

Mise en perspective de nos contributions

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

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

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

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

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

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

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

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

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

3.1 Approche classique de reconnaissance de personnes dans les vidéos

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

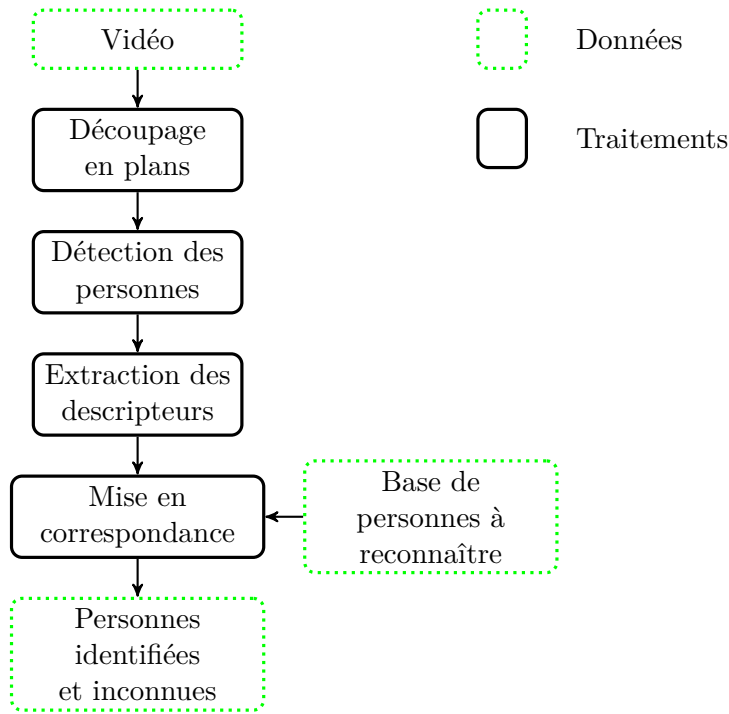


FIGURE 3.1 – Schéma présentant une approche classique de reconnaissance de personnes dans les flux vidéo.

