Apprentissage en ligne d'une signature audiovisuelle pour la ré-identification de personne

François-Xavier Decroix[13] Frédéric Lerasle[23] Julien Pinquier[1] Isabelle Ferrané[1]

[1] Université de Toulouse, UPS, IRIT {decroix, pinquier, ferrane}@irit.fr
[2] CNRS, LAAS, lerasle@laas.fr [3] Université de Toulouse, UPS, LAAS

Résumé

L'intelligence ambiante pose entre autres le problème de la détection des activités humaines, l'enjeu est par exemple la gestion automatique de l'énergie ainsi que l'analyse des interactions entre les usagers partageant le lieu. Pour caractériser les interactions entre individus ou entre un individu et l'infrastructure d'un bâtiment, une tâche de ré-identification des usagers du lieu lors de leur déplacement est nécessaire et l'utilisation de modèles multimodaux permet clairement de robustifier cette ré-identification. Dans cet article, nous proposons une méthode de fusion audiovisuelle, introduisant un nouvel indice de confiance de zones de saillance audio-vidéo, pour l'apprentissage d'une signature audiovisuelle d'une personne.

Mots Clef

Signature audiovisuelle, fusion audio-vidéo, ré-identification de personne.

Abstract

In intelligent environments, activity detection is a necessary pre-processing step for adaptive energy management and interaction with humans. To characterize the interactions between individuals or between an individual and the infrastructure of a building, a re-identification process is required and using multimodal models improves its robustness. In this paper, we propose a method for audiovisual fusion, introducing a novel confidence index of audio-video salience zones, for training an audiovisual signature of a person.

Keywords

Audiovisual signature, audio-video fusion, person re-identification

1 Introduction

Le processus de ré-identification est défini comme l'association de nouvelles observations d'une personne détectée par un capteur ou un réseau de capteurs et d'observations antérieures du même individu. Contrairement à l'identification de personne, aucune information préalable n'est nécessaire et cette tâche peut être poursuivie en ligne.

Du fait de l'expansion des capteurs ambiants dans les domaines professionnels et privés, cette tâche peut alimenter plusieurs applications en surveillance [1], en extraction d'information dans le multimédia et en détection d'activité. Afin de tirer profit de la complémentarité des percepts des capteurs, des approches multimodales couplant des caméras RGB-D et thermiques [2] on encore des données RFID [3] ont émergé, améliorant la robustesse de la signature d'une personne, l'apparence n'étant pas toujours discriminante. Cependant, combiner des informations basées sur l'audio et sur la vidéo demeure complexe compte tenu de la dissemblance de ces deux modalités. Simulant l'activité et les interactions humaines, intrinsèquement multimodales, la fusion audiovisuelle s'est étendue à des applications allant des interfaces homme machine aux véhicules intelligents [4] ou encore aux maisons intelligentes [5].

Cet article propose une nouvelle méthode de fusion audiovisuelle pour la construction d'une signature bimodale dans un contexte de détection d'activité en intérieur dans des salles où les capteurs installés sont à champs partiellement joints pour limiter l'instrumentalisation du lieu. Nous décrivons un nouvel indice de confiance dans le domaine audiovisuel joint pour la construction d'une signature audiovisuelle d'une personne, et le validons sur notre propre base de données audiovisuelles.

L'article est structuré comme suit. Une signature audiovisuelle pour une personne est présentée dans la section 2. Nous décrivons ensuite notre approche de fusion audiovisuelle dans la section 3 et les expérimentations et évaluations associées ainsi que la conclusion sont données dans les sections 4 et 5 respectivement.

2 Signatures sonores et visuelles

L'architecture globale du système est représentée en Figure 1. Le sous-synoptique en bleu est détaillé dans cette section. Les deux autres parties, en vert et en rouge sont traitées dans la section 3.

