Suivi de gestes temps réel par traitement d'images couleur

Real Time tracking of human gestures using color image processing

Thierry Chateau, Antoine Vacavant

Lasmea, UMR6602 du CNRS/Université Blaise Pascal,
63172 Aubière Cedex, France
thierry.chateau@lasmea.univ-bpclermont.fr

Manuscrit reçu le 15 juin 2004

Résumé et mots clés

Cet article adresse le problème de la détection et du suivi de zones colorées par traitement d'images couleur. L'application finale concerne le suivi 2D sans marqueurs de la tête et des deux mains d'une personne. Nous proposons une méthode originale de localisation de zones colorées, appliquée à l'issue d'une phase d'extraction de la personne par rapport au fond. L'appartenance d'un pixel de l'image au fond est modélisée par une densité de probabilités exprimée dans un espace couleur. Nous montrons que l'utilisation de mixtures de gaussiennes permet d'approcher cette densité de probabilités. La prise en compte de l'évolution temporelle du système est assurée par un filtre à particules, réputé robuste aux occultations partielles et aux modèles non gaussiens.

Imagerie couleur, suivi de zones colorées, soustraction de fond, filtre à particules.

Abstract and key words

This paper deals with areas detection and tracking using color based image processing. The application adressed concerns human head and hands tracking. We propose an original method in order to locate, for an foreground image, colored areas. Foreground image is provided by a probabilistic way using Gaussian Mixture Models (GMM) of probability density fonctions. The temporal tracking is then achieved by a particle filter, which is well adapted to partial occlusions and non gaussian models.

Colour image, colored patch tracking, background substraction, particle filter.

Introduction

L'imagerie couleur [19] est un domaine de recherche trés actif. De plus en plus de méthodes de suivi de motifs dans des séquences d'images intègrent des mesures couleur [5]. Ces dernières sont également particulièrement bien adaptées dans le cas de la détection de personnes ou de suivi de gestes [20, 17], où les propriétés colorimétriques des pixels sont prises en compte pour extraire de l'image les personnes à suivre.

Le suivi de gestes par vision artificielle est une tâche indispensable dans la conception d'un grand nombre de systèmes. Parmis ces derniers, Gravilla [10] cite, entre autres, les applications suivantes :

- les mondes virtuels intéractifs,
- les jeux vidéo,
- la surveillance des magasins,
- les interfaces hommes-machines,
- le langage des signes pour les mal-entendants,
- l'analyse de gestes sportifs (tennis, golf).

Il est alors possible de classer ces applications selon deux catégories principales. Celle où il est nécessaire d'avoir une approche en trois dimensions (c'est par exemple le cas pour l'analyse de gestes sportifs) et celle où une approche en deux dimensions est suffisante (comme par exemple certaines interfaces homme/machine). Les problèmes se rapportant à la seconde catégorie sont de manière générale bien mieux traités que ceux se rapportant à la première, pour des raisons de simplicité. Dans certaines applications [4], il est possible de se contenter de la mesure de la trajectoire image des deux mains et de la tête d'une personne. Dans cet article, nous adressons le problème du suivi 2D sans marqueurs de la tête et des deux mains d'une personne, par vision couleur monoculaire. Les sorties de cette méthode sont les trajectoires, dans le plan image, des trois zones suivies. Le synoptique présenté figure 1 montre le découpage du système en trois parties principales :

1. Une méthode d'extraction de fond est tout d'abord appliquée afin de conserver uniquement les pixels appartenant au premier plan de l'image. Cette dernière est basée sur une modélisation de la densité de probabilité de la couleur de chaque pixel par une mixture de gaussiennes. Nous comparons, pour trois espaces couleur, les performances de cet algorithme d'extraction de fond.
2. Dans la deuxième sous-partie, la recherche de trois amas de pixels de couleur peau est effectuée, à partir de l'image de premier plan. Les centres de gravité des trois amas détectés correspondent aux centres de gravité de la tête et des deux mains.
3. L'évolution temporelle de ces trois positions est alors prise en compte à l'aide d'un filtre à particules. D'autre part, ce dernier permet de recaler les positions des trois zones d'une image sur l'autre.

