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 1998 [1]. 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 boitier 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 Internet [2]. De plus, l'accès haut débit [3] 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 smartphones [4]. 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 3h24 [5] 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 Google [6], pour les années 2011, 2012 et 2013, 5 requêtes parmis 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 Internet [7] 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 personnalités [8]. 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 bruit [9],

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, fourni 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 reconnaitre 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 est 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 consine 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'appelation 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 librarie OpenCV (http ://docs.opencv.org).

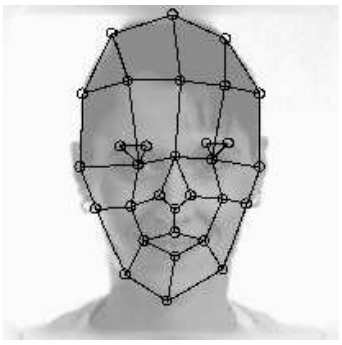


FIGURE 2.3 – Exemple de masque correspondant aux points d'intérêt du visage positionné par la méthode de Wiskott et al. [110].

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 nœud 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 glissant sur un axe vertical (avec un peu de recouvrement). Chaque position est modélisée par une transformée 2D en cosinus discrète

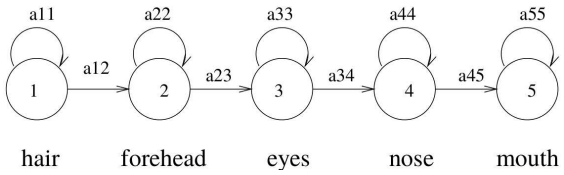


FIGURE 2.4 – Illustration de la structure des états du visage modélisé par un HMM, ainsi que des probabilités de transition. (Illustration tirée de [81]).

(*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 précédement 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 différentiables 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 voisins [2], après une binarisation [3] 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 *LBPH* (LBP histogram).

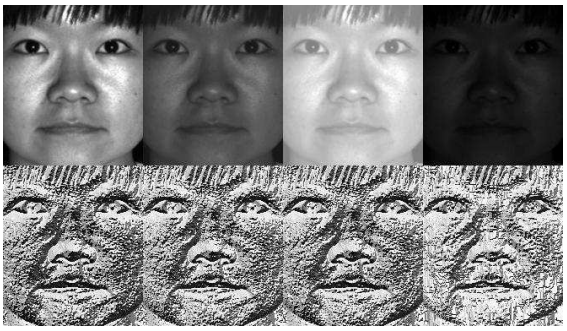


FIGURE 2.5 – Exemple de représentation d'un visage à partir de LBP, en variant la luminosité de l'image source de façon uniforme. Exemple tiré de la documentation en ligne de la librarie OpenCV (http ://docs.opencv.org).

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 particulaire [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'illumination 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'illumination, la MSM a été étendue par la *constrainted 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 constrainted 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 reconnaitre 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 corréler 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 entrainé 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 :

$$h(b) =< n_b >, \quad b = 1, \ldots, B \tag{2.1}$$

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 :

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

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 :

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

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 :

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

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 :

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

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.

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

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 des 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).

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

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. $\hat{\Sigma}_i^{-1}$ est l'estimateur de leur covariance :

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

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 :

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

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" (*pairewise 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 :

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

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 :

$$\gamma(C_1, C_2) = \left\lfloor \frac{\max(|C_1|, |C_2|)}{2} \right\rfloor - 1 \tag{2.11}$$

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

	h	i	s	t	o	g	r	a	m	m	e
s			•								
p											
a								•			
t				•							
i		•									
o					•						
g						•					
r							•				
a								•			
m									•	•	
m									•	•	
e											•

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 :

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

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 séries temporelles, sous certaines restrictions. Les séries 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 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 créé 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

