Nommage de groupes

6.1 Introduction

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

6.2 Nommage d'une occurrence à partir de ses trames

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

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

de visages qui associe une indentité à une trame :

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

$$f \rightarrow \iota \tag{6.2}$$

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

6.2.1 Utilisabilité d'une trame

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

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

6.2.2 Reconnaissance basée sur une trame unique

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

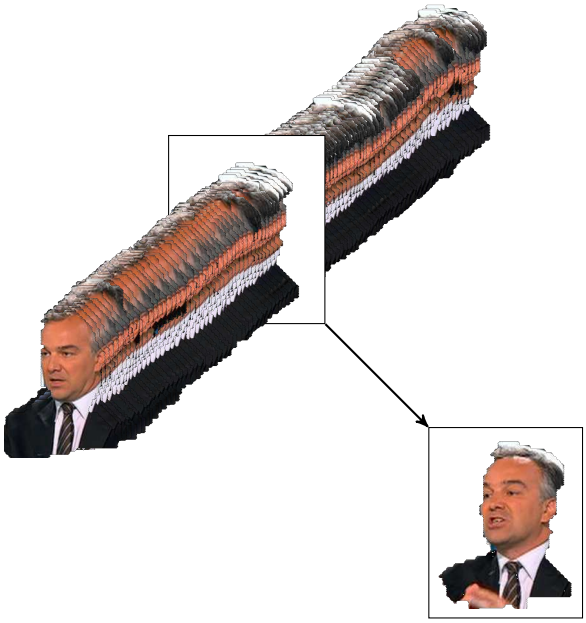

FIGURE 6.1 – Reconnaissance à partir de la trame centrale d'une occurrence vidéo de personne.

Choix de la trame centrale

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

Critère de couleur moyenne

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

one de mouvement minimal

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

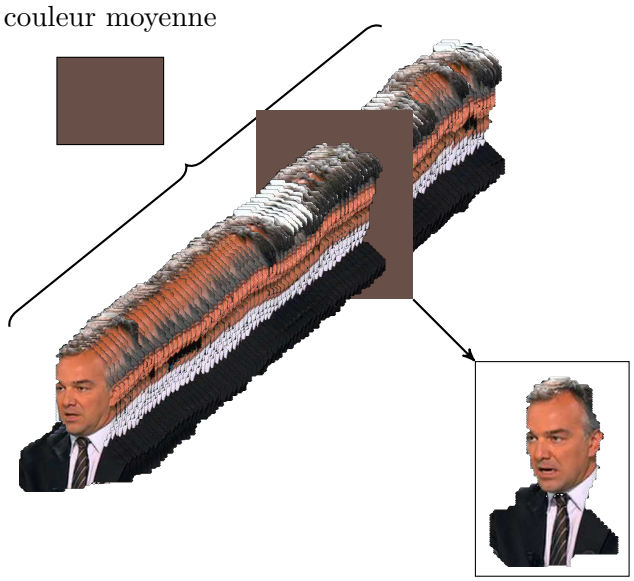

FIGURE 6.2 – Sélection de la trame la plus représentative, en termes de couleur moyenne, d'une occurrence vidéo de personne.

