Modélisation de visage par fusion d'information couleur dans le cadre de la théorie de l'évidence et suivi par filtrage particulaire

Robust face tracking using color fusion and particle filter

F. Faux F. Luthon[1]

[1] Laboratoire LIUPPA (EA 3000)

IUT Informatique, Château Neuf, Place Paul Bert, 64100 Bayonne, France
faux@iutbayonne.univ-pau.fr

Résumé

Cet article présente un système de détection et de suivi en temps réel de visages dans une séquence vidéo. L'approche proposée consiste en la modélisation de la teinte chair du visage par un processus de fusion pixel de trois informations couleur, dans le cadre de la théorie de l'évidence de Dempster-Shafer. Pour cela, deux phases sont utilisées. Une phase d'initialisation simple et rapide, prend en compte au plus près la réalité terrain. L'utilisateur sélectionne manuellement sur une image une zone ombrée, une zone surexposée puis une zone d'intensité moyenne du visage. La teinte chair du visage est alors modélisée grâce au processus de fusion. Une phase de suivi sur la séquence vidéo utilise l'idée fréquemment rencontrée que les contours extérieurs d'un visage sont approximés par une ellipse englobant la région de teinte chair. Une méthode directe d'ajustement par les moindres carrés rapide et robuste aux bruits permet de dimensionner l'ellipse dans le contexte de fond complexe où certains contours issus de fausses détections perturbent le dimensionnement de la courbe. Ensuite, les paramètres de l'ellipse (centre, axe mineur, grand axe et orientation) sont utilisés par un algorithme de filtrage particulaire afin de gérer de manière robuste le suivi dynamique de la position, de la taille et de la pose du visage. L'originalité de la méthode réside en la modélisation de la teinte chair par la fusion de trois sources indépendantes du point de vue cognitif. De plus, les fonctions de masse (théorie de Dempster-Shafer) sont déterminées à partir de modèles a priori intégrant des données contextuelles spécifiques au visage. Ainsi, la particularité du visage de présenter des zones ombrées (cou) ou surexposées (nez, front) est prise en compte et la sensibilité face aux conditions d'éclairement est diminuée.

Les résultats de la modélisation de la teinte chair, de la fusion, de l'ajustement par une ellipse et du suivi seront illustrés et commentés dans cet article. Les limites ainsi que les évolutions de la méthode sont aussi développées en conclusion.

Mots Clefs

Détection de visage, Dempster-Shafer, fusion, suivi de visage, condensation, couleur.

Abstract

This paper describes a real time face detection and tracking system. The method consists in modelling the skin face by a pixel fusion process of three colour sources within the framework of the Demster-Shafer theory. The algorithm is composed of two phases. In a simple and fast initialising stage, the user selects successively on an image, a shadowy, an overexposed and a zone of mean intensity of the face. Then the fusion process models the face skin colour. Next, on the video sequence, a tracking phase uses the key idea that the face exterior edges are well approximated as an ellipse including the skin colour blob resulting from the fusion process. As ellipse detection get easily distracted in clutterd environnements by edges caused by non-faces objects, a simple and fast efficient least square method for fitting ellipse is used. The ellipse parameters (center, minor axe, major axe, orintation) are taken into account by a stochastic algorithm using a particle filter in order to realise a robust face tracking in position, size and pose.The originality of the method consists in modelling the skin face by a pixel fusion process of three cognitive independant colour sources. Moreover, mass sets are determined from a priori models taking into account contextual variables specific to the face under study. Hence, the face particularity which is to present shadowy (neck) and overexposed zones (nose, front) is considered, so sensitivity to lighting conditions decreases. Results of skin face modelling, fusion, ellipse fitting and tracking are illustrated and discussed in this paper. The limits of the method and future works are also commented in conclusion.

Keywords

Face detection, Dempster-Shafer, fusion, face tracking, condensation, skin hue.

1 Introduction

La détection d'un visage dans une image ou une séquence vidéo est nécessaire à de nombreuses applications telles que l'interaction homme machine (IHM), la reconnaissance, l'identification, la visioconférence, la robotique ou la télésurveillance.

Plus de 150 méthodes de détection [1, 2], allant des techniques dites de bas niveau utilisant des primitives telles que la texture, les contours, le mouvement, la couleur [3, 4] jusqu'aux approches de haut niveau telles que les modèles d'apparence, les réseaux de neurones ou les SVM (Support Vector Machines) [5, 6, 7, 8], sont proposées dans la littérature.

Cependant, il demeure très difficile de réaliser un algorithme insensible aux occlusions, aux variations de pose, d'échelle, d'orientation, à la présence d'un fond complexe, aux modifications des conditions d'éclairement.

Le visage est un objet non rigide dont la particularité est de présenter des zones ombrées (cou) ou surexposées (front, nez) dont la localisation est variable essentiellement en raison du mouvement dans une séquence vidéo.

Dans cet article, ces caractéristiques comportementales sont prises en compte en intégrant des variables contextuelles dans un processus de fusion de trois informations couleur afin de modéliser fidèlement la teinte chair du visage.

En effet, la fusion de données combine l'information issue de différentes sources dans le but de prendre une décision.

Ainsi, l'utilisation conjointe de plusieurs sources, données ou connaissances partielles, permet une meilleure compréhension du phénomène observé à condition que ces mesures soient fiables même si elles sont peu précises.

Cependant les informations à combiner ne sont jamais parfaites et les imperfections prennent différentes formes incluant principalement l'ambiguïté, l'imprécision, l'incertitude ou l'incomplétude.

Analyser et exploiter ces imperfections doit permettre d'optimiser la détection de visage.

C'est pourquoi la modélisation proposée dans cet article utilise la fusion au niveau pixel de données hétérogènes de teinte chair dans le cadre de la théorie de l'évidence de Dempster-Shafer [9, 10].

En effet, les informations couleur sont incontestablement pertinentes [11] pour l'analyse de visage en vidéo mais les transformations couleur standards demeurent sensibles aux conditions d'éclairement et très bruitées dans les zones ombrées. Afin de pallier ces inconvénients les informations de couleur de peau sont représentées dans l'espace logarithmique couleur LUX [12]. L'algorithme comprend deux phases. Lors d'une étape d'apprentissage simple et rapide prenant en compte au plus près la réalité terrain, l'utilisateur sélectionne manuellement sur une image une zone ombrée, une zone surexposée puis une zone d'intensité moyenne du visage. La teinte chair du visage est alors modélisée grâce au processus de fusion.

La fiabilité des capteurs par rapport au contexte est modélisée par des degrés de confiance.

Cette modélisation permet de segmenter l'image en régions de teinte chair ("blobs").

