de représentation des couleurs RGB. Cela peut être dû aux informations de textures qui à l'échelle du corps complet de la personne sont moins pertinentes pour distinguer les personnes.

De plus, le descripteur LBP nécessite une étape d'extraction préalable, sur chaque trame, avant qu'un histogramme spatio-temporel puisse être construit dessus. Cette étape de calcul introduit un coût calculatoire important. Ainsi les histogrammes spatio-temporels basés sur le descripteur LBP sont non seulement moins précis pour mettre en correspondance des occurrences vidéo de personnes mais leur coût calculatoire est supérieur aux approches basées sur les espaces de couleurs.

5.5.4 Comparaison des mesures de similarités

Après avoir étudié l'évolution de la précision en fonction du nombre de partitions et de l'espace de représentation des couleurs. Nous allons maintenant comparer la précision mesurée en fonction de la mesure de similarité exploitée. Ainsi, nous allons nous intéresser aux résultats des combinaisons Bhattacharyya-Mahalanobis (BM) et χ^2 -Mahalanobis (χ^2M). En premier lieu, nous rappelons que la complexité des deux mesures est équivalente, comme nous l'avons mentionné dans la Section 4.5. Dans un premier temps nous allons faire varier l'espace de couleur utilisé, en gardant le nombre de partitions constant, fixé à 350.

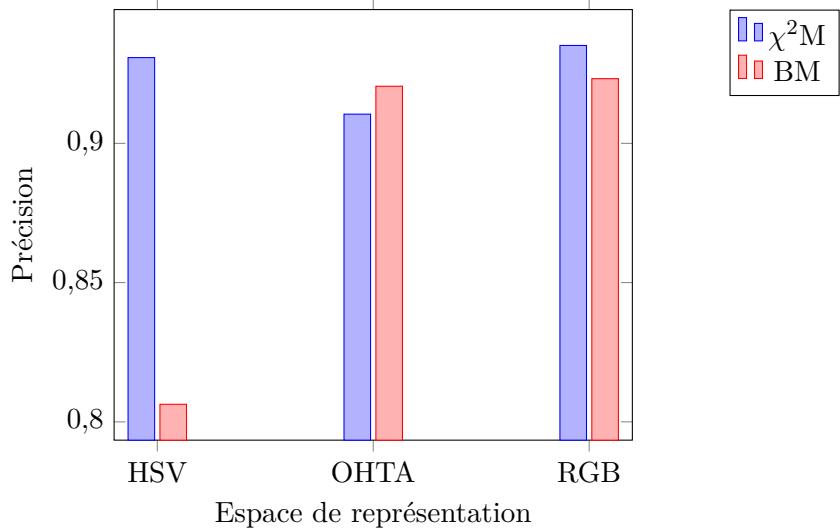


FIGURE 5.5 – Précision selon la mesure utilisée en fonction des espaces de représentation des couleurs différents et un nombre de partitions fixé à 350.

La Figure 5.5 présente la précision mesurée, en fonction de l'espace de couleur et de la mesure de similarité considérée, pour un nombre de partitions constant, fixé à 350. Nous remarquons que la combinaison Bhattacharyya-Mahalanobis produit une précision inférieure à la combinaison χ^2 -Mahalanobis dans la plupart des cas. Cela est d'autant plus flagrant dans le cas de l'espace de couleur HSV qui voit la précision mesurée se dégrader drastiquement. Il n'y a que pour l'espace de représentation OHTA que la précision mesurée augmente. Cette dernière reste pour autant inférieure à celle mesurée sur l'espace de couleur RGB. Les partitions des histogrammes spatio-temporels sont plus homogènes avec l'espace de représentation OHTA du fait de l'absence de corrélation des canaux. La mesure χ^2 pondère les différences selon la taille des partitions, si les partitions sont homogènes, cela n'apporte rien. Cela peut expliquer pourquoi la distance de Bhattacharyya produit de meilleurs résultats par rapport à la mesure du χ^2 .

Nous allons maintenant faire varier le nombre de partitions utilisées pour construire nos histogrammes spatio-temporels en utilisant l'espace de représentation RGB.

La Figure 5.6 montre que la précision de la mesure de similarité χ^2 -Mahalanobis est plus élevé que celle de la mesure Bhattacharyya-Mahalanobis jusqu'au seuil de 1.500 partitions. À partir de ce seuil, la précision de la mesure χ^2 -Mahalanobis se stabilise avant de commencer à diminuer. La précision de la mesure Bhattacharyya-Mahalanobis

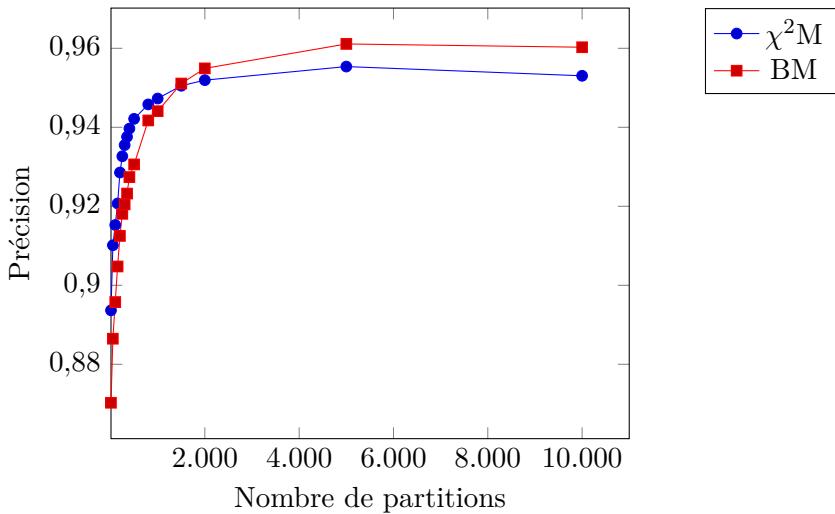


FIGURE 5.6 – Évolution de la précision en fonction du nombre de partitions des HST, entre 10 et 10.000 partitions, selon la mesure de similarité considérée.

continue de progresser jusqu'à atteindre un plafond autour de 5.000 partitions pour une précision d'environ 0,96.

Le comportement de l'évolution de la précision observée avec la mesure de similarité Bhattacharyya-Mahalanobis s'explique par le fait qu'un trop petit nombre de partitions fera diminuer la précision liée à la mesure de Bhattacharyya (cf Section 2.2.2) en surestimaant la région de recouvrement. Un trop grand nombre de partitions fera diminuer la précision liée à la mesure de Bhattacharyya en créant des partitions vides de membres. De plus, l'information spatio-temporelle, mesurée par la distance de Mahalanobis, se retrouve de diluée dans plusieurs partitions quand le nombre de celles-ci devient trop grand. De ce fait, la précision diminue à partir d'un certain seuil. La mesure du χ^2 est moins sensible à cela. La mesure de similarité basée sur la distance de Bhattacharyya ne se compare favorablement à celle basée sur le χ^2 que quand le nombre de partitions est grand (> 1.500). On observe que pour un nombre de partitions plus faible, mis en évidence par la Figure 5.7, la distance basée sur le χ^2 est supérieure à l'autre. Cette différence de précision est maximale pour 50 partitions avec 2,4 points de précision de différence. Elle est quasiment nulle à 1.500 partitions. Néanmoins, la précision maximale atteinte avec la distance de Bhattacharyya n'est qu'un peu supérieure à celle atteinte avec la distance basée sur le χ^2 .

5.5.5 Stratégie de construction

Dans la Section 4.3, nous avons proposé de plusieurs stratégies de construction pour les histogrammes spatio-temporels. Nous nous intéressons maintenant à la précision qu'il est possible d'atteindre avec chaque stratégie. Pour cela, nous comparons les résultats obtenus par les différentes stratégies de construction en utilisant l'espace de représentation RGB. Nous avons aussi fixé le nombre de partitions utilisées dans la construction à 350 pour les mêmes raisons que dans les expérimentations précédentes.

Nous comparons les précisions obtenues par la mise en correspondance des histo-

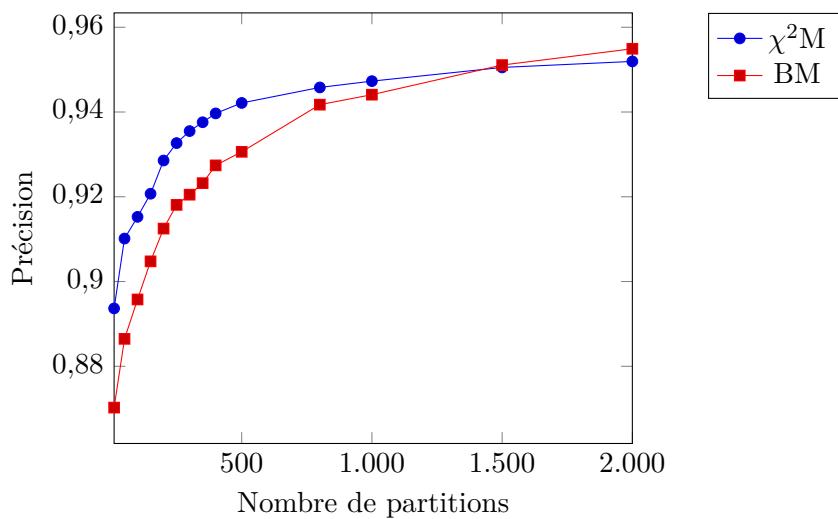


FIGURE 5.7 – Évolution de la précision en fonction du nombre de partitions des HST entre 10 et 2.000 partitions, selon la mesure de similarité considérée.

grammes spatio-temporels construits par cumul de l'information sur toutes les trames ("cm") avec celle obtenue par une fenêtre glissante ("slide") et une fenêtre sautante ("jump"). La mesure de similarité basée sur le χ^2 et la distance de Mahalanobis est utilisée pour comparer ces histogrammes spatio-temporels.

Enfin, la stratégie qui consiste à construire un histogramme spatio-temporel par canal de l'espace de couleur a été aussi comparé ("3d").

Pour les comparaisons où l'on a plus d'un histogramme spatio-temporel pour représenter une occurrence, la mesure de similarité est la même que celle proposée précédemment. Seule la mesure de similarité maximale obtenue en comparant chaque histogramme spatio-temporel d'une occurrence à tous les autres de l'autre occurrence, est conservée.

Comme le montre le diagramme de la Figure 5.8, la stratégie qui consiste à représenter toute l'occurrence vidéo d'une personne par un seul histogramme spatio-temporel est celle qui, de loin, donne les meilleurs résultats en termes de précision. La construction par fenêtre sautante donne des résultats inférieurs à ceux obtenus par la construction par fenêtre glissante. Pour rappel, le but de cette construction était de réduire le coût calculatoire de la construction par fenêtre glissante, en acceptant une perte de précision. La stratégie qui consiste à construire un histogramme spatio-temporel par canal de l'espace de couleur produit les résultats les plus faibles. Les canaux de l'espace de couleur RGB sont fortement corrélés. Le fait de séparer ces canaux dégrade l'information d'apparence. Cela explique que cette approche donne des résultats inférieurs aux autres. Il ressort de notre étude que les stratégies de construction d'histogrammes spatio-temporels par fenêtre ne permettent pas d'augmenter la précision. De plus, la construction des histogrammes spatio-temporels ainsi que leur comparaison est bien plus coûteuse à calculer (cf. Section 4.5). Bien que ces stratégies de construction puissent être intéressantes en termes de précision dans le cas de vidéos très longues, sans découpages en plans, le coût calculatoire devient rapidement prohibitif avec l'augmentation de la durée des vidéos. L'approche de construction par fenêtre est donc à éviter dans tous les cas.

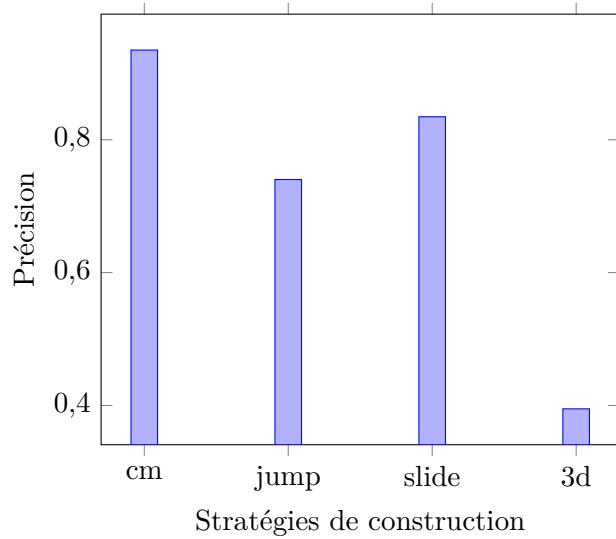


FIGURE 5.8 – Comparaison de la précision de plusieurs stratégies de construction de HST sur les OVP pour 350 partitions et l'espace de couleur RGB.

5.6 Significativité de la mesure de similarité

Afin de valider la pertinence de nos propositions et de nos résultats, nous avons réalisé un test statistique pour chacun d'entre eux. Ces tests nous permettent de vérifier que nous arrivons bien à discriminer les différentes occurrences de personnes différentes. En revanche, le test statistique ne nous permet pas de comparer les propositions.

5.6.1 Test de Student

Pour cela nous avons réalisé un test de Student basé sur les mesures de similarités obtenue sur le corpus présenté précédemment. Nous avons deux séries de mesures de similarités, une contenant les mesures de similarités obtenues entre des occurrences vidéo de personnes portant la même identité (S_1) et une autre obtenues à partir d'occurrences vidéo de personnes portant des identités différentes (S_2) tel que :

$$S_1 = \{s(hst_{o_i}, hst_{o_j}) \mid id(o_i) = id(o_j), i \neq j\} \quad (5.1)$$

$$S_2 = \{s(hst_{o_i}, hst_{o_j}) \mid id(o_i) \neq id(o_j), i \neq j\} \quad (5.2)$$

Notre test statistique permet de vérifier l'hypothèse nulle H_0 :

$$H_0 : \mu_1 = \mu_2 \quad (5.3)$$

où μ_1 est la moyenne de S_1 et μ_2 celle de S_2 .

En d'autres termes, on part de l'hypothèse qu'il n'existe pas de différence significative entre les deux séries de données. Dans notre cas, cela revient à supposer que notre approche ne permet pas de discriminer les différentes personnes de la base REPERE.

Si la p-valeur (ou p-value) du test de Student est inférieure à 0,005 alors le test a rejeté l'hypothèse H_0 et aura démontré qu'il existe une différence entre les deux séries de données et que cette différence est significative.

La p-valeur indique la probabilité que les résultats obtenus soient dûs au hasard. Ainsi une p-valeur inférieure à 0,05 indique qu'il y a moins de 5% de chance que les résultats soit obtenus de façon aléatoire. Plus la p-valeur est faible plus la différence est significative. Le seuil de 0,05 a longtemps été estimé par la communauté scientifique comme étant suffisant. Récemment, ce seuil a fait débat dans la communauté [57]. Bien qu'un seuil de 5% soit suffisant pour déterminer la significativité des résultats, ce seuil ne permet pas de garantir la reproductibilité de l'expérimentation. Ainsi, de nombreux scientifiques souhaitent abaisser le seuil de significativité d'un facteur d'au moins 10. Nous utiliserons ainsi le seuil de 0,005 (0,5%), recommandé dans [57].

Plusieurs versions du test de Student existent, dont une qui suppose que les deux échantillons comparés ont la même variance, et une autre qui suppose que leurs variances sont différentes. Après avoir testé que les variances des deux échantillons étaient bien semblables (p-value égale à $2,2e^{-16}$), nous avons appliqué le test de Student correspondant sur toutes les données expérimentales.

5.6.2 Séries de données testées

Les différentes stratégies de constructions d'histogramme spatio-temporel ont été aussi vérifiées. Afin d'éviter tout biais, les similarités obtenues quand un élément est comparé à lui-même (valeur de 1) ont été retirées des séries de données S_1 .

5.6.3 Significativité de la similarité

Il ressort que **toutes** les configurations mentionnées précédemment permettent de discriminer entre les personnes, la p-value de chaque test est à chaque fois très proche de zéro ($2,2e^{-16}$), ce qui correspond au meilleur score de significativité possible avec le test de Student implémenté dans R³. Nous avons aussi vérifié que dans chaque configuration, la moyenne des valeurs de S_1 était significativement plus grande que la moyenne des S_2 . Encore une fois, **tous** les tests effectués vérifient cette propriété avec une p-valeur très proche de zéro.

5.6.4 Significativité de l'augmentation du nombre de partitions

Nous avons aussi voulu vérifier qu'augmenter le nombre de partitions d'un histogramme pour une configuration donnée augmente la moyenne des valeurs de S_1 et diminue la moyenne des valeurs de S_2 .

Dans tous les tests la valeur moyenne de mise correspondance est significativement plus élevé. En revanche, la moyenne des valeurs de S_2 ne diminue pas de façon significative et cela pour tous les tests.

Ainsi, augmenter le nombre de partitions permet de mieux mettre en correspondance les personnes, mais ne permet pas d'améliorer le rejet. En d'autres termes, augmenter le nombre de partitions augmente la similarité d'occurrences de même identité, mais ne diminue pas la similarité entre occurrences d'identité différentes.

5.6.5 Résumé de la significativité de nos résultats

En conclusion, les différents tests statistiques effectués sur la mise en correspondance d'occurrences vidéo de personnes nous confirment que notre approche permet, de façon

3. The R Project for Statistical Computing : <http://www.r-project.org/>

très significative, de discriminer les différentes personnes. Cela est vrai quel que soit le nombre de partitions et l'espace de couleur utilisés ou le type d'histogrammes utilisés (parmi les histogrammes spatio-temporels, les spatiogrammes et les histogrammes de couleurs) pour décrire les occurrences vidéo. La p-valeur des tests statistiques réalisés garantit à la fois la significativité de notre approche, ainsi que la reproductibilité de nos résultats. En revanche, les tests statistiques ne nous permettent pas de comparer la qualité des résultats entre les différentes approches.

5.7 Qualité du regroupement

Après avoir étudié les résultats en termes de précision des histogrammes spatio-temporels selon les paramètres utilisés, nous allons maintenant nous intéresser aux résultats obtenus lors du regroupement. Pour cela, les matrices de similarités calculées lors des expérimentations précédentes vont être utilisées afin de regrouper les occurrences vidéo d'une même personne. L'objectif est d'obtenir un seul groupe pour chaque identité. Nous allons ainsi évaluer la qualité du regroupement en utilisant différents indices appropriés. Chacun permet d'évaluer un aspect particulier du regroupement.

5.7.1 Regroupement hiérarchique ascendant

L'approche utilisée pour effectuer le regroupement est un clustering hiérarchique ascendant que nous avons présenté dans la Section 4.8. Le regroupement est initialisé en mettant dans un même groupe les occurrences vidéo de personnes avec une similarité supérieure ou égale à 0,99. Pour rappel, cette valeur nous permet de mettre en correspondance deux occurrences vidéo sans faire d'erreur. Pour plus de détails à ce sujet sont présentés dans la Section 4.8.

Nous avons appliquée cette méthode de regroupement sur des histogrammes spatio-temporels construits sur différents espaces de couleurs, avec un nombre différent de partitions et des mesures de similarités différentes. Cela permet, en outre, de vérifier si une configuration avec une précision plus faible dans une tâche de recherche offre les mêmes performances lors du regroupement d'occurrences.

Nous notons Ω notre regroupement, il s'agit d'un ensemble de groupes ω contenant des occurrences vidéo de personnes o tel que :

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_{|\Omega|}\} \quad (5.4)$$

et

$$\omega_i = \{o_1^i, o_2^i, \dots, o_{|\omega_i|}^i\} \quad (5.5)$$

Dans notre regroupement, pour l'évaluation, on s'intéresse au sous-ensemble C des identités \mathbb{I} présentes dans celui-ci :

$$C = \{\iota_i \in \mathbb{I} \mid \exists o \in \mathbb{O}, id(o) = \iota_i\} \quad (5.6)$$

5.7.2 Mesure de pureté

La pureté [48] est une mesure de l'homogénéité des groupes. Autrement dit, cette mesure vérifie si les éléments au sein d'un groupe appartiennent à une même classe. Pour calculer la pureté, on commence par attribuer à chaque groupe l'étiquette de l'identité la

plus fréquente parmi ses membres. Ensuite, il s'agit d'un simple calcul de précision dont voici la formule :

$$\text{purete}(\Omega, C) = \frac{1}{\sum_{k=0}^{|\Omega|-1} |\omega_k|} \sum_{k=0}^{|\Omega|-1} \max_j |\omega_k \cap id^{-1}(\iota_j)| \quad (5.7)$$

Comme toutes les occurrences vidéo font partie du regroupement, cette définition peut être simplifié en :

$$\text{purete}(\Omega, C) = \frac{1}{|\Omega|} \sum_{k=0}^{|\Omega|-1} \max_j |\omega_k \cap id^{-1}(\iota_j)| \quad (5.8)$$

Le cas particulier où plusieurs classes auraient la même fréquence n'influence pas le résultat : on obtient le même résultat quel que soit l'étiquette sélectionnée. Par ailleurs, on peut noter qu'une pureté parfaite de 1 peut être obtenue si chaque élément est dans un cluster différent (i.e. il y a autant de clusters que d'éléments).