Pose frontale

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

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

6.2.3 Reconnaissance basée sur plusieurs trames

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

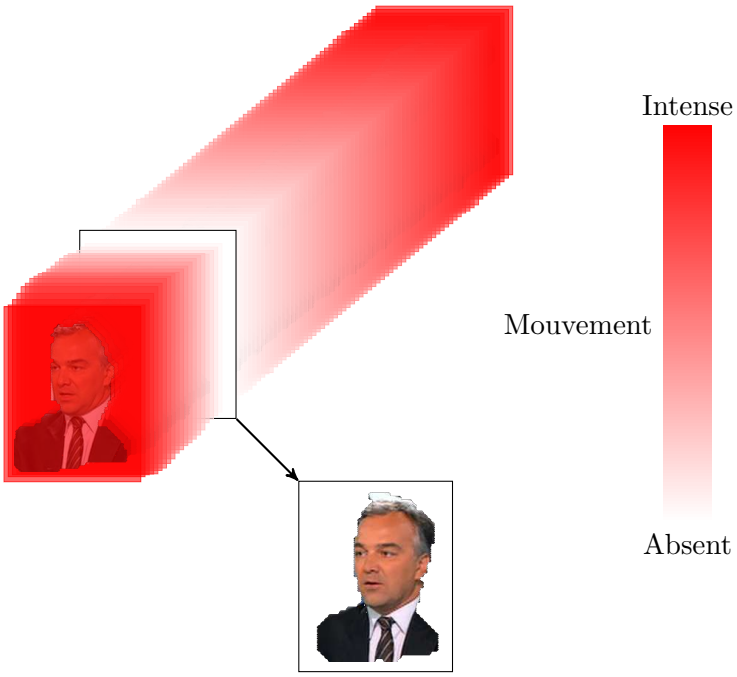

FIGURE 6.3 – Sélection de la trame située dans la zone avec un minimum de mouvement d'une occurrence vidéo de personne.

Choix des trames

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

Combinaison des résultats

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

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

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

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

rapport entre cette fréquence et le nombre de trames votantes :