Une seconde phase de suivi utilise l'idée fréquemment rencontrée [13, 14] que les contours extérieurs d'un visage sont approximés par une ellipse englobant la région de teinte chair. Une méthode directe d'ajustement par les moindres carrés [15] rapide et robuste aux bruits permet de dimensionner l'ellipse dans le contexte de fond complexe où certains contours issus de fausses détections perturbent le dimensionnement de la courbe.

Ensuite, les paramètres de l'ellipse (centre, axe mineur, grand axe et orientation) sont pris en compte par un algorithme de filtrage particulaire afin de gérer de manière robuste le suivi dynamique de la position, de la taille et de la pose du visage.

L'article est organisé comme suit. Le chapitre 2 est consacré à un rappel sur la théorie de l'évidence, le chapitre 3 développe le processus de modélisation et de fusion, le chapitre 4 présente les résultats de cette modélisation. Après avoir rappelé le formalisme du filtrage particulaire, le chapitre 5 décrit son application au suivi de visage et présente les résultats expérimentaux. Les évolutions et perspectives de la méthode font office de conclusion au chapitre 6.

2 Théorie de l'évidence

La théorie de l'évidence de Dempster-Shafer (DS) a été introduite par Dempster et formalisée par Shafer. Elle représente à la fois l'imprécision et l'incertitude à l'aide de fonctions de masse m, de plausibilité Pls et de croyance Bel. Cette théorie se décompose en trois étapes : la définition des fonctions de masse, la combinaison d'informations et la décision.

2.1 Définition des fonctions de masse

L'ensemble des hypothèses pour une source (typiquement une classe dans un problème de classification multisource) est défini sur l'espace Ω appelé espace de discernement.

Posons $\Omega = \{\omega_1, \omega_2, ..., \omega_k, ...\omega_N\}$ où ω_k désigne une hypothèse en faveur de laquelle une décision peut être prise.

Les fonctions de masse sont définies sur tous les sous-ensembles de l'espace Ω et non seulement sur les singletons comme dans les probabilités.

Une fonction de masse m est définie comme une fonction de 2^Ω dans [0,1]. En général on impose $m(\emptyset) = 0$ et une normalisation de la forme :

$$\sum_{A \subseteq \Omega} m(A) = 1$$

Une fonction de croyance Bel est une fonction totalement croissante de 2^Ω dans [0,1] définie par :

$$\forall A_1 \in 2^\Omega, ..., A_K \in 2^\Omega,$$
$$Bel(\cup_{i=1...K} A_i) \geq \sum_{I \subseteq \{1...K\}, I \neq 0} (-1)^{|I|+1} Bel(\cap_{i \in I} A_i)$$

où $|I|$ désigne le cardinal de I et $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$. Etant donné une fonction de masse m, la fonction Bel définie par :

$$\forall A \in 2^{\Omega}, Bel(A) = \sum_{B \subseteq A, B \neq 0} m(B)$$

est une fonction de croyance.
Inversement, à partir d'une fonction de croyance Bel, on peut définir une fonction de masse m par :

$$\forall A \in 2^{\Omega}, m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} Bel(B)$$

Une fonction de Plausibilité Pls est également une fonction de 2^{Ω} dans [0,1] définie par :

$$\forall A \in 2^{\Omega}, Pls(A) = \sum_{B \cap A \neq \emptyset} m(B)$$

La plausibilité mesure la confiance maximum que l'on peut avoir en A.
La possibilité d'affecter des masses aux hypothèses composées et donc de travailler sur 2^{Ω} plutôt que sur Ω constitue un des avantages de cette théorie. Elle permet une modélisation très riche et très souple, en particulier de l'ambiguïté ou de l'hésitation entre classes.

2.2 Combinaison évidentielle

En présence de plusieurs capteurs ou de plusieurs informations provenant d'un même capteur, il devient intéressant de combiner les connaissances de chaque source pour en extraire une connaissance globale et d'y appliquer une règle de décision.
Dans la théorie de DS, les masses sont combinées par la somme orthogonale de Dempster.
Soit m_j la fonction de masse associée à la source j, pour un sous-ensemble A de Ω on obtient :

$$(m_1 \oplus ... \oplus m_l)(A) = \frac{\sum_{B_1 \cap ... \cap B_l = A} m_1(B_1)...m_l(B_l)}{1 - \sum_{B_1 \cap ... \cap B_l = \emptyset} m_1(B_1)...m_l(B_l)} \quad (1)$$

Ce type de combinaison qui n'est pas idempotente suppose l'indépendance cognitive des sources plutôt que l'indépendance statistique [16].
Le mode de combinaison disjonctif est aussi possible en remplaçant l'intersection dans la formule (1) par une opération ensembliste :

$$(m_1 \oplus_{\cup} ... \oplus_{\cup} m_l)(A) = \sum_{B_1 \cup ... \cup B_l = A} m_1(B_1)...m_l(B_l) \quad (2)$$

2.3 Processus de décision

Contrairement à la théorie Bayésienne où le critère de décision est très souvent le maximum de vraisemblance, la théorie de l'évidence propose de nombreuses règles. Les plus utilisées sont le maximum de crédibilité, le maximum de plausibilité, les règles basées sur l'intervalle de confiance, le maximum de probabilité pignistique [17] et la décision par maximum de vraisemblance.

3 Description de la méthode

3.1 Sources et champ de discernement

Les informations couleur de peau sont représentées dans l'espace logarithmique couleur LUX. Cet espace couleur non linéaire est basé sur une transformation logarithmique (modèle LIP de Jourlin et al.) [12].
Les expressions des composantes LUX à partir de l'espace couleur RGB codé sur 3×8 bits sont données par :

$$L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1$$

$$U = \begin{cases} 128\left(\frac{L+1}{R+1}\right) \text{ pour } R > L \\ 256 - 128\left(\frac{R+1}{L+1}\right) \text{ sinon} \end{cases}$$

$$X = \begin{cases} 128\left(\frac{L+1}{B+1}\right) \text{ pour } B > L \\ 256 - 128\left(\frac{B+1}{L+1}\right) \text{ sinon} \end{cases}$$

La procédure de fusion utilise 3 sources ("capteurs" couleurs) notées S_j, $(j = 1, 2, 3)$ telles que :

$$S_1 = U,\ S_2 = X \text{ et } S_3 = 0.5(L+U) = W \quad (3)$$

La source S_3, prenant en compte la composante de luminance très riche du point de vue sémantique, est ajoutée à la source S_1 afin de mieux caractériser les variations de teinte chair dues aux conditions d'éclairement. Chaque source S_j fournit une mesure notée M_j.
Le champ de discernement est défini par 2 hypothèses : $\Omega = \{\omega_1, \omega_2\}$. ω_1 représente l'hypothèse visage et ω_2, complément de ω_1, symbolise le fond ($\omega_2 = \bar{\omega}_1$).

