Traitement du signal audio-visuel et visiophone personne libre

J. E. Viallet, M. Collobert, R. Feraud, O. Bernier,
Y. Mahieux*, G. Le Tourneur*, A. Gilloire* et D. Collobert

DLT/DLI/TNT, * DIH/CMC
France Télécom - Centre National d'Etudes des Télécommunications
BP 40 - 22307 Lannion Cedex - France
E-mail: viallet@lannion.cnet.fr

RÉSUMÉ

Les informations visuelles et acoustiques sont au coeur de la (télé)communication entre les personnes. Le visage est la principale source d'information. Des techniques de détection du mouvement et de la teinte de la peau délimitent des régions d'intérêt où peuvent se trouver des visages. Un réseau de neurones détecte le visage et fournit la position et l'échelle du visage. Le visage repéré est suivi, en temps réel, par une caméra motorisée et par une antenne acoustique qui génère un lobe orientable. La prise de vue et la prise de son sont ainsi centrées en permanence sur l'utilisateur qui est libre de se déplacer et libre de tout équipement spécifique. Le traitement du signal audio-visuel sont intégrées à LISTEN, démonstrateur du visiophone "personne libre".

ABSTRACT

Both visual and acoustical informations provide effective means of telecommunication between persons. In this context, the face provides the most important visual and acoustical informations. Regions of interest obtained by movement and skin color detection delimit parts of an image where faces can be found. A neural network detects the face position and scale. The detected face is then tracked, in real time, by a motorized camera and by the steerable beam of a microphone array. Thus video and audio acquisition are always focussed on the user which is free to move and free from specific equipments. These processing techniques of the audio-visual signal are the kernel of LISTEN, a "free person" videophone.

1 Introduction

L'image a un rôle déterminant pour les nouveaux services de télécommunications. Les prises de vue et de son des visiophones actuels sont fixes et obligent l'utilisateur à se maintenir dans les champs visuels et acoustiques.

Nous proposons que ce ne soit plus l'utilisateur qui s'adapte à son terminal, mais l'inverse : l'utilisateur doit pouvoir communiquer librement là où il se trouve. Pour cela, il est nécessaire de localiser la ou les personnes présentes. Or, les personnes et la pièce dans laquelle ils se trouvent constituent une scène visuelle et acoustique a priori complexe à analyser. La localisation acoustique des personnes est difficile et peu précise en raison de la multiplicité des sources et des chemins acoustiques et de l'intermittence de l'activité de parole. Par contre, une personne est toujours caractérisée par son image, même si son visage est plus ou moins de profil. C'est pourquoi nous privilégions la détection visuelle des visages pour localiser les personnes [2].

La collaboration de techniques de traitement de l'image et du son permet d'améliorer des services tels que le visiophone et la visioconférence. Ces améliorations sont intégrées dans un démonstrateur de visiophone "personne libre", développé à Lannion : LISTEN (Locating Individual Speakers and Tracking ENvironment) [5].

2 Traitement de l'image

Pour suivre une personne qui se lève de sa chaise, se déplace dans une pièce, se rend à un tableau, il est nécessaire de savoir détecter rapidement son visage. Or, la plupart des nombreuses méthodes ne localisent rapidement que les visages vus de face et en gros plan [3].

C'est pourquoi nous décomposons le traitement. D'une part, des modules de détection du mouvement et de la teinte de la peau, modules de bas niveau, analysent en temps réel le flux vidéo et délimitent des régions d'intérêt. D'autre part, seules les régions d'intérêt sont alors explorées par un mélange conditionnel de réseaux de neurones, pour détecter des visages qui peuvent être de profil et éloignés de la caméra. Une région d'intérêt est une partie de l'image, connexe après un traitement spécifique (figure 1.d). Ainsi, pour une personne, la détection de la teinte de la peau fournit typiquement trois régions d'intérêt : le visage et les deux mains.

2.1 Détection de mouvement

La fonction de visiophonie "personne libre" permet à une personne, libre de se mouvoir, de communiquer. Cette faculté laissée à l'utilisateur se traduit par une propriété dont nous tirons parti. Une différence entre deux images successives permet de désigner à d'autres modules de traitement des parties actives de l'image.