La première partie de cet article adresse l'approche retenue pour la sélection des points appartenant à l'avant plan. La classification des pixels appartenant à la peau et la détection des trois amas correspondant à la tête et aux deux mains est abordé dans la deuxième partie. La troisième partie traite de la prise en compte de l'aspect temporel du suivi par l'utilisation d'un filtre à particules. Une conclusion et des perspectives sont données dans la quatrième partie.

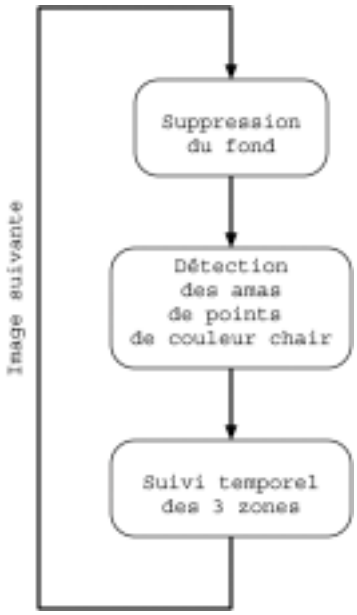

Figure 1. Synoptique général de la solution proposée.

1. Suppression du fond

Dans le cas où la scène est observée à l'aide d'une caméra fixe, l'utilisation de méthodes de soustraction de fond permet d'extraire facilement les informations de premier plan de l'image. Dans le cas d'application de suivi de gestes, il s'agit de la silhouette de la personne à suivre. Parmi les méthodes les plus utilisées, on distingue les méthodes dîtes non-adaptatives et les méthodes dîtes adaptatives. Les premières sont utilisées dans le cas d'un éclairage de la scène constant. Dans les applications de suivi de geste, ce n'est pas toujours le cas. Nous avons donc choisi une approche adaptative. Cette dernière est inspirée des travaux de Stauffer et Grimson [18], repris dans [11]. Le but est ici d'extraire la silhouette de la personne filmée. Un des avantages de cette approche réside dans le fait qu'elle est capable de modéliser des fonds dans lesquels le pixel du fond peut avoir deux couleurs distinctes ; par exemple dans le cas d'une image d'arbre, certains pixels peuvent être verts (feuille) ou bleu (ciel), en fonction du vent. De plus, ce modèle adaptatif prend en compte des nouveaux objets et supporte les variations d'intensité lumineuse.

1.1. Principe de l'algorithme

Les valeurs prises par les pixels du fond sont modélisées par plusieurs gaussiennes. Cela permet de prendre en compte des valeurs différentes pour chacun d'entre eux. La figure 2 montre les valeurs prises par un pixel de l'image dans le cas où un nouvel objet est inséré dans la scène. La courbe présente alors deux modes disctints. Ce cas de figure ne peut être pris en compte de manière correcte à l'aide d'un modèle utilisant une seule gaussienne. Il faut donc choisir une représentation plus complète. Dans ce cas, une représentation sous la forme de mixture de gaussiennes est préférée.

À chaque pixel est associée une somme pondérée de N gaussiennes (dit *GMM*). On notera que $2 \leq N \leq 5$, et N sera ajusté suivant la complexité du modèle désiré. En théorie, le nombre de gaussiennes n'est pas limité. Par contre, nos expériences se sont limitées à un nombre maximum de 5 Gaussiennes. D'après [18], il n'est pas utile d'aller au delà. Dans le cas d'une implantation temps réel (15 images par seconde), on choisira par exemple $N = 2$.