L'apparence visuelle et le timbre de la voix sont deux modalités décorrélées : il est en effet impossible de prédire le comportement de l'une par l'observation de l'autre et réciproquement. Des informations de genre et d'âge sont exploitables dans certains cas, mais ne sont pas assez discriminantes pour de la ré-identification en environnement

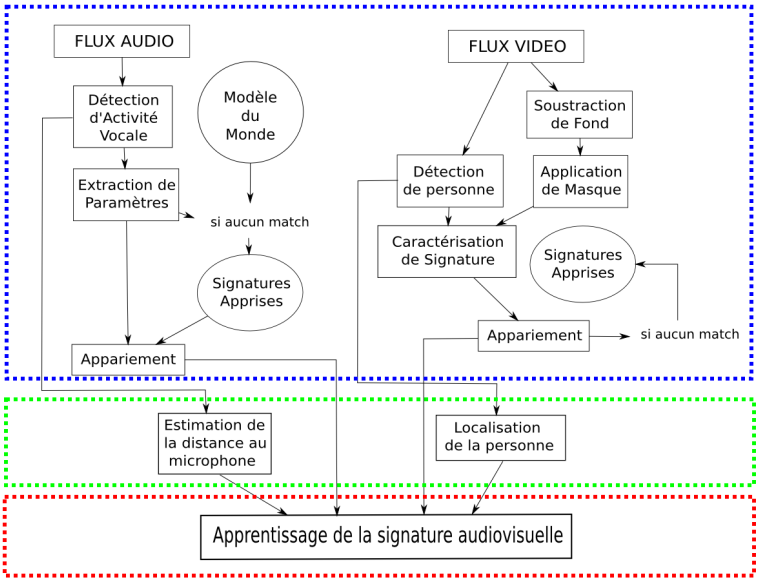

FIGURE 1 – Synoptique de notre système

non contrôlé. Ainsi une signature visuelle et une signature audio seront générées séparément et ensuite associées dans une étape de fusion tardive (voir la section 3).

2.1 Signature audio

Pour construire une représentation du locuteur indépendante du texte, nous utilisons des Modèles de Mélanges de lois Gaussiennes (GMM), approche largement utilisée et qui montre de bonnes performances sur les bases de données publiques de parole [6]. Un GMM est décrit ainsi :

$$p(x|\lambda_{aud}) = \sum_{i=1}^{M} w_i g(x_i|\mu_i, \Sigma_i) \qquad (1)$$

avec x notre vecteur de paramètres et w_i le poids associé à chaque composante gaussienne de densité $g(x_i|\mu_i, \Sigma_i)$; où μ_i est sa moyenne et Σ_i sa matrice de covariance. Notre signature audio pour un locuteur sera alors définie par :

$$\lambda_{aud} = \{w_i, \mu_i, \Sigma_i\} \qquad i = 1, ..., M. \qquad (2)$$

Les paramètres du vecteur x sont fondés sur une analyse cepstrale, en utilisant l'échelle perceptive MEL. Suivant le paramétrage décrit en [7], le vecteur de paramètres est composé de 50 valeurs : 19 coefficients statiques (MFCC), 19 dérivées premières, 11 dérivées secondes ainsi que la dérivée de l'énergie.

Plutôt que de créer le modèle de chaque locuteur à partir de rien, ce qui exigerait une grande quantité de données d'apprentissage pour chaque locuteur, nous apprenons un modèle du monde (Universal Background Model - UBM) à l'aide d'un ensemble de données de parole provenant de nombreux locuteurs. Dans notre étude, nous avons utilisé le corpus BREF [8] composé de 90 locuteurs différents (50 femmes et 40 hommes), pour un total de 167359 phrases.

Le modèle du monde donne une représentation générale du locuteur « moyen ». Ce modèle est estimé par l'algorithme d'espérance-maximisation (EM) [9] et est adapté par estimation du Maximum *A Posteriori* (MAP) pour obtenir le modèle du locuteur ciblé avec une quantité limitée de données d'apprentissage [10], en général 2 à 3 minutes par locuteur sont suffisantes.

Soit un vecteur de paramètres y d'une observation d'un segment Y de parole, le score de similarité est calculé comme le ratio de log-vraisemblance de l'hypothèse que Y soit prononcé par le locuteur appris vs. Y prononcé par un autre, plus proche du modèle du monde :

$$LLR(y) = log(p(y|\lambda_{GMM_x})) - log(p(y|\lambda_{UBM})) \qquad (3)$$

2.2 Signature vidéo

La ré-identification visuelle de personne a été un domaine de recherche prisé ces dernières années, s'appliquant à la robotique, au multimédia et en particulier à la vidéo-surveillance, résultant de l'émergence de nombreuses approches [11]. Cette tâche soulève plusieurs difficultés, en particulier dans les cas de réseaux de caméras à champs disjoints, comme des occultations partielles, des changements de luminosité, de posture, de point de vue, de réponse colorimétrique ou de scenarii non contrôlés.