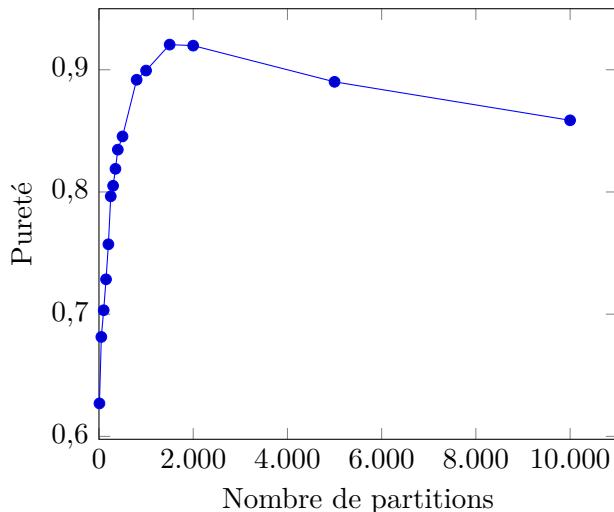


FIGURE 5.9 – Pureté du regroupement, en fonction du nombre de partitions compris entre 10 et 10.000 de l'histogramme spatio-temporel sur l'espace de couleur RGB.

On observe que la pureté de notre regroupement, pour un nombre relativement faible de partitions, Figures 5.9 et 5.10, augmente avec le nombre de partitions. Cela reste vrai jusqu'à un seuil légèrement inférieur à 2.000 partitions, à partir duquel la pureté diminue graduellement.

Ce comportement suit celui observé lors du calcul de la précision dans une tâche de recherche que nous avons présenté précédemment. En effet, nous remarquons qu'à partir d'un certain seuil (au-delà de 1.500 partitions), le trop grand nombre de partitions dilue l'information et ne permet plus, de façon aussi efficace, de discriminer entre les personnes.

5.7.3 Mesure de fragmentation

La fragmentation quantifie la dispersion de chaque identité dans différents clusters. La formule de la fragmentation est [48] :

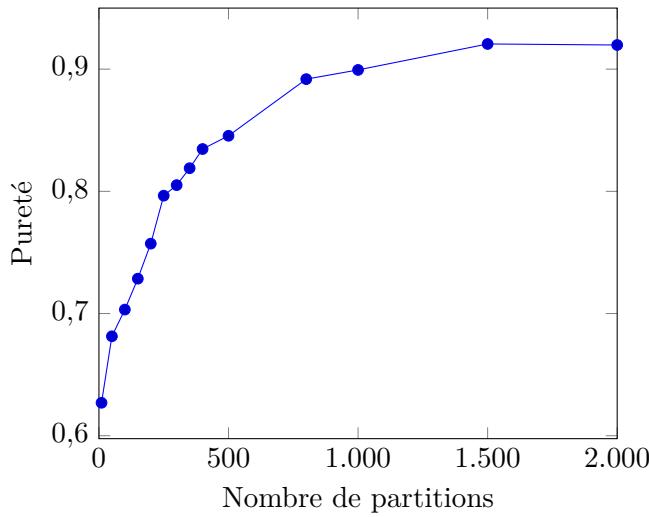


FIGURE 5.10 – Pureté du regroupement, en fonction du nombre de partitions compris entre 10 et 2.000 de l'histogramme spatio-temporel sur l'espace de couleur RGB.

$$frag(\Omega, C) = \frac{\sum_{i=0}^{|C|-1} |\{\omega \in \Omega | \exists o \in id^{-1}(\iota_i), o \in \omega\}|}{|C|} \quad (5.9)$$

La fragmentation mesure, en moyenne, dans combien de groupes apparaît chaque identité du regroupement. Ainsi, si chaque identité est représentée par un cluster dédié, la fragmentation est de 1. Il est important de noter qu'une fragmentation inférieure à 1 est obtenue si le nombre de clusters est inférieur au nombre d'identités.

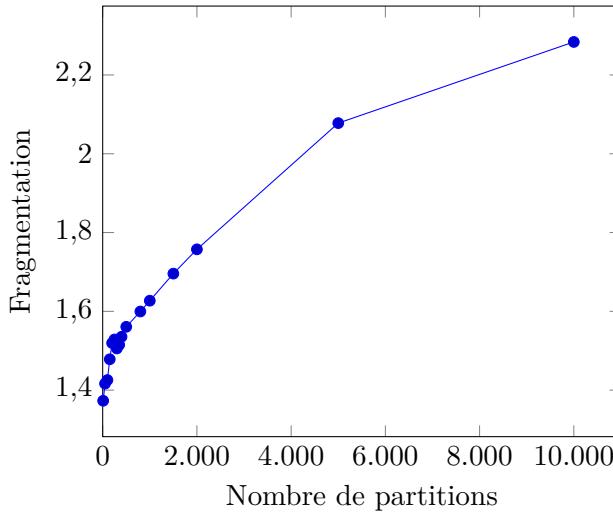


FIGURE 5.11 – Fragmentation en fonction du nombre de partitions compris entre 10 et 10.000 de l'histogramme spatio-temporel sur l'espace de couleur RGB.

On observe dans notre regroupement que l'indice de fragmentation (cf. Figures 5.11 et 5.12) diminue en augmentant le nombre de partition. Ceci indique que chaque identité

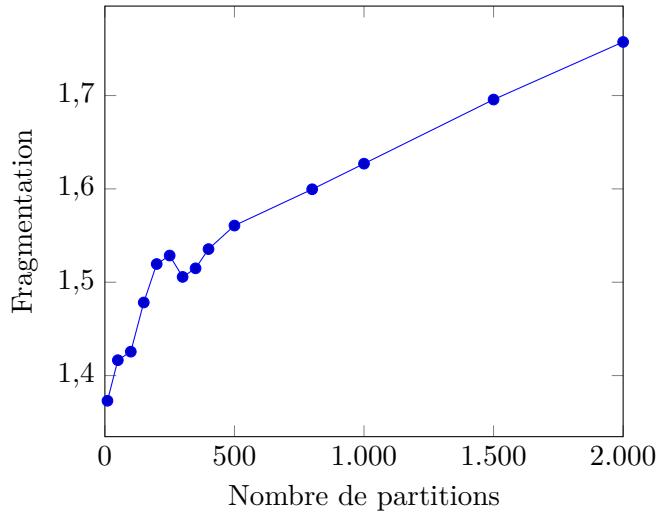


FIGURE 5.12 – Fragmentation en fonction du nombre de partitions compris entre 10 et 2.000 de l'histogramme spatio-temporel sur l'espace de couleur RGB.

est représentée par plusieurs clusters. Ainsi, de plus en plus de clusters sont créés avec l'augmentation du nombre de partitions des histogrammes spatio-temporels.

Nous interprétons cela par le fait qu'en augmentant le nombre de partitions, la mesure de similarité devient moins tolérante. Nous avons observé lors des tests statistiques que la mesure de similarité, entre les occurrences associées à des identités différentes, ne diminue pas avec l'augmentation du nombre de partitions. Cependant, la similarité des occurrences de même identité augmente. Ainsi, les éléments étant en moyenne plus similaires entre eux, de moins en moins de clusters peuvent être regroupés par le clustering hiérarchique. Cela est d'ailleurs confirmé par la pureté, vue précédemment, qui augmente avec le nombre de partitions (jusqu'à un certain seuil).

5.7.4 Vrais/faux positifs/négatifs

Pour évaluer un regroupement, il est courant de se baser sur le nombre de "vrais positifs" (TP), "vrais négatifs" (TN), "faux positifs" (FP) et "faux négatifs" (FN). Ceux-ci sont définis de la façon suivante :

$$TP = |\{(o, o') \in \mathbb{O}^2 | id(o) = id(o'), o \in \omega_i, o' \in \omega_i\}| \quad (5.10)$$

$$TN = |\{(o, o') \in \mathbb{O}^2 | id(o) \neq id(o'), o \in \omega_i, o' \in \omega_j, i \neq j\}| \quad (5.11)$$

$$FP = |\{(o, o') \in \mathbb{O}^2 | id(o) \neq id(o'), o \in \omega_i, o' \in \omega_i\}| \quad (5.12)$$

$$FN = |\{(o, o') \in \mathbb{O}^2 | id(o) = id(o'), o \in \omega_i, o' \in \omega_j, i \neq j\}| \quad (5.13)$$

$TP + TN$ peut être vu comme le nombre d'occurrences correctement regroupées et $FP + FN$ comme le nombre d'erreurs de regroupement commises.

5.7.5 Indice de Rand

L'indice de Rand [91] est une mesure de similarité entre deux partitions d'un ensemble. Il est principalement utilisé en catégorisation automatique. Son principe est, pour chaque paire d'objets, de voir si elle a été classée de la même façon (ensemble ou séparément) dans les deux partitions.

L'indice de Rand (RAND) se définit de la façon suivante :

$$\text{RAND} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.14)$$

L'indice de Rand mesure d'une certaine façon le taux de réussite du clustering. En effet, le nombre de classifications correctes ($TP + TN$) est divisé par le nombre total de classifications.

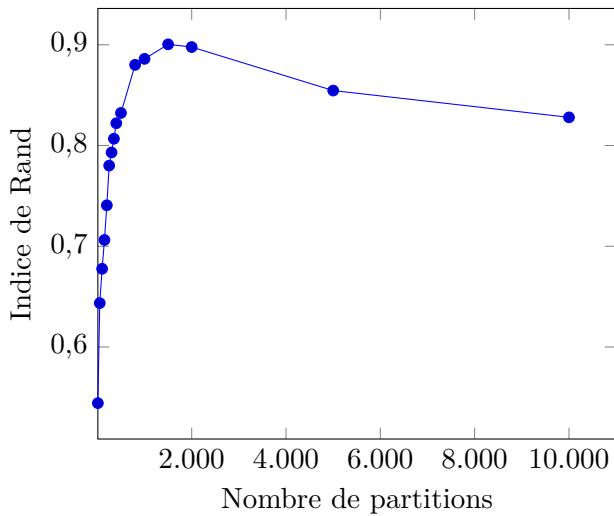


FIGURE 5.13 – Indice de Rand en fonction du nombre de partitions, compris entre 10 et 10.000, de l'histogramme spatio-temporel construit sur l'espace de couleur RGB.

On observe, dans notre regroupement, Figures 5.13 et 5.14, que l'indice de Rand progresse fortement en augmentant le nombre de partitions. Cette progression s'arrête autour de 2.000 partitions où l'indice de Rand entame une diminution progressive. Cette évolution correspond à celle observée précédemment lors du calcul de la pureté ou encore de la précision dans une tâche de recherche.

5.7.6 Rappel/Précision

Les deux mesures les plus emblématiques de l'évaluation de résultats sont probablement la précision et le rappel. La précision est le nombre de résultats pertinents retrouvés rapporté au nombre de résultats total proposés. Le rappel est défini par le nombre de résultats pertinents retrouvés au regard du nombre de résultats pertinents possibles. La précision P et le rappel R sont définis de la façon suivante :

$$P = \frac{TP}{TP + FP} \quad (5.15)$$

$$R = \frac{TP}{TP + FN} \quad (5.16)$$

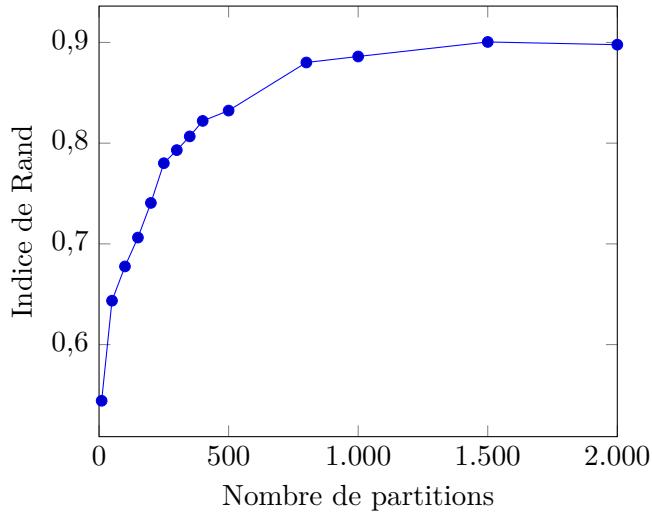


FIGURE 5.14 – Indice de Rand en fonction du nombre de partitions, compris entre 10 et 2.000, de l'histogramme spatio-temporel construit sur l'espace de couleur RGB.

5.7.7 F-mesure

La F-mesure [76] est utilisée pour équilibrer la contribution des faux négatifs à la mesure en pondérant le rappel à travers un paramètre $\beta \geq 0$. Cela nous permet de calculer la F-mesure en utilisant la formule suivante :

$$F_\beta = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R} \quad (5.17)$$

Notez que quand $\beta = 0$, $F_0 = P$. En d'autres termes, le rappel n'a aucun impact sur la F-mesure quand $\beta = 0$. Augmenter β confère un poids de plus en plus important au rappel dans la F-mesure. Dans notre regroupement, on étudie l'évolution de la F1-Mesure en fonction du nombre de partitions. Cette mesure prend en compte la précision et le rappel sans pondération particulière entre les deux indices.

On observe, dans les Figures 5.15 et 5.16, que la mesure F1 augmente avec le nombre de partitions jusqu'au seuil de 2.000 partitions où elle entame une diminution progressive. L'évolution de la mesure F1 correspond en tous points à celle observée avec les autres indices.

5.7.8 Indice Fowlkes–Mallows

L'indice de Fowlkes–Mallows [37] est une méthode d'évaluation externe qui est utilisée pour comparer la similarité entre deux regroupements. Cette mesure de similarité peut être soit entre deux clustering hiérarchique soit entre un clustering et une classification servant de vérité terrain. Une valeur élevée de l'indice Fowlkes–Mallows indique une similarité élevée entre les deux regroupements.

L'indice Fowlkes–Mallows (FM) est défini de la façon suivante :

$$FM = \sqrt{\frac{TP}{TP + FP} * \frac{TP}{TP + FN}} \quad (5.18)$$

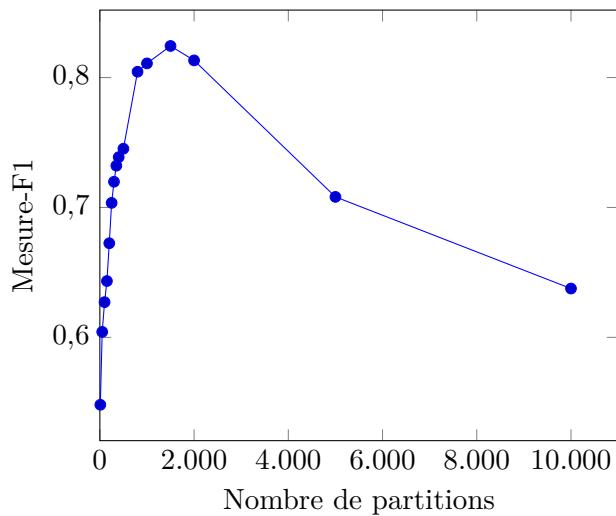


FIGURE 5.15 – Mesure F1 en fonction du nombre de partitions, compris entre 10 et 10.000, de l'histogramme spatio-temporel sur l'espace de couleur RGB.

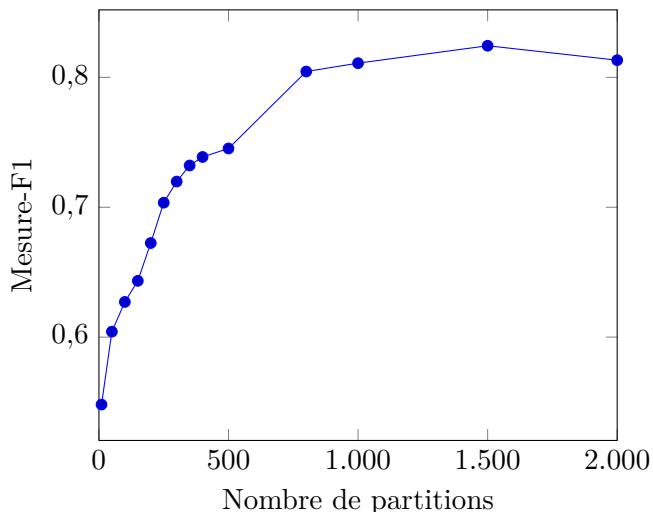


FIGURE 5.16 – Mesure F1 en fonction du nombre de partitions, compris entre 10 et 2.000, de l'histogramme spatio-temporel sur l'espace de couleur RGB.

Où TP est le nombre de vrais positifs, FP est le nombre de faux positifs et FN est le nombre de faux négatifs. Cette équation peut aussi s'écrire, plus simplement :

$$FM = \sqrt{P * R} \quad (5.19)$$

Où P est la précision et R le rappel.

L'évolution de l'indice de Fowlkes-Mallows en fonction du nombre de partitions, Figures 5.17 et 5.18 est similaire à l'évolution des autres indices où de la précision dans une tâche de recherche.

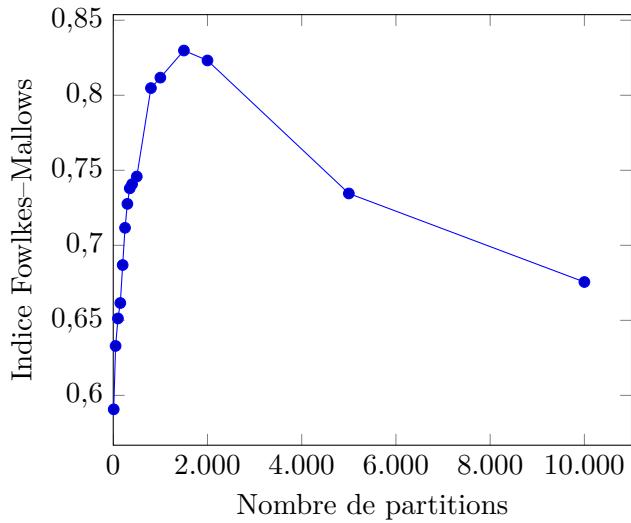


FIGURE 5.17 – Indice Fowlkes-Mallows en fonction du nombre de partitions, compris entre 10 et 10.000, de l’histogramme spatio-temporel sur l’espace de couleur RGB.

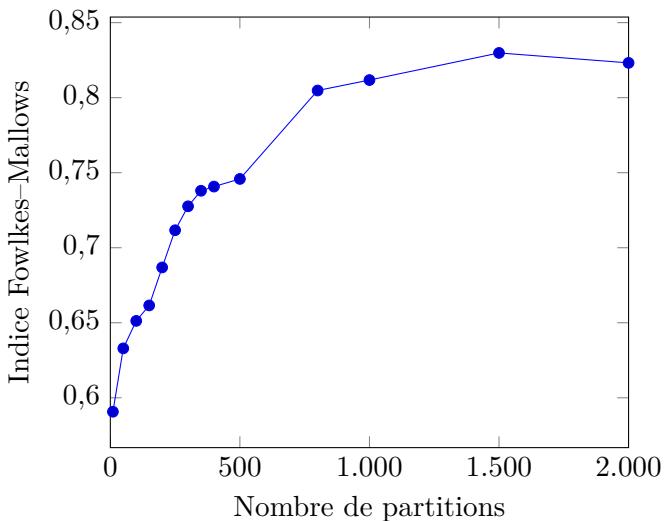


FIGURE 5.18 – Indice Fowlkes-Mallows en fonction du nombre de partitions, compris entre 10 et 2.000, de l’histogramme spatio-temporel sur l’espace de couleur RGB.

5.7.9 Résumé de la qualité du regroupement

Dans les différents indices mesurant la qualité du regroupement, nous observons une rapide progression entre 10 et 2.000 partitions avant une diminution progressive. Nous expliquons ce comportement par le fait que dans un premier temps, augmenter le nombre de partitions permet de mieux caractériser l’information spatio-temporelle. Les informations non corrélées sont rangées dans une même partition ce qui renforce le pouvoir descriptif de l’ensemble. Quand le nombre de partitions commence à être trop grand, les informations commencent à être séparées dans des partitions différentes. Les histogrammes spatio-

temporels deviennent trop discriminants et ne permettent plus de mesurer la similarité entre deux occurrences vidéo d'une même personne dans des conditions trop différentes. Ainsi, au-delà d'un certain seuil, augmenter le nombre de partitions donne des résultats inférieurs tout en augmentant la quantité de calculs à réaliser. Il est ainsi important d'identifier ce seuil dans les différentes applications où les histogrammes spatio-temporels interviennent.

Dans notre application, sur des émissions audiovisuelles, nous avons observé que ce seuil se situe à 1.500 partitions. Nous supposons que dans différentes applications ce seuil devrait varier, bien que de manière limitée car notre corpus propose des émissions très différentes. Ainsi, il peut être intéressant de chercher ce seuil, par exemple par une approche dichotomique, entre 1.000 et 2.000 partitions.