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

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

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

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

6.3 Nommage d'un groupe à partir de ses occurrences

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

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

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

6.3.1 Sélection d'une occurrence unique

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

Centre du groupe

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

Choix basé sur l'indice de confiance

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

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

Choix aléatoire

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

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

6.3.2 Choix du nombre d'occurrences

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

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

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

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

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

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

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

Combinaison des résultats

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

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

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

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

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

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

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

6.4 Résumé sur le nommage des groupes d'occurrences

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

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

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

Chapitre 7

Validation des approches de nommage des personnes

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

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

7.1 Expérimentations

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

7.1.1 Présentation des données

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

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

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

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

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

7.1.2 Utilisation des données

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

Normalisation des visages

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

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

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

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

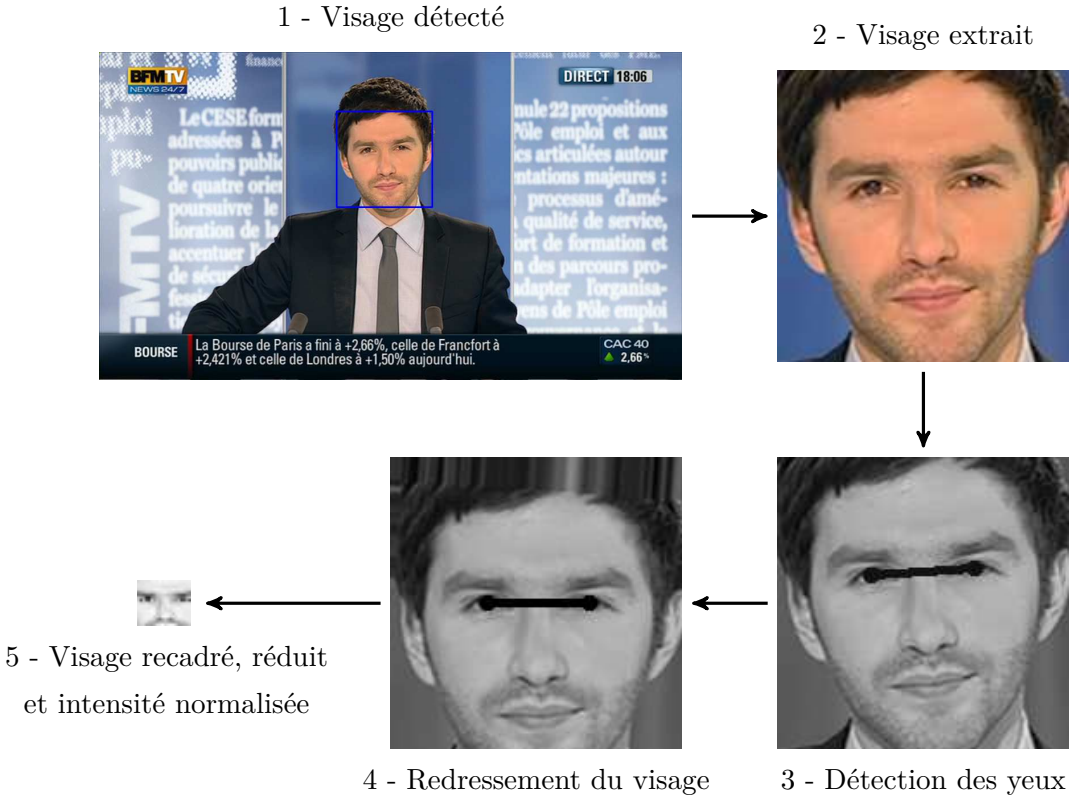

FIGURE 7.1 – Exemple de normalisation d'un visage par l'approche [28] adaptée à la reconnaissance.

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

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

Entraînement et utilisation du modèle

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

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

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

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

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

7.2 Taux de reconnaissance de référence

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

7.2.1 Calcul de la précision

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

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

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

7.2.2 Résultat de référence

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

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

7.3 Identification des occurrences vidéo de personnes

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

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

Vote à la majorité relative

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

Les résultats dans nos expérimentations seront exprimés avec la notation suivante :
— les prédictions d'identités correctes sont notées T (*true*),
— les prédictions d'identités incorrectes sont notées F (*false*).

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

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

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

Vote à la majorité absolue

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

Nous complétons la notation précédente avec :

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

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

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

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

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

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

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

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

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

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

7.4 Variation de la proportion de visages considérés

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

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

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

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

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

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

7.4.1 Selon l'ordre de la séquence

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

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

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

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

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

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

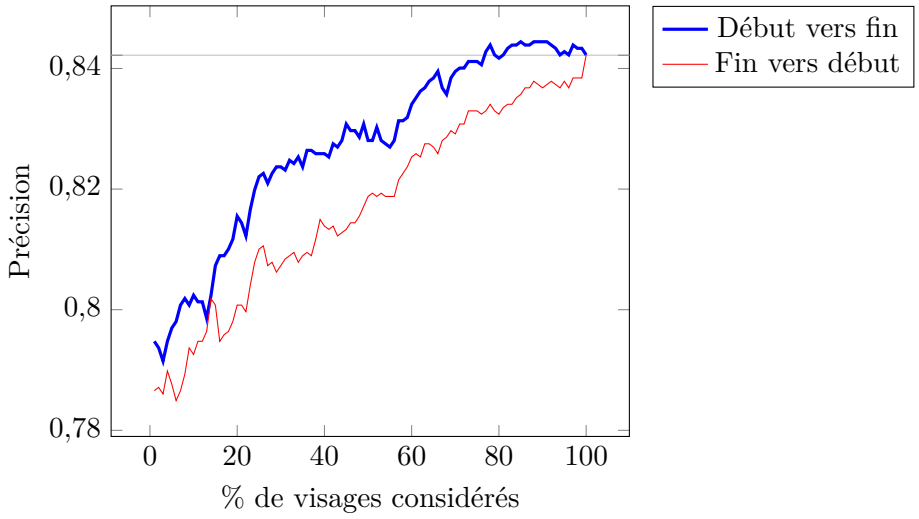

FIGURE 7.2 – Comparaison de la précision moyenne d'identification d'une OVP en fonction de la proportion de visages utilisés choisis selon l'ordre d'apparition et selon l'ordre inverse.

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

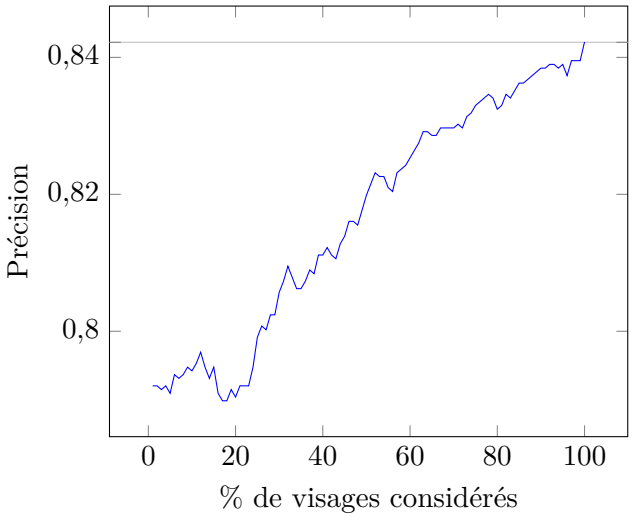

FIGURE 7.3 – Précision d'identification d'une OVP en fonction de la proportion de visages utilisés choisis du milieu de l'OVP vers les extrémités.