Dans notre approche, la soustraction de fond, qui consiste à segmenter les régions appartenant au sujet en mouvement, est traitée par un clustering GMM en ligne du fond, détaillé en [12].

Pour la détection de personne, nous utilisons des histogrammes de gradient orienté [13]. La silhouette de la personne cible est caractérisée par la distribution locale de gradients d'intensité ou des directions des contours. Une Machine à Vecteurs de Support (SVM) linéaire est ensuite apprise pour la classification personne/non-personne.

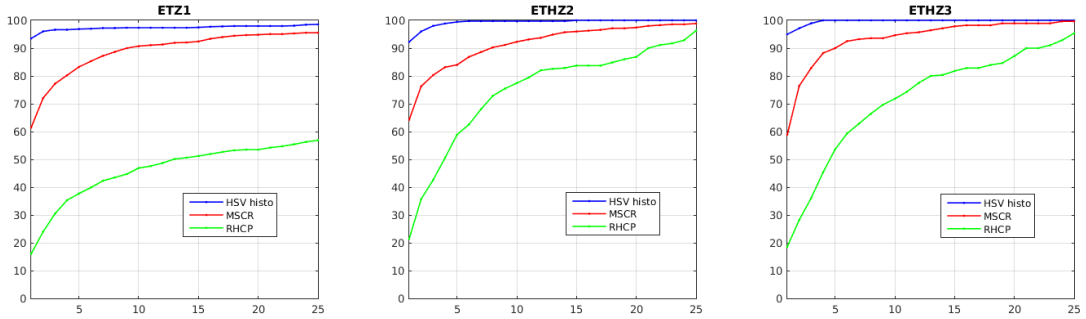

FIGURE 2 – CMC des histogrammes HSV en bleu, MSCR en rouge et RHSP en vert pour 3 sous-ensembles de la base publique ETHZ

Une fois le masque du premier plan appliqué sur la détection, nous extrayons des paramètres afin de construire des descripteurs discriminants de la personne. La conception de paramètres est un domaine vastement exploré et une revue de ceux-ci est présentée en [14]. La plupart d'entre eux suivent un modèle partitionné. Dans notre approche, à l'instar de la méthode SDALF décrite en [15], la silhouette est divisée en tête, torse et jambes par l'estimation des axes horizontaux séparant des régions de grandes différences d'apparence et d'aire similaire.

Le descripteur SDALF atteint d'excellentes performances dans l'état de l'art sur les bases publiques connues, notamment pour des images basse résolution, présentant des occultations, des changements de posture, de point de vue et d'éclairage. Il est généré par la combinaison des trois paramètres suivants, extraits des parties du corps :

- Histogrammes HSV : plus robustes aux changements de luminosité que ceux dans l'espace RGB par la séparation de l'information de couleur et d'intensité. Ils sont également pondérés par un noyau gaussien centré sur l'axe de symétrie pour mettre en valeur l'information contenue dans les pixels centraux, plus pertinente.
- Maximally Stable Color Region (MSCR) : introduit dans [16], ce descripteur évalue des distances de couleur entre des pixels d'une image pour en trouver les régions homogènes, alors modélisées par des ellipses. Ces dernières sont ensuite représentées par leur aire, centroïde, matrice de second moment et leur couleur.
- Recurrent High-Structured Patches (RHSP) : présenté dans [15] pour caractériser l'information de texture dans des zones à haute entropie en analysant les invariances de portions d'images tirées aléatoirement.

La Figure 2 présente les Courbes Cumulatives de Correspondance (CMC) pour les 3 paramètres sur trois sous-ensembles de la base de données ETHZ [17] comprenant un total de 8580 imagettes. Bien que les trois paramètres précédemment présentés portent des informations complémentaires, les histogrammes HSV seuls font preuve d'une efficacité satisfaisante, pour un coût de calcul bien plus faible que celui du descripteur RHSP en particulier. Dans notre contexte applicatif, un descripteur compact et à faible coût étant requis, nous n'allons donc plus prendre en compte les MSCR et RHCP par la suite.

Dans le but d'obtenir une signature visuelle compacte, nous appliquons un partitionnement via l'algorithme k-means sur un ensemble de descripteurs d'une personne unique, puis les k plus proches sont concaténés pour constituer une signature comportant de grandes variations de vue et de posture.

$$\lambda_{vid} = \{histo_i\} \qquad i = 1, ..., k. \tag{4}$$

Cette tâche permet à la signature visuelle de gagner en compacité en éliminant les images redondantes.