5.8 Précision et clustering

Nous avons remarqué que l'évolution de la précision et des différents indices de qualité du regroupement sont semblables. Nous avons voulu comparer la précision avec ces différents indices pour vérifier s'il était possible d'estimer les propriétés du clustering à partir du calcul de la précision à n_i sur la matrice de similarités.

Pour comparer la précision aux différents indices, nous les avons affichés l'un en fonction de l'autre et avons calculé une droite de régression linéaire.

5.8.1 Précision et pureté

Dans un premier temps nous comparons la précision à la pureté en affichant la pureté en fonction de la précision.

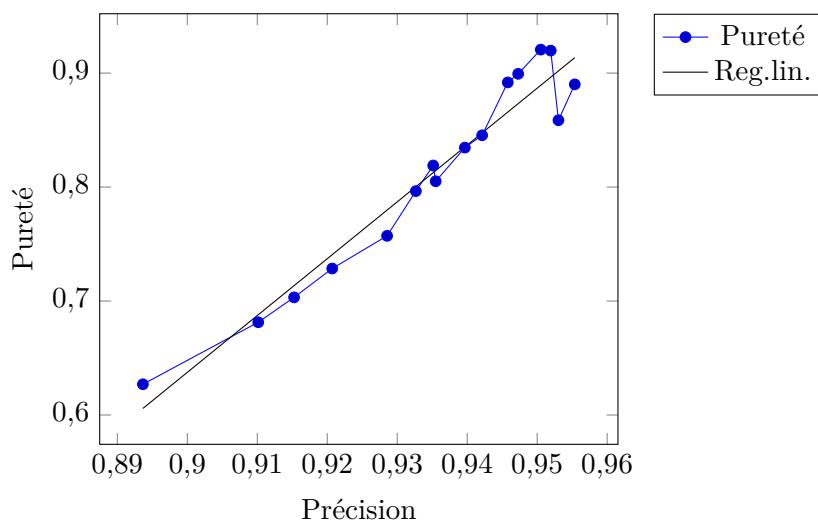


FIGURE 5.19 – Pureté du clustering en fonction de la précision mesurée.

Dans la Figure 5.19, la courbe présentant la pureté en fonction de la précision suit de près la droite de régression linéaire. On peut donc en conclure que dans notre application aux émissions audiovisuelles, la pureté du clustering peut être correctement estimée à

partir de la précision. Il est ainsi possible de fixer le paramétrage des histogrammes spatio-temporels afin de répondre à un objectif de pureté du regroupement de personnes.

5.8.2 Précision et fragmentation

D'une façon similaire, nous avons voulu comparer la précision à la fragmentation du clustering. Dans nos expérimentations, nous avons observé que la fragmentation était croissante avec l'augmentation du nombre de partitions.

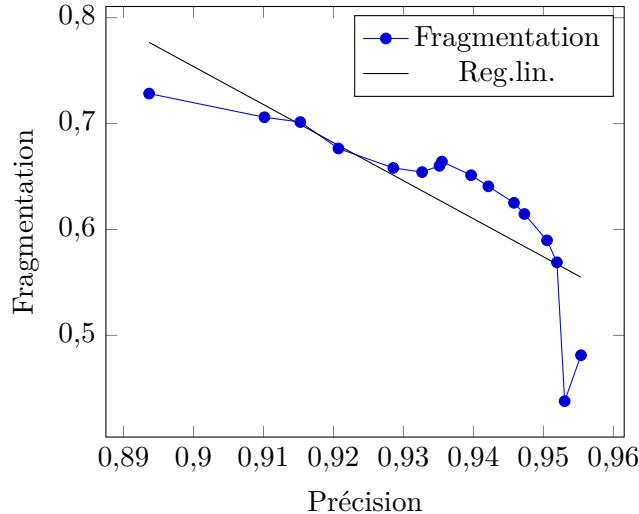


FIGURE 5.20 – Fragmentation du clustering en fonction de la précision mesurée.

Cependant, quand nous affichons la fragmentation en fonction de la précision Figure 5.20 nous observons que la courbe présentant la fragmentation en fonction de la précision est loin de former une droite. De ce fait, elle ne peut pas être estimée correctement par une droite de régression linéaire. Ainsi, la fragmentation peut difficilement être estimée à partir du taux de précision mesuré.

5.8.3 Précision et Rand

Nous avons enfin voulu voir si la précision permettait d'estimer l'indice de Rand. Pour cela, nous avons affiché l'indice de Rand en fonction de la précision et calculé et affiché la droite de régression linéaire.

Nous observons dans la Figure 5.21 que la courbe présentant l'indice de Rand en fonction de la précision suit de très près la droite de régression linéaire. L'indice de Rand peut donc être estimé de façon relativement précise à partir du taux de précision mesuré pour choisir les paramètres des histogrammes spatio-temporels.

5.8.4 Résumé de la mesure de précision pour l'évaluation du clustering

Nous avons vu que, dans la plupart des cas, les indices d'évaluation du clustering pouvaient être estimés à partir de la précision à n d'une tâche de recherche. Ce résultat est particulièrement utile à prendre en considération lors du paramétrage des histogrammes spatio-temporels. Il est ainsi facile de prédire les qualités du regroupement à partir de

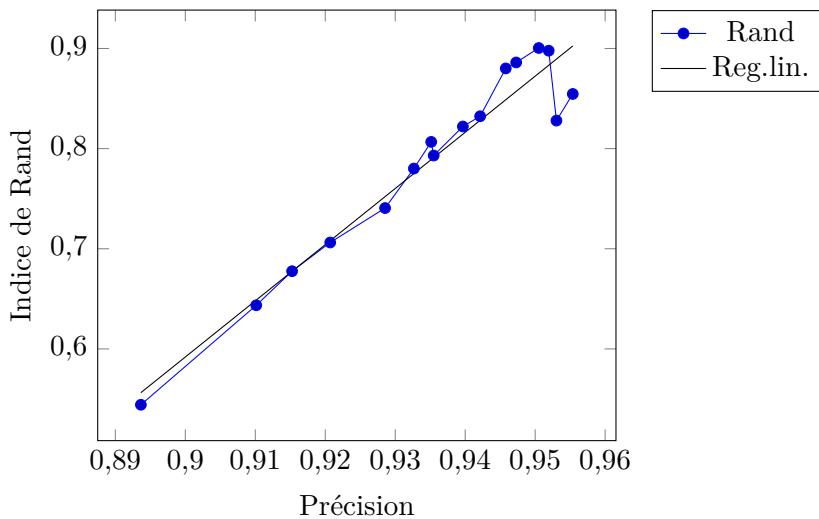


FIGURE 5.21 – Indice de Rand du clustering en fonction de la précision mesurée.

la simple mesure de précision. Il n'y a que la fragmentation qui semble difficile à prédire simplement.

Dans ces conditions, il est possible de se fixer des objectifs de qualité du clustering. Inversement, il est possible de respecter des contraintes de coût calculatoire tout en prédisant l'impact de celles-ci sur le regroupement.

5.9 Comparaison avec des méthodes existantes

Après avoir étudié le fonctionnement des histogrammes spatio-temporels nous allons les comparer avec d'autres descripteurs de l'état de l'art en termes de performances et de coût calculatoire. Pour cela, nous comparons les résultats obtenus par les histogrammes spatio-temporels à ceux obtenus par les histogrammes de couleur et les spatiogrammes qui sont les principales approches concurrentes à la nôtre.

5.9.1 Précision

Dans un premier temps, nous comparons l'évolution de la précision des différentes approches en faisant varier le nombre de partitions. Cela nous permet de vérifier si les différentes approches évoluent de la même façon. De plus, cela nous permet de comparer les résultats en termes de précision entre les différentes approches.

On remarque dans les Figures 5.22 et 5.23 que la précision évolue de façon parallèle entre les différentes approches. La précision des différentes approches progresse rapidement pour un petit nombre croissant de partitions (entre 10 et 1.000). La précision atteint un maximum pour commencer à diminuer à partir de 5.000 partitions. Notre approche basée sur les histogrammes spatio-temporels obtient une meilleure précision que les approches basées sur les histogrammes de couleur ou les spatiogrammes. Ceci confirme notre hypothèse selon laquelle l'information spatio-temporelle est importante pour distinguer les occurrences vidéo de personnes.

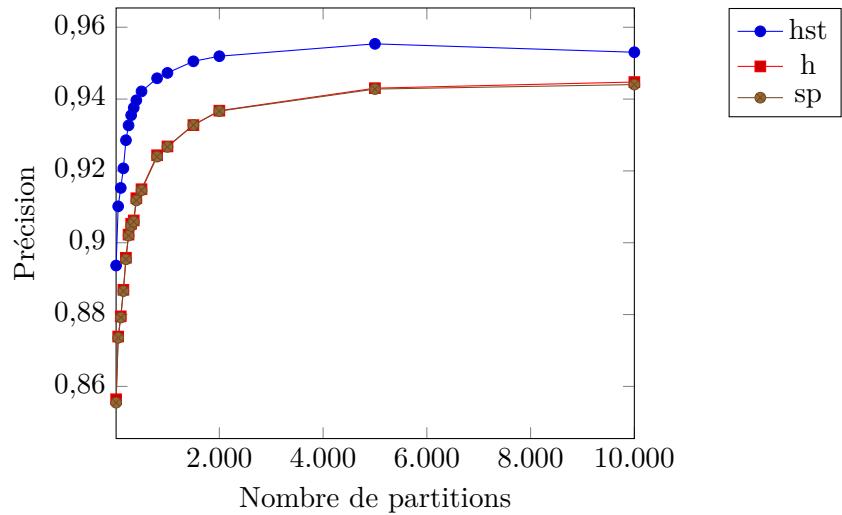


FIGURE 5.22 – Évolution de la précision en fonction du nombre de partitions du modèle entre 10 et 10.000 partitions. Les courbes correspondants à l'histogramme et au spatiogramme se chevauchent ici.

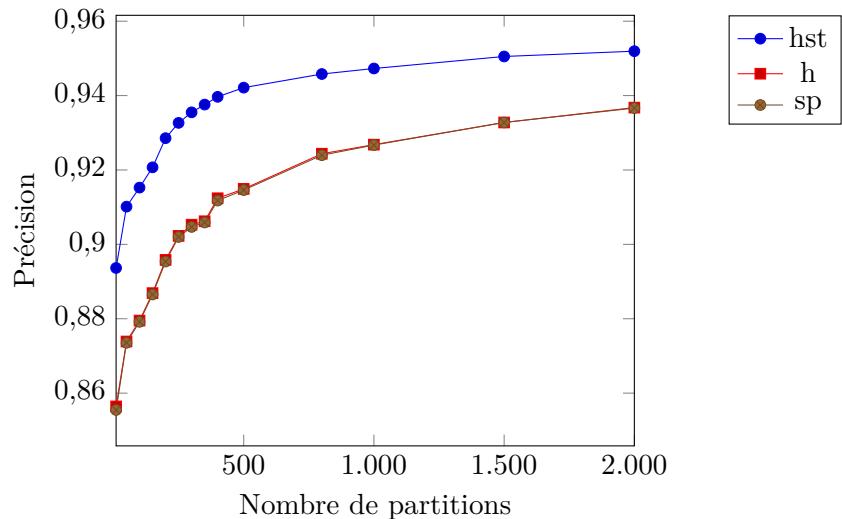


FIGURE 5.23 – Évolution de la précision en fonction du nombre de partitions du modèle entre 10 et 2.000 partitions. Les courbes correspondants à l'histogramme et au spatiogramme se chevauchent ici.

Il est intéressant d'observer que la précision des spatiogrammes est quasiment identique à celle des histogrammes de couleur. Il semble ainsi que l'information spatiale seule ne soit pas pertinente pour discriminer entre les occurrences vidéo de personnes et que la couleur seule donne de bons résultats.

5.9.2 Efficience

Après avoir comparé la précision de chaque approche, nous allons comparer leur efficience, pour cela nous avons calculé l'efficience relative de chaque approche. Celle-ci se calcule en divisant le gain en précision par le gain en nombre de partitions entre deux mesures successives. Elle compare ainsi le gain en précision en fonction du gain en complexité, représenté ici par l'augmentation du nombre de partitions. Cela permet de mesurer jusqu'à quel point il est utile d'augmenter le nombre de partitions de chaque approche.

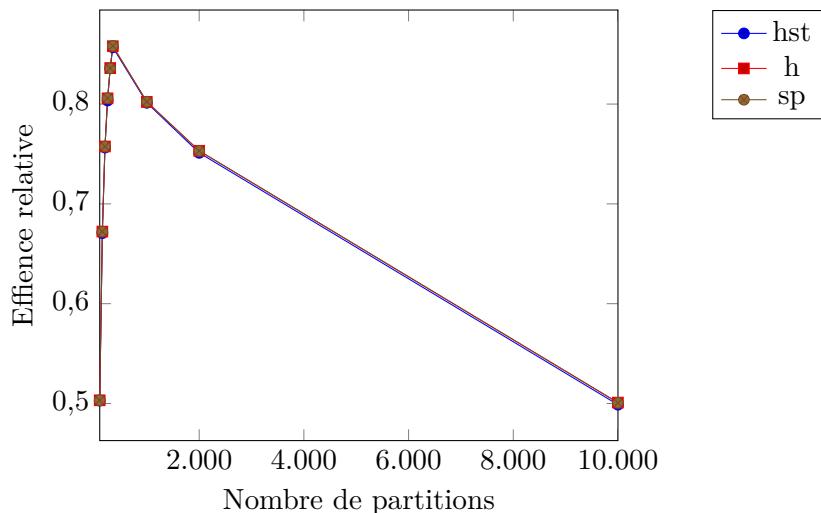


FIGURE 5.24 – Efficience relative de la précision par rapport à l'augmentation entre 10 et 10.000 du nombre de partitions utilisées pour la construction. Les trois courbes se chevauchent ici.

Dans les Figures 5.24 et 5.25, on observe que l'efficience relative des différentes approches est identique et est croissante jusqu'à un seuil de partitions de 500, à partir de ce point, l'efficience commence à diminuer de plus en plus rapidement. Cela signifie que pour un nombre de partitions supérieur à 500, la complexité augmente plus rapidement que la précision. Ainsi, à partir de 500 partitions, il est de moins en moins intéressant d'augmenter ce nombre de partitions. Cependant, dépasser ce seuil permet d'augmenter la précision mais chaque gain en précision à un coût de plus en plus élevé. Ceci est mis en évidence par le calcul de l'efficience absolue qui regarde le coût total de la précision.

Les Figures 5.26 et 5.27 montrent l'efficience absolue de la précision par rapport au nombre de partitions. On remarque de nouveau que les trois approches comparées ont une efficience absolue quasiment identique. Cette efficience absolue est rapidement décroissante entre 10 et 2.000 partitions, quelque soit l'approche considérée et continue de diminuer, mais moins rapidement à partir de ce point. L'efficience absolue est très faible à partir de 5.000 partitions et peut être considérée comme nulle à 10.000 partitions.

Cette efficience absolue nous indique qu'il n'est pas utile d'utiliser plus de 2.000 partitions, et que pour un nombre supérieur de partitions, le système perd complètement en efficience. On remarque en effet que la précision des différentes approches cesse d'augmenter à partir de ce seuil pour, au contraire, perdre en précision (voir Figure 5.22). Dans notre application aux émissions audiovisuelles, un nombre de partitions compris autour

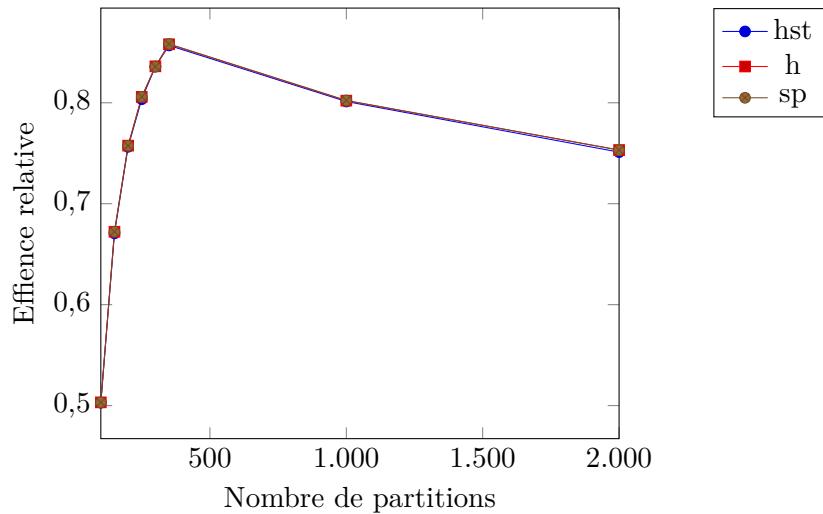


FIGURE 5.25 – Efficience relative de la précision par rapport à l'augmentation entre 100 et 2.000 du nombre de partitions utilisées pour la construction. Les trois courbes se chevauchent ici.

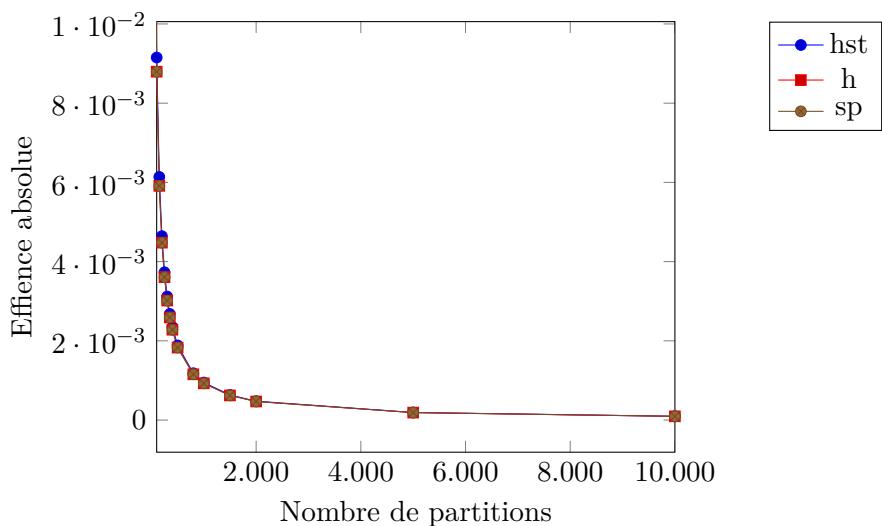


FIGURE 5.26 – Efficience absolue de la précision par rapport au nombre de partitions, compris entre 100 et 10.000, utilisées pour la construction. Les trois courbes se chevauchent ici.

de 1.500, pour les histogrammes spatio-temporels, permet d'atteindre un bon rapport précision/coût calculatoire.

5.9.3 Précision à coût mémoire constant

Afin de confirmer la pertinence de notre approche, nous avons voulu comparer les différentes approches en fonction de leur coût mémoire. Comme nous l'avons étudié dans

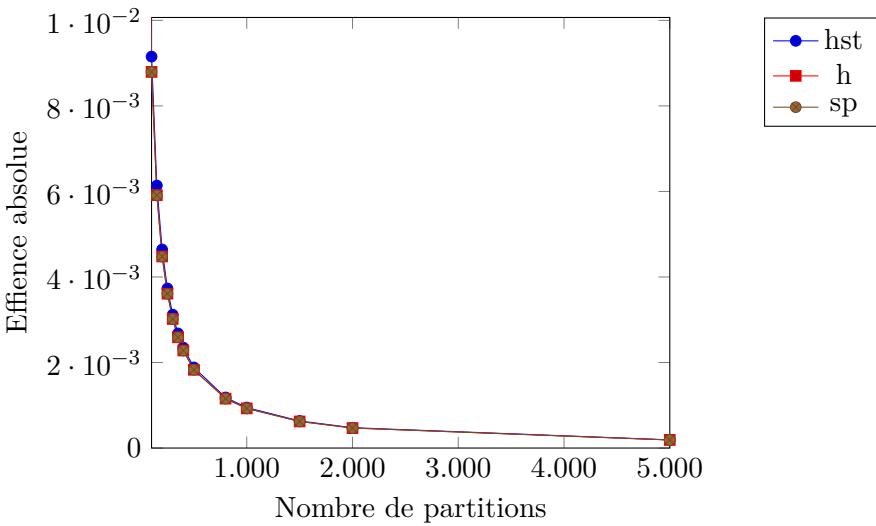


FIGURE 5.27 – Efficience absolue de la précision par rapport au nombre partitions, compris entre 100 et 5.000, utilisées pour la construction. Les trois courbes se chevauchent ici.

la Section 4.5, les histogrammes spatio-temporels, les spatiogrammes et les histogrammes de couleurs construit par cumul ont des complexités similaires pour leur construction et leur comparaison. Seul le coût mémoire occupé par ces descripteurs varie de façon significative. Ainsi, nous allons comparer la précision des différentes approches en prenant en compte cette différence de coût mémoire.