3.2 Modèle a priori

Afin de déterminer le modèle a priori, lors d'une phase d'initialisation l'utilisateur sélectionne respectivement trois zones caractéristiques du visage : une zone ombrée, la zone d'éclairement d'intensité moyenne et une zone surexposée.
Trois variables contextuelles z_i , $(i = 1, 2, 3)$ sont prises en considération : zone ombrée z_1 ; zone moyenne z_2 ; zone surexposée z_3.
Les histogrammes calculés sur chaque zone sélectionnée et pour chaque mesure M_j permettent de déterminer les densités de probabilité conditionnelles $p(M_j/\omega_1, z_i)$. Ces dernières sont ensuite approximées par des fonctions gaussiennes $N_{ij}(\mu_{ij}, \sigma_{ij})$ (Fig. 3).
μ_{ij} et σ_{ij} sont respectivement la moyenne et l'écart type de la mesure M_j sur la zone z_i.
En plus des informations statistiques sur la zone, la répartition spatiale des données (M_1, M_2, M_3) dans l'espace couleur (U, X, W), c'est à dire le domaine couleur noté D_i, contribue de manière implicite à l'élaboration du modèle.

Pour chaque contexte z_i, 3 couples de coefficients α_{ij} et β_{ij} génèrent 6 segments de droite $(\mu_{ij}+\beta_{ij}-\alpha_{ij}\sigma_{ij};\mu_{ij}+\beta_{ij}+\alpha_{ij}\sigma_{ij})$ qui permettent de synthétiser un parallélépipède.

Ces coefficients sont calibrés tels que l'enveloppe du parallélépipède englobe au mieux le domaine couleur D_i (Fig. 4).

Dès lors le modèle a priori est défini par une fonction notée $skin_{ij}$ telle que :

$$skin_{ij} = \begin{cases} \frac{N_{ij}(\mu_{ij},\sigma_{ij})}{\max(N_{ij}(\mu_{ij},\sigma_{ij}))} \\ \text{si } \mu_{ij}+\beta_{ij}-\alpha_{ij}\sigma_{ij} \leq M_j \leq \mu_{ij}+\beta_{ij}+\alpha_{ij}\sigma_{ij} \\ 0 \text{ sinon.} \end{cases} \tag{4}$$

3.3 Fonctions de masse

Appriou [18] a suggéré une approche qui consiste à introduire chaque densité de probabilité a priori $p(M_j/\omega_1, z_i)$ et son degré de confiance d_{ij} correspondant dans une fonction de masse $m_{ij}(.)$. Cet ensemble est défini par une approche axiomatique dans l'ensemble de discernement Ω.

Dans la démarche développée ici les densités de probabilités sont remplacées par les fonctions $skin_{ij}$ qui prennent en considération implicitement la répartition volumique des données dans l'espace couleur (U, X, W).

Les éléments focaux associés à $m_{ij}(.)$ sont ω_1, ω_2 et Ω.

Les fonctions de masses sont définies par :

$$\begin{aligned} m_{ij}(\omega_1) &= \frac{d_{ij}R_i skin_{ij}}{1+R_i skin_{ij}} \\ m_{ij}(\omega_2) &= \frac{d_{ij}}{1+R_i skin_{ij}} \\ m_{ij}(\Omega) &= 1-d_{ij} \end{aligned} \tag{5}$$

Les coefficients de pondération R_i traduisent la prise en compte des données $skin_{ij}$ pour chaque zone i.

3.4 Degrés de confiance

Pour les sources $S_1 = U$ et $S_2 = X$ l'ambiguïté entre classes est faible. Dès lors une modélisation probabiliste est utilisée d'où $d_{i1} = d_{i2} \approx 1$.

Par contre la source S_3 dépend de la luminance. La fiabilité de cette source pour la classe visage s'avère maximale ($d_{i3} = 1$) uniquement pour le niveau de gris moyen (μ_{i3}) de la zone modélisée.

L'ambiguïté entre classes croît ($m(\Omega) > 0$) lorsque M_3 s'écarte de μ_{i3}.

C'est pourquoi une fonction d'appartenance floue (Eq. 6) est utilisée pour caractériser la fiabilité de la source S_3 pour le contexte z_i (Fig. 1).

Sachant que $\mu_{13} \leq \mu_{23} \leq \mu_{33}$ on obtient :

$$\begin{aligned} &\text{pour } i=1 : \\ &d_{13} = \begin{cases} \frac{M_3+2\mu_{23}-3\mu_{13}}{2\mu_{23}-2\mu_{13}} \text{ si } 3\mu_{13}-2\mu_{23} \leq M_3 \leq \mu_{13} \\ \frac{M_3-\mu_{23}}{\mu_{13}-\mu_{23}} \text{ si } \mu_{13} \leq M_3 \leq \mu_{23} \\ 0 \text{ sinon} \end{cases} \\ &\text{pour } i=2 : \\ &d_{23} = \begin{cases} \frac{M_3-\mu_{13}}{\mu_{23}-\mu_{13}} \text{ si } \mu_{13} \leq M_3 \leq \mu_{23} \\ \frac{M_3-\mu_{33}}{\mu_{23}-\mu_{33}} \text{ si } \mu_{23} \leq M_3 \leq \mu_{33} \\ 0 \text{ sinon} \end{cases} \\ &\text{pour } i=3 : \\ &d_{33} = \begin{cases} \frac{M_3-\mu_{23}}{\mu_{33}-\mu_{23}} \text{ si } \mu_{23} \leq M_3 \leq \mu_{33} \\ \frac{M_3+2\mu_{23}-3\mu_{33}}{2\mu_{23}-2\mu_{33}} \text{ si } \mu_{33} \leq M_3 \leq 3\mu_{33}-2\mu_{23} \\ 0 \text{ sinon} \end{cases} \end{aligned} \tag{6}$$

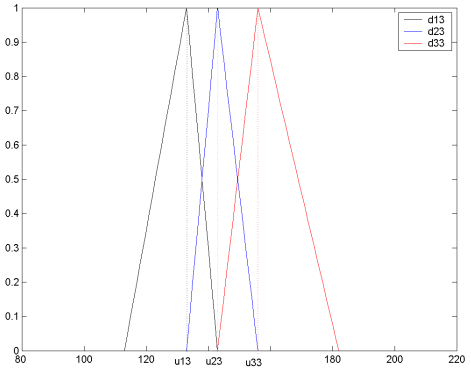


FIG. 1 – Degrés de confiance d_{i3} en fonction de M_3.