7.4.2 Du milieu vers les extrémités

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

7.4.3 De façon aléatoire

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

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

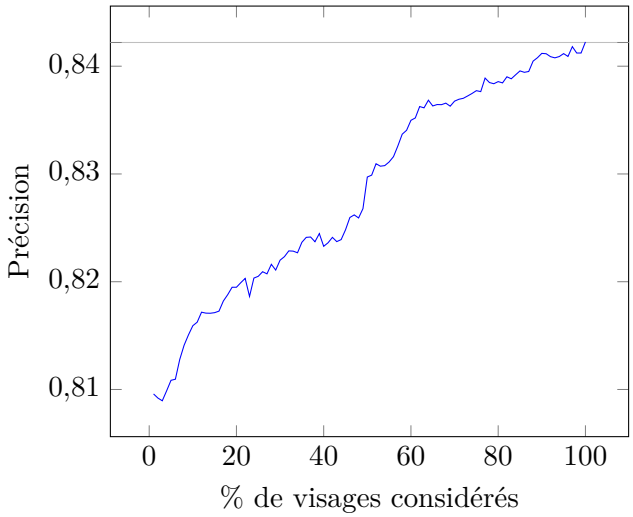

FIGURE 7.4 – Précision d'identification d'une OVP en fonction de la proportion de visages utilisés choisis aléatoirement.

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

7.4.4 Discussion sur la proportion de visages à considérer

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

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

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

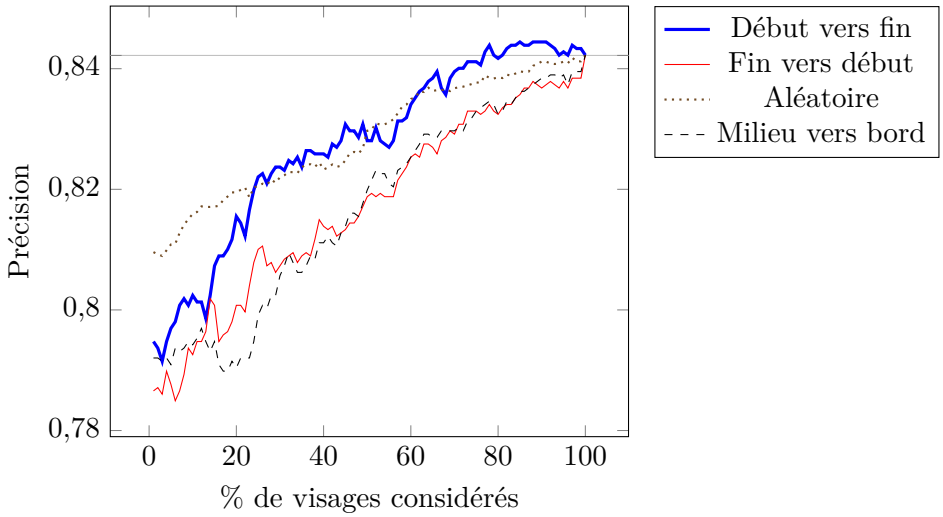

FIGURE 7.5 – Comparaison de la précision d'identification d'une OVP en fonction de la proportion de visages utilisés et de la stratégie de sélection.

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

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

7.5 Propagation d'identités à partir d'OVP nommées

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

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

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

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

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

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

majorité absolue.