Nous utilisons comme base les histogrammes de couleurs avec un coût mémoire de 1 par partition (la donnée de comptage). Pour rappel, les spatiogrammes ont un coût mémoire de 6 par partition qui comprend les données de comptage, les positions moyennes \bar{x} , \bar{y} et les covariances $cov(x, x)$, $cov(x, y)$ et $cov(y, y)$. Les histogrammes spatio-temporels ont un coût mémoire de 9 par partition qui comprend les données des spatiogrammes en ajoutant la position moyenne \bar{t} ainsi que les covariances associées au temps $cov(x, t)$, $cov(y, t)$ et $cov(t, t)$.

Les Figures 5.28 et 5.29 montre la précision des différentes approches en fonction du coût mémoire relatif aux histogrammes de couleur. On remarque que pour un coût mémoire inférieur à 4.500, les histogrammes de couleurs donnent la meilleure précision, bien qu'ils soient progressivement rattrapés par les histogrammes spatio-temporels. Pour un coût de 4.500, les histogrammes spatio-temporels et les histogrammes de couleurs ont des précisions équivalentes. Cela signifie qu'un histogramme spatio-temporel de 500 partitions est équivalent à un histogramme de couleurs de 4.500 partitions. Au-delà d'un coût mémoire de 4.500, les histogrammes spatio-temporels ont une précision croissante avec le coût mémoire, alors que la précision des histogrammes de couleurs diminue. Ainsi, il n'existe pas d'histogrammes de couleurs pouvant atteindre la précision des histogrammes spatio-temporels quand ceux-ci possèdent plus de 500 partitions.

Les spatiogrammes sont en retrait en termes de précision en fonction du coût mémoire relatif à celui des histogrammes de couleur. Ce n'est qu'à partir d'un coût mémoire de 30.000 que la précision des spatiogrammes arrive à égaler puis à dépasser celle des histogrammes de couleurs. Ainsi, un spatiogramme de 5.000 partitions est équivalent

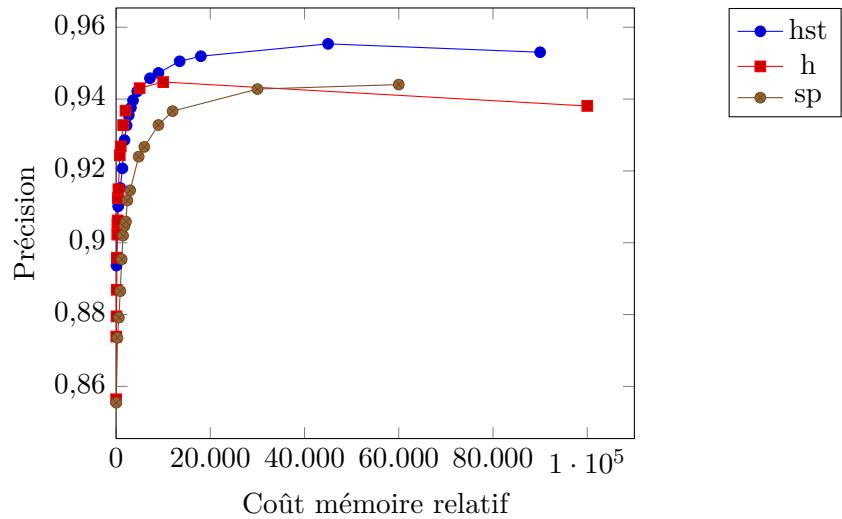


FIGURE 5.28 – Évolution de la précision en fonction du coût mémoire, relatif aux histogrammes de couleur.

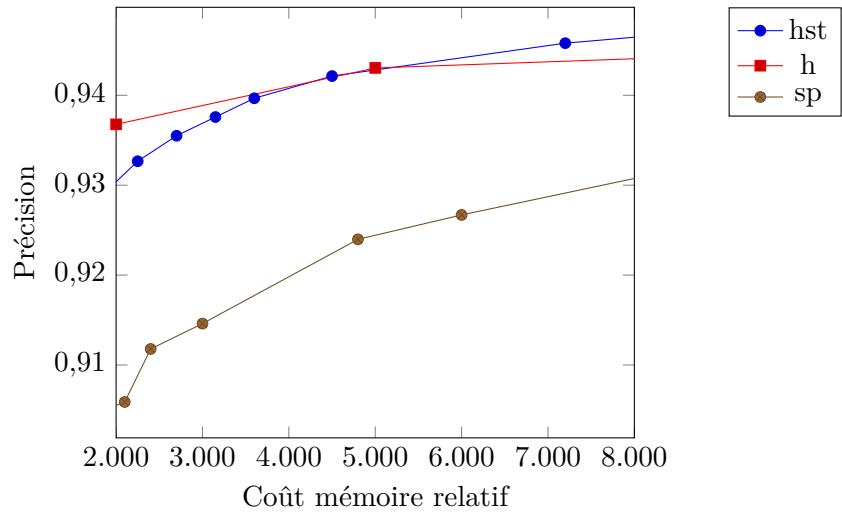


FIGURE 5.29 – Évolution de la précision en fonction du coût mémoire, relatif aux histogrammes de couleur.

à un histogramme de couleurs de 30.000 partitions. La précision des spatiogrammes ne semble pas rattraper celle des histogrammes spatio-temporels, elles semblent évoluer parallèlement l'une à l'autre.

En résumé, les histogrammes spatio-temporels surpassent, pour des coûts mémoire équivalents, les performances des histogrammes de couleurs et des spatiogrammes. Ces résultats mettent bien en évidence la contribution de la composante temporelle prise en compte dans les histogrammes spatio-temporels pour mettre en correspondance des occurrences vidéo de personnes. Cela valide donc notre hypothèse que l'aspect temporel des vidéos est porteur d'une information qui, couplée à l'information spatiale, permet

de distinguer les personnes. De plus, cela valide de façon expérimentale notre approche. Il est également important de noter que l'information spatiale seule ne permet pas de mettre en correspondance des occurrences vidéo de personnes de façon plus efficace qu'en utilisant de simples histogrammes de couleurs.

5.10 Résumé des résultats des expérimentations

Nous avons proposé une approche qui consiste à mettre en correspondance des occurrences vidéo de personnes afin de les regrouper par personnes. Dans les expérimentations, nous avons validé expérimentalement chaque étape de notre approche (présentée dans le Chapitre 4). Pour cela, nous avons appliqué notre approche à des émissions audiovisuelles réelles, issues de BFMTV et LCP.

De plus, nous avons testé, les paramètres optimaux des histogrammes spatio-temporels. Cela nous a permis de montrer que l'espace de couleur RGB permet de mieux discriminer les occurrences vidéo de personnes que les autres espaces de couleurs testés. En concordance avec nos hypothèses, il est effectivement possible de prédire les propriétés du regroupement obtenu à partir de la précision mesurée lors du paramétrage des histogrammes spatio-temporels. Il est ainsi possible de fixer des objectifs en termes de qualité du regroupement et de choisir de cette façon les paramètres. Inversement, il est possible de fixer la complexité du système et de prédire l'impact que cela aura lors du regroupement de personnes.

Enfin, nous avons comparé notre approche à d'autres approches basées sur les histogrammes de couleurs ou les spatiogrammes. Ainsi, à complexité égale, les histogrammes spatio-temporels permettent d'obtenir de meilleurs résultats que les spatiogrammes ou que les histogrammes de couleurs. Ceci confirme que la composante temps complémentaire de façon très significative la composante spatiale pour mettre en correspondance des occurrences vidéo de personnes et donc de les regrouper. Cette contribution est d'autant plus importante quand le nombre de classes est petit, avec plus de 200 points de base de précision gagnés, par rapport à d'autres approches utilisant pourtant du cumul d'information, pour les plus petits nombres de classes. De plus, les histogrammes spatio-temporels sont meilleurs en performances absolues, mais aussi relativement à leur coût.

Troisième partie

Nommage des personnes

Chapitre 6

Nommage de groupes

6.1 Introduction

Nous avons présenté et validé dans les Chapitres 4 et 5 une méthode pour regrouper les occurrences vidéo de personnes par le biais d'histogrammes spatio-temporels construits à partir de ces occurrences. Les groupes ainsi constitués sont composés d'occurrences visuellement similaires (au sens de la similarité d'histogramme spatio-temporel), sous l'hypothèse que ces groupes permettent de séparer les identités. Dans le cas idéal, les groupes et les identités sont en bijection. Cela nécessite, dans un premier temps, de pouvoir décider de l'identité d'une occurrence en se basant sur les résultats de reconnaissance individuelle pour les trames qui composent l'occurrence. Pour ce faire, différentes stratégies sont envisagées : décision d'identité basée sur une unique trame sélectionnée dans l'occurrence, ou bien sur un sous-ensemble de trames. Dans ce contexte, nous discutons de l'utilisabilité d'une trame pour la reconnaissance faciale. Dans un deuxième temps, il s'agit d'étiqueter (nommer) les groupes selon les identités des occurrences qui les composent. Pour cela, nous proposons différentes stratégies : décision basée sur une seule occurrence sélectionnée dans le groupe selon différents critères, ou bien à partir d'un sous-ensemble des occurrences qui le composent. Nous discutons du coût calculatoire de chaque approche. Enfin, nous concluons ce chapitre sur une synthèse de nos propositions.

6.2 Nommage d'une occurrence à partir de ses trames

Nous nous intéressons à la manière de décider d'une occurrence pour déterminer son identité à partir des trames qui la composent. Pour cela, plusieurs stratégies pour nommer une occurrence vidéo à partir de ses trames sont envisageables. Nous en présentons différentes en déclinant les avantages et inconvénients de chacune.

Comme présenté dans la Section 2.1, la reconnaissance de visages se base la plupart du temps sur une image fixe (reconnaissance statique). Les techniques actuelles donnent de très bons résultats dès lors que les conditions de prise de vue sont contrôlées (i.e. pose frontale, expression neutre, pas d'occultation, éclairage maîtrisé). En revanche, les performances peuvent rapidement se dégrader dans le cas contraire. Peu d'algorithme de reconnaissance exploitent réellement la vidéo (reconnaissance dynamique). Une vidéo étant composée d'une séquence d'images, il est ainsi possible d'appliquer un algorithme de reconnaissance statique sur les images qui la composent. Si on appelle \mathbb{F} l'ensemble des trames des vidéos du corpus, nous pouvons définir la fonction \hat{id}_f de reconnaissance

de visages qui associe une identité à une trame :

$$\hat{id}_f : \mathbb{F} \rightarrow \mathbb{I} \quad (6.1)$$

$$f \rightarrow \iota \quad (6.2)$$

Les algorithmes de reconnaissance sont coûteux en temps de calcul et les appliquer sur toutes les trames d'une vidéo interdirait un passage à l'échelle. Ainsi, il est nécessaire de définir une stratégie afin d'exploiter au mieux cette séquence d'images dans le cadre d'une approche dynamique de la reconnaissance de personnes dans les occurrences vidéo. Dans un premier temps, nous allons discuter de l'utilisabilité d'une trame avant de nous intéresser aux méthodes de sélections d'une trame, pour ensuite généraliser nos travaux au choix de plusieurs trames.

6.2.1 Utilisabilité d'une trame

Avant tout, il est important de noter que toutes les trames ne sont pas exploitables par les algorithmes de reconnaissance. Nous avons vu dans l'état de l'art sur la reconnaissance (cf Section 2.1) que ces différents algorithmes présentent des contraintes d'utilisation très fortes et nécessitent des conditions particulières pour produire de bons résultats. En effet, ils nécessitent que les images soient normalisées de façon à reproduire ces conditions de façon homogène pour toutes les trames de la séquence vidéo. Cette normalisation nécessite souvent de déterminer des points particuliers du visage servant de référence pour la normalisation. Les points les plus utilisés sont généralement situés sur les yeux, le nez et la bouche. La localisation de ces points d'intérêt peut être problématique dans de nombreux cas (occultations, expressions faciales, clignement des yeux, artefacts de compression, etc.), rendant l'image inexploitable pour la reconnaissance. Les expérimentations présentées par la suite, dans le Chapitre 7, montrent qu'une part importante des images de visages (environ 60%) est inexploitable pour ces raisons. Dans le cas de ces images, le résultat de l'identification est indéterminé : $\hat{id}_f(f) = \emptyset$, avec \emptyset l'identité inconnue.

Il est donc important, pour les approches qui ne considèrent qu'une seule trame de l'occurrence vidéo de personne, de choisir une trame exploitable. Dans les stratégies que nous envisageons dans les sections suivantes, nous considérons exclusivement les trames exploitables pour la reconnaissance de personnes : $\{f | \hat{id}_f(f) \neq \emptyset\}$.

6.2.2 Reconnaissance basée sur une trame unique

La première stratégie que nous considérons consiste à utiliser une unique trame pour décider de l'identité de l'occurrence vidéo de personne. Dans un premier temps, nous envisageons de sélectionner la trame située au centre de la séquence vidéo. Nous allons ensuite considérer le choix de la trame la plus représentative selon un critère de couleur moyenne, c'est-à-dire la plus proche en termes de similarité de couleur à la moyenne calculée sur l'ensemble des trames de la séquence. Nous considérons ensuite la sélection de la trame affichant une différence minimale avec ses voisines (zone de mouvement minimal de la séquence). Enfin, nous considérons le choix d'une trame dans laquelle le sujet adopte la pose frontale la plus favorable à la reconnaissance.

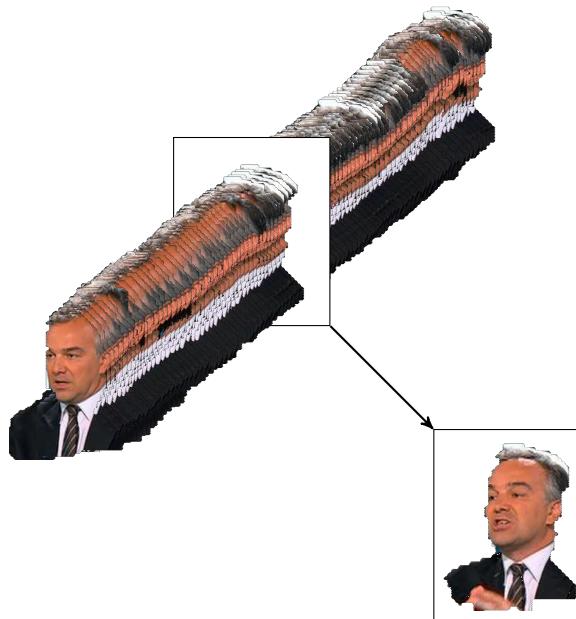


FIGURE 6.1 – Reconnaissance à partir de la trame centrale d'une occurrence vidéo de personne.

Choix de la trame centrale

Sans a priori sur la séquence de trames, le choix de n'importe quelle trame peut convenir. Cependant, dans la pratique, les premières et les dernières trames d'une séquence sont susceptibles de contenir des effets de transition, fondu enchaîné, traveling ou autre. Ainsi, l'avantage du choix de la trame centrale est qu'il s'agit de celle située le plus loin possible des extrémités de la vidéo. L'inconvénient de cette approche est que la trame située au milieu de la séquence n'offre aucune garantie d'être représentative de l'ensemble de la séquence vidéo.

Critère de couleur moyenne

Une alternative au choix de la trame centrale consiste à sélectionner la trame la plus représentative de la séquence en termes de couleur moyenne. Pour ce faire, la couleur moyenne de chaque trame est utilisée pour déterminer la couleur moyenne de la séquence. La trame retenue est la trame dont la couleur moyenne est la plus proche de la couleur moyenne de la séquence.

Trame de mouvement minimal

Une autre possibilité pour sélectionner une trame est de retenir la trame affichant le moins de différence par rapport à ses trames voisines. Cette approche permet d'éviter les flous de mouvement parfois présents à l'image, et amplifiés par la compression de la vidéo. L'algorithme du flot optique permet de déterminer la quantité de mouvements au sein de la vidéo, afin de sélectionner une trame dans la zone de mouvement minimal.

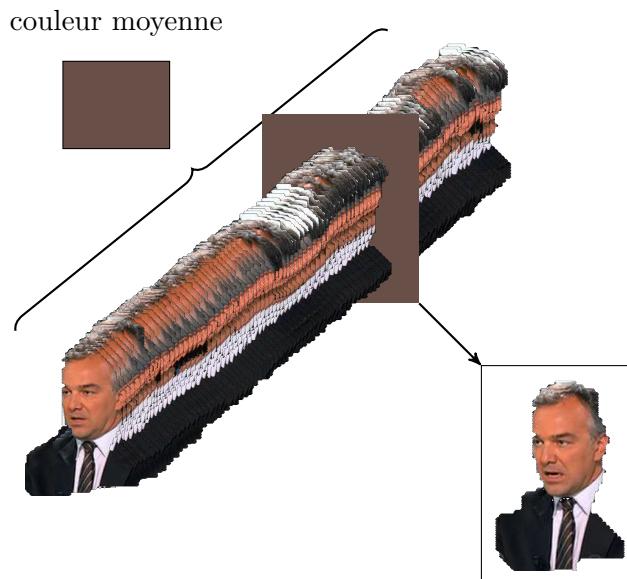


FIGURE 6.2 – Sélection de la trame la plus représentative, en termes de couleur moyenne, d'une occurrence vidéo de personne.

Pose frontale

Les algorithmes statiques de reconnaissance faciale produisent de meilleurs résultats quand les conditions de prise de vues sont contrôlées. Dans le cadre de d'émissions télévisées, il n'est pas possible de contrôler la prise de vue. En revanche, il est possible de rechercher la trame qui offre les meilleures conditions, notamment la trame affichant la pose la plus frontale (aucune rotation de la tête en roulis, lacet ou tangage). Il est nécessaire de recourir à un algorithme d'estimation de la pose de la tête, afin de déterminer la pose de la tête de la personne dans chaque trame de la séquence, et ainsi de sélectionner la trame affichant la meilleure pose.

Nous avons décrit quatre stratégies pour sélectionner une trame en vue de la reconnaissance faciale, dans le but déterminer l'identité de l'occurrence. La section suivante s'intéresse à l'utilisation de plusieurs trames afin de combiner les résultats obtenus pour les rendre plus robustes.

6.2.3 Reconnaissance basée sur plusieurs trames

Après avoir vu comment sélectionner une trame parmi toutes celles de la vidéo, on s'intéresse maintenant à la sélection de plusieurs trames. Pour cela trois problèmes se posent, le premier est de déterminer le nombre de trames à considérer. La reconnaissance étant très coûteuse en temps de calcul, il est utile de limiter son utilisation à un nombre restreint de visages. Ce problème est développé dans les expérimentations présentées dans le chapitre suivant (Chapitre 7). Le deuxième problème est la sélection de ces trames en vue de la reconnaissance. La question de la fusion des résultats obtenus sur chaque trame considérée se pose.

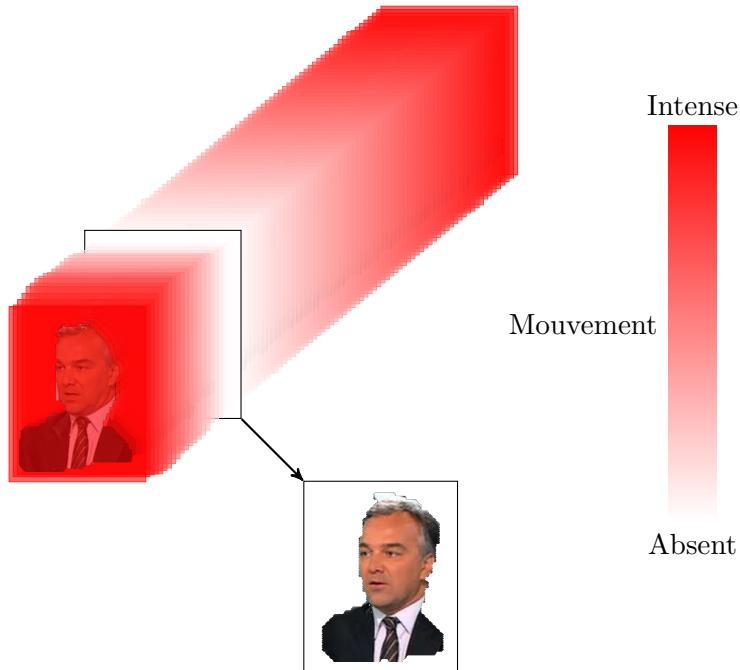


FIGURE 6.3 – Sélection de la trame située dans la zone avec un minimum de mouvement d'une occurrence vidéo de personne.

Choix des trames

Pour le choix de n trames, il semble naturel de suivre les stratégies évoquées précédemment (dans la Section 6.2.2) portant sur la sélection d'une trame. Une stratégie immédiate consiste à échantillonner (aléatoirement ou uniformément) les trames de la séquence. Alternativement, les trames peuvent être tirées selon une des stratégies évoquées dans la section précédente (distance au centre, similarité couleur, mouvement, posture), dans un ordre croissant pour ensuite en sélectionner les n premières. Une fois la reconnaissance de visages appliquée aux trames sélectionnées, le problème de la fusion des résultats se pose.

Combinaison des résultats

La fusion des résultats revient à un problème de nommage d'un ensemble à partir de ses éléments. Nous avons présenté dans la Section 2.3 la manière dont ce problème était abordé dans la littérature. Le choix d'un vote à la majorité sur les identités proposées semble ainsi indiqué à notre cas. Celui-ci répond à notre problème et propose un score de confiance associé au résultat.