3.5 Décision

A chaque pixel en le site $s(x, y)$ de composantes couleur (M_1 ; M_2 ; M_3)(Eq. 3) sont associées trois masses contextuelles $m_i(s)$ par la règle de combinaison orthogonale normalisée de Dempster-Shafer (Eq. 1) :

$$m_i(s) = \oplus m_{ij}(s) \tag{7}$$

La règle de fusion disjonctive (Eq. 2) combine les masses contextuelles $m_i(s)$ (Eq. 5) afin d'associer une masse unique $m(s)$ à chaque pixel.

Cependant, les résultats expérimentaux montrent qu'une pondération des masses $m_i(s)$ par les degrés de confiance $d_{i3}(M_3(s))$ améliore la qualité de la segmentation. On obtient dès lors :

$$\begin{aligned} m(s) &= d_{13}(M_3(s)).m_1(s) \oplus_\cup d_{23}(M_3(s)).m_2(s) \\ &\quad \oplus_\cup d_{33}(M_3(s)).m_3(s) \end{aligned} \tag{8}$$

4 Résultats de la modélisation

4.1 Modèle a priori zone ombrée

La phase d'initialisation étant identique pour chaque zone, seule la modélisation de la zone ombrée z_1 est présentée en détail ci-après.
Dans un premier temps, l'utilisateur enregistre son visage dans une image puis sélectionne manuellement sur cette dernière une zone ombrée du visage (Fig. 2).

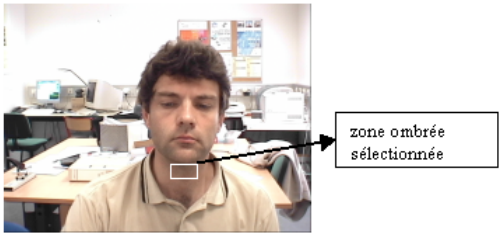
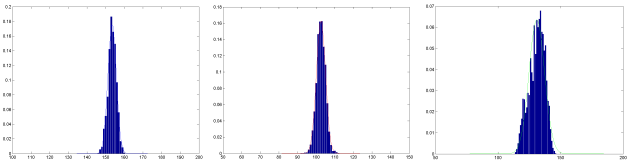


FIG. 2 – Phase d'apprentissage zone visage ombrée

Les densités de probabilité sont approximées pour chaque mesure M_j par une fonction gaussienne $N_{1j}(\mu_{1j}, \sigma_{1j})$ (Fig. 3).



a) b) c)

FIG. 3 – Densités de probabilité pour zone ombrée z_1 : a) $p(M_1/\omega_1, z_1)$; b) $p(M_2/\omega_1, z_1)$; c) $p(M_3/\omega_1, z_1)$.

Les segments de droite $(\mu_{1j} + \beta_{1j} - \alpha_{1j}\sigma_{1j}\,;\, \mu_{1j} + \beta_{1j} + \alpha_{1j}\sigma_{1ij})$ déterminés en faisant varier α_{1j} et β_{1j} sur chaque voie, génèrent un parallélépipède qui englobe le domaine couleur D_1.

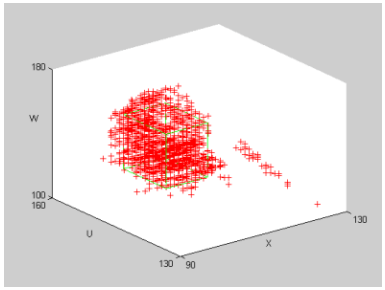
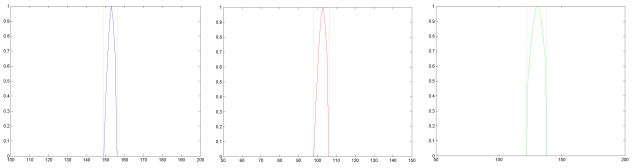


FIG. 4 – Représentation dans l'espace couleur (U, X, W), des composantes couleur des pixels (domaine D_1) et enveloppe parallélépipédique associée.

Ici $\alpha_{11} = \alpha_{12} = 2$; $\alpha_{13} = 1.7$; $\beta_{11} = \beta_{12} = \beta_{13} = 0$; et $\sigma_{11} = 2.3$; $\sigma_{12} = 2.55$; $\sigma_{13} = 6.5$.

Ces coefficients α_{1j} et β_{1j} bornent dès lors les densités de probabilités et déterminent les fonctions $skin_{1j}$ (cf. Eq. 4 et Fig. 5).

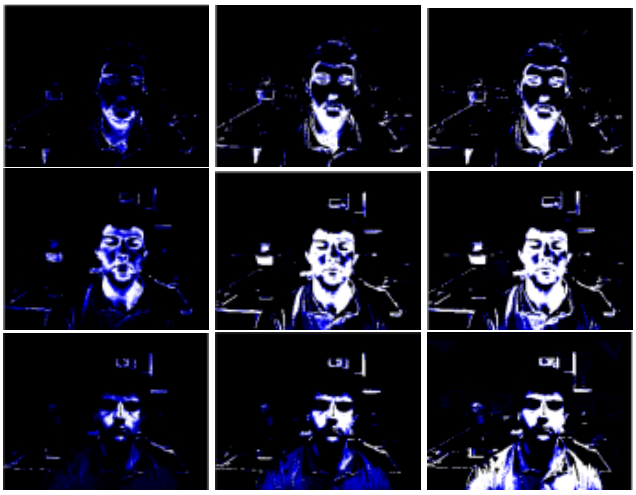


a) b) c)

FIG. 5 – Fonctions $skin_{1j}$: a) $skin_{11}$; b) $skin_{12}$; c) $skin_{13}$

4.2 Détection de la zone ombrée

Les fonctions de masses $m_{1j}(\omega_1) = \frac{d_{1j} R_1 skin_{1j}}{1+R_1 skin_{1j}}$ (Eq. 5) dépendent du paramètre R_1 qui pondère l'importance aux données $skin_{1j}$ caractérisant la classe visage. La fusion pixel conjonctive normalisée $m_1(s) = \oplus m_{1j}(s)$ (Eq. 7) affecte par indexation à chaque pixel en le site $s(x, y)$ de composantes $(M_1\,;\, M_2\,;\, M_3)$ (Eq. 3) une masse $m_1(s)$.
Les 3 images du haut de la Fig. 6 présentent les résultats expérimentaux de modélisation pour différentes valeurs de R_1. Pour $R_1 = 1$ la modélisation est satisfaisante car seules les parties teinte chair ombrées (en blanc) présentent sous le menton,le contour des yeux et sous les cheveux sont détectées. Pour $R_1 = 5$ ou $R_1 = 10$ la zone de modélisation s'élargit et apparaissent des fausses détections.



a) b) c)

FIG. 6 – Détection réalisée à partir des fonctions de masse m_i, $(i = 1, 2, 3)$ pour différentes valeurs de R_i : a) $R_i = 1$; b) $R_i = 5$; c) $R_i = 10$. En haut : zone ombrée $(i = 1)$; Au milieu : zone moyenne $(i = 2)$; En bas : zone surexposée $(i = 3)$.