Plus grand est le nombre de vues différentes depuis les caméras, plus importante sera la valeur de k nécessaire. Dans le cas d'une caméra unique, installée à un angle d'inclinaison $\alpha = 25°$, empiriquement $k = 6$ clusters encodent correctement les variabilités de l'apparence.

Les similarités entre histogrammes sont évaluées par la distance de Bhattacharyya [18].

3 Vers une signature audiovisuelle

Évoqué en préambule, le but est d'associer les modèles audio et vidéo par fusion tardive. Par conséquent nous cherchons les zones environnementales où les signatures sont fonctionnelles et ainsi partitionner l'espace en zones de détection audio et/ou vidéo eu égard aux positionnements relatifs des caméras et microphones ambiants.

3.1 Localisation Audio

La localisation de source sonore est ordinairement réalisée à l'aide d'indices binauraux comme la différence de niveaux inter-auraux et la différence de temps inter-auraux. À l'image de la perception humaine [19] ces indices fournissent une estimation de l'azimut horizontal, pendant que l'élévation peut être inférée à partir du filtrage provoqué par les réflexions du son dans le pavillon [20].

Dans notre contexte de microphone unique, cependant, il est impossible d'évaluer de tels indices mais des para-

mètres acoustiques pour l'estimation de distance à un auditeur proche peuvent être extraits. Outre des descripteurs naïfs comme l'intensité sonore (la pression sonore subit une perte de 6dB en doublant la distance dans des espaces ouverts), cette distance présente une corrélation avec le ratio du son direct au son réverbéré (Direct-to-Reverberant energy Ratio - DRR) [21]. Une étude récente d'indices acoustiques pour la perception auditive de la distance chez les humains est présentée en [22]. Pour des signaux sonores de synthèse, avec un spectre et une enveloppe temporelle connus, ce ratio peut être estimé à travers la corrélation croisée inter-aurale, la variance spectrale, l'enveloppe spectrale, ou encore l'analyse de l'attaque et de la descente [23]. Ces indices sont cependant mis en défaut pour des signaux de parole, du fait de leur non-stationnarité, nous exploitons alors une mesure d'intelligibilité introduite en [24] comme le ratio de modulation d'énergie de parole sur la réverbération (Speech to Reverberation Modulation energy Ratio - SRMR).

Le spectre de modulation d'un signal de parole (le spectre de son enveloppe temporelle) dans des conditions dites « studio », a des composantes distribuées autour de 2-16Hz, avec des pics autour de 4Hz, la fréquence syllabique standard. L'ajout de réverbération blanchit le spectre de modulation et des composantes dans des bandes de fréquence de modulation plus élevées apparaissent comme le montre la Figure 3.

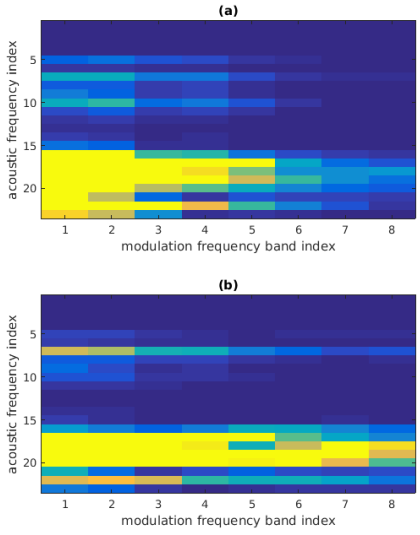

FIGURE 3 – Spectre de modulation de la même trame pour du signal propre (a) et réverbéré ((b)

La fraction jaune du spectre de modulation dépeint ses valeurs les plus importantes. Dans le cas du signal propre en (a), nous pouvons observer que la modulation d'énergie est restreinte aux 4 premières bandes de fréquence de modulation et qu'elle s'étend sur les 8 bandes pour de la parole réverbérée (voir le tableau 1 pour les correspondances index/fréquences de modulation).

Après filtrage du signal de parole par 23 filtres gammatones, le SRMR est défini comme le ratio de l'énergie dans les bandes de basses fréquences de modulation sur l'énergie dans les bandes de hautes fréquences de modulation :

$$SRMR = \frac{\sum_{k=1}^{4} \overline{\varepsilon}_k}{\sum_{k=5}^{8} \overline{\varepsilon}_k} \tag{5}$$

avec $\overline{\varepsilon}_k$ l'énergie de modulation moyenne dans la bande k.