L'identité résultat est l'identité la plus fréquente dans l'occurrence o , ce qui est donné par la formule :

$$\iota_o = \arg \max_{\iota \in \mathbb{I}} |\{f_k \in o | \hat{id}_f(f_k) = \iota\}| \quad (6.3)$$

Pour déterminer le score de confiance $\text{conf}(o, \iota)$ associé à l'identité ι_o , il suffit de calculer la fréquence de cette identité au sein de l'occurrence vidéo, et de calculer le

rapport entre cette fréquence et le nombre de trames votantes :

$$\text{conf}(o, \iota) = \frac{|\{f_k \in o | \hat{id}_f(f_k) = \iota\}|}{|\{f_k \in o | \hat{id}_f(f_k) \neq \emptyset\}|} \quad (6.4)$$

Ce score permet de donner un indice sur la confiance qu'on peut attribuer à l'identité proposée. Un score inférieur à 0,5 indique que l'identité proposée représente moins de la moitié des identités reconnues, tout en étant l'identité la plus fréquente. Rejeter cette identité reviendrait à réaliser un vote à la majorité absolue et à donner lui attribuer l'étiquette inconnue \emptyset . Ce vote à l'avantage de rejeter les fausses identités car on peut supposer qu'une identité, qui représente moins la moitié des trames sélectionnées n'est pas fiable.

6.2.4 Synthèse du nommage d'une occurrence à partir de ses trames

Nous avons proposé plusieurs stratégies pour assigner une identité à une occurrence à partir des trames qui la composent. Tout d'abord, nous avons mis en évidence que toutes les trames ne sont pas exploitables pour la reconnaissance. Il convient donc de choisir des trames adaptées et favorables à la reconnaissance. Afin de sélectionner ces trames, nous avons proposé différents critères : la position des trames dans la séquence, la couleur moyenne des trames, la zone de mouvement minimal et la posture de la tête du sujet de l'occurrence vidéo. Nous avons discuté du nombre de trames à sélectionner pour la reconnaissance de visage. Choisir plusieurs trames présente l'avantage de rendre plus robuste les résultats produits par une seule trame au détriment du temps de calcul. Dans le cas où plusieurs trames sont sélectionnées, nous avons présenté une méthode de fusion des résultats de reconnaissance issus de ces différentes trames, nous proposons de mettre en œuvre un vote à la majorité. L'avantage de ce dernier est qu'il fournit un score permettant de mesurer la confiance à accorder à l'identité assignée. Nous nous intéressons maintenant aux façons de propager ces résultats de reconnaissance pour assigner une étiquette aux groupes.

6.3 Nommage d'un groupe à partir de ses occurrences

Dans cette section, nous proposons des stratégies pour propager les identités des occurrences vidéo aux groupes. Ce problème s'apparente fortement au précédent, du fait qu'il s'agit de nommer un ensemble à partir de ces éléments membres. Un des objectifs de la propagation est de limiter le recours aux algorithmes de reconnaissances de personnes à certaines occurrences pour minimiser le temps de calculs.

Pour assigner une identité à un groupe d'occurrences vidéo de personnes à partir de l'identité de ses membres ($\text{id}(o)$ pour toutes les occurrences du groupe), il convient de définir une stratégie utilisant au mieux le regroupement afin de propager correctement l'identité. Ainsi, dans l'ensemble des occurrences de chaque groupe, il s'agit de sélectionner celles à utiliser pour la reconnaissance.

De manière analogue à la sélection des trames pour déterminer l'identité d'une occurrence (cf. Section 6.2), nous envisageons différentes stratégies de sélection d'occurrences représentatives du groupe. Nous distinguons ici encore l'utilisation d'une occurrence unique et l'utilisation d'occurrences multiples.

6.3.1 Sélection d'une occurrence unique

Dans cette section, nous décrivons différentes stratégies pour choisir une occurrence représentative dans un groupe en vue d'assigner une identité à ce groupe.

Centre du groupe

Une stratégie intuitive pour sélectionner une occurrence représentative d'un groupe est de choisir l'occurrence la plus centrale. En se basant sur la matrice de similarités construite pour l'algorithme de regroupement, cette stratégie consiste à sélectionner l'occurrence du groupe qui présente la similarité la plus élevée en moyenne avec les autres occurrences du groupe. Le calcul de ces moyennes est simple et réutilise les similarités générées pour le regroupement. Cette stratégie offre un compromis entre la représentativité de l'occurrence et la complexité mise en œuvre. La sélection de l'élément du groupe ayant en moyenne la plus forte similarité avec les autres éléments offre une garantie certaine de représentativité et permet d'éviter de sélectionner un *outlier* et devrait augmenter la qualité globale de l'approche. Toutefois, sélectionner plusieurs occurrences pose un nouveau problème : il est possible que l'on obtienne plusieurs identités.

Choix basé sur l'indice de confiance

Nous avons vu que la reconnaissance de l'identité d'une occurrence à partir de ces trames (cf. Section 6.2) associe un score de confiance $\text{conf}(o, \iota)$ à chaque identité. Ce score peut permettre de sélectionner les occurrences dont l'identité porte un score de confiance maximal, idéalement supérieur à 0,5 pour garantir la fiabilité du choix de l'identité.

La limitation de cette stratégie est qu'il est nécessaire d'assigner une identité à toutes les occurrences pour ne garder que celles qui offrent une confiance dans leur identité suffisante. La propagation exploitant l'indice de confiance ne répond à l'objectif de minimiser le recours aux algorithmes de reconnaissance de personnes.

Choix aléatoire

Une façon simple de procéder serait de sélectionner aléatoirement cette occurrence. Cette approche présente l'avantage d'être très simple à mettre en œuvre. Cependant, elle ne garantit pas que l'occurrence sélectionnée soit représentative de l'ensemble, il peut éventuellement s'agir d'un *outlier*¹. Cette stratégie est susceptible d'être plus efficace dans le cas où les groupes sont compacts.

Nous avons vu trois stratégies sélectionner une unique occurrence pour propager son identité à l'ensemble du groupe. La section suivante s'intéresse à la sélection de plusieurs occurrences et la fusion des résultats afin de les rendre plus robustes.

6.3.2 Choix du nombre d'occurrences

Après avoir vu comment sélectionner une occurrence parmi toutes celles du groupe, on s'intéresse maintenant à la sélection de plusieurs occurrences afin d'affiner les résultats de la propagation. Néanmoins, considérer l'ensemble des occurrences du groupe réduit l'intérêt du regroupement. Toutefois, même en considérant toutes les occurrences

1. Un *outlier* est une donnée aberrante, il s'agit d'une observation qui se trouve "loin" des autres observations [79].

d'un groupe, celui-ci apporte l'information que ces occurrences sont supposées porter la même identité. Ainsi, le groupe permet de prendre en considération chaque élément qu'il contient afin de décider, par un vote, une identité pour le groupe. Il est possible de pondérer le vote par le score de confiance attribué à chaque identité.

Le problème est de déterminer le nombre d'occurrences à considérer pour réaliser ce vote. Les contraintes que l'on cherche à respecter, notamment en ce qui concerne le coût calculatoire, doivent influencer cette décision. Précisons que pondérer le vote par un score de confiance réduit les possibilités d'avoir une égalité malgré un nombre pair de votes. Cependant, elle peut faire aboutir le vote sur une égalité malgré un nombre impair de votes – ce cas étant très peu probable.

Pour déterminer le nombre d'occurrences à considérer, nous proposons plusieurs stratégies. Premièrement, il s'agit de considérer un nombre aléatoire d'occurrences tirées de chaque groupe.

La deuxième stratégie se fonde sur la théorie des échantillons en statistique pour considérer un quartile des occurrences de chaque groupe. L'avantage de cette proportion est qu'elle permet de conserver un coût calculatoire inférieur à celui de la reconnaissance appliquée à chaque occurrence tout en prenant en compte un nombre significatif d'occurrences.

Enfin, nous proposons simplement de fixer le nombre d'occurrences à considérer pour atteindre un objectif de coût calculatoire précis. Cette approche est particulièrement utile dans le cadre d'un système avec de forte contrainte de temps.

Combinaison des résultats

La fusion des résultats revient de nouveau à un problème de nommage d'un ensemble à partir de ses éléments. Dans le cas général, on souhaite propager une seule identité pour l'ensemble d'un groupe. Hors, il est possible que plusieurs identités soient proposées par les différentes occurrences. Dès lors, il faut choisir comment combiner les réponses pour choisir la plus pertinente.

L'approche la plus simple pour résoudre ce problème est de considérer l'identité de chaque occurrence comme un vote de manière à attribuer l'identité la plus fréquente à l'ensemble du groupe. Les occurrences ont pour identités $\hat{id}(o)$, qui peuvent éventuellement être indéterminées (\emptyset). Nous nous inspirerons des différentes propositions visant à déterminer l'identité d'une occurrence vidéo de personne à partir de plusieurs trames (cf. Section 6.2).

L'identité ι_Ω d'un groupe est donnée par la formule suivante :

$$\iota_\Omega = \arg \max_{\iota \in \mathbb{I}} |\{o \in \Omega | \hat{id}(o) = \iota\}| \quad (6.5)$$

Le score de confiance $\text{conf}(\Omega, \iota)$ associé à l'identité est :

$$\text{conf}(\Omega, \iota) = \frac{|\{o \in \Omega | \hat{id}(o) = \iota\}|}{|\{o \in \Omega | \hat{id}(o) \neq \emptyset\}|} \quad (6.6)$$

Ainsi, nous proposons de mettre en œuvre un vote majoritaire pondéré par un score de confiance pour nommer un groupe à partir des occurrences qui le constituent.

6.4 Résumé sur le nommage des groupes d'occurrences

Nous avons proposé différentes stratégies portant autant sur le nommage des groupes que des occurrences vidéo de personnes qui les constituent pour identifier et propager cette identité à l'ensemble du groupe.

Nous avons identifié plusieurs critères qui permettent de sélectionner les trames d'une occurrence vidéo de personne nécessaires à son identification. De plus, nous avons proposé différents critères pour déterminer le nombre d'occurrences identifiées à considérer pour propager leur identité à l'ensemble du groupe. Dans le cas du nommage d'occurrences et du nommage de groupes, nous fusionnons les identités proposées à l'aide d'un vote majoritaire pondéré.

Le chapitre suivant décrit une validation expérimentale de ces propositions concernant le nommage des groupes d'occurrences.

Chapitre 7

Validation des approches de nommage des personnes

Dans la Partie II, nous avons proposé une méthode de regroupement des occurrences de personnes basée sur leur apparence globale. Celle-ci permet de ranger les différentes occurrences d'une même identité dans un même groupe. Dans le Chapitre 6 nous avons présenté différentes stratégies portant sur le nommage des groupes et ses occurrences vidéo de personnes qui les constituent pour identifier et propager cette identité à l'ensemble d'un groupe.

Dans ce chapitre, nous validons nos propositions expérimentalement. Dans un premier temps nous présentons les expérimentations qui vont nous servir à déterminer un taux de reconnaissance de référence. Celui-ci nous servira à évaluer les performances des approches pour déterminer l'identité des occurrences vidéo à partir de leurs trames. Après avoir assigné une identité à certaines occurrences, nous propagons cette identité à l'ensemble du groupe. Le taux de reconnaissance de référence permet d'évaluer les performances de la propagation selon le nombre d'occurrences vidéo de personnes considérées.

7.1 Expérimentations

Dans cette section nous évaluons les différentes propositions faites dans le Chapitre 6. Pour cela nous allons de nouveau utiliser le corpus du projet REPERE, qui est étiqueté en fonction du nom des personnes, comme vérité terrain pour notre évaluation. Nous présentons les données du corpus ainsi que les prétraitements des visages permettant d'obtenir une base d'identités à reconnaître, qui permet l'apprentissage d'un classifieur SVM. Nous présentons ensuite la mesure d'évaluation de nos expérimentations qui est la précision.

7.1.1 Présentation des données

Les données que nous utilisons pour entraîner notre modèle de reconnaissance faciale sont issues des visages annotés dans le corpus REPERE que nous avons présenté précédemment.

Les prédictions utilisent les données annotées manuellement, utilisées précédemment lors du regroupement d'occurrences vidéo de personnes.

Dans nos expérimentations, nous utilisons les résultats de ré-identification obtenus à l'aide des histogrammes spatio-temporels, exploitant 1.500 partitions et l'espace de couleur RGB. La mesure de similarité utilisée est celle décrite dans l'Équation 4.7, qui combine une distance de Mahalanobis avec celle du χ^2 .

Le corpus du projet REPERE contient environ 20.000 têtes annotées. Parmi celles-ci, seules 9.017 sont des visages de face sans occultation, dans lesquels les deux yeux de la personne sont détectés. Elles représentent 209 identités différentes, avec entre 9 et 595 exemples de visages par identité. Les présentateurs des émissions du corpus sont encore une fois mieux représentés que les autres personnes. Ces données nous servent pour l'apprentissage du modèle de reconnaissance faciale. Ainsi, toutes les images sont normalisées et linéarisées (cf. Section 2.1).

Les données de tests proviennent des occurrences vidéo utilisées précédemment pour évaluer le regroupement de personnes (cf. Section 5.2). Les trames composant ces occurrences vidéo de personnes subissent les mêmes traitements. On obtient ainsi 73.028 visages, issus de 2.316 occurrences de personnes, appartenant à 53 identités. En moyenne chaque identité est représentée par 1.378 visages.

7.1.2 Utilisation des données

Les vidéos doivent subir plusieurs traitements afin de pouvoir être utilisées pour l'apprentissage des personnes et pour leur reconnaissance. Dans un premier temps nous présentons la normalisation des visages. Elle permet d'obtenir des visages présentant des propriétés homogènes en termes de couleurs, de position et d'échelle. Les visages sont ensuite linéarisés de façon à pouvoir être utilisés par un classifieur SVM à noyau gaussien.

Normalisation des visages

Avant de pouvoir projeter un visage dans un SVM [25], plusieurs étapes doivent être réalisées pour normaliser le visage. Ceci est fait pour que tous les visages d'une même personne présentent des conditions homogènes facilitant ainsi leur reconnaissance. En nous inspirant de la normalisation proposée par Danisman et al. [28], nous appliquons les différentes étapes de ce processus de normalisation des visages, illustré dans la Figure 7.1. Dans les différentes occurrences vidéo, pour chaque trame, les visages sont détectés à l'aide du détecteur de Viola & Jones [106]. Les visages sont, dans un premier temps, convertis en niveaux de gris. Les yeux sont détectés grâce à un réseau de neurones artificiels. Dans le but de rendre horizontal l'axe reliant les yeux, une rotation est appliquée aux visages. Le centre de rotation est défini comme le milieu du segment reliant les deux yeux.

L'image est ensuite recadrée sur le visage en utilisant des paramètres géométriques permettant de conserver le front et de supprimer la bouche et les bords du visage (dont les oreilles) : la bouche présente trop de variabilités du fait de l'expression du sujet, ce qui perturbe la modélisation et la reconnaissance des différents sujets [116].

Enfin, les pixels de l'image, sont normalisés pour obtenir des valeurs entre 0 et 1 et l'image est ensuite linéarisée pour former un vecteur (cf. Section 2.1).

Normaliser les visages est relativement long car cela nécessite de détecter les yeux dans toutes les trames contenant un visage détecté. Ainsi pour détecter les yeux dans les centaines de milliers de visages (visages annotés et visages détectés dans les occurrences vidéo) qui composent notre corpus complet nous a pris environ deux jours complet. La

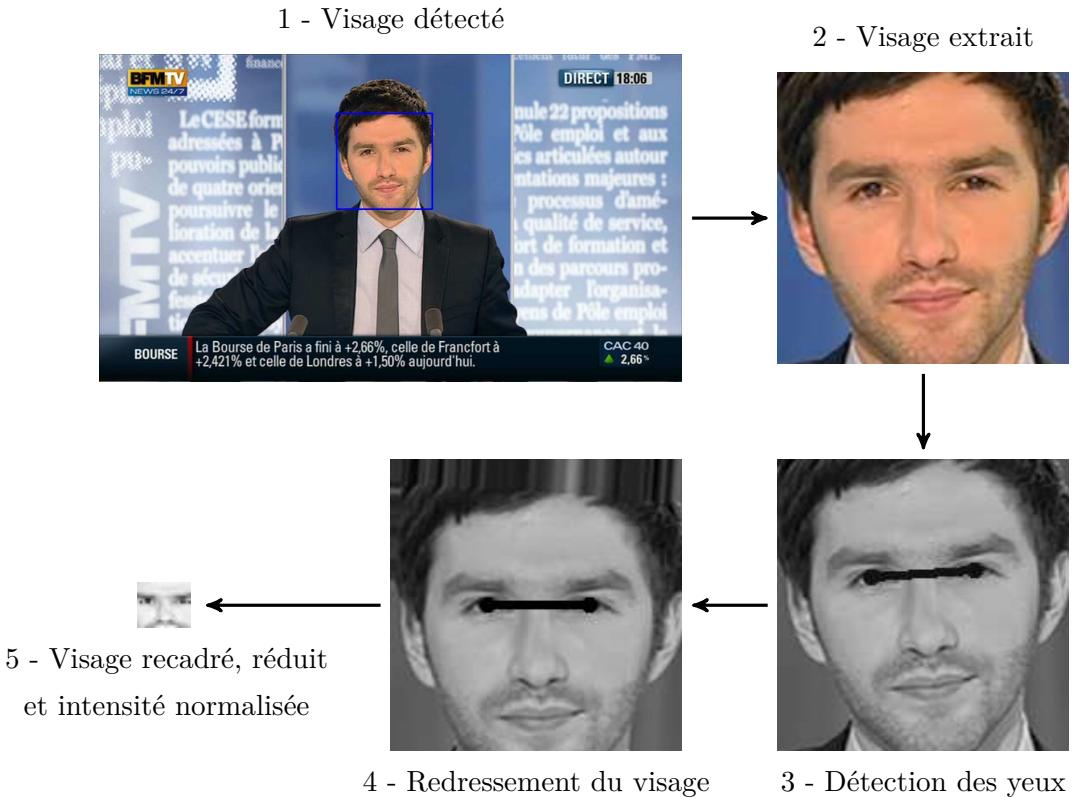


FIGURE 7.1 – Exemple de normalisation d'un visage par l'approche [28] adaptée à la reconnaissance.

linéarisation des 90.000 visages normalisés de notre corpus a pris environ une vingtaine de minutes sur un serveur de calcul. Pour information, il s'agit d'un serveur Linux équipé de deux processeurs Intel Xeon CPU E5620 (16 cœurs de calcul) cadencés à 2,40GHz et de 16GiB de mémoire vive (DDR3).

Toutes les occurrences vidéo de personnes n'ont pas au moins un visage exploitable. 484 occurrences vidéo n'ont pas été normalisées car le réseau de neurones artificiels n'a pas réussi à détecter les deux yeux de ces visages. Ainsi seulement 1.832 occurrences vidéo de personnes peuvent être nommées par la reconnaissance de visages. Les autres devront être nommées par la propagation des identités des occurrences qui auront été nommées.

Entraînement et utilisation du modèle

Les différents visages sont préfixés de leur identifiant. Un SVM, avec un noyau gaussien, est ensuite entraîné sur l'ensemble des visages. Nous utilisons dans nos expérimentations l'implémentation des SVM fourni dans la librairie libSVM. Ses paramètres C (coût) et γ (noyau gaussien) sont déterminés automatiquement par un *grid search* [25], itérant sur différentes valeurs à la recherche de ceux qui maximise la précision de la validation croisée. Cette étape étant coûteuse en temps de calcul, elle est réalisée sur GPU par la version CUDA de la libSVM.

Pour effectuer la reconnaissance d'un nouveau visage, lui aussi doit être normalisé puis linéariser avant d'être présenté au classifieur qui prédit la classe (l'identité) correspondant à ce visage. La prédiction étant relativement rapide, nous l'effectuons sur CPU avec l'implémentation standard de la libSVM.

Nous avons retenu cette approche car elle présente plusieurs propriétés intéressantes pour notre approche :

- bien que l'apprentissage d'un modèle SVM soit relativement long du fait de la validation croisée, la prédiction est très rapide,
- ce classifieur permet de considérer un très grand nombre de classes, comme nous l'avons vu dans la Section 2.1 ce n'est pas le cas de nombreuses approches,
- cette approche utilise les pixels bruts du visage, il n'est pas nécessaire de calculer un descripteur ce qui pourrait s'avérer coûteux.

Nous avons donc décidé d'utiliser cette approche pour reconnaître les visages. Pour mémoire, notre objectif n'est pas de proposer un algorithme de reconnaissance de visages mais des stratégies utilisant ces résultats pour nommer les occurrences puis les groupes d'occurrences. Ainsi, la méthode de reconnaissance de visages peut facilement être remplacée par une autre de l'état de l'art plus performante.