4.3 Synthèse de la modélisation

La Fig. 6 (milieu et bas) présente également les résultats de l'étape de la fusion conjonctive pour les zones moyenne

et saturée, dont la démarche est similaire à celle présentée aux paragraphes 4.1 et 4.2.
La modélisation détecte assez fidèlement les parties ombrées et surexposées du visage (Fig. 6 haut et bas). Le modèle moyen (Fig. 6 milieu). bien que plus représentatif du visage présente des défauts dans les zones ombrées ou surexposées.
C'est pourquoi la fusion disjonctive (Eq. 8) est mise en œuvre pour mettre en concordance les différents modèles et optimiser le résultat de la détection (Fig. 7).

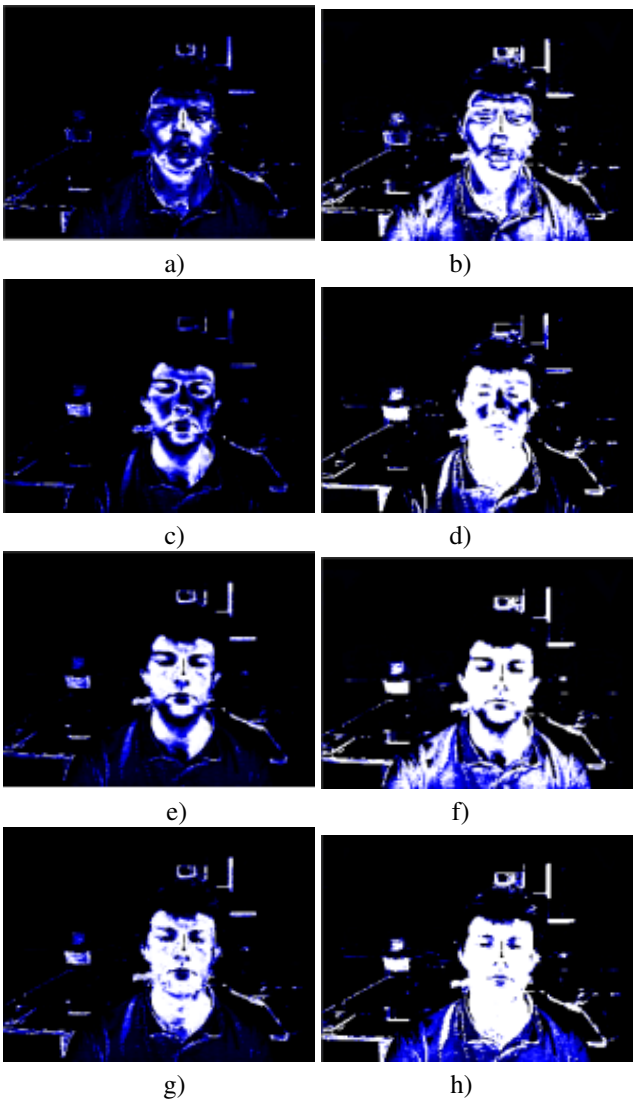


FIG. 7 – Image issue de fusion pour différentes combinaison de masses : a) $R_1 = 1$, $R_2 = 0$, $R_3 = 0$; b) $R_1 = 5$, $R_2 = 0$, $R_3 = 5$; c) $R_1 = 1$, $R_2 = 1$, $R_3 = 0$; d) $R_1 = 5$, $R_2 = 5$, $R_3 = 0$; e) $R_1 = 0$, $R_2 = 1$, $R_3 = 1$; f) $R_1 = 0$, $R_2 = 5$, $R_3 = 5$; g) $R_1 = 1$, $R_2 = 1$, $R_3 = 1$; h) $R_1 = 5$, $R_2 = 5$, $R_3 = 5$.

Lorsque uniquement deux masses sont combinées la modélisation est insuffisante et présente des defauts soit :
– sur les parties du visage d'intensité moyenne (Fig. 7 a,b).
– sur les parties du visage surexposées (Fig. 7 c,d).
– sur les parties ombrées (Fig. 7 e,f).
La modélisation est optimale lorsque les 3 masses sont combinées ($R_1 = 1$, $R_2 = 1$, $R_3 = 1$) (Fig. 7 g). Cependant une pondération trop importante du paramètre R_i détériore légèrement la qualité de la modélisation (Fig. 7 f).
Ainsi malgré un fond complexe composé d'éléments de couleur proche de celle de la chair (table, pull, affiche), le processus de fusion assure une assez bonne segmentation et s'avère aussi robuste aux occlusions et aux variations de pose (Fig. **??**).

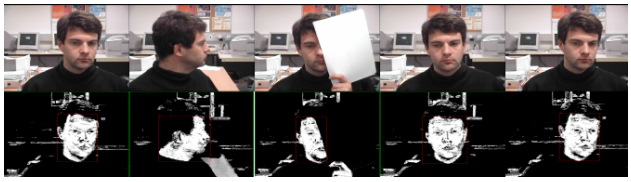


FIG. 8 – a) image 10 ; b) image 11 ; c) image 18 ; d) image 29 ; e) image 31.

5 Etape de suivi par un algorithme de condensation

5.1 Formalisme du filtrage particulaire

L'algorithme de filtrage particulaire initialement conçu pour des problèmes de traitement du signal [16], a été étendu en vision par ordinateur sous le nom d'algorithme de "condensation" [20].
Si l'on se place dans le cas d'un objet unique, le vecteur X_t représentant l'état caché de l'objet d'intérêt suit la loi d'évolution (Eq. 9) et est observé par le vecteur Y_t à des instants discrets selon l'équation (Eq. 10) :

$$X_t = F_t(X_{t-1}, V_t) \quad (9)$$
$$Y_t = H_t(X_t, W_t) \quad (10)$$

Aucune hypothèse n'est faite sur les fonctions F_t et H_t et les deux processus V_t et W_t sont des bruits blancs pas nécessairement gaussiens indépendants entre eux et indépendants de la condition initiale X_0.
Le but est d'estimer récursivement la distribution de probabilité a posteriori $B_t = p(X_t/Y_{1:t})$ de X_t à l'instant t conditionnellement aux observations cumulées jusqu'à cet instant, ainsi que toute fonction de l'état $g(X_t)$ par l'espérance conditionnelle $E[g(X_t)]$.
Le filtrage comprend deux étapes :