À un temps t, on considère que le modèle $\mathbf{M_t}$ généré pour chaque pixel à partir des mesures images $\{\mathbf{Z_0}, \mathbf{Z_1}, \ldots, \mathbf{Z_{t-1}}\}$ est correct. On évalue alors la probabilité que ce pixel appartienne au fond :

$$P(\mathbf{Z_t}|\mathbf{M_t}) = \sum_{n=1}^{n=N} \alpha_n \mathcal{N}(\mu_\mathbf{n}, \Sigma_\mathbf{n}) \tag{1}$$

$$\mathcal{N}(\mu_\mathbf{n}, \Sigma_n) = \frac{1}{(2\pi)^{d/2}|\Sigma_n|^{1/2}} e^{-\frac{1}{2}(\mathbf{Z_t}-\mu_\mathbf{n})^T \Sigma^{-1}(Z_t-\mu_\mathbf{n})} \tag{2}$$

où d est la dimension de l'espace de couleur de la mesure $\mathbf{Z_t}$.

On peut noter que chaque gaussienne n est distinguée par :
- une moyenne $\mu_\mathbf{n}$,
- une matrice de covariance Σ_n,
- un poids α_n, avec $\sum_n \alpha_n = 1$.

Dans le cas où la matrice de covariance est quelconque, sa forme est définie symétrique positive. Si on considère que paramètres du vecteurs de mesure sont statistiquement indépendants (les 3 plans de l'espace couleur sont indépendants), l'écriture de la matrice Σ_n peut se simplifier :

$$\begin{aligned}\Sigma_n &= \begin{pmatrix} {}_r\upsilon_n & 0 & 0 \\ 0 & {}_g\upsilon_n & 0 \\ 0 & 0 & {}_b\upsilon_n \end{pmatrix} \\ &= \begin{pmatrix} {}_r\sigma_n^2 & 0 & 0 \\ 0 & {}_g\sigma_n^2 & 0 \\ 0 & 0 & {}_b\sigma_n^2 \end{pmatrix}\end{aligned} \tag{3}$$

Pour chaque pixel de l'image, trois écart types sont associés. Ces derniers n'ont aucune raison d'être identiques. Neanmoins, nous avons choisi de ramener l'expression de Σ_n à un scalaire, afin d'avoir une application qui fonctionne en temps réel. Dans le but d'alléger les calculs pour un traitement temps réel, on considère que les termes de la diagonale de Σ_n sont identiques. Cette simplification, qui revient à biaiser l'estimation de chaque variance, conduit à une détérioration des performances de la méthode. Neanmoins, cette détérioration reste acceptable. L'expression précédente s'écrit alors :

$$\Sigma_n = \sigma_n^2 I_3 \tag{4}$$

où I_3 est la matrice identité d'ordre 3.

Le fond de la scène peut être modélisé par une ou plusieurs gaussiennes, et le premier plan par des gaussiennes supplémentaires.

On cherche tout d'abord à sélectionner les gaussiennes expliquant le fond de la scène dans la GMM. Pour ce faire, l'ensemble des gaussiennes est ordonné par ordre décroissant de $\frac{\alpha_n}{\sigma_n}$. Ainsi, les gaussiennes qui ont à la fois le plus de poids – c'est-à-dire qu'elles expliquent le plus le fond – et une variance faible sont mises en valeur.

Les C gaussiennes expliquant le fond sont alors sélectionnées :

$$C = \arg\min_c \left(\sum_{n=1}^{n=c} \alpha_n > T\right) \tag{5}$$

où T est un seuil fixé à l'avance. T représente le rapport minimum des données que l'on peut associer au fond de la scène. En général, T est petit si le fond est unimodal et grand sinon.

Figure 2. Modèle de couleur d'un pixel dans le cas où un objet (chaise) est ajouté à la scène. On représente ici les composantes rouge et verte prises par un pixel. On remarque alors que la distribution est bimodale.

La mesure actuelle $\mathbf{Z_t}$ est associée à une gaussienne de la GMM si la condition suivante est vérifiée :

$$||\mathbf{Z_t} - \mu_\mathbf{n}|| < K\sigma_n \tag{6}$$

où K vaut 2 ou 3. L'opérateur $<$ est vrai si toutes les composantes du vecteur à gauche sont inférieures à $K\sigma_n$.
L'indice n' de la gaussienne qui réussit le test indique si le pixel $\mathbf{Z_t}$ représente le fond ou non :
- Si $n' < C$, $\mathbf{Z_t}$ appartient au fond,
- sinon, $\mathbf{Z_t}$ appartient au premier plan.
L'adaptativité du modèle est assurée en mettant à jour les paramètres de cette gaussienne en utilisant un filtre passe-bas :