7.2 Taux de reconnaissance de référence

Dans cette section nous présentons la mesure d'évaluation pour évaluer le nommage des occurrences vidéo de personnes. Nous appliquons la reconnaissance faciale à tous les visages exploitables de notre corpus de test. Le taux de précision que nous obtenons ainsi servira de référence dans les expérimentations suivantes.

7.2.1 Calcul de la précision

La précision de la reconnaissance (P) est utilisée comme critère pour évaluer les performances de nos différentes propositions. Elle est calculée comme le nombre d'identifications correctes sur le nombre total de prédictions :

$$P = \frac{|\{o \in \mathbb{O} | \hat{id}(o) = id(o)\}|}{|\mathbb{O}|} \quad (7.1)$$

où \mathbb{O} est l'ensemble des occurrences vidéo de personnes, $id(o)$ est l'identité de l'occurrence o dans la vérité terrain et $\hat{id}(o)$ est l'identité prédite. Cette précision va nous permettre de déterminer un taux de reconnaissance de référence obtenu par le SVM qui nous sert à quantifier le gain en précision qu'il est possible d'obtenir en mettant en œuvre nos propositions.

7.2.2 Résultat de référence

Nous avons prédit l'identité de chaque visage de notre corpus de test à l'aide d'un SVM entraîné sur les visages annotés du corpus REPERE. Nous utilisons la mesure de précision donnée précédemment pour évaluer la précision de la reconnaissance et déterminer le taux de référence que nous utilisons dans les expérimentations suivantes. Le SVM a obtenu un taux de précision de la prédiction des identités des visages sur le corpus de tests de 83%. Cette valeur sera le taux de reconnaissance de référence dans les expérimentations

suivantes. Pour information, toutes les prédictions ont été calculées en 200 secondes sur le serveur de calcul mentioné précédemment. La version CPU de la libSVM [23] a été utilisée pour réaliser ces prédictions.

7.3 Identification des occurrences vidéo de personnes

Nous évaluons maintenant le vote à la majorité relative et le vote à la majorité absolue pour nommer une occurrence vidéo de personne à partir des identités déterminées sur tous les visages exploitables des occurrences vidéo (cf. Section 6.2). L'objectif est d'évaluer les performances de ces deux votes selon la précision et le nombre d'occurrences qu'ils permettent de nommer. Les résultats de ce filtrage servent de référence pour évaluer les résultats obtenus en ne considérant qu'un sous-ensemble des trames de chaque occurrence vidéo.

Ainsi, nous utilisons tous les visages exploitables de chaque occurrence pour déterminer l'identité qui émerge de toute l'occurrence suite à ces votes. Dans ces votes, chaque identité prédite par le SVM sur une trame compte comme une voix.

Vote à la majorité relative

Dans cette première expérimentation, nous appliquons un vote majoritaire. Ainsi, nous attribuons l'identité ayant reçu le plus de voix à l'occurrence (cf. Équation 6.5).

Les résultats dans nos expérimentations seront exprimés avec la notation suivante :

- les prédictions d'identités correctes sont notées T (*true*),
- les prédictions d'identités incorrectes sont notées F (*false*).

	T	F
Quantité	1.543	289
Proportion	84,22%	15,78%

TABLE 7.1 – Résultats de l'attribution d'identité par vote majoritaire sur la reconnaissance faciale.

Les résultats de l'attribution d'identité par vote majoritaire, affichés dans le Tableau 7.1. Dans le cas du vote à la majorité relative, toutes les occurrences vidéo de personnes sont nommées. On remarque que la précision est légèrement meilleure que le taux de reconnaissance faciale de référence mesuré sur les prédictions du SVM dans la Section 7.2. Ainsi, un simple vote à la majorité permet d'améliorer sensiblement la précision de la reconnaissance faciale. Ceci s'explique par le fait que le SVM, pour certains visages, échoue dans l'attribution de l'identité. Un vote majoritaire permet de solutionner ces cas en propageant l'identité majoritaire de l'occurrence.

Vote à la majorité absolue

Pour éviter de propager des identités incorrectes, il faut filtrer les résultats et donc rejeter de telles identités. Nous considérons que les identités n'ayant pas reçu suffisamment de voix sont probablement incorrectes et doivent être rejetées. Dans cette optique, nous appliquons un vote à la majorité absolue.

Nous complétons la notation précédente avec :

- les identités acceptées sont notées P (*positive*),
- les identités rejetées sont notées N (*negative*).

Les résultats peuvent maintenant être présentés sous la forme de combinaison d'identité correcte/incorrecte (T/F) et acceptée/rejetée (P/N).

Ainsi, toute identité majoritaire qui ne reçoit pas au moins la moitié des votes est rejetée. Dans ce cas, l'occurrence vidéo n'est pas nommée et se voit attribuer l'identité "inconnue" (\emptyset).

	TP	FN	TN	FP
Quantité	1.521	140	22	149
Proportion	83,03%	7,64%	1,20%	8,13%

TABLE 7.2 – Résultats de l'attribution d'identité par vote à la majorité absolue sur la reconnaissance faciale.

Dans nos résultats, affichés dans le Tableau 7.2, on remarque que 162 occurrences n'ont pas été nommées ($TN+FN$) car leur identité a été rejetée. Parmi c'est occurrences, une très faible partie a été rejeté à tort. Le nombre d'identités correctement attribuées est proche du nombre de identités correctes (TP est proche de P). Le vote à la majorité absolue a ainsi rejeté principalement des identités qui étaient effectivement mal reconnues par le SVM.

Vote à la majorité	occurrences nommées	précision
relative	100%	84,22%
absolue	91,16%	91,08%

TABLE 7.3 – Comparaison des résultats du filtrage par le vote à la majorité relative et celui à la majorité absolue.

En comparant les résultats du filtrage par les deux votes (cf. Tableau 7.3), on remarque que le vote à la majorité absolue permet d'améliorer de façon importante la précision de l'identification des occurrences vidéo de personnes. Cette stratégie est utile car elle améliore de façon importante le taux de reconnaissance par rapport à celui mesuré sur l'ensemble des prédictions faites par le SVM sur tous les visages.

Ainsi, nous avons vu que le vote, qu'il soit à la majorité relative comme absolue, permet de corriger une partie des identités prédites de façon erronée par l'algorithme de reconnaissance des visages, en l'occurrence le classifieur SVM. Dans cette expérimentation, nous avons utilisé tous les visages des occurrences vidéo de personnes. Nous étudions dans la section suivante l'évolution de la précision du nommage des occurrences en ne considérant qu'un sous-ensemble des visages de chaque occurrence vidéo de personne.

7.4 Variation de la proportion de visages considérés

Dans cette section, nous nous intéressons au nombre de visages utilisés afin de déterminer l'identité d'une occurrence vidéo de personne. En effet, considérer tous les visages de toutes les occurrences vidéo est très coûteux en temps de calcul (de l'ordre de plusieurs jours de calculs sur notre serveur). Ainsi, cette approche ne permet pas un passage à l'échelle. L'objectif est donc de déterminer le nombre idéal de visages à considérer limiter l'usage de la reconnaissance de visages.

Nous avons pour cela proposé et présenté dans le chapitre précédent plusieurs stratégies. Soulignons que les occurrences vidéo n'ont pas toutes la même durée et que tous les visages ne sont pas exploitables. Environ 60% des visages qui composent une occurrence sont normalisables.

Dans cette expérimentation, un vote à la majorité absolue n'est pas adapté. En rejetant certaines identités, ce vote rendrait l'étude de l'évolution de la précision difficile car le nombre d'occurrences considérées évoluerait. Ainsi, dans les expérimentations présentées ici, le taux de précision moyen de référence est de 84,22%, il correspond à la précision obtenue par le vote à la majorité relative considérant tous les visages exploitables de toutes les occurrences vidéo de personnes (cf. Tableau 7.1). Concernant ce taux de référence, il est possible de le dépasser avec un choix de visages particulièrement adapté. Ce taux est ainsi symbolisé dans les différentes figures qui suivent par une ligne horizontale.

En faisant évoluer la proportion de visages considérés, nous allons mesurer la précision moyenne en prenant en compte tous les pourcentages entiers (jusque 100%) avec un pas de 1.

Dans le Chapitre 6, nous avons supposé que les trames situées aux extrémités d'une occurrence pourraient être moins pertinente pour déterminer l'identité de l'occurrence. Nous étudions ce point expérimentalement en considérant plusieurs ordres pour sélectionner les visages d'une occurrence vidéo test selon leurs positions :

1. dans l'ordre de la séquence (du début à la fin)
2. dans l'ordre inverse (de la fin au début)
3. du milieu de la séquence vers les extrémités
4. de façon aléatoire.

7.4.1 Selon l'ordre de la séquence

Dans la Figure 7.2, la courbe présente la précision moyenne de l'identification en fonction du nombre de visages considérés dans l'ordre de la vidéo. Nous observons que la précision moyenne augmente globalement tout en suivant une courbe à la progression irrégulière. La précision initiale en considérant 1% des visages de l'occurrence est de 79,48%. En considérant une proportion comprise entre 77% et 92% des visages de l'occurrence, la précision dépasse le taux de précision de référence (de 84,22%) : en considérant 90% des visages pris dans l'ordre la précision est de 84,44%.

L'évolution irrégulière peut s'expliquer par les deux points suivants. On considère un vote à la majorité, celui-ci peut, pour un petit nombre de votants, faire basculer le vote d'une identité à l'autre. De plus, pour les occurrences vidéo composées de peu de visages (< 100), augmenter la proportion de visages considérés d'un point n'augmente pas nécessairement le nombre total de visages considérés.

Les variations tendent à diminuer avec l'augmentation de la proportion de visages considérés. Cela est dû au fait que pour un grand nombre de votants, une voix supplémentaire influence peu le résultat du vote. De plus, la diminution de la précision moyenne lors des derniers 8% semble indiquer que considérer ces trames dégrade la précision de la reconnaissance.

La Figure 7.2 présente également l'évolution de la précision moyenne en considérant le deuxième ordre de sélection des visages (ordre inverse). La précision moyenne initiale en utilisant 1% des visages de l'occurrence, est de 78,66%, soit presque un point de moins que dans l'ordre précédent. La précision moyenne évolue ensuite globalement de façon

croissante, également de manière irrégulière. On note qu'il faut attendre de considérer l'intégralité des visages pour atteindre la valeur de référence.

Ainsi, on remarque de nouveau que les derniers visages de la vidéo (c'est-à-dire les premiers considérés ici) sont moins pertinents pour déterminer l'identité d'une occurrence vidéo de personne. Ceci s'explique en partie par le fait que les journalistes sont très représentés dans les occurrences vidéo de notre corpus. Les journalistes adoptent un comportement très standardisé dans leur manière de présenter une émission. Ainsi, au début d'un plan, le journaliste fait face à la caméra en la regardant pour présenter une information. Quand plusieurs personnes sont présentes sur le plateau de l'émission, le journaliste va parcourir du regard ces autres personnes pour passer la parole à l'une d'entre elles. Dans tous les cas, la fin du propos d'un journaliste est marquée par un changement de plan. Les invités ont une façon moins formelle de présenter un propos dans une émission. Toutefois, les différentes coupures lors du montage ont tendance à marquer la fin d'un propos à l'aide d'un changement de plan (marquant ainsi la fin de l'occurrence vidéo de personne). Ainsi, la plupart des occurrences vidéo débutent avec une personne faisant face à la caméra, le regard fixé sur celle-ci et finissent sur le journaliste qui affiche une pose non frontale. La situation du début est idéale pour la reconnaissance du visage de la personne, ce qui explique que la précision soit meilleure au début d'une occurrence vidéo qu'à la fin de celle-ci.

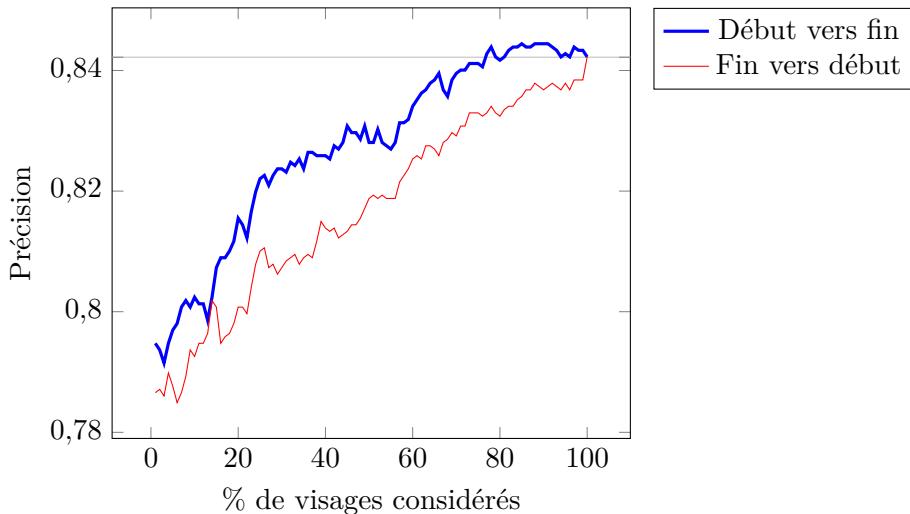


FIGURE 7.2 – Comparaison de la précision moyenne d'identification d'une OVP en fonction de la proportion de visages utilisés choisis selon l'ordre d'apparition et selon l'ordre inverse.

En plus de ce décalage initial entre l'ordre naturel et l'ordre inverse, on remarque que l'écart en termes de précision est maximal entre 30% et 50% de visages considérés. Nous pouvons donc supposer que la première moitié de chaque vidéo montre plus de visages correctement identifiés que la deuxième moitié.

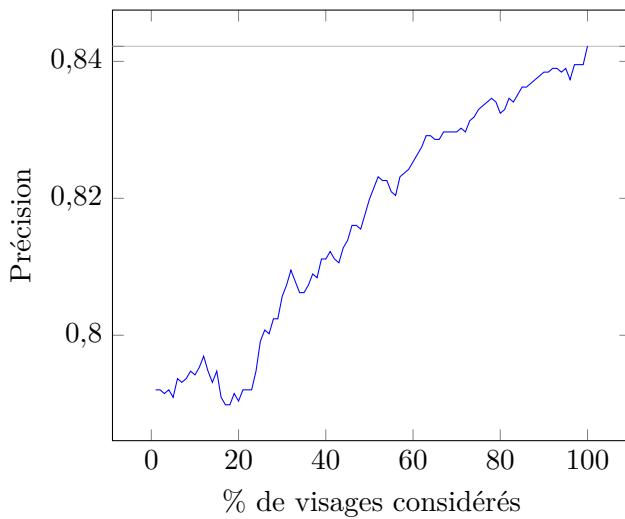


FIGURE 7.3 – Précision d’identification d’une OVP en fonction de la proportion de visages utilisés choisis du milieu de l’OVP vers les extrémités.

7.4.2 Du milieu vers les extrémités

La Figure 7.3 présente la précision moyenne en fonction de la proportion de visages considérés en partant du milieu de chaque séquence pour aller vers les extrémités (le début et la fin). La précision initiale en considérant 1% des visages est de 79,2%. Les premiers visages situés au milieu de la vidéo sont donc pertinents pour identifier les occurrences vidéo de personnes. La précision est ensuite croissante entre 1% et 15% des visages du milieu de la vidéo. La précision moyenne diminue fortement pour stagner autour de 79% pour entre 16% et 24% des visages situés au milieu de la séquence. Il semble donc que les visages situés dans cet intervalle sont les moins pertinents pour identifier les occurrences vidéo de personnes. Il n’y a pas de raison évidente pour expliquer ce point, qui est vraisemblablement dû aux données considérées.

7.4.3 De façon aléatoire

Nous avons jusque-là supposé que l’ordre temporel dans lequel apparaissaient les visages était important à considérer pour identifier les occurrences vidéo de personnes. Nous allons maintenant nous abstraire de cet ordre temporel et considérer les visages sans ordre particulier. Pour cela, nous choisissons aléatoirement les visages, en considérant chaque visage une seule fois (tirage sans remise). Afin d’éviter tout biais engendré par le tirage aléatoire, l’expérimentation est réalisée 100 fois et la moyenne des précisions moyennes pour chaque proportion est calculée.

La Figure 7.4 présente ces résultats. On remarque qu’initialement, en considérant 1% des visages, la précision moyenne est de 80,69%, ce qui est plus élevé que dans les autres cas que nous avons étudiés. La précision moyenne est globalement croissante et toujours de manière irrégulière. L’amplitude des variations est beaucoup plus faible que dans les autres expérimentations. Ceci s’explique par le fait de moyenner les résultats de plusieurs itérations, ce qui a pour effet de lisser des valeurs. On remarque de plus que la moyenne des précisions n’est à aucun moment supérieure au taux de précision de référence. Cela

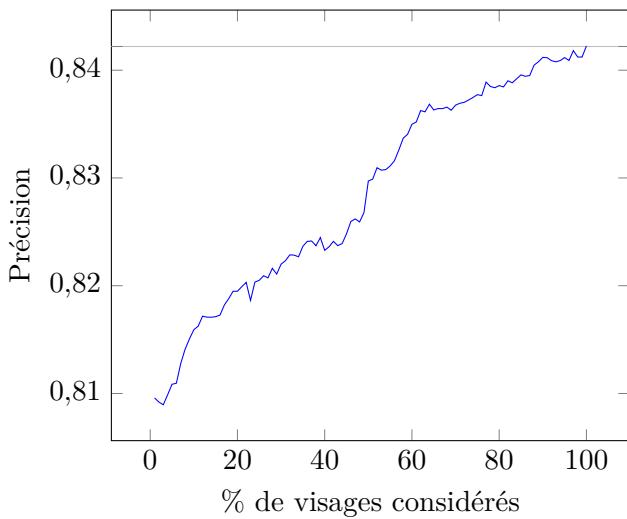


FIGURE 7.4 – Précision d’identification d’une OVP en fonction de la proportion de visages utilisés choisis aléatoirement.

peut s’expliquer par le fait que des visages sélectionnés aléatoirement sont représentatifs de l’occurrence vidéo mais ne présentent pas de conditions particulièrement favorables à leur reconnaissance. Ainsi, aucune configuration aléatoire prise individuellement n’a dépassé cette valeur.

7.4.4 Discussion sur la proportion de visages à considérer

En résumé, on constate que la précision moyenne de l’identification d’occurrences vidéo de personnes est globalement liée à la proportion de visages considérés. En comparant les différentes approches (Figure 7.5), on constate lorsqu’on utilise un petit échantillon de visages pour identifier les occurrences vidéo, il est plus efficace de choisir cet échantillon de façon aléatoire. Cet échantillon récolté est supposé être représentatif de l’ensemble de la vidéo. En revanche, les trames ainsi sélectionnées ne présentent pas des conditions particulièrement favorables à la reconnaissance.

Bien que, sous certaines conditions, il soit possible d’obtenir une précision moyenne supérieure à celle obtenue en considérant tous les visages de la vidéo, ces conditions demandent des a priori sur les données. Dans notre cas, nous avons observé qu’en prenant uniquement les 90% premiers visages, il était possible d’obtenir une précision de 84,44%, légèrement supérieure à la précision de référence de 84,22%. Ce résultat reste difficilement généralisable et est probablement lié à notre corpus de données.

Il est intéressant de noter que les différentes expérimentations montrent que la précision moyenne varie relativement peu en fonction du nombre de visages considérés. Entre considérer 1% des visages choisis aléatoirement et tous les visages, l’écart est de 3 points de précision moyenne. En considérant la moitié des visages, on obtient une précision seulement un point plus faible que la référence. Soulignons qu’une différence d’un point de précision correspond, dans notre expérimentation, à environ 700 visages. Il est important de mettre en perspective les performances par rapport au temps de calcul. Ainsi, réduire de moitié le nombre de visages à considérer revient à diminuer d’environ 30h le temps

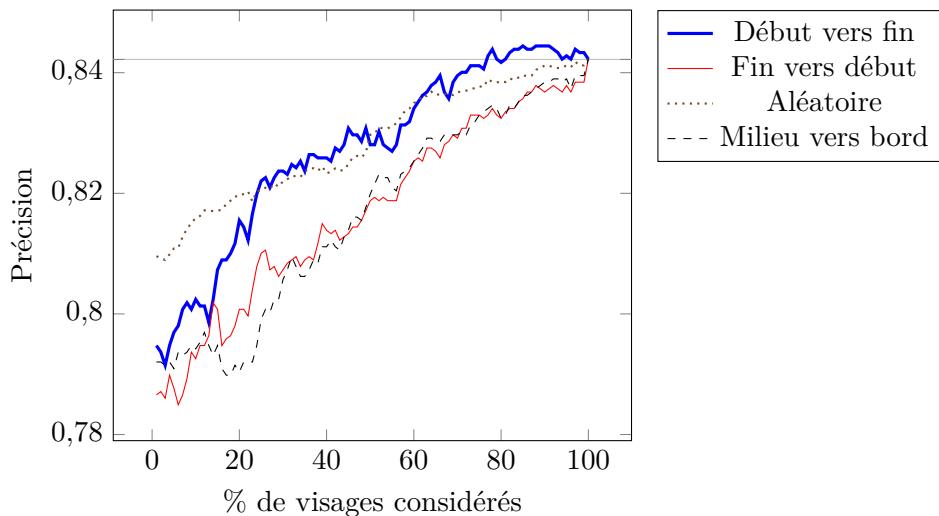


FIGURE 7.5 – Comparaison de la précision d’identification d’une OVP en fonction de la proportion de visages utilisés et de la stratégie de sélection.

de traitements (sur notre serveur) pour une perte en termes de précision relativement petite.