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Proposition générale

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

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

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

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

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

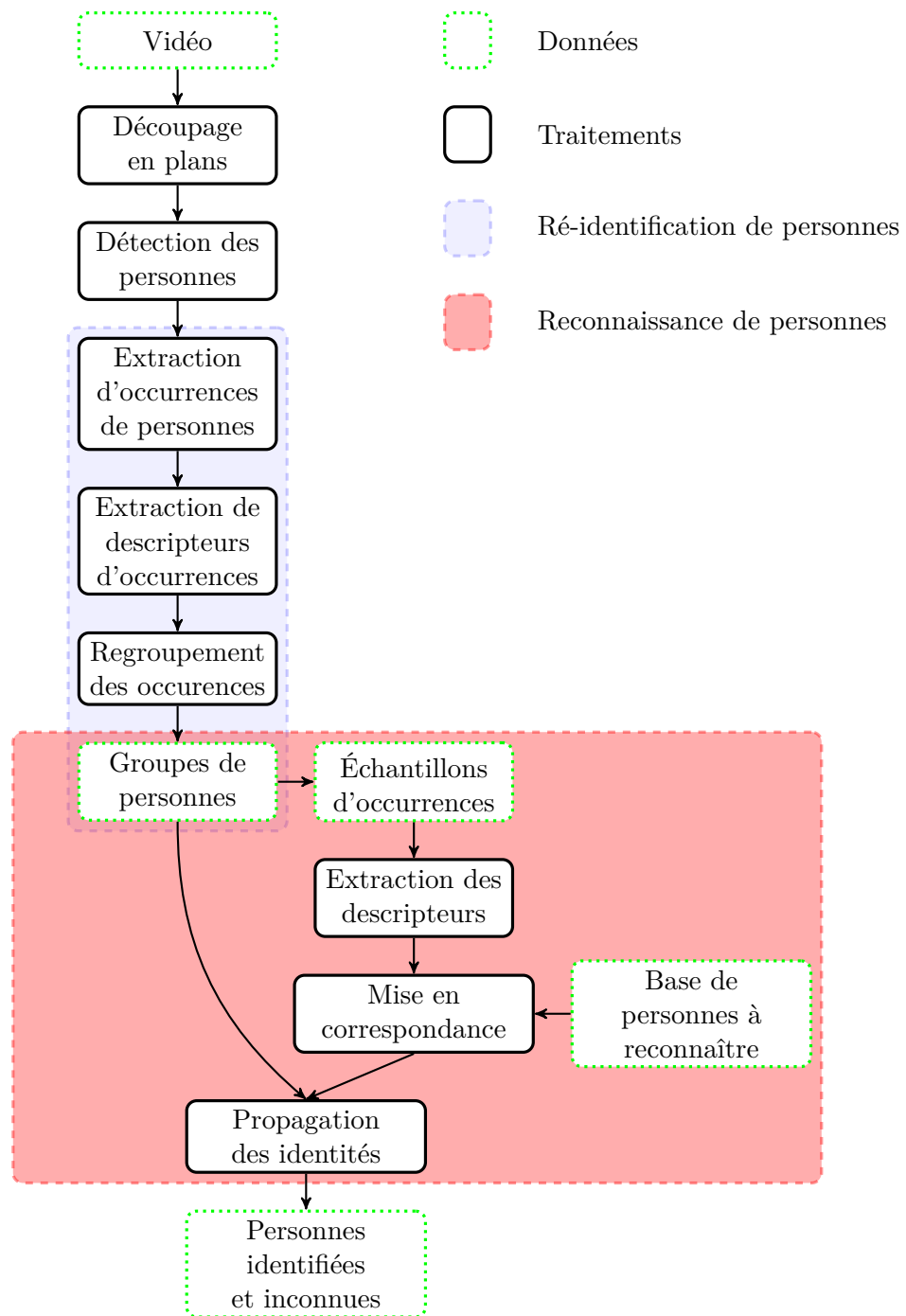


FIGURE 3.2 – Schéma présentant notre approche globale pour la reconnaissance dynamique de personnes dans les flux vidéo.

Deuxième partie

Regroupement des occurrences vidéo de personnes

Chapitre 4

Approche proposée pour le regroupement des OVP

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

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

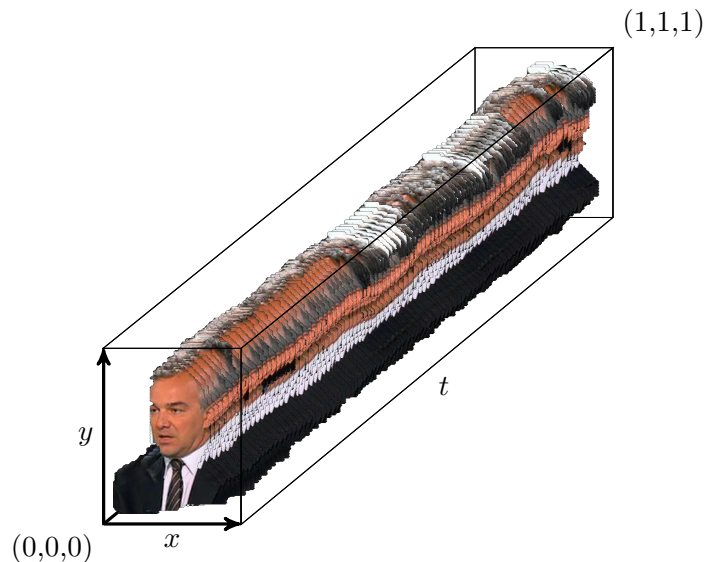


FIGURE 4.1 – Illustration de la notion d'occurrence vidéo de personne (OVP) (*person-track*).

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

4.1 Histogrammes spatio-temporels

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

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

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

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

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

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

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

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

4.2 Interprétations

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

4.2.1 Illustration

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

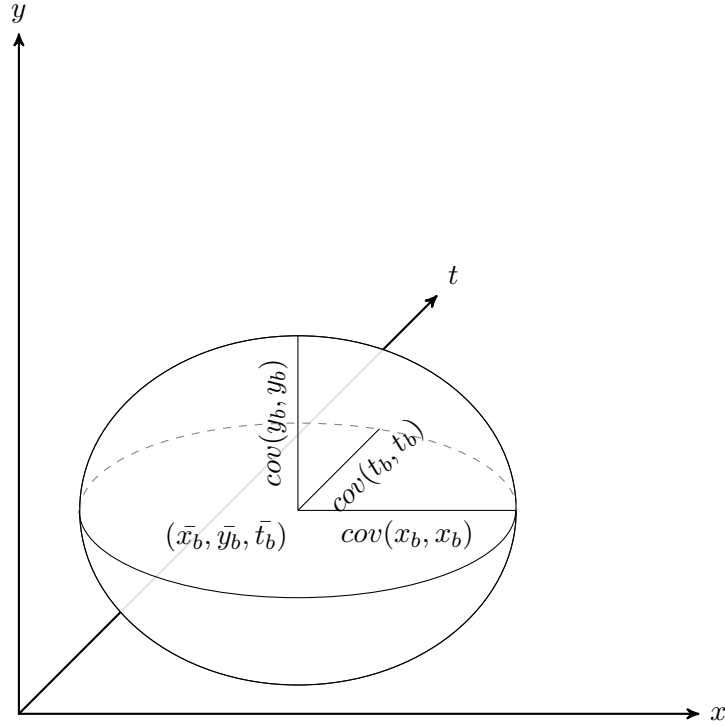


FIGURE 4.2 – Illustration de la représentation d'une partition b dans le volume tridimensionnel de l'occurrence vidéo à l'aide d'un ellipsoïde.