$$\alpha'_n \longleftarrow (1-\delta)\alpha'_n + \delta \tag{7}$$

$$\mu'_\mathbf{n} \longleftarrow (1-\delta)\mu'_\mathbf{n} + \delta\mathbf{Z_t} \tag{8}$$

$$\sigma^{2'}_n \longleftarrow (1-\delta)\,\sigma^{2'}_n + \delta(\mathbf{Z_t} - \mu'_\mathbf{n})^T(\mathbf{Z_t} - \mu'_\mathbf{n}) \tag{9}$$

où δ est le coefficient d'apprentissage. Il permet d'ajuster la constante de temps d'adaptation du modèle. Pour les gaussiennes d'indice $n \neq n'$, la moyenne et la variance reste inchangées ; par contre, le poids est normalisé :

$$\alpha_n \longleftarrow (1-\delta)\alpha_n \tag{10}$$

Si le test effectué dans l'equation (6) échoue pour toutes les gaussiennes de la GMM, le pixel est associé au premier plan. La gaussienne de plus petit poids est alors réinitialisée avec la mesure courante :

$$\alpha_n = \delta \tag{11}$$

$$\mu_\mathbf{n} = \mathbf{Z_t} \tag{12}$$

$$\sigma_n^2 = \bar{\sigma}^2 \tag{13}$$

où $\bar{\sigma}^2$ est une variance élevée. Par exemple, on peut la fixer à 128 dans le cadre d'un traitement RGB sur 256 niveaux.
Les mêmes affectations sont également utilisées lors de l'initialisation des gaussiennes, pour la première image de la séquence.
On notera par la suite I_p la liste des points de la silhouette extraite (premier plan) de l'image I.

1.2. Résultats obtenus

1.2.1. Comportement de la GMM

La méthode développée permet d'accepter de nouveaux objets au bout d'un certain temps, s'ils sont immobiles (figure 3).
La GMM est mise à jour selon la cohérence entre les distributions du modèle et le pixel actuellement traité. Dès qu'une gaussienne explique de manière satisfaisante un pixel, elle est mise à jour (test 6). Son poids augmente alors dans la GMM, puisque les autres gaussiennes voient leur poids diminué (par la normalisation).
La figure 3, montre un exemple d'évolution de la mixture de gaussiennes (dans l'espace rouge vert) dans le cas de l'insertion d'un objet dans la scène (une chaise). Pour l'image (a), le fond est vide. Les deux gaussiennes sont initialisées avec les mêmes mesures. La variance est importante ($\bar{\sigma}^2$) et les poids identiques (0.5).
Ensuite, la première gaussienne est mise à jour progressivement (image (b)). Son poids augmente, sa variance diminue et sa moyenne reflète parfaitement la couleur du mur. La deuxième gaussienne n'explique plus le fond.
Lorsque la chaise est posée (image (c)), la deuxième gaussienne est initialisée et le poids associé à la première gaussienne diminue.
Au bout d'un certain temps (image (d)), le poids associé à la deuxième gaussienne devient supérieur à celui associé à la première gaussienne et la chaise fait partie du fond.
Dans l'image (e), la chaise est complètement intégrée au fond (fin de séquence). Seule la deuxième gaussienne explique la couleur du fond (la chaise).
La valeur de $\frac{\alpha_n}{\sigma_n}$ permet de représenter le poids d'une gaussienne par rapport à l'autre. Ce facteur est utilisé afin de les ordonner dans l'approche proposée et révèle le rapport de l'explication du fond d'une gaussienne sur l'ensemble de la GMM (voir figure 4).
Dès le début de la séquence, la première gaussienne (en pointillés) représente le fond et son poids augmente dans la GMM. Dès que la chaise est introduite dans la scène (point 1), la valeur de cette gaussienne diminue. La deuxième gaussienne représente le fond à partir du point 2. La chaise n'est plus considérée dès lors comme premier plan mais comme appartenant au fond.