TABLE 1 – Fréquences Centrales (f_c) et Bandes Passantes (BP) du Spectre de Modulation, en Hz

Index de Bande de Fréquence de Modulation							
1	2	3	4	5	6	7	8
f_c							
4,0	6,5	10,7	17,6	28,9	47,5	78,1	128,0
BP							
1,9	3,4	5,9	9,8	15,9	26,4	43,2	70,8

Une mise à jour de cette métrique est introduite en [25] où l'indépendance au texte et à la fréquence fondamentale est améliorée. Cette métrique est ostensiblement corrélée à la distance comme sur la Figure 4. Le SRMR est affiché en bleu et est évalué toutes les secondes à des distances dont la valeur inverse est affichée en rouge. Plus un locuteur est proche (autour de 50s dans notre exemple), plus il est intelligible. Le SRMR sera alors interprété comme un indice de confiance de la proximité au microphone.

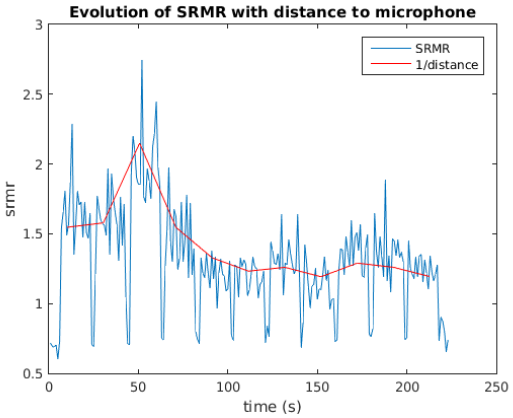

FIGURE 4 – SRMR de 11 itérations d'un segment de parole pour différentes distances et une longueur de trame de 1s

3.2 Localisation Vidéo

Contrairement aux percepts audio, les projections dans le plan du sol des détections visuelles peuvent être aisément inférées depuis la vue d'une caméra. Nous calibrons au

préalable le plan du sol relativement à la caméra par l'extraction de grille placée sur le plan du sol comme illustré sur la Figure 5.

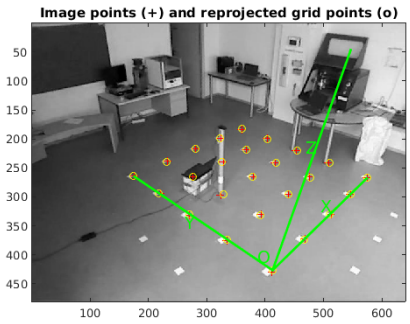

FIGURE 5 – Calibrage du plan du sol relativement à la caméra

3.3 Fusion Audiovisuelle

Comme introduit précédemment, nous traitons la fusion audiovisuelle comme une recherche de régions saillantes pour les deux modalités à chaque instant. La pièce est instrumentée par plusieurs caméras C et microphones M, placés de manière à avoir un sous-ensemble $\{M_K\}$ de M dans le champ de vue d'au moins une caméra.
Pour une caméra et un microphone M_k, considérons $p_{aud}(t, M_k)$ et $p_{vid}(t, M_k)$ les modèles des percepts audio et vidéo respectivement, à l'instant t. Leurs composantes sont :

- le meilleur candidat $\lambda_{aud,i}$ donné par la vérification de locuteur et le SRMR comme Indice de Confiance Audio (ICA) associé ;
- le meilleur candidat $\lambda_{vid,j}$ donné par la réidentification visuelle et l'inverse de la distance euclidienne au microphone M_k comme Indice de Confiance Visuel (ICV) associé ;

$$p_{aud}(t, M_k) = \begin{cases} \lambda_{aud,i} \\ ICA_{t,M_k} = SRMR_{t,M_k} \end{cases} \quad (6)$$

$$p_{vid}(t, M_k) = \begin{cases} \lambda_{vid,j} \\ ICV_{t,M_k} = \frac{1}{\sqrt{(x-x_M k)^2+(y-y_M k)^2}} \end{cases} \quad (7)$$