1. une étape de prédiction donnée par l'équation suivante :
$$p(X_t/Y_{1:t-1}) = \int p(X_t/X_{t-1})p(Y_t/X_{t-1})dX_{t-1}$$
2. la donnée d'une observation Y_t permet par la règle de Bayes de corriger cette prédiction :
$p(X_t/Y_{1:t}) \propto p(Y_t/X_t).p(X_t/Y_{1:t-1})$

La récursivité de l'algorithme nécessite la spécification d'un modèle dynamique décrivant l'évolution de l'état, $p\left(X_t/X_{t-1}\right)$, et d'un modèle d'adéquation des données à l'état $p(Y_t/X_t)$ avec aussi les hypothèses d'indépendance conditionnelle :

$$X_t \perp Y_{1\,:\,t-1}/X_{t-1} \text{ et } Y_t \perp Y_{1\,:\,t-1}/Y_t.$$

L'équation de ce filtre est obtenue simplement mais il est en général impossible de la résoudre sauf dans le cas particulier où les fonctions F_t et G_t sont linéaires et les bruits W_t et V_t sont gaussiens.
Elle se ramène alors aux équations du filte de Kalman-Bucy.
Dans le cas de modèles non-linéaires ou non gaussiens, l'idée générale du filtre particulaire consiste à chercher une approximation de la distribution de probabilité conditionnelle $p(X_t/Y_{1\,:\,t})$ sous la forme d'une combinaison linéaire pondérée de masses de Dirac appelées particules.
L'algorithme (Tab. 1) consiste à faire évoluer le nuage de particules $\Sigma_t = \{\xi_t^n, q_t^n, n = 1....N\}$ où les positions $\{\xi_t^n, n = 1...N\}$ sont des éléments de l'espace d'état et où les poids $\{q_t^n, n = 1...N\}$ ont des valeurs comprises entre 0 et 1 et de l'utiliser pour estimer : $B_t = p(X_t/Y_{1\,:\,t})$ par $\mathrm{B}_{\Sigma_t} = \sum_{n=1}^{N} q_t^n \delta_{\xi_t^n}$ avec $\sum_{n=1}^{N} q_t^n = 1$. Afin d'éviter la dégénérescence du nuage un rééchantillonnage est effectué lorsque le nombre de particules effectives, estimé par N_{est} est inférieur à un seuil fixé N_{seuil}.

TAB. 1 – Algoritme du filtrage particulaire

Initialisation : $\left\{ \begin{array}{l} \xi_0^n \sim p(X_0) \\ q_0^n = 1/N \end{array} \right.$ n = 1,...,N

Pour $t = 1, ..., T$:

Prédiction : $\left\{ \begin{array}{l} v_t^n \sim p(V_t) \\ \xi_{t/t-1}^n = F_t(\xi_{t/t-1}^n, v_t^n) \end{array} \right.$ n = 1...N

Correction : $q_t^n = q_{t-1}^n \frac{b_t(y_t;\xi_{t/t-1}^n)}{\sum_{n=1}^{N} b_t(y_t;\xi_{t/t-1}^n) q_{t-1}^n}$

Estimation : $\hat{\mathrm{E}}\left[g(X_t)\right] = \sum_{n=1}^{N} q_t^n g(\xi_{t/t-1}^n)$

Ré-échantillonnage si :

$N_{est} = \frac{1}{\sum_{n=1}^{N} (q_t^n)^2} < N_{seuil} \left\{ \begin{array}{l} \xi_t^n \sim \sum_{k=1}^{N} q_t^k \delta_{\xi_{t/t-1}^n} \\ q_t^n = 1/N \end{array} \right.$

5.2 Application de la condensation au suivi de visage

Dans le cadre de suivi de visage dont le mouvement propre n'est pas prévisible et change fréquemment de direction le suivi par filtrage particulaire trouve toute sa justification.

De plus, une idée fréquemment utilisée est d'approximer la forme du visage par une ellipse.
Ainsi, les contours extérieurs du visage sont approximés par une ellipse de centre noté (x_c, y_c), d'axe mineur ℓ, de grand axe GA et d'orientation θ. L'ensemble de ces paramètres est regroupé dans le vecteur d'état $X_t = [x_{c_t}, y_{c_t}, \ell_t, GA_t, \theta_t]$.
Le but est donc d'estimer de manière robuste les paramètres de X_t de l'ellipse représentant la forme du visage que l'on souhaite suivre.
La dynamique du vecteur d'état est décrite par le modèle suivant [21] :

$$p(X_t/X_{t-1}) = (1 - \beta_u) N(X_t/X_{t-1}, \Sigma) + \beta_u U_x(X_t)$$

où $N(./\mu, \Sigma)$ représente la distribution gaussienne de moyenne μ et de covariance Σ, $U_X(.)$ représente la distribution uniforme sur l'ensemble X. Le coefficient β_u , $0 \leq \beta_u \leq 1$ pondère la distribution uniforme et $\Sigma = diag(\sigma_{xc}^2, \sigma_{yc}^2, \sigma_\ell^2, \sigma_{GA}^2, \sigma_\theta^2)$ est la matrice diagonale composée des variances des composantes du vecteur d'état. Cette modélisation suppose que les composantes du vecteur d'état évoluent suivant des modèles gaussiens aléatoires mutuellement indépendants. L'introduction d'une légère composante uniforme gère les rares mouvements erratiques perçus comme des sauts dans la séquence vidéo. Elle aide aussi l'algorithme à se verrouiller après une période partielle ou totale d'occlusion.
Après une phase d'initialisation, la méthode de suivi procède en deux phases. Dans un premier temps l'algorithme détermine le centre (x_c, y_c) du modèle puis estime ensuite la taille et l'orientation (pose) de l'ellipse.

Initialisation. Afin de sélectionner le visage quelle que soit sa position initiale dans l'image, les particulessont réparties en position suivant une loi uniforme alors que les paramètres de taille et d'orientation sont fixées à une valeur constante ($\ell_0 = 20$; $GA_0 = 25$; $\theta_0 = 0$) (Fig. 9). Le poids des particules vaut 1/N avec N nombre de particules N=100 ;

Estimation du centre du modèle. Dans cette phase, afin de déterminer la position du vecteur d'état à l'instant t les paramètres de taille et d'orientation sont fixes. Il s'agit des composantes du vecteur d'état estimées à l'instant $t-1$. Le modèle de mesure pondère une particule n plus ou moins fortement en fontion de la quantité d'information de couleur teinte chair contenue à l'intérieur de celle-ci. Il s'agit donc de sommer l'ensemble des niveaux de gris contenus à l'intérieur de l'ellipse n. En effet la mesure $Y_{n,t}$ est choisie proportionnelle à la somme des niveaux de gris au carré à l'intérieur de la particule n. La probabilité d'observation est dès lors définie par :

$$p(Y_t/X_t) = \prod_n p(Y_{n,t}/X_{n,t}) \text{ où } Y_{n,t} \propto \sum s(x,y)^2$$