7.5.1 Identités issues d'un vote à la majorité relative

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

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

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

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

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

7.5.2 Identités issues d'un vote à la majorité absolue

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

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

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

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

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

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

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

7.5.3 Discussion sur la détermination des identités initiales

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

7.6 Variation de la proportion d'occurrences utilisées

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

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

7.6.1 Propagation par sélection aléatoire

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

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

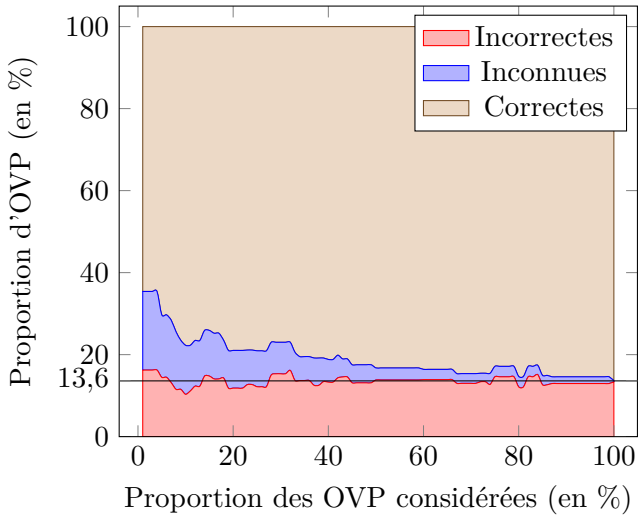

FIGURE 7.6 – Résultats de la propagation d'identités en utilisant un pourcentage des OVP choisies de façon aléatoire dans chaque groupe.

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

7.6.2 Propagation par ordre de similarité

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

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

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

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

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

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

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

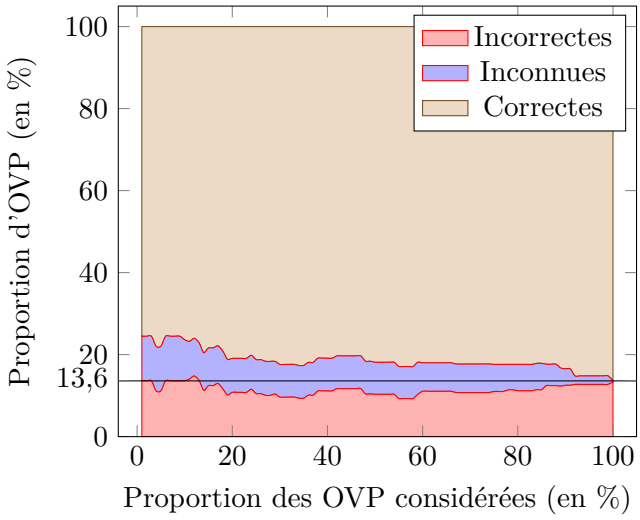

FIGURE 7.7 – Résultats de la propagation d'identités en fonction de la proportion d'OVP choisies par ordre décroissant de similarité à la moyenne dans chaque groupe.

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

7.6.3 Propagation par score de confiance

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

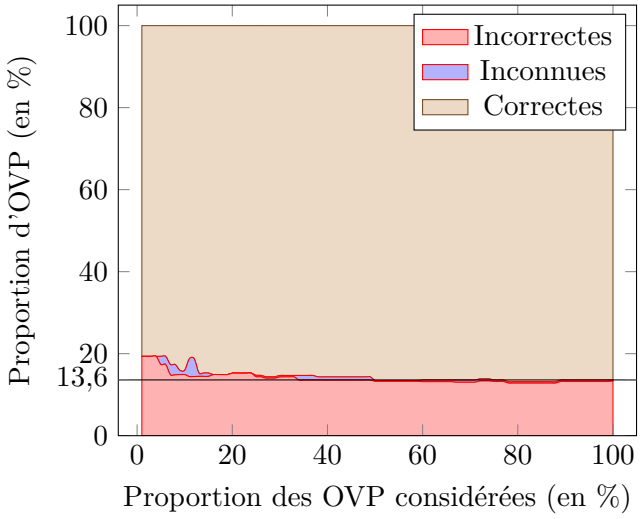

FIGURE 7.8 – Résultats de la propagation d'identités en fonction de la proportion d'OVP choisies par ordre décroissant de confiance dans chaque groupe.

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

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

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

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

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

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

7.6.4 Discussion sur les stratégies de propagation

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

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

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

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

7.7 Conclusion

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

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

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

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

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

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

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