Notre mesure de saillance audiovisuelle jointe est alors définie par l'Indice de Confiance Audio Vidéo suivant :

$$ICAV_{t,M_k} = ICA_{t,M_k} * ICV_{t,M_k} \quad (8)$$

Pour des valeurs d'$ICAV_{t,M_k}$ supérieures à un seuil th, fonction de la confiance minimale autorisée, les percepts audio et vidéo sont joints pour définir la personne détectée comme le couple : $\lambda_{av,ij} = (\lambda_{aud,i}, \lambda_{vid,j})$.
Dans le cas où plusieurs détections sont présentes à l'instant t_1, chaque couple (i, j) potentiel est estimé par la probabilité :

$$p(\lambda_{av,ij}, t_1) = \frac{1}{N(t_1)} \quad (9)$$

avec $N(t_1)$ le nombre de couples potentiels à l'instant t_1. Pour chaque couple (i, j) potentiel observé de nouveau à un instant t_2, cette probabilité est ainsi mise à jour :

$$p(\lambda_{av,ij}, t_1 + t_2) = p(\lambda_{av,ij}, t_1) + p(\lambda_{av,ij}, t_2) - p(\lambda_{av,ij}, t_1) * p(\lambda_{av,ij}, t_2) \quad (10)$$

L'analyse spatio-temporelle permet de lever les ambiguïtés d'association et les couples les plus probables sont alors formés au terme de la phase d'apprentissage.

4 Expérimentations et évaluations

4.1 Mise en œuvre

Nous n'avons trouvé aucune base de données publique correspondant au contexte de notre problématique, par conséquent nous avons acquis des données vidéo et audio pour nos propres évaluations.
Dans une salle de réunion typique, de taille 6m par 6m, 2 capteurs sont placés, une caméra dans un coin de la pièce, à 2,5m de hauteur et d'inclinaison approximativement $\alpha = 25°$. Un microphone USB MXL-AC 404, utilisé généralement dans des vidéoconférences est placé au centre de la pièce (dans le champ de vision).
La base de donnée est composée de 3 participants se déplaçant de manière à couvrir l'espace de la pièce de manière exhaustive, diffusant dans le même temps 81 itérations d'un segment de parole propre issu de BREF, un corpus de parole lue en français [8], à l'aide d'une enceinte Bluetooth. Les données sont alors divisées en un ensemble d'apprentissage de 486 segments de parole avec 544 trames visuelles avec lequel est apprise la zone de verrouillage audiovisuel, et un ensemble de test de 243 segments de parole et 222 trames visuelles. La durée totale de la base de données est de 1h34.
Pour générer notre signature audio, nous utilisons ALIZE, une plate-forme open-source pour la reconnaissance de locuteur [7] ainsi que sa boîte à outils LIA RAL. Les paramètres audio sont calculés par la boîte à outil open-source SPro. Le modèle du monde est composé de 512 lois gaussiennes avec des matrices de covariance diagonales. Le détecteur HOG, ainsi que la soustraction de fond et les descripteurs visuels sont générés à l'aide de la librairie OpenCV, réunis dans un environnement MATLAB.
Les détections visuelles étant sujettes à de nombreuses fausses alarmes, nous ajoutons une étape de post-traitement pour les filtrer. Des fausses détections de personnes induites par l'arrière plan sont observées de manière récurrente. Elles présentent cependant des masques de premier plan nuls. Considérons le résultat des détections pour un participant. Sur 302 détections, 182 sont des fausses alarmes. Nous filtrons alors les détections en considérant que le masque doit comporter au moins un pourcentage minimum de pixels associés à des zones mobiles détectées, rejetant les autres. A partir d'un autre sous-ensemble de données de détections annotées, nous apprenons une Machine à Vecteurs de Support (SVM) en une dimension sur

la somme des pixels de premier plan pour évaluer le pourcentage minimum $p = 16\%$. 175 fausses détections sont ainsi éliminées, les 7 restantes présentent des détections partielles (torse ou jambes) qui admettent alors une proportion importante de pixels de premier plan.
La Figure 6 présente la répartition des pixels de premier plan sur les masques des détections. Nous distinguons aisément deux agglomérats, un correspondant aux fausses alarmes, en rouge, et un autre aux vrais positifs, en vert.