Le critère de maximum de vraisemblance permet alors de sélectionner l'ellipse la plus significative et son centre définit les composantes de position du vecteur d'état. Suite

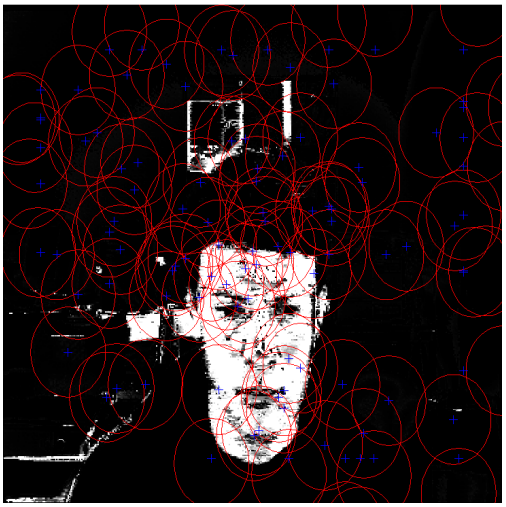
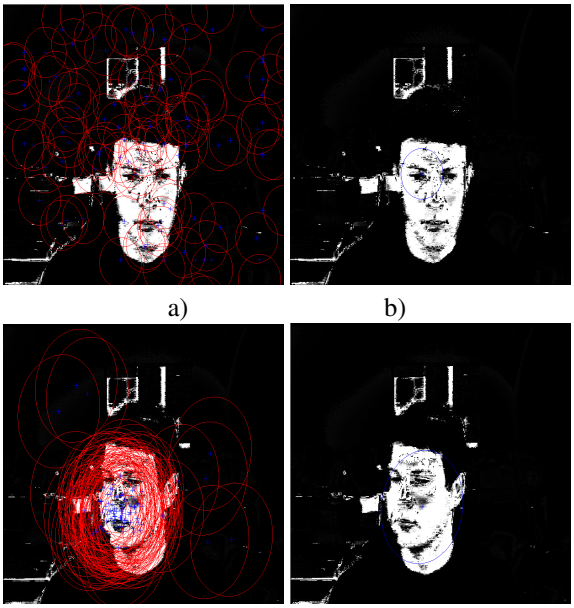


FIG. 9 – initialisation des particules



a) b) c) d)

FIG. 10 – a) initialisation des particules ; b) position pour l'image 1 ; c) particules pour l'image 13 ; d) position pour l'image 13

à l'initialisation, la position estimée (Fig. 10 b) ne correspond pas à celle du centre du visage en raison du dimensionnement incorrect en taille et orientation de l'ellipse. Au bout de quelques images, ces paramètres étant corrigés (voir ci-après), l'algorithme estime avec une meilleure précision le centre de l'ellipse (Fig. 10 d).

Estimation taille et pose. A partir du centre défini à l'étape précédente, une opération de remplissage puis de binarisation est réalisée. Dès lors les contours de la forme du visage sont extraits puis approximés par une ellipse appelée ellipse de mesure.(Fig. 11 b,e). Or la littérature sur l'ajustement d'ellipses se divise en deux grandes catégories :

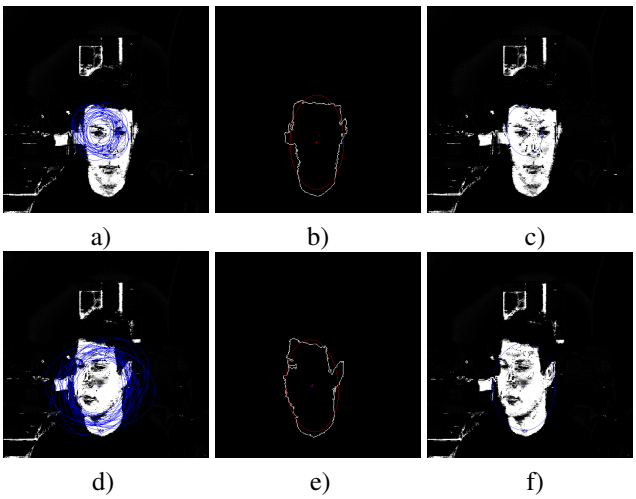
– les méthodes de regroupement "clustering" parmi lesquelles celles utilisant la transformation de Hough .
– les méthodes d'ajustement par moindres carrés.

Ces dernières ont pour but de déterminer un ensemble de paramètres entre les points de données et l'ellipse.

L'algorithme présenté par Fitzgibbon et al. [15] réalise un compromis entre rapidité et précision pour l'ajustement d'ellipse et est très robuste aux bruits.

Ainsi dans le contexte de fond complexe où certains contours issus de fausses détections perturbent le dimensionnement de la forme cette approche a été choisie. Elle fournit une mesure d'ellipse de paramètres $\hat{X}_t = \left[\hat{x}_{c_t}, \hat{y}_{c_t}, \hat{\ell}_t, \hat{GA}_t, \hat{\theta}_t\right]$.

Dès lors la densité de probabilité de l'observation à l'instant t conditionnée sur le fait que les composantes de taille et d'orientation du vecteur d'état se réalisent en la particule $X_{n,t}$, est choisie proportionnelle à la distance entre l'ellipse prédite et l'ellipse mesurée. La densité de probabilité minimale permet alors d'isoler la particule ayant la forme la plus vraisemblable.



a) b) c) d) e) f)

FIG. 11 – a) particules image 1 ; b) mesure de l'ellipse image 1 ; c) résultat filtrage image 1 ; c) particules image 13 ; b) mesure de l'ellipse image 13 c) résultat filtrage image 13