1.2.2. Qualité de l'extraction

L'espace de couleur dans lequel l'extraction de fond est effectuée peut influer sur la qualité des résultats obtenus. Sur la figure 5, une extraction a été réalisée avec trois espaces de couleurs:
- l'espace RGB (Rouge Vert Bleu),
- l'espace HS (Teinte, Saturation),
- l'espace YUV (format délivré par la caméra utilisée).

On remarque que les trois espaces de couleurs donnent des résultats proches. On peut noter que dans le cas de l'utilisation de l'espace teinte saturation (HS), les densités de probabilité sont exprimées dans un espace de dimension deux, en comparaison avec les deux autres espaces qui sont d'ordre trois. La conséquence de cette réduction de dimension est une diminution du nombre de calculs effectués. D'autre part, on peut noter l'apparition de mauvaises détections au niveau de zones d'ombre (ombre portée par la personne).

1.2.3.Résultats sur une séquence d'extérieu}

La figure 6 montre les résultats obtenus dans le cas d'une séquence acquise en milieu extérieur. Une chaise est ajoutée à l'arrière plan. Sur les sous-figures (a), (b), (c), et (d), la chaise

est ajoutée. Elle appartient encore au premier plan. Dans les sous-figures (e), (f), (g) et (h), la chaise est progressivement insérée dans le modèle du fond ($\gamma = 0{,}01$ pour cette séquence). D'autre part, des zones appartenant à la personne sont classées comme fond. En fait, ces pixels correspondent à des zones sombres de la personne, avec un arrière plan lui aussi sombre.

1.3. Détails de l'implantation

Il faut que la solution dévéloppée fonctionne en temps réel vidéo (environ 15 images par seconde). Dans ce cadre, l'algorithme de soustraction de fond a été codé de manière à pouvoir agir sur les temps d'exécution. La soustraction peut être effec-