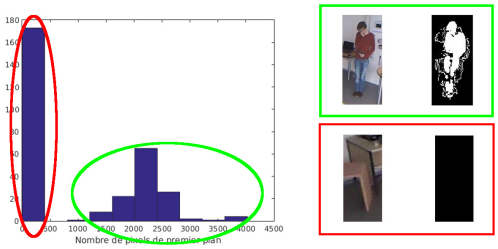

FIGURE 6 – Histogramme des pixels des masques, en rouge les fausses alarmes

4.2 Évaluations

Apprentissage audiovisuel. Pour les percepts vidéo et audio, les signatures des 3 sujets sont correctement apprises, toutes les détections sont correctement regroupées selon 3 modèles distincts. Les représentations sont conçues pour être robustes dans des conditions plus extrêmes que celles correspondant à notre protocole, d'où leur efficacité. En effet, ce protocole contient un nombre limité de sujets relativement dissemblants, ce qui peut être observé sur la Figure 7. Cette figure présente les imagettes correspondant aux histogrammes porteurs des signatures visuelles des 3 participants, où les variabilités internes de pose et de luminosité sont bien décrites par le regroupement k-means.

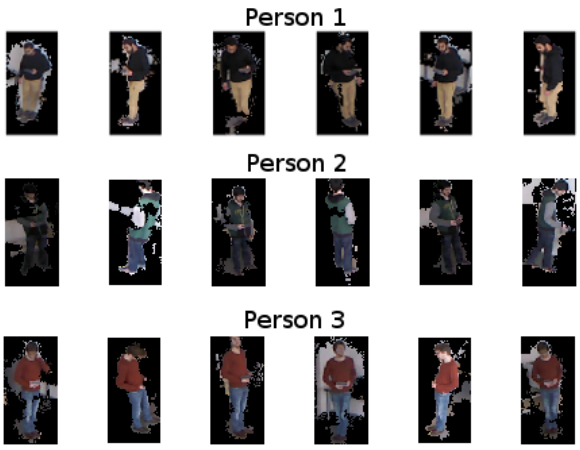

FIGURE 7 – Signatures visuelles des 3 participants

Les valeurs de l'ICAV sont évaluées à partir des percepts audio et vidéo du corpus d'apprentissage pour les 3 participants. Dans les zones aveugles, l'Indice de Confiance Visuel est fixé à zéro et l'ICAV n'est pas calculé à l'emplacement du microphone (position centrale), le locuteur et le microphone ne pouvant être confondus.
La zone de fusion audiovisuelle est délimitée par les valeurs de l'ICAV supérieures à un seuil th.
Pour évaluer son contour, nous apprenons une SVM à noyau gaussien sur les valeurs de l'ICAV des données d'apprentissage et évaluons le taux d'erreur de classification en fonction de th, comme le montre la FIgure 8.

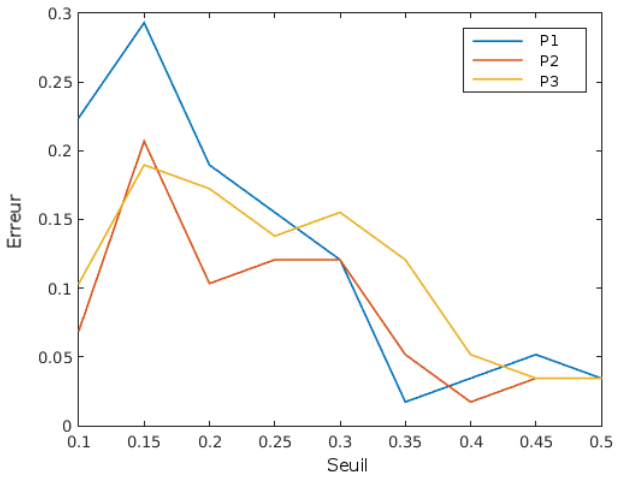

FIGURE 8 – Taux d'erreur de classification

Le seuil est fixé à : $th = 0.4$, pour lequel le taux d'erreur de classification est inférieur à 10% pour les 3 participants et les contours de la zone de fusion audiovisuelle est affichée sur la Figure d).
The threshold is fixed at : $th = 0.4$, for which the classification error rate is lower than 10% for the 3 participants and the contour of the audiovisual fusion zone displayed in Figure 9 d).