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

4.2.2 Information de mouvement

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

4.2.3 Interprétation de la dimension temporelle des HST

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

4.3 Stratégies de construction

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

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

4.3.1 Accumulation des pixels

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

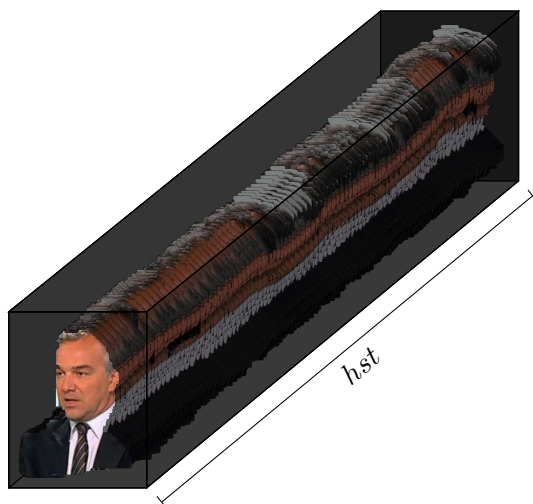


FIGURE 4.3 – Schéma illustrant la construction d'un histogramme spatio-temporel par accumulation des pixels de toutes les trames de la séquence vidéo.

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

4.3.2 Fenêtre glissante

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

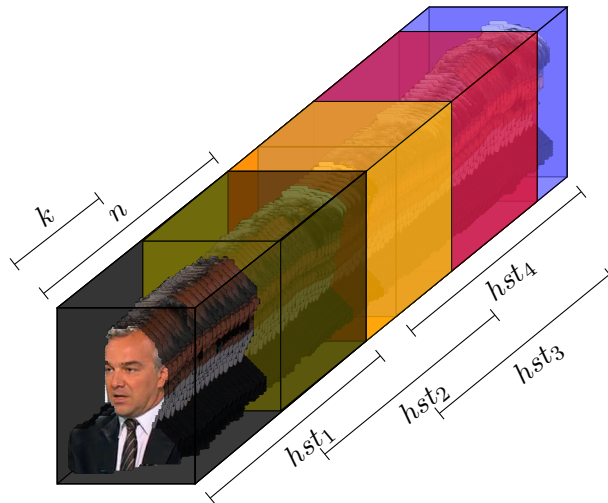


FIGURE 4.4 – Schéma illustrant la construction de plusieurs histogrammes spatio-temporels sur une fenêtre de n trames glissant le long de l'occurrence vidéo avec un décalage de k trames.

longueur, contrairement à la construction par cumul.

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

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

4.3.3 Fenêtre sautante

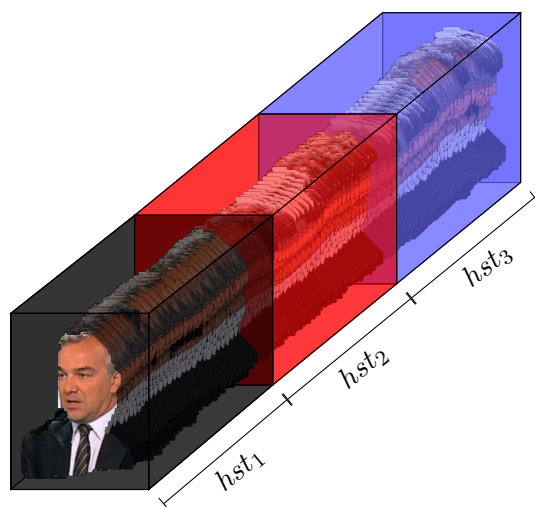


FIGURE 4.5 – Schéma illustrant la construction de plusieurs histogrammes spatio-temporels sur une fenêtre sautante, de taille n trames, le long de l'occurrence vidéo.