En conclusion, il est possible de régler ce paramètre pour obtenir un rapport complexité/performance adéquat pour une application donnée. Dans un contexte sans contrainte particulière, il est possible de maximiser la précision en considérant tous les visages.

7.5 Propagation d’identités à partir d’OVP nommées

Nous avons présenté la manière choisir un nombre réduit de trames pour identifier une occurrence vidéo. Le problème qui se pose maintenant est celui de la propagation au sein des groupes des identités des occurrences vidéo nommées à celles qui ne l’ont pas été. Dans le Chapitre 6, nous avons proposés plusieurs stratégies de propagation que nous évaluons dans cette section.

Notre approche a permis d’identifier 1.832 occurrences vidéo de personnes. 484 occurrences n’ont pas été identifiées car elles ne contiennent pas de visages exploitables pour la reconnaissance. Elles sont néanmoins dans un groupe contenant d’autres occurrences nommées.

En utilisant le résultat du regroupement (cf. Section 5.7), nous propageons l’identité des occurrences nommées aux autres. Ce processus de propagation permet :

- d’identifier des occurrences vidéo de personnes sans identité (marquée \emptyset pour inconnue),
- d’assigner une identité à l’ensemble du groupe,
- de corriger les identités attribuées par groupe de sorte que toutes les occurrences d’un même groupe aient la même identité.

Dans nos expérimentations, nous réalisons cette propagation en utilisant un vote à la majorité absolue (cf. Section 7.3).

Nous étudions l’existence d’une différence importante entre la propagation des identités déterminées par un vote à la majorité relative et celle déterminées par un vote à la

majorité absolue.

7.5.1 Identités issues d'un vote à la majorité relative

Dans un premier temps, nous réalisons cette propagation en utilisant les identités des 1.832 occurrences déterminées par le vote à la majorité relative.

		Corrects	Incorrects	Inconnus
Avant (1.832 OVP)	Quantité	1.543	289	
	Proportion	84,22%	15,77%	
Après (2.316 OVP)	Quantité	2.001	310	5
	Proportion	86,40%	13,39%	0,21%

TABLE 7.4 – Résultats avant et après la propagation des identités des occurrences vidéo de personnes, déterminées par vote majoritaire simple.

Les résultats sont présentés dans le Tableau 7.4. Nous remarquons que le taux de précision augmente sensiblement, passant de 84,22% (cf. Tableau 7.1) à 86,40%, les 5 inconnus ne sont pas comptés comme des erreurs. Ces derniers correspondent à un groupe particulier dont les identités qui le composent n'ont pas permis d'atteindre un consensus pour la propagation.

La propagation nous a permis d'augmenter le nombre total d'occurrences nommées de 26%, en passant de 1.832 occurrences nommées à 2.311.

7.5.2 Identités issues d'un vote à la majorité absolue

Dans cette section, l'identité (déterminée par un vote à la majorité absolue) est propagée. Nous utilisons 1.670 occurrences identifiées pour nommer les 2.316 occurrences.

		Corrects	Incorrects	Inconnus
Avant (1.832 OVP)	Quantité	1.521	149	162
	Proportion	83,02%	8,13%	8,84%
Après (2.316 OVP)	Quantité	2.002	290	24
	Proportion	86,44%	12,52%	1,04%

TABLE 7.5 – Résultats avant et après la propagation des identités des occurrences vidéo de personnes, déterminées par vote à la majorité absolue.

Les résultats de cette propagation, présentés dans le Tableau 7.5, sont semblables à ceux obtenus dans l'expérimentation précédente. Ainsi 86,44% des occurrences vidéo de personnes sont correctement nommées, 12,52% reçoivent une mauvaise identité et 1,04% des occurrences vidéo de personnes ne reçoivent aucune identité. Ainsi la précision de l'identité des occurrences nommées par le système (correctes + incorrectes) passe de 91,08% à 87,34% tout en nommant 27,8% d'occurrences vidéo supplémentaires (de 1.670 à 2.292 occurrences).

Le nombre d'inconnus est un peu plus élevé en utilisant des identités issues d'un vote majoritaire. Cela s'explique par deux raisons. La première est que, avant la propagation, moins d'occurrences portent une identité. Ainsi, certains groupes n'ont pas d'identité à

propager. Deuxièmement, les identités de certains groupes n'ont pas atteint la majorité absolue permettant la propagation.

Le principal avantage de propager des identités issues du vote à la majorité absolue est de commettre moins d'erreurs. Cependant, cette approche permet d'identifier moins d'occurrences pour un gain en précision négligeable par rapport la propagation des identités issues d'un vote à la majorité absolue.

7.5.3 Discussion sur la détermination des identités initiales

En comparant les résultats des deux expérimentations, on constate que la différence de la précision avant la propagation s'estompe après celle-ci. Ainsi, dans les deux cas, le même nombre d'occurrences vidéo de personnes se voient attribuer la bonne identité. La différence réside au niveau du nombre d'erreurs commises. En effet, la propagation réalisée à partir des identités issues d'un vote à la majorité absolue permet de rejeter un plus grand nombre de fausses attributions en passant de 310 erreurs à 290 erreurs avec le même nombre de bonne identité. Déterminer les identités des occurrences vidéo de personnes avec un vote à la majorité absolue n'est pas utile, car cela n'offre pas un gain en précision significatif après la propagation.

7.6 Variation de la proportion d'occurrences utilisées

Dans cette expérimentation, nous faisons varier la proportion d'occurrences utilisées pour identifier chaque groupe du regroupement (cf. Chapitre 5). Les identités, issues de la propagation par vote à la majorité relative, sont sélectionnées selon différents critères (cf. Section 6.3). Notre premier critère est la sélection d'une quantité croissante d'identités choisies aléatoirement parmi chaque groupe. Le deuxième critère est la sélection d'un nombre grandissant d'identités choisies par ordre décroissant de similarité moyenne. Le dernier critère consiste à sélectionner les identités par ordre décroissant du score de confiance associé à celle-ci.

Dans les Figures 7.6 à 7.8, nous représentons la proportion d'occurrences vidéo correctement identifiées, la proportion d'occurrences incorrectement identifiées et la proportion d'occurrence non identifiées (car aucune identité n'a obtenu la majorité absolue). Une ligne horizontale, placée à la valeur 13,6%, symbolise la somme des proportions d'occurrences inconnues 0,21% ($TN + FN$) et incorrectes 13,39% (FP) obtenues en considérant 100% des identités obtenues par le vote à la majorité relative (cf. Section 7.5). Les identités sont ensuite utilisées lors d'un vote à la majorité absolue afin de choisir l'identité à propager à l'ensemble du groupe.

7.6.1 Propagation par sélection aléatoire

Dans ce premier cas, un nombre croissant d'identités, sélectionnées de façon aléatoire (sans remise), est considéré (voir la Figure 7.6).

On observe une majorité d'occurrence correctement nommées. Initialement, en ne considérant que 1% des identités portée par des occurrences vidéo, la proportion des occurrences vidéo correctement identifiées après la propagation est de 64,6%, 16,2% sont incorrectement nommées et 19,17% ne sont pas identifiées. Quand la proportion d'occurrences vidéo considérées augmente, le nombre d'occurrences inconnues diminue rapidement au profit d'occurrences correctement identifiées. Le nombre d'occurrences portant

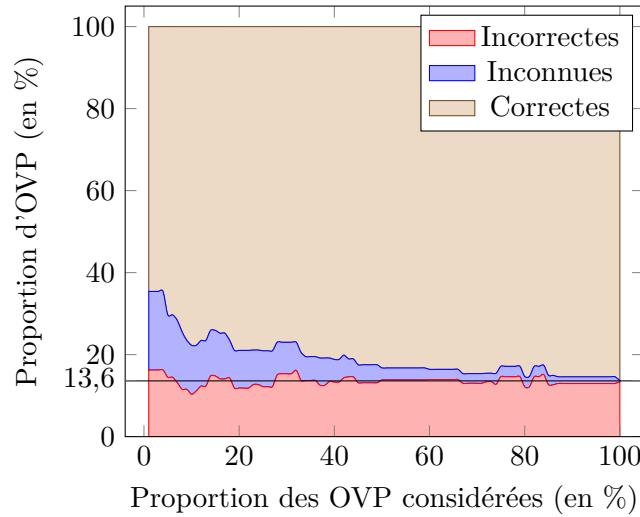


FIGURE 7.6 – Résultats de la propagation d’identités en utilisant un pourcentage des OVP choisies de façon aléatoire dans chaque groupe.

une identité incorrecte reste relativement stable. Cette approche permet ainsi d’augmenter le nombre d’occurrences vidéo identifiées, tout en conservant le taux d’erreur constant.

7.6.2 Propagation par ordre de similarité

Dans ce deuxième cas, la similarité moyenne des occurrences vidéo identifiées a été calculée avec toutes les autres occurrences vidéo d’un même groupe (cf. Section 6.3.1). Les occurrences vidéo sont triées dans l’ordre décroissant selon cette similarité moyenne. Ainsi, les occurrences situées au centre du groupe sont considérées en premier, puis celle situé de plus en plus loin de ce centre ; cela est fait dans le but de donner la priorité aux occurrences jugées les plus représentatives.

Les résultats sont reportés dans la Figure 7.7 : on observe que le nombre d’occurrences non identifiées diminue progressivement avec l’augmentation du nombre d’occurrences vidéo considérées. La proportion d’occurrences incorrectement identifiées reste stable quelque soit la quantité d’occurrences considérées pour la propagation.

La situation initiale, en considérant 1% des identités portées par des occurrences vidéo est comparé dans le Tableau 7.6 avec la situation initiale du cas précédent. On remarque

Critère	Correctes	Incorrectes	Inconnues
Aléatoire	64,6%	16,2%	19,2%
Similarité	75,5%	13,6%	10,9%

TABLE 7.6 – Comparaison de la situation initiale, considérant 1% des identités, selon le critère de similarité et l’aléatoire.

que la situation initiale du cas prenant en compte les occurrences selon leur similarité est bien plus avantageuse que celle les sélectionnant aléatoirement.

Ainsi, le fait de considérer les occurrences dans l’ordre de leur similarité moyenne au sein d’un groupe permet effectivement de sélectionner les identités les plus représentatives

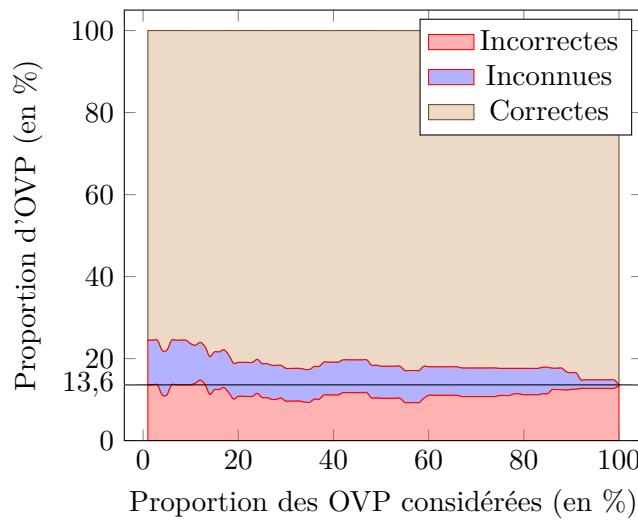


FIGURE 7.7 – Résultats de la propagation d’identités en fonction de la proportion d’OVP choisies par ordre décroissant de similarité à la moyenne dans chaque groupe.

du groupe. Le taux d’erreur reste stable quelque soit le nombre d’occurrences utilisées pour la propagation. Cette dernière approche permet d’identifier les occurrences avec une précision constante et des conditions initiales plus favorables que dans le cas précédent.

7.6.3 Propagation par score de confiance

Dans ce dernier cas, nous sélectionnons un nombre croissant d’identités en nous basant sur le score de confiance associé à ces identités. Pour mémoire, ce score est obtenu à partir du résultat du vote ayant déterminé cette identité.

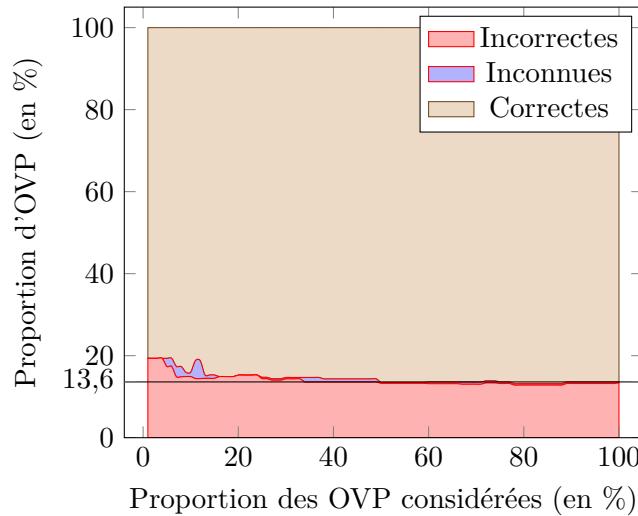


FIGURE 7.8 – Résultats de la propagation d’identités en fonction de la proportion d’OVP choisies par ordre décroissant de confiance dans chaque groupe.

La Figure 7.8 présente les résultats de la propagation en fonction de la proportion d'identités considérées lorsqu'elles sont rangées par score de confiance décroissant. On observe qu'en considérant plus d'identités, la proportion d'incorrectes diminue très progressivement. La proportion d'occurrences non identifiées varie de manière marginale. Globalement, le nombre d'occurrences correctement identifiées augmente.

Critère	Correctes	Incorrectes	Inconnues
Aléatoire	64,6%	16,2%	19,2%
Similarité	75,5%	13,6%	10,9%
Confiance	80,7%	19,3%	0%

TABLE 7.7 – Comparaison de la situation initiale, considérant 1% des identités, selon le score de confiance, la similarité et l'aléatoire.

La situation initiale, en considérant 1% des identités portées par des occurrences vidéo est comparée dans le Tableau 7.7 avec les situations initiales des cas précédents. On remarque que la proportion d'occurrences correctement nommées est de 80,7%, ce qui est le taux initial le plus élevé par rapport aux deux autres cas. Le taux d'occurrences incorrectement nommées est aussi le plus élevé des trois cas avec 19,3% d'erreurs. En revanche, dans cette situation initiale, **toutes** les occurrences vidéo se voient attribuer une identité. De plus, le taux d'erreur converge rapidement vers le taux final attendu en considérant toutes les identités disponibles.

Il est intéressant de noter que cette stratégie permet, en considérant seulement 50% des identités d'obtenir les mêmes résultats qu'en considérant toutes les identités. De plus les résultats obtenus en considérant entre 50% et 99% des identités permet d'obtenir de meilleurs résultats qu'en les considérant toutes. Le maximum est atteint en considérant 78% des occurrences vidéo où 86,8% des occurrences vidéo sont correctement nommées, 12,7% se voient attribuer une mauvaise identité et seulement 0,35% restent inconnues.

Ainsi, la stratégie qui consiste à considérer les identités à propager selon leur score de confiance permet de nommer l'ensemble des occurrences vidéo en utilisant uniquement la moitié des occurrences vidéo de chaque groupe.

7.6.4 Discussion sur les stratégies de propagation

Nous avons comparé plusieurs stratégies de sélection d'identités à propager à l'ensemble du groupe. Nous avons étudié comment évoluent les résultats de propagation en fonction du nombre d'identités considérées. Nous pouvons affirmer que la meilleure stratégie est celle qui consiste à considérer les identités selon leur score de confiance. Elle permet de sélectionner les identités les plus fiables pour la propagation.

Cette propagation semble la plus adaptée, bien qu'elle nécessite le calcul du score de confiance associé à chaque identité. Ainsi, il est nécessaire d'avoir identifié toutes les occurrences afin de propager les plus fiables. Cette stratégie ne permet pas de limiter l'utilisation de la reconnaissance de visages ce qui est contradictoire avec les objectifs de la propagation.

La stratégie consistant à sélectionner les identités selon leur similarité à l'ensemble du groupe des occurrences vidéo offre donc un avantage sur ce point. En effet, la similarité moyenne a déjà été calculée pour réaliser le regroupement en utilisant la matrice de similarités. La précision de l'algorithme de reconnaissance de visage est prédictible.

Dans le cas des SVM, elle est donnée lors de la validation croisée. Il est ainsi possible, en utilisant cette stratégie basée sur la matrice de similarités de choisir le nombre d'occurrences à nommer (avec la précision prédictive de l'algorithme) avant d'obtenir les identités. Elle permet de ne considérer qu'un nombre restreint d'occurrences vidéo sur lesquelles appliquer l'algorithme de reconnaissance et ainsi d'éviter de lourds calculs.

7.7 Conclusion

Nous avons réalisé différentes expérimentations pour identifier et étudier les meilleures stratégies à adopter pour nommer les occurrences vidéo regroupées.

Pour déterminer l'identité d'une occurrence vidéo de personne, nous avons vu que la stratégie qui consiste à considérer environ 50% des visages exploitables choisis aléatoirement sur l'ensemble de la vidéo donne les meilleurs résultats. La plupart des algorithmes de reconnaissance nécessitent que les visages soient normalisés selon certains critères, différents en fonction de l'approche mise en œuvre. Dans la plupart des cas, il est nécessaire de pouvoir détecter les deux yeux du visage. Nous avons vu que peu de visages permettaient cela. Dans nos données utilisant des vidéos issues d'émissions audiovisuelles, seules 20% des trames contiennent un visage normalisable.

Ainsi, cela consiste à sélectionner aléatoirement une trame et tenter de normaliser le visage qu'elle contient pour prédire son identité. Cette opération est à répéter jusqu'à obtenir l'équivalent de 30% des trames de la vidéo, cela représente environ 50% des visages exploitables (cf. Section 7.4).

Pour déterminer l'identité de chaque visage exploitable, nous avons mis en œuvre l'approche basée sur un SVM (noyau gaussien), car celle-ci présentait des propriétés avantageuses à notre approche.

Nous avons mis en évidence que pour nommer les occurrences vidéo de chaque groupe, il suffit d'identifier 50% de ces occurrences pour propager leur identité au reste du groupe. Pour sélectionner ces occurrences, l'approche la plus efficace consiste à calculer la similarité moyenne de chaque occurrence avec toutes les autres du groupe. La similarité moyenne est calculée pour réaliser le regroupement, cette stratégie ne requiert pas de surcoût notable.

Une fois que 50% des occurrences les plus similaires à l'ensemble du groupe sont sélectionnées, il faut propager leur identité au groupe. Pour cela, un vote à la majorité absolue donne les meilleurs résultats.

Il est possible d'ajuster le coût calculatoire en modifiant les paramètres des différentes étapes. Le cas extrême pour minimiser le coût serait de considérer une unique occurrence vidéo de personne de chaque groupe (celle située près du centre de celui-ci) et de prédire son identité en appliquant le SVM à partir du premier visage exploitable choisi aléatoirement.

Quatrième partie

Conclusion et perspectives

Chapitre 8

Synthèse de nos contributions et perspectives

Nous nous sommes intéressés dans ce travail de thèse à la reconnaissance dynamique de personnes dans les émissions audiovisuelles. Nous avons proposé une contribution en deux volets. Le premier volet consiste à isoler toutes les occurrences de personnes au sein d'une émission, et à les regrouper en clusters en se basant sur les histogrammes spatio-temporels. Le second volet propose différentes stratégies de reconnaissance pour assigner une identité aux occurrences de personnes selon les trames qui composent la séquence, et pour propager les identités au sein des groupes selon leurs membres.

8.1 Synthèse de nos contributions

Notre étude de l'état de l'art des approches de reconnaissance de personnes a mis en évidence que la plupart sont coûteuses à mettre en œuvre. Elles sont pour la plupart sensibles aux conditions de prises de vue et à la pose des personnes. Ainsi, les approches de l'état de l'art produisent de bons résultats quand les conditions sont favorables. Dans les émissions audiovisuelles, le visage des personnes est rarement propice à sa reconnaissance par les approches classiques. Dès lors, il est utile de faire le regroupement (clustering) des personnes selon leurs identités pour les propager depuis les occurrences vidéo de personnes reconnaissables par les approches classiques vers celle qui ne le sont pas. Le problème de mettre en correspondance les personnes selon leur identité est abordé par les approches de ré-identification de personnes. Pour être efficace, une approche de ré-identification nécessite de distinguer les personnes dans les vidéos. Cependant, peu d'approches prennent en compte l'aspect temporel des vidéos dans la description qu'elles donnent des personnes. De plus, la plupart des approches ne permettent que de ré-identifier un nombre limité de personnes.