FIG. 1 — [a] : Image d'origine issue d'une séquence. [b] : Détection de la teinte de la peau : visage et main de la personne et de la Joconde. [c] : Détection du mouvement entre deux images : la personne et l'écran (non synchronisé) sont détectés. [d] : Détection des zones d'intérêts : parties connexes de teinte de la peau en mouvement : la Joconde immobile n'est pas retenue.

La détection de mouvement (figure1.c) permet d'éliminer les parties du décor de teinte de la peau ainsi que des illustrations fixes de visages (tableaux, affiches) . Un seuil prend en compte le bruit des capteurs CCD. La détection de mouvement ne se fait que lorsque la caméra est immobile, sans changement de focale et de mise au point. Des variations brutales de l'éclairage limitent la portée de ce type de traitement.

2.2 Détection de la teinte de la peau

La teinte de la peau est caractéristique de tous les individus et est rapide à évaluer. Plusieurs auteurs [7], [10] ont montré que la teinte de la peau est un bon indicateur de la présence du visage mais également des mains (figure1.b). La représentation des couleurs dans l'espace YUV présente plusieurs avantages. La surface U.V est relativement robuste aux changements d'éclairage et est peu sensible à l'origine des personnes (la variation principale porte alors sur Y). De plus, la représentation YUV, disponible en sortie de la carte de numérisation, évite toute conversion, pénalisante pour un système temps réel, vers un autre espace de couleurs. Le module de détection de la teinte de la peau est adapté semi-automatiquement à chaque caméra et à chaque pièce.

2.3 Détection de visage

Le détecteur de visage par réseau de neurones repère un visage, de face ou de profil, jusqu'à 4 mètres de la caméra (pour un champ de 60 degrés et une image au format CIF). Son taux de fausses alarmes particulièrement faible garantit que ce qui est détecté est un visage et non une main.

Deux premiers réseaux de neurones évaluent la probabilité qu'une imagette de 20*15 pixels de large (après rehaussement et lissage) soit respectivement un visage vu de face, soit un visage vu de profil. Les réseaux autoassociatifs réalisent une réduction de dimensionnalité non-linéaire. Lors de l'apprentissage, afin que la réduction effectuée par le réseau ne génère que des images de visages et que l'algorithme s'oriente vers une solution non linéaire, des contres-exemples sont contraints à être reconstruits comme le voisinage du visage le plus proche. Le réseau évalue une distance entre l'imagette et l'ensemble des visages. Des exemples de non-visages sont aléatoirement sélectionnés parmi les fausses alarmes obtenus sur des images sans visage. Le réseau est entraîné avec ces exemples et le processus est répété jusqu'à ce que le taux de fausse alarme soit suffisamment faible. Un troisième réseau,

de type porte, calcule la probabilité conditionnelle des deux précédents réseaux [6].

L'ensemble des réseaux détecte des visages orientés jusqu'à 50 degrés. Le taux de détection de l'ordre de 87% et le taux de fausse alarme de l'ordre de 10^{-6}, obtenus sur la base de test A du CMU [4], améliorent les résultats obtenus par l'auteur de cette base [9]. Le taux de détection atteint 98% sur les images de *usenix face database*[4].

Une heuristique explore les régions d'intérêt à différentes échelles. La détection d'un visage au sein de la plus grande région d'intérêt se fait typiquement en moins d'une demi seconde.

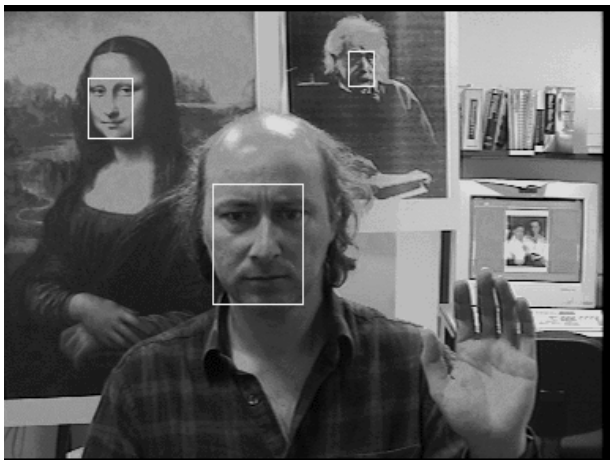

FIG. 2 — En analysant l'ensemble de la figure 1.a, les visages de la personne, de la Joconde et d'Einstein sont détectés. Seul le visage de la personne aurait été détecté si le réseau de neurones avait été appliqué aux seules zones d'intérêt de la figure 1.d.