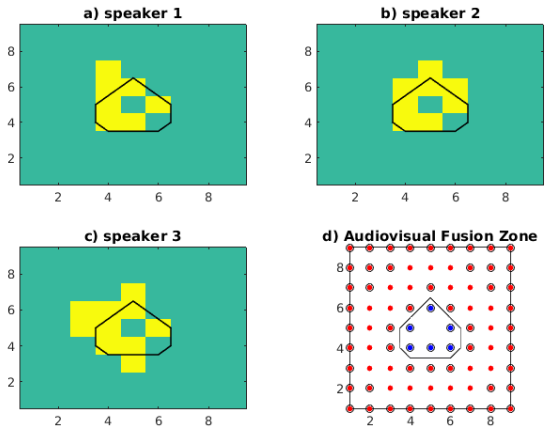

FIGURE 9 – a), b) and c) : Cartes binaires par ICAV pour les trois participants, d) zone de fusion (bleu)

Vérification conjointe ID/zone. La zone de fusion audiovisuelle est alors confrontée à nos données de test. A partir des données ICAV calculées sur ce corpus et du seuil appris th, des cartes binaires sont extraites et montrées en Figure 9 a), b) et c). Les résultats de la classification conjointe ID/zone sont présentés dans la Table 2. Les erreurs décrivent les mauvaises associations audiovisuelles ainsi que les fausses estimations de zone.

	Emplacements	Zones aveugles	Taux d'erreur
$P1$	81	23	0.051
$P2$	81	23	0.069
$P3$	81	23	0.103

TABLE 2 – Résultats de classification

5 Conclusion

Dans cet article, nous avons présenté une méthode d'apprentissage d'une signature audiovisuelle d'une personne en couplant des approches de l'état de l'art pour les deux modalités évaluées séparément et fusionnées tardivement.
La contribution principale de cet article réside dans l'utilisation d'un nouvel indice de confiance audiovisuel pour la recherche de régions de saillance pour les percepts vidéo et audio simultanément. Il est basé sur la cohérence de la localisation visuelle et l'estimation de la distance auditive à plusieurs instants de la phase d'apprentissage.
Nos travaux futurs se concentreront sur l'analyse spatio-temporelle des trajectoires audio-vidéo pour la fusion, puis sur le transport des signatures dans un réseau éparse de capteurs. À l'aide de ces signatures nous projetons d'effectuer une ré-identification multimodale multi-cibles dans une pièce afin d'en inférer l'activité humaine en temps réel pour des réunions, des conférences ou des groupes de travail à partir des interactions entre les participants dominants.

Remerciements

Les auteurs souhaitent remercier l'opération neOCampus [1] pour son support financier.

Références

[1] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recogn. Lett.*, vol. 33, no. 14, pp. 1828–1837, Oct. 2012.

[2] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapes, and S. Escalera, "Tri-modal person re-identification with rgb, depth and thermal features," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 301–307.

[3] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat, "Vision and {RFID} data fusion for tracking people in crowds by a mobile robot," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 641 – 651, 2010, special Issue on Multi-Camera and Multi-Modal Sensor Fusion.

[4] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, June 2010, pp. 174–178.

[5] C. Busso, S. Hernanz, C.-W. Chu, S. il Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, "Smart room : participant and speaker localization and identification," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 2, March 2005, pp. ii/1117–ii/1120 Vol. 2.

[6] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1-2, pp. 91–108, Aug. 1995.

[7] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, March 2005, pp. 737–740.

[8] L. F. Lamel, J. luc Gauvain, M. Eskenazi, and M. E. Limsi-cnrs, "Bref, a large vocabulary spoken corpus for french," pp. 505–508.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[11] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270 – 286, 2014.

[12] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, Aug 2004, pp. 28–31 Vol.2.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1.

[14] R. Satta, "Appearance descriptors for person re-identification : a comprehensive review," *CoRR*, 2013.

1. https://www.irit.fr/neocampus/

[15] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 2360–2367.

[16] P.-E. Forssen, "Maximally stable colour regions for recognition and matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[17] W. Schwartz and L. Davis, "Learning Discriminative Appearance-Based Models Using Partial Least Squares," in *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[18] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhya : The Indian Journal of Statistics (1933-1960)*, vol. 7, no. 4, pp. 401–406, 1946.

[19] L. Rayleigh, "On our perception of sound direction," *Philosophical Magazine Series 6*, vol. 13, no. 74, pp. 214–232, 1907.

[20] A. Saxena and A. Ng, "Learning sound location from a single microphone," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, May 2009, pp. 1737–1742.

[21] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415.

[22] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans : a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, pp. 1–23, 2015.

[23] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.

[24] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766–1774, Sept 2010.

[25] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014.