Pour résoudre ces différents problèmes et permettre d'identifier les personnes dans les émissions audiovisuelles, nous avons apporté plusieurs contributions. Nous avons défini un descripteur original, l'histogramme spatio-temporel, qui permet de décrire les occurrences vidéo de personnes en utilisant l'aspect visuel (l'apparence), spatial (positions dans l'image), ainsi que l'aspect temporel (le mouvement) de celles-ci. Les histogrammes spatio-temporels ont été évalués expérimentalement à l'aide d'un corpus de données réelles contenant des émissions TV issues des chaînes BFMTV et LCP. Elles ont montré que le descripteur que nous avons proposé permet de distinguer les personnes dans

les émissions audiovisuelles, sous l'hypothèse que cette apparence ne varie pas au cours de l'émission. Les expérimentations ont montré que la construction par accumulation des pixels, représentés dans l'espace de couleurs RGB, permettait de mieux distinguer les personnes. Notre approche est, de plus, robuste aux variations de prise de vue ainsi qu'à la pose des personnes. Elle produit de meilleurs résultats de ré-identification que les approches basées sur les histogrammes de couleurs ou les spatiogrammes pour un coût calculatoire inférieur à ceux-ci. Nous avons ainsi été capables de regrouper efficacement un très grand nombre d'occurrences vidéo de personnes selon leurs identités.

Ce regroupement nous a permis de sélectionner un nombre limité d'occurrences vidéo de personnes à identifier pour les propager à l'ensemble des autres occurrences vidéo. Nos contributions incluent pour cela des stratégies pour sélectionner un nombre limité d'occurrences vidéo pertinentes afin de nommer l'ensemble des occurrences vidéo de personnes. Notre approche propose des stratégies pour nommer efficacement les occurrences vidéo ainsi sélectionnées. Nous contribuons au nommage des occurrences vidéo de personne en proposant des stratégies permettant un recours minimal à une approche de reconnaissance faciale de l'état de l'art. Notre approche est ainsi indépendante de l'approche de reconnaissance faciale utilisée, et peut éventuellement être remplacée par une autre approche de l'état de l'art. Les expérimentations faites sur les émissions audiovisuelles ont montré que la propagation que nous proposons permet ainsi de nommer plus de personnes, tout en corrigeant certaines erreurs d'identifications. En conclusion, notre approche permet de reconnaître dans les émissions audiovisuelles plus de personnes, avec une plus grande précision, pour un coût calculatoire relativement faible, que les autres approches de l'état de l'art.

8.2 Mise en œuvre de nos contributions

Une partie de ces travaux de thèse a été mise en œuvre dans le consortium PERCOL, dans le cadre du défi REPERE. La matrice de similarités présentée dans l'Équation 4.15 a été utilisée pour réaliser le regroupement multimodal (c'est-à-dire basé sur des informations visuelles, audio et textuelles) de personnes dans les émissions télévisuelles. Ce regroupement a été utilisé dans une fusion multimodale pour propager les identités déterminées à l'aide des autres modalités. Les matrices de similarités ont eu un impact significatif sur l'évaluation du regroupement de personnes par la mesure *Estimated Global Error Rate* (EGER) [59], proposée comme la mesure principale d'évaluation pour le défi REPERE. Cette mesure (à minimiser) correspond à un taux d'erreurs entre les annotations fournies pour l'évaluation et celles proposées par les différents systèmes. Sans matrices de similarités basées sur l'apparence visuelle des personnes, le système de regroupement de PERCOL (lors de l'évaluation intermédiaire) obtient un score EGER de 41,1. Ce regroupement est réalisé à partir des autres modalités de la vidéo. En ajoutant à ce système les matrices de similarités d'histogrammes spatio-temporels (de 1.500 partitions dans l'espace de couleurs RGB), nous obtenons un meilleur score EGER de 32,8. Cette différence significative met en valeur l'intérêt des histogrammes spatio-temporels pour distinguer les personnes dans les émissions audiovisuelles, et met en valeur les contributions de cette thèse concernant le regroupement, présentées dans la Partie II.

8.3 Perspectives de travail

À court terme, nous envisageons d'étudier plus en profondeur les possibilités offertes par les histogrammes spatio-temporels. En filtrant les partitions des histogrammes spatio-temporels pour ne garder que celles qui témoignent de mouvements sur l'axe du temps, nous pourrons séparer les phénomènes visuels qui varient dans le temps de ceux qui sont constants et ainsi séparer ces différents phénomènes pour les étudier. Par exemple, en ne conservant que les partitions contenant du mouvement, il serait éventuellement possible de reconnaître des événements, des actions ou des émotions.

Nous envisageons également l'application de notre approche à d'autres types de vidéos. Par exemple, les histogrammes spatio-temporels sont susceptibles d'être appliqués directement aux vidéos issues de caméra de surveillance, afin de pouvoir ré-identifier les personnes dans un contexte multi-caméras.

Parmi les pistes de travail qu'il serait intéressant d'explorer, nous envisageons la création d'un espace de couleurs ad-hoc pour les émissions audiovisuelles, en nous inspirant des travaux réalisés sur l'espace de couleur OHTA. Cela permettrait d'obtenir un espace de couleurs moins corrélées et pourrait éventuellement contribuer à mieux distinguer les différentes personnes pour améliorer la qualité du clustering.

Jusqu'ici, nous n'avons considéré les performances qu'à partir du résultat produit sur chaque trame de la vidéo. Il pourrait être intéressant d'utiliser une technique de super-résolution, qui combine les différentes trames afin d'obtenir une image de qualité supérieure pour améliorer la reconnaissance de personnes. L'augmentation de la qualité de l'image pourrait bénéficier aux approches locales se basant sur des descripteurs très précis pour reconnaître les personnes.

Dans un cadre plus large, nous envisageons de coupler notre approche avec des modalités complémentaires de la vidéo, de manière analogue aux travaux réalisés dans le consortium PERCOL. Les modalités envisagées incluses notamment l'audio, le texte issu de la reconnaissance optique de caractères (OCR) et les métadonnées en provenance, par exemple, de guides électroniques de programmation (*electronic program guide*, EPG). L'objectif à long terme est la proposition d'un modèle multimodal unifié pour la reconnaissance de personnes dans les émissions audiovisuelles.

Bibliographie

- [1] Nasir Ahmed, T. Natarajan, and Kamisetty R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1) :90–93, 1974.
- [2] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*, pages 469–481. Springer, 2004.
- [3] William R. Aiken. Cathode ray tube, June 11 1957. US Patent 2,795,731.
- [4] Ognjen Arandjelović and Roberto Cipolla. An information-theoretic approach to face recognition from face motion manifolds. *Image and Vision Computing*, 24(6) :639–647, 2006.
- [5] Rémi Auguste. Space-Time Histograms for person re-identification, comparison with lower order Histograms on news videos. In *7th Multitel Spring workshop on video analysis*, Mons, Belgique, June 2012.
- [6] Rémi Auguste, Amel Aissaoui, Jean Martinet, and Chabane Djeraba. Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés. In *Compression et Représentation des Signaux Audiovisuels (CORESA)*, page article30, Lille, France, May 2012. 6 pages.
- [7] Rémi Auguste, Amel Aissaoui, Jean Martinet, and Chabane Djeraba. Ré-identification de personnes dans les journaux télévisés basée sur les Histogrammes spatio-temporels. In Yves Lechevallier, Guy Melançon, and Bruno Pinaud, editors, *Revue des Nouvelles Technologies de l'Information (RNTI), 2012*, volume RNTI-E-23, pages 547–548, Bordeaux, France, January 2012. Hermann-Éditions.
- [8] Talis Bachmann. Identification of spatially quantised tachistoscopic images of faces : How many pixels does it take to carry identity ? *European Journal of Cognitive Psychology*, 3(1) :87–103, 1991.
- [9] Werner Backhaus, Reinhold Kliegl, and John S. Werner. *Colour Vision : Perspectives from Different Disciplines*. Walter de Gruyter, 1998.
- [10] Michel Balinski and Rida Laraki. A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences*, 104(21) :8720–8725, 2007.
- [11] Gregory V. Bard. Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. In *Proceedings of the fifth Australasian symposium on ACSW frontiers*, volume 68, pages 117–124. Australian Computer Society, Inc., 2007.
- [12] Martin Bauml, Keni Bernardin, Mika Fischer, Hazim Kemal Ekenel, and Rainer Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 441–447. IEEE, 2010.

- [13] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf : Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.
- [14] Frédéric Béchet, Rémi Auguste, Stéphane Ayache, Delphine Charlet, Géraldine Damnati, Benoit Favre, Corinne Fredouille, Christophe Levy, Georges Linares, and Jean Martinet. Percol0 - un système multimodal de détection de personnes dans des documents vidéo. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, volume 1, pages 553–560, Grenoble, France, June 2012. ATALA/AFCP.
- [15] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4) :270–286, 2014.
- [16] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7) :711–720, 1997.
- [17] Meriem Bendris, Benoit Favre, Delphine Charlet, Géraldine Damnati, Rémi Auguste, Jean Martinet, Gregory Senay, et al. Unsupervised face identification in tv content using audio-visual sources. *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI 2013)*, 2013.
- [18] Favre Benoit, Géraldine Damnati, Frédéric Béchet, Meriem Bendris, Delphine Charlet, Rémi Auguste, Stéphane Ayache, Benjamin Bigot, Alexandre Delteil, Richard Dufour, Corinne Fredouille, Georges Linares, Jean Martinet, Gregory Senay, and Pierre Tirilly. PERCOLI : a person identification system for the 2013 REPERE challenge. In *SLAM Proceedings 2013*, pages –, Marseille, France, August 2013. ANR 2010-CORD-102-01.
- [19] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā : The Indian Journal of Statistics*, pages 401–406, 1946.
- [20] Irving Biederman and Peter Kalocsai. Neural and psychophysical analysis of object and face recognition. In *Face Recognition*, pages 3–25. Springer, 1998.
- [21] Nathaniel D. Bird, Osama Masoud, Nikolaos P. Papanikopoulos, and Aaron Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2) :167–177, June 2005.
- [22] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293. Association for Computational Linguistics, 2000.
- [23] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2 :27 :1–27 :27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [24] Kyong Chang, Kevin W. Bowyer, Sudeep Sarkar, and Barnabas Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(9) :1160–1165, 2003.
- [25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.
- [26] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1 of *CVPR '05*, pages 886–893, Washington, DC, USA, 2005. IEEE.

- [27] Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3) :171–176, March 1964.
- [28] Taner Danisman, Ioan Marius Bilasco, Nacim Ihaddadene, and Chabane Djeraba. Automatic facial feature detection for facial expression recognition. In *Proceedings of the Fifth International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 407–412, 2010.
- [29] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3) :197–208, 2000.
- [30] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.
- [31] Xiaomeng Wu Duy-Dinh Le and Shin’ichi Satoh. *Encyclopedia of multimedia*, chapter Face Detection, tracking, and recognition for broadcast video, pages 228–238. Springer-Verlag New York Inc, 2008.
- [32] Mady Elias and Jacques Lafait. *La couleur : Lumière, vision et matériaux*, chapter La connaissance de la couleur. Collection Échelles. Belin, 2006.
- [33] Hicham G. Elmongui, Mohamed F. Mokbel, and Walid G. Aref. Spatio-temporal histograms. *Advances in Spatial and Temporal Databases*, pages 19–36, 2005.
- [34] Hans G. Feichtinger and Thomas Strohmer. *Gabor analysis and algorithms : Theory and applications*. Springer, 1998.
- [35] Graham David Finlayson. *Coefficient color constancy*. PhD thesis, Citeseer, 1995.
- [36] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2) :179–188, 1936.
- [37] Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383) :553–569, 1983.
- [38] Kazuhiko Fukui, Björn Stenger, and Osamu Yamaguchi. A framework for 3d object recognition using the kernel constrained mutual subspace method. In *Asian Conference of Computer Vision (ACCV)*, pages 315–324. Springer, 2006.
- [39] Kazuhiko Fukui and Osamu Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Robotics Research*, pages 192–201. Springer, 2005.
- [40] Vineet Gandhi and Remi Ronfard. Detecting and naming actors in movies using generative appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3706–3713. IEEE, 2013.
- [41] Niloofar Gheissari, Thomas B. Sebastian, Richard Hartley, Peter H. Tu, and Jens Rittscher. Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1528–1535. IEEE, 2006.
- [42] Adrian J. Gibbs and George A. McIntyre. The diagram, a method for comparing sequences. *European Journal of Biochemistry*, 16(1) :1–11, 1970.
- [43] Dmitry O. Gorodnichy. On importance of nose for face tracking. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 181–186, 2002.
- [44] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International workshop on performance evaluation of tracking and surveillance*. Citeseer, 2007.

- [45] Peter Grosche, Meinard Muller, and Frank Kurth. Cyclic tempogram— a mid-level tempo representation for music signals. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5522–5525. IEEE, 2010.
- [46] John Guild. The colorimetric properties of the spectrum. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 149–187, 1932.
- [47] Ziad M. Hafed and Martin D. Levine. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3) :167–188, 2001.
- [48] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3) :107–145, 2001.
- [49] Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6. ACM/IEEE, 2008.
- [50] Richard W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2) :147–160, 1950.
- [51] John A. Hartigan. *Clustering algorithms*. Wiley series in probability and mathematical statistics : Applied probability and statistics. John Wiley & Sons, 1975.
- [52] John A. Hartigan and Manchek A. Wong. Algorithm as 136 : A k-means clustering algorithm. *Journal of the Royal Statistical Society*, pages 100–108, 1979.
- [53] Martin Hirzer, Csaba Beleznai, PeterM. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In Anders Heyden and Fredrik Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 91–102. Springer Berlin Heidelberg, 2011.
- [54] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.
- [55] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406) :414–420, 1989.
- [56] Matthew A. Jaro. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7) :491–498, 1995.
- [57] Valen E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48) :19313–19317, 2013.
- [58] Kai Jungling, Christoph Bodensteiner, and Michael Arens. Person re-identification in multi-camera networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 55–61. IEEE, 2011.
- [59] Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carré, Aude Giraudel, and Philippe Joly. A presentation of the repere challenge. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2012.
- [60] Takeo Kanade. *Computer recognition of human faces*, volume 47. Birkhauser Verlag, Basel und Stuttgart, 1977.
- [61] Michael David Kelly. Visual identification of people by computer. Technical report, DTIC Document, 1970.

- [62] Vera Kettnaker and Ramin Zabih. Bayesian multi-camera surveillance. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2. IEEE, 1999.
- [63] Michael Kirby and Lawrence Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(1) :103–108, 1990.
- [64] Solomon Kullback. *Information theory and statistics*. Courier Dover Publications, 1997.
- [65] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79–86, 1951.
- [66] Frank Kurth, Thorsten Gehrman, and Meinard Müller. The cyclic beat spectrum : Tempo-related audio features for time-scale invariant audio identification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 35–40, 2006.
- [67] Henry Oliver Lancaster and Eugene Seneta. *Chi-Square Distribution*. Wiley Online Library, 1969.
- [68] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–313. IEEE, 2003.
- [69] Vladimir I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707, 1966.
- [70] Mu Li, Yang Zhang, Muhua Zhu, and Ming Zhou. Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1025–1032. Association for Computational Linguistics, 2006.
- [71] Stan Z. Li and Anil K. Jain. *Handbook of face recognition*. Springer, 2011.
- [72] Shang-Hung Lin, Sun-Yuan Kung, and Long-Ji Lin. Face recognition/detection by probabilistic decision-based neural network. *IEEE Transactions on Neural Networks*, 8(1) :114–132, 1997.
- [73] Xiaoming Liu, Tsuhan Chen, and Susan M. Thornton. Eigenspace updating for non-stationary process and its application to face recognition. *Pattern Recognition*, 36(9) :1945–1959, 2003.
- [74] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110, 2004.
- [75] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2 :49–55, 1936.
- [76] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
- [77] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- [78] Baback Moghaddam and Alex P. Pentland. Face recognition using view-based and modular eigenspaces. In *International Symposium on Optics, Imaging, and Instrumentation*, pages 12–21. International Society for Optics and Photonics, 1994.

- [79] David S. Moore and George P. McCabe. *Introduction to the Practice of Statistics*. WH Freeman/Times Books/Henry Holt & Co, 1999.
- [80] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9) :1997–2006, September 2003.
- [81] Ara V. Nefian and Monson H. Hayes III. Hidden markov models for face recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 2721 –2724 vol.5, may 1998.
- [82] Thanh Duc Ngo, Duy-Dinh Le, Shin’ichi Satoh, and Duc Anh Duong. Robust face track finding in video using tracked points. *IEEE International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, pages 59–64, 2008.
- [83] Yu-Ichi Ohta, Takeo Kanade, and Toshiyuki Sakai. Color information for region segmentation. *Computer graphics and image processing*, 13(3) :222–241, 1980.
- [84] Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition*, volume 1, pages 582–585. IEEE, 1994.
- [85] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.
- [86] Ofir Pele and Michael Werman. The quadratic-chi histogram distance family. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 749–762. Springer Berlin Heidelberg, 2010.
- [87] P. Jonathon Phillips. *Support vector machines applied to face recognition*, volume 285. Citeseer, 1998.
- [88] P. Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5) :295–306, 1998.
- [89] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference (BMVC)*, volume 1, page 5, 2010.
- [90] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3) :205–229, 1989.
- [91] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850, 1971.
- [92] Austin Roorda, Andrew B. Metha, Peter Lennie, and David R. Williams. Packing arrangement of the three cone classes in primate retina. *Vision research*, 41(10) :1291–1306, 2001.
- [93] Assumpta Sabater and Federico Thomas. Set membership approach to the propagation of uncertain geometric information. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 2718–2723. IEEE, 1991.
- [94] Mohamed Ibrahim Saleh. *Using Ears for Human Identification*. PhD thesis, Virginia Polytechnic Institute and State University, 2007.

- [95] David Sankoff and Joseph B. Kruskal. Time warps, string edits, and macromolecules : the theory and practice of sequence comparison. *Reading : Addison-Wesley Publication*, 1 :167–168, 1983.
- [96] Shin’ichi Satoh. Comparative evaluation of face sequence matching for content-based video access. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 163–168. IEEE, 2000.
- [97] William Schwartz, Huimin Guo, and Larry Davis. A robust and scalable approach to face identification. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *European Conference on Computer Vision (ECCV)*, volume 6316 of *Lecture Notes in Computer Science*, pages 476–489. Springer Berlin / Heidelberg, 2010.
- [98] Gregory Shakhnarovich, John W. Fisher, and Trevor Darrell. Face recognition from long-term observations. In *European Conference on Computer Vision (ECCV)*, pages 851–865. Springer, 2002.
- [99] Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition (CVPR)*, pages 593–600. IEEE, 1994.
- [100] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A (JOSA A)*, 4(3) :519–524, 1987.
- [101] Daniel Swets and John Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(8), 1996.
- [102] Dung Truong Cong, Catherine Achard, Louahdi Khoudour, and Lounis Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In Pasquale Foggia, Carlo Sansone, and Mario Vento, editors, *Image Analysis and Processing (ICIAP)*, volume 5716 of *Lecture Notes in Computer Science*, pages 179–189. Springer Berlin / Heidelberg, 2009.
- [103] Dung Truong Cong, Louahdi Khoudour, and Catherine Achard. Approche spectrale et descripteur "couleur-position statistique" pour la ré-identification de personnes à travers un réseau de caméras. *Groupe d'Etudes du Traitement du Signal et des Images (GRETSI)*, 2009.
- [104] Arne Valberg. *Light Vision Color*. Wiley, 2007.
- [105] Taras K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1) :52–57, 1968.
- [106] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2) :137–154, 2002.
- [107] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2) :275–309, 2013.
- [108] Xiaoyue Wang, Lexiang Ye, Eamonn Keogh, and Christian Shelton. Annotating historical archives of images. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 341–350. ACM, 2008.
- [109] William E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359. ERIC, 1990.

- [110] Laurenz Wiskott, J-M Fellous, N Kuiger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7) :775–779, 1997.
- [111] David Wright. A re-determination of the trichromatic coefficients of the spectral colours. *Transactions of the Optical Society*, 30(4) :141, 1929.
- [112] Osamu Yamaguchi, Kazuhiro Fukui, and Ken-ichi Maeda. Face recognition using temporal image sequence. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 318–323, 1998.
- [113] Jian Yang, Jing-yu Yang, and Alejandro F. Frangi. Combined fisherfaces framework. *Image and Vision Computing*, 21(12) :1037–1044, 2003.
- [114] Xuemei Zhang and David H. Brainard. Estimation of saturated pixel values in digital color imaging. *JOSA A*, 21(12) :2301–2310, 2004.
- [115] Wenyi Zhao, Rama Chellappa, and P. Jonathon Phillips. *Subspace linear discriminant analysis for face recognition*. Citeseer, 1999.
- [116] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. Face recognition : A literature survey. *ACM Computing Surveys (CSUR)*, 35(4) :399–458, 2003.
- [117] Wenyi Zhao, Arvindh Krishnaswamy, Rama Chellappa, Daniel L. Swets, and John Weng. Discriminant analysis of principal components for face recognition. In *Face Recognition*, pages 73–85. Springer, 1998.
- [118] Shaohua Zhou, V. Krueger, and Rama Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1-2) :214–245, 2003.