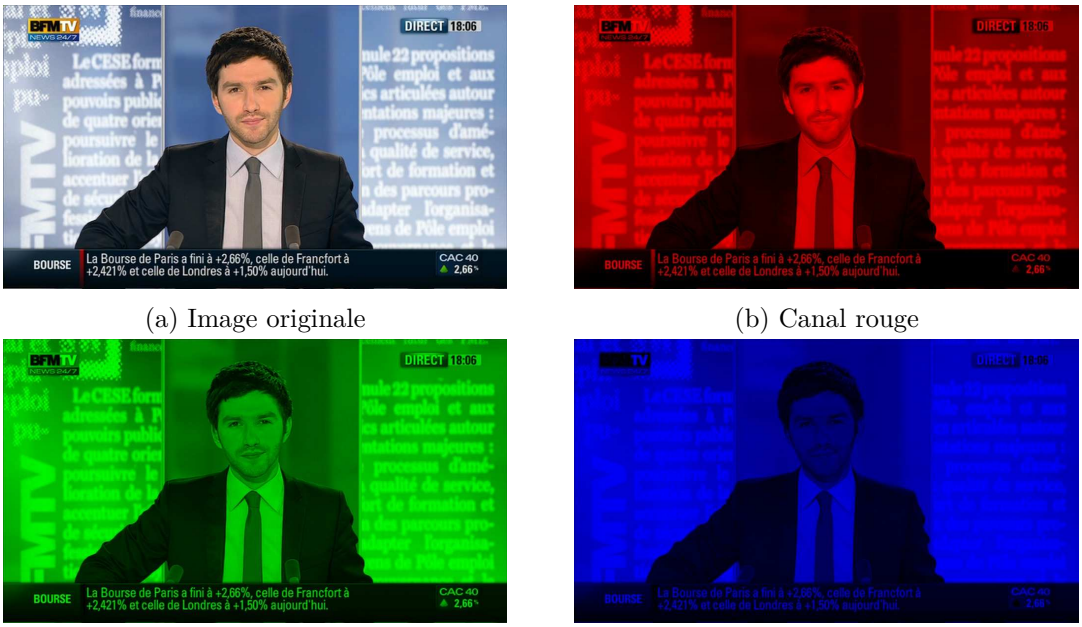
La représentation Rouge-Vert-Bleu (RVB, RGB en anglais) est la représentation la plus répandue des images. Les quantités de rouge, de vert et de bleu de chaque pixel sont exprimées indépendamment. Cependant, il a été montré qu'il existe une forte corrélation entre les différentes valeurs pour rouge, 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,



(a) Image originale (b) Canal rouge

(c) Canal vert (d) Canal bleu

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 :

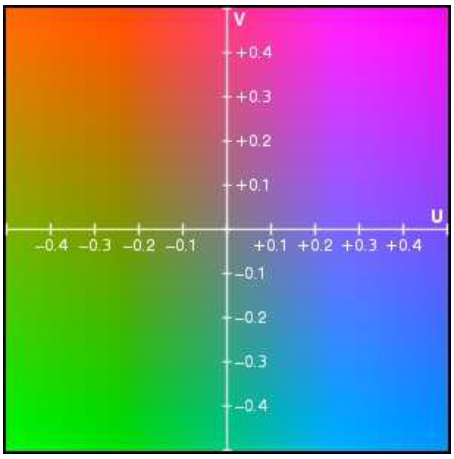


FIGURE 2.8 – Exemple d'une plage U-V, où Y = 0,5, représentée à l'intérieur de la gamme de couleurs RVB ; en noir et blanc, seule Y est utilisée, toutes ces couleurs rendent donc le même gris.

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

$$Y = 0,2989 \times R + 0,5866 \times G + 0,1145 \times B \quad (2.16)$$
$$Cb = -0,1688 \times R - 0,3312 \times G + 0,5000 \times B \quad (2.17)$$
$$Cr = 0,5000 \times R - 0,4184 \times G - 0,0816 \times B \quad (2.18)$$