2.4 Poursuite, prise de vue et contrôle de la caméra

Plusieurs personnes peuvent être dans la pièce. LISTEN considère que c'est la personne la plus proche de la caméra (ou dont l'image est la plus grande) qui est suivie. La région d'intérêt qui correspondant au visage est alors suivie en analysant en permanence les régions d'intérêt.

L'image a deux fonctions : la communication et l'analyse de la scène. L'analyse, nécessaire pour suivre la personne qui se déplace, est facilitée par un champ large et des mouvements rapides de caméra. La communication impose que l'image soit centrée sur l'utilisateur et que les mouvements de caméra soient souples et agréables à regarder. Ces fonctions d'analyse et de prise de vue, bien que générant des contraintes opposées, sont réalisées cycliquement avec une seule caméra en activant et inhibant les modules de détection et de commande. Une caméra motorisée, commandée en site et azimut, suit la personne. En contrôlant la focale, le cadre obtenu rend compte de la gestuelle de la personne et conserve une même taille au visage alors que la personne s'éloigne ou se rapproche de la caméra. Un contrôle de l'exposition (diaphragme et gain) compense les variations d'éclairage.

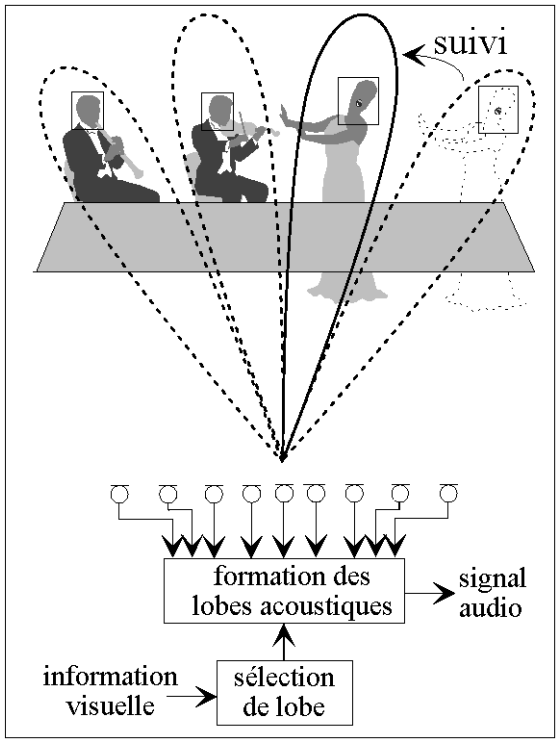

FIG. 3 — Schéma de principe de l'antenne acoustique.

3 Prise de son par antenne acoustique

Une prise de son directive est réalisée par le lobe orientable d'une antenne acoustique développée au CNET. L'antenne acoustique, composée de neuf microphones unidirectionnels, est organisée en quatre sous-antennes [8]. Chaque sous-antenne, formée de cinq microphones et caractérisée par un espacement inter-capteur, est associée à une bande de fréquence au moyen d'un filtre basse-bande. Ainsi, un comportement homogène de l'antenne est obtenue sur une large plage de fréquence.

Le lobe de l'antenne est orientable, au moyen de filtres d'interpolation, dans treize directions uniformément réparties sur le demi-plan horizontal face à l'antenne $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Afin d'optimiser la directivité spatiale de l'antenne, les filtres de chaque sous-antenne sont adaptés pour chacune des directions. Selon la direction pointée, la largeur du lobe est comprise entre 20 et 30 degrés (à 1 kHz). Le traitement est réalisé par un processeur en virgule fixe. Les paramètres de chaque lobe (filtres d'interpolation et filtres de sous-antenne) sont stockés dans une mémoire associée. La commutation du lobe dans la direction fournie par le localisateur visuel s'effectue instantanément. Le rythme de mise à jour du lobe dépend de la fréquence à laquelle la position du visage de la personne est connue. La cadence de traitement de LISTEN (image PAL au format CIF), proche de 25 Hz, améliore la performance de 5 Hz rapportée par [1].