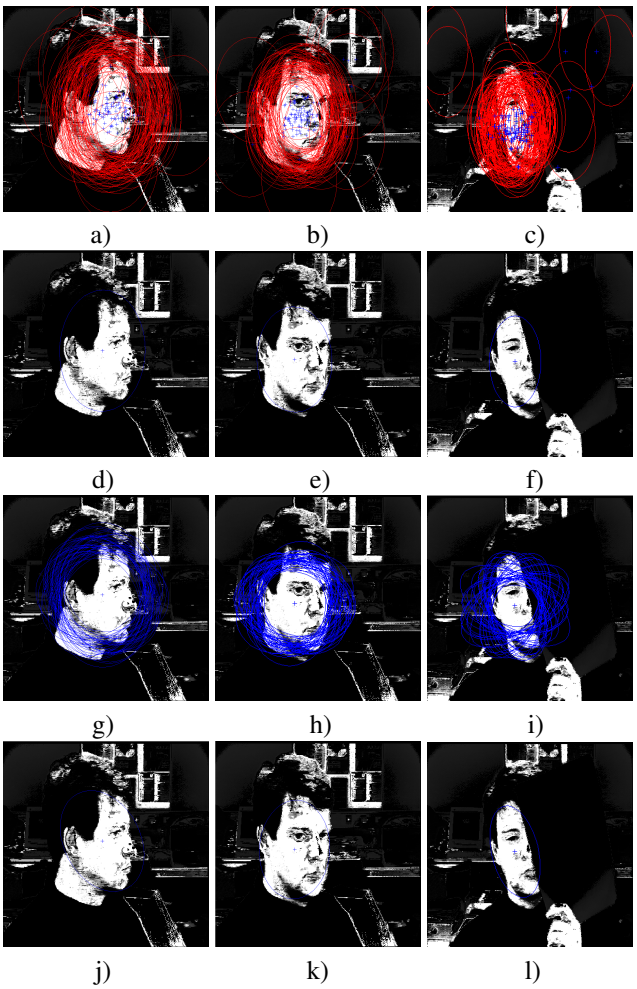


FIG. 12 – a) particules pour position image 15 ; b) particules pour position image 18 ; c) particules pour position image 27 ; d) position image 15 ; e) position image 18 ; f) position image 27 ; g) particules pour taille et pose image 15 ; h) particules pour taille et pose image 18 ; i) particules pour taille et pose image 27 ; j) résultat filtrage image 15 ; k) résultat filtrage image 18 ; l) résultat filtrage image 27

Les résultats sur la séquence (Fig. 12) traduisent le comportement de l'algorithme dans le contexte de visage de côté (image 15) ou d'occlusion partielle (image 27).

6 Conclusion et perspectives

La méthode présentée ici est la première étape d'une nouvelle approche du problème de détection et suivi de visages. L'originalité de la méthode réside en la modélisation de la teinte chair par la fusion de trois sources indépendantes du point de vue cognitif. De plus, les fonctions de masse (théorie de Dempster-Shafer) sont déterminées à partir de modèles a priori intégrant des données contextuelles spécifiques au visage. Ainsi, la particularité du visage de présenter des zones ombrées (cou) ou surexposées (nez, front) est prise en compte et la sensibilité face aux conditions d'éclairement est diminuée.

Cependant la fusion uniquement d'informations couleur est parfois insuffisante et l'utilisation d'autres données transversales telles que le mouvement, la texture ou les contours doit améliorer la détection. Au niveau du suivi une estimation en une seule étape des paramètres de position , taille et pose de l'ellipse par l'algorithme de filtrage particulaire pourra être envisagée. Une régulation sera intégrée afin d'adapter de manière dynamique les paramètres du processus de fusion et ainsi optimiser la qualité des informations (précision,incertitude ...).

L'objectif est donc de réaliser un système de vision active où l'image est vue comme un tout. Il n'est dès lors pas nécessaire d'avoir une segmentation optimale à partir d'un seul critère (ici couleur chair) mais la collaboration dynamique des informations doit contribuer à une meilleure robustesse de la segmentation et du suivi.

Références

[1] E. Hjelmås and B.K. Low, Face detection : A survey, *Computer Vision and Image Understanding*, Vol. 83, No. 3, pp. 236-274, Sept. 2001.

[2] Ming-Hsuan Yang, David Kriegman and Narendra Ahuja, Detecting faces in images : A survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, pp. 34-58, 2002.

[3] Prem Kuchi, Prasad Gabbur, P. Subbanna Bhat and Sumam David S., Human Face Detection and Tracking using Skin Color Modeling and Connected Component Operators, *IETE Journal of Research*, Vol. 38, No. 3&4, pp. 289-293, May-Aug 2002.

[4] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse, Blob Analysis of the Head and Hands : A Method for Deception Detection, *International Conference on System Science* (HICSS'05), Hawaii, 2005.

[5] C.Garcia and M.Delakis. A neural architecture for fast and robust face detection. *Proc. of the IEEE-IAPR International Conference on Pattern Recognition* (ICPR'02), 2002.

[6] F. Yang et M. Paindavoine. Implementation of an RBF neural network on embedded systems : real-time face tracking and identity verification. *IEEE Trans. on Neural Networks*, 2003.

[7] C.C. Chiang, W.N. Tai, M.T. Yang, Y.T. Huang, and C.J. Huang. A novel method for detecting lips, eyes and faces in real time. *Real-Time Imaging*, 9, 2003.

[8] P. Viola and M. Jones, Fast and Robust Classification Asymmetric AdaBoost and a Detector Cascade, *Advances in Neural Information Processing System* 14, MIT Press, Cambridge, MA, 2002.

[9] Dempster, A. P., A generalisation of Bayesian inference, *Journal of the Royal Statistical Society*, pp. 205-247, 1968.

[10] Shafer, G. *A Mathematical Theory of Evidence.* Princeton University Press,1976.

[11] Michael J. Swain and Dana H. Ballard, Color indexing, *International Journal of Computer Vision*, Vol.7, No.1, p.11-32, Nov. 1991

[12] M. Liévin, F. Luthon, Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of faces in video. *IEEE Trans. on Image Processing*, Vol. 13, No. 1, Jan. 2004, pp. .

[13] Stan Birchfield, Elliptical Head Tracking Using Intensity Gradients and Color Histograms, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, pages 232-237, June 1998

[14] Nanda, H. and Fujimura, K., A robust elliptical head tracker, Automatic Face and Gesture Recognition, 2004. *Proceedings. Sixth IEEE International Conference on*, May 2004,pp. 469-474.

[15] A. Fitzgibbon, M. Pilu and R. Fisher, Direct Least Squares Fitting of Ellipses, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 5, May 1999.

[16] I. Bloch, H. Maître, Data fusion in 2D and 3D image processing : An overview. http ://mirror.impa.br/sibgrapi97/anais/pdf/bloch.pdf.

[17] Ph. Smets. Constructing the Pignistic Probability Function in a Context of Uncertainty. *Uncertainty in Artificial Intelligence*, 5 :29-39, 1990.

[18] A. Appriou, DTIM Multisensor signal processing in the framework of the theory of evidence, *NATO/RTO Lecture Series 216 on Application of Mathematical Signal Processing Techniques to Mission Systems,* Nov 1999.

[19] N. Gordon, D. Salmond and A. Smith. Novel approach to non linear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2) :107-113,1993.

[20] M. Isard and A. Blake. CONDENSATION-conditional density propagation for visual tracking *Int. J. Computer Vision,* 29(1) :5-28,1998.

[21] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of IEEE*, 92(3) :495-513, 2004.