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

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

4.3.4 Séparation des canaux colorimétriques

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

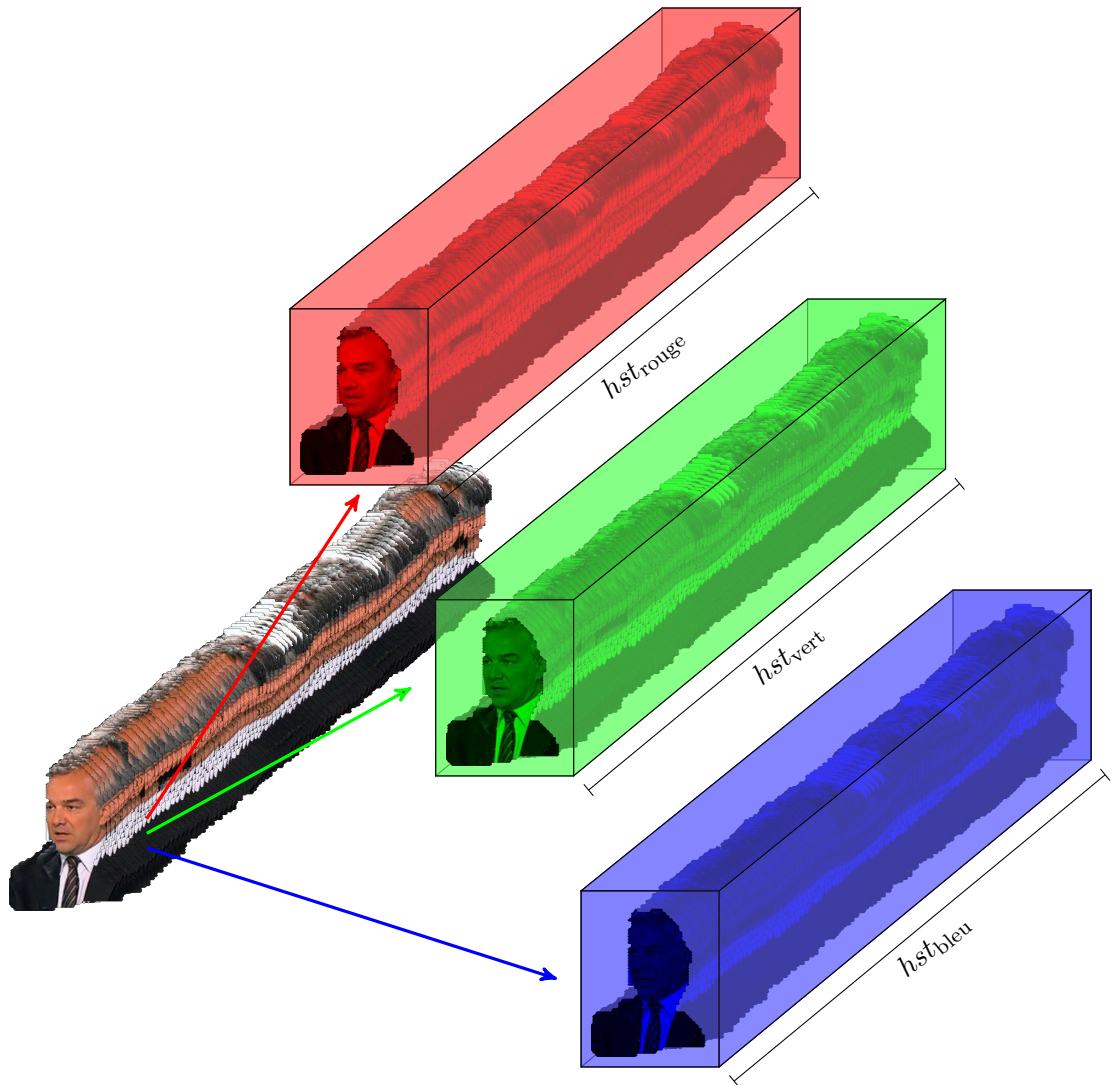


FIGURE 4.6 – Schéma illustrant la construction d'histogrammes spatio-temporels pour chaque canal et par accumulation des pixels sur les trames de la vidéo

le coût calculatoire par le nombre de canaux considérés.

4.4 Mesure de similarité

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

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

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

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

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

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

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

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

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

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

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

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

4.5 Complexité

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

4.5.1 Complexité de la construction

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

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

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

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

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

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

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

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

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

4.5.2 Coût mémoire des descripteurs

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

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

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

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

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

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

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

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

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

4.5.3 Complexité des mesures de similarités

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

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

Ce coût est similaire pour les spatiogrammes et les histogrammes.

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

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

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

4.6 Matrices de similarités

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

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

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

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

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

4.7 Choix du nombre de partitions des histogrammes spatio-temporels

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

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

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

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

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

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

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

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

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

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

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

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

La précision à n_i pour o_i est donné par P_i :

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

Ensuite, nous calculons la moyenne pondérée des précisions P_i :

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

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

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

4.8 Clustering d'histogrammes spatio-temporels

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

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

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

4.9 Résumé des propositions

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

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

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