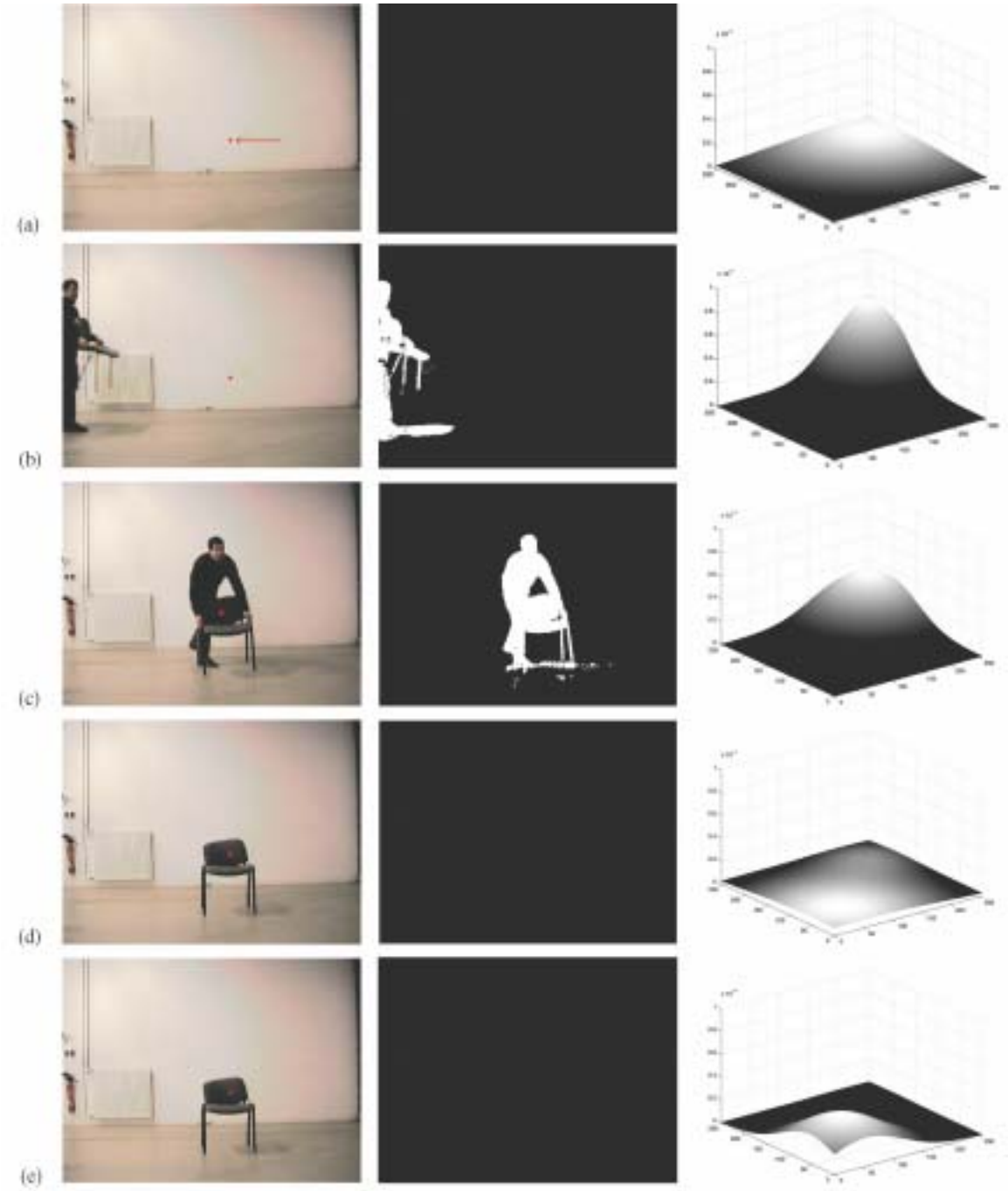

Figure 3. Exemple d'évolution de la mixture de gaussiennes (dans l'espace rouge vert) dans le cas de l'insertion d'un objet dans la scène (une chaise). (a) : initialisation ; l'écart type de chaque gaussienne est important. (b) : l'écart type de la gaussienne expliquant le fond diminue. (c) : insertion d'une chaise dans la scène. Cette dernière est classée au premier plan (d) : la chaise est vue comme un nouvel objet. (e) : l'écart type de la gaussienne initialisée sur le nouvel objet diminue et ce dernier va être considéré comme appartenant au fond.

tuée uniquement sur un sous-échantillonnage de l'image. Dans ce cas, on considérera que les pixel voisins au pixel traité sont dans le même état (avant plan ou arrière plan).
Avec un PC classique, une acquisition 640×480 et 2 gaussiennes dans la GMM, nous obtenons les résultats exposés dans la figure 7.
Dans le cas d'un traitement effectué sur un point sur quatre, la cadence vidéo est d'environ six images par secondes.

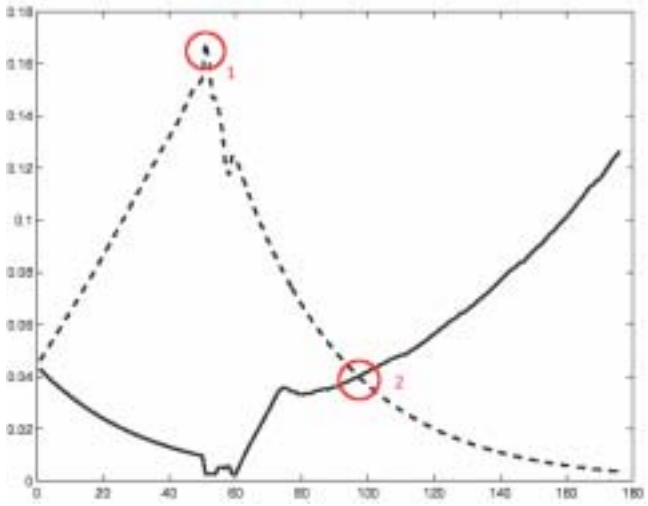

Figure 4. Comparaison des valeurs de $\frac{\alpha}{\sigma}$ des deux gaussiennes de l'exemple figure 3.