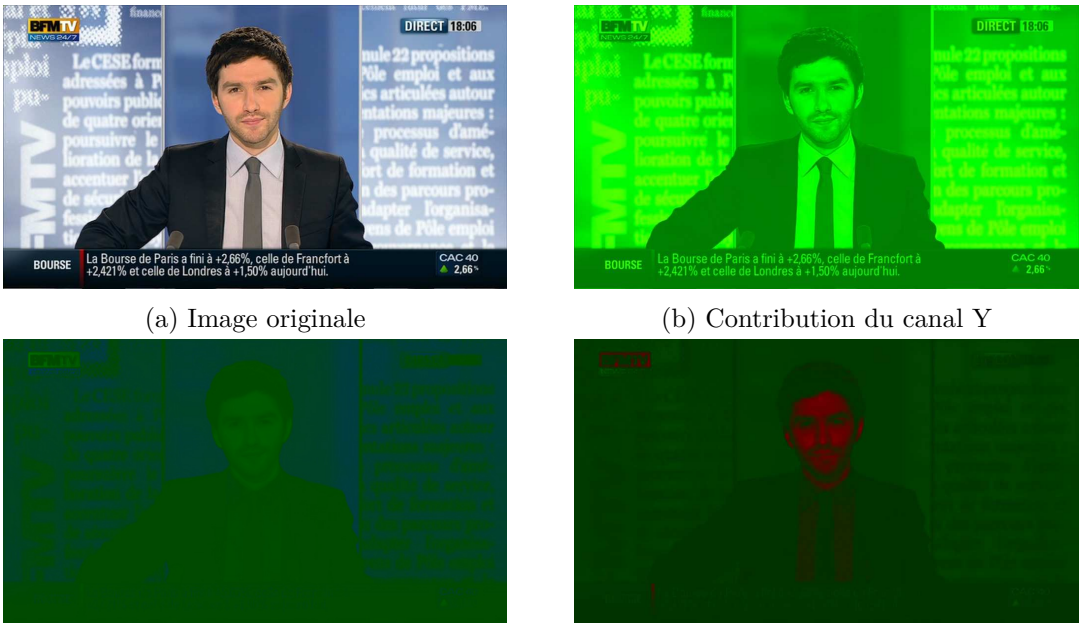
Représentation HSV

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 dénomé 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



(a) Image originale

(b) Contribution du canal Y

(c) Contribution du canal U

(d) Contribution du canal V

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

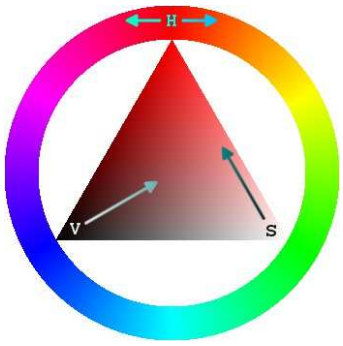


FIGURE 2.10 – Une roue de couleurs HSV permet à l'utilisateur de sélectionner une multitude de couleurs.

(a) Image originale

(b) Contribution du canal L

(c) Contribution des canaux H et L

(d) Contribution des canaux S et L

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.

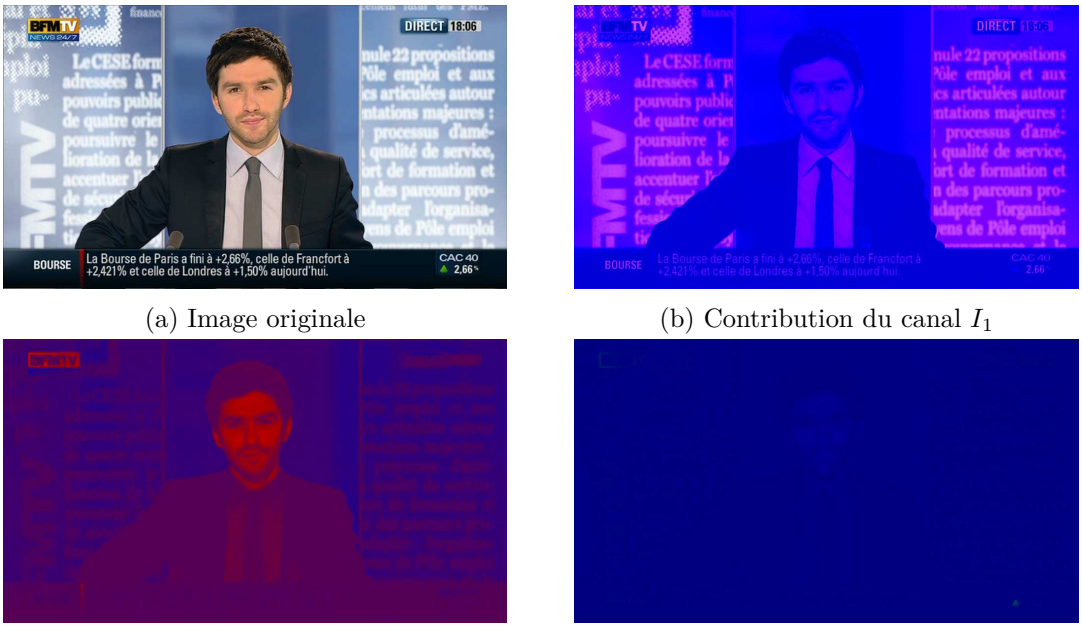
De cette ACP, les trois composantes principales I_1, I_2 et I_3 ont été conservées. Ces dernières permettent d'approximer toutes les couleurs. L'espace de couleur ainsi généré est une transformation linéaire simple de l'espace RVB.

Cette transformation se fait grâce aux équations suivantes :

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

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

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



(a) Image originale

(b) Contribution du canal I_1

(c) Contribution du canal I_2

(d) Contribution du canal I_3

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 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 tri-stimulus 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 chromacies [4] de celle-ci.

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

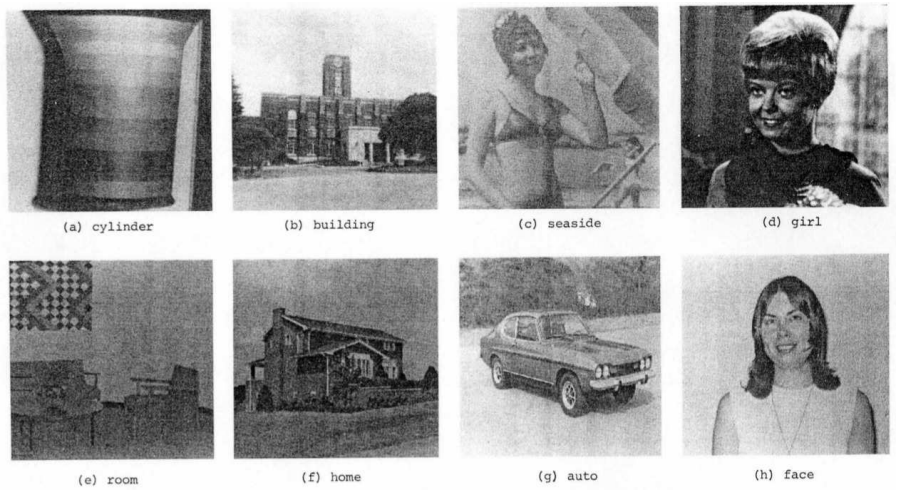
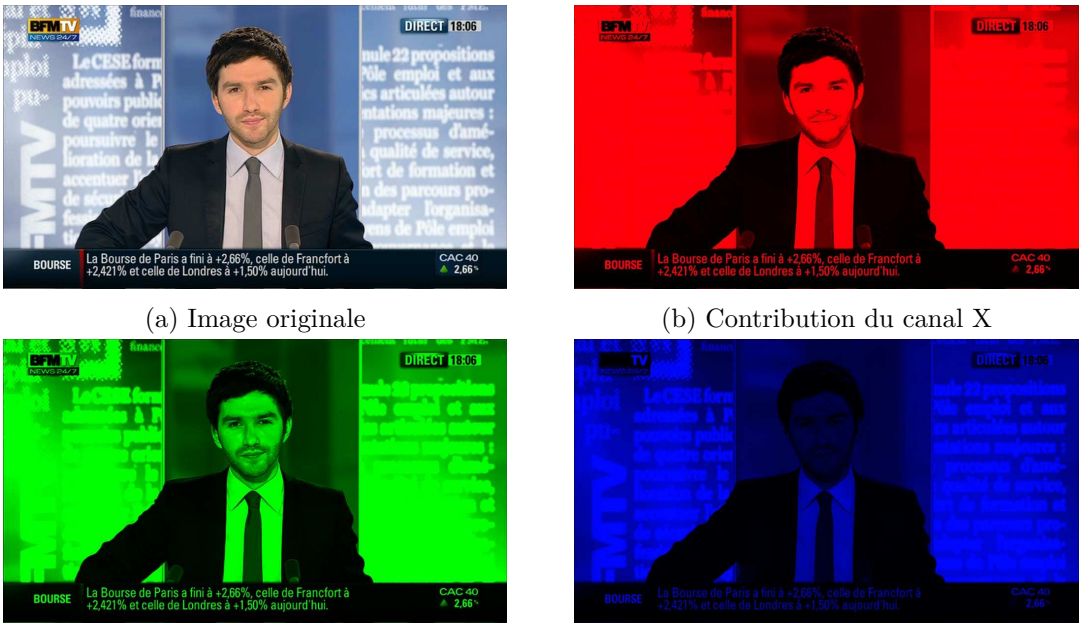


FIGURE 2.14 – Collection d'images ayant servi pour la construction de l'espace de représentation OHTA. Ces images sont issues de l'article [83]. Les images utilisées étaient en couleur, mais seules les versions en niveaux de gris sont facilement disponibles aujourd'hui.

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.



(a) Image originale

(b) Contribution du canal X

(c) Contribution du canal Y

(d) Contribution du canal Z

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.

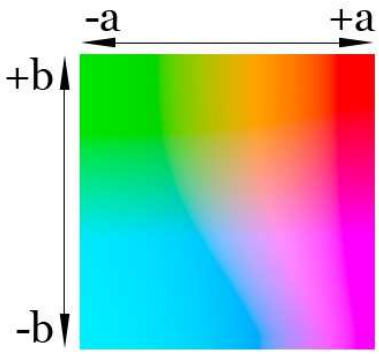


FIGURE 2.17 – Représentation des valeurs des canaux a* et b* pour une luminance de 75% dans l'espace de représentation des couleurs L*a*b*.

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

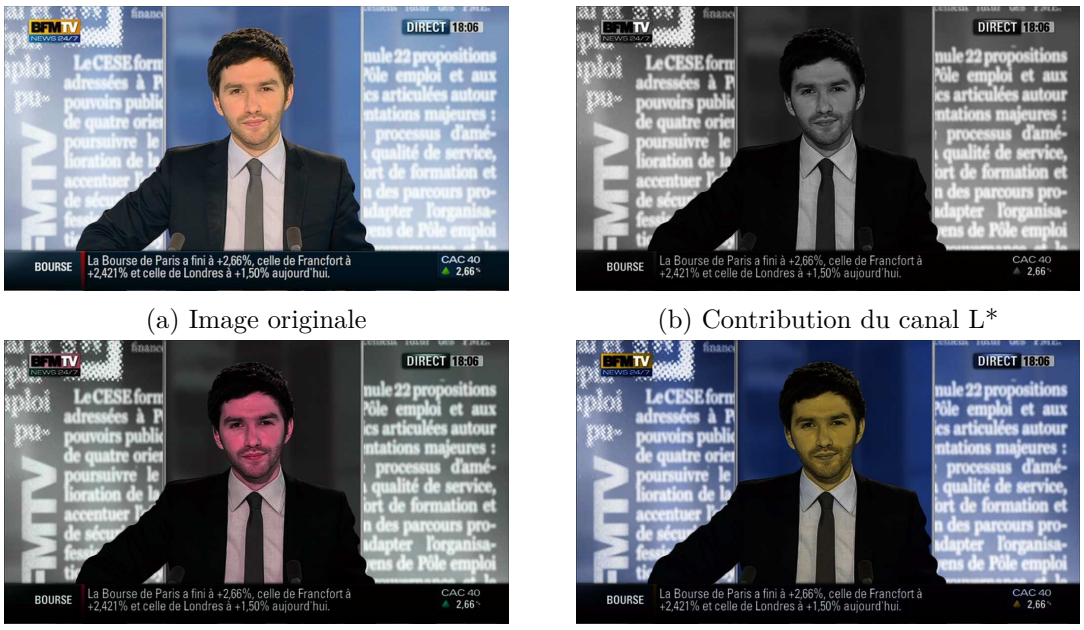
est nécessaire d'effectuer une conversion. Cette conversion peut avoir un coût important lorsque l'on considère une vidéo, car il est nécessaire de décoder chaque trame en RGB, puis de la convertir dans l'autre espace de représentation. Le coût de cette conversion dépend du nombre de trames, de leur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



(a) Image originale

(b) Contribution du canal L*

(c) Contribution des canaux L*a*

(d) Contribution des canaux L*b*

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 reconnaitre 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 exploitables.

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 reconnaitre 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 maitrisé par les équipes de tournage, dans l'environnement contrôlé qu'est le studio de télévision.