FIG. 4 — Le visiophone “personne libre” LISTEN pilote l’antenne acoustique à partir des informations visuelles fournies par la caméra motorisée. ©France Télécom-CNET M. Le Gal.

4 LISTEN un visiophone personne libre

Les techniques de traitement de l’image et du son pour la visiophonie “personne libre” autorisent une prise de vue et une prise de son centrées en permanence sur la personne. La personne est libre de se déplacer, libre de tout équipement. Elle dialogue avec l’image de son interlocuteur et ne s’adresse plus à une caméra et un microphone fixes. Ces techniques sont intégrées au sein de LISTEN, démonstrateur de visiophone “personne libre”. L’évaluation de l’impact de ces nouvelles techniques sur la télécommunication est examinée au moyen de deux démonstrateurs LISTEN, opérationnels depuis un an. L’un se trouve dans une salle de réunion, où plusieurs personnes peuvent communiquer, se déplacer au tableau, ... L’autre est utilisé dans le cadre d’un poste de travail individuel.

5 Perspectives et Conclusion

Lorsque plusieurs personnes se trouvent initialement dans le champ de la camèra, notre détecteur de visage est capable de les localiser. Mais avec une seule caméra motorisée et un seul lobe acoustique, il n’est possible de suivre qu’une seule personne à la fois. Pour suivre simultanément plusieurs personnes il est souhaitable d’analyser la scène avec une caméra fixe disposant d’un champ très large. La prise de vue est alors assurée par une ou plusieurs caméras motorisées. Une antenne acoustique effectue la prise de son en orientant un lobe sur chacune des personne détectées par le localisateur visuel. Les intervenants seraient, par exemple, déterminés par les lobes dont l’activité acoustique sera la plus importante. L’image et la parole des intervenants pourront alors être transmis jusqu’à ce que d’autres participants interviennent. Couplées avec la reconnaissance de la parole, les techniques de visiophonie “personne libre” permettront le développement d’interfaces hommes machines innovantes.

Références

[1] S. Basu, M. Casey W. Gardner, A. Azarbayejani, and A. Pentland, “Vision-Steered Audio for Interactive Environments”, *IMAGE’COM 96*, Bordeaux, France, Mai 1996.

[2] U. Bub, M. Hunke, and A. Waibel, “Knowing Who to Listen To in Speech Recognition : Visually guided beamforming”, in *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, Michigan, 1995.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, “Human and Machine Recognition of Faces : A Survey”, *Proceedings of the IEEE*, Vol. 83(5), Mai 1995.

[4] CMU test sets et Usenix face database, *http ://www.cs.rug.nl/∼peterkr/FACE/face.html*.

[5] M. Collobert, R. Feraud, G. Le Tourneur, O. Bernier, J. E. Viallet, Y. Mahieux, D. Collobert, “LISTEN : A System for Locating and Tracking Individual Speakers”, *Proceedings of the second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, Octobre 1996.

[6] R. Feraud, O. Bernier, J. E. Viallet,M. Collobert, D. Collobert, “ A Conditional Ensemble of Neural Networks for Face Detection, Applied to Locating and Tracking an Individual Speaker”, à paraitre *Proceedings of CAIP*, Kiel, Septembre 1997.

[7] M. Hunke, and A. Waibel, “Face Locating and Tracking for Human-Computer Interaction”, *Proceedings of the 28th Asimolar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, Novembre 1994.

[8] Y. Mahieux, G. Le Tourneur, and A. Saliou, “A Microphone array for Multimedia Workstations”, *Journal of the AES*, Vol. 44(5), pp. 365-372, 1996.

[9] H. Rowley, S. Baluja, and T. Kanade, “Human Faces Detection in Visual Scenes”, *Advances in Neural Information Processing Systems 8*, 1995.

[10] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder : Real-Time Tracking of the Human Body”, *SPIE Photonics East*, Vol. 2615, pp. 89-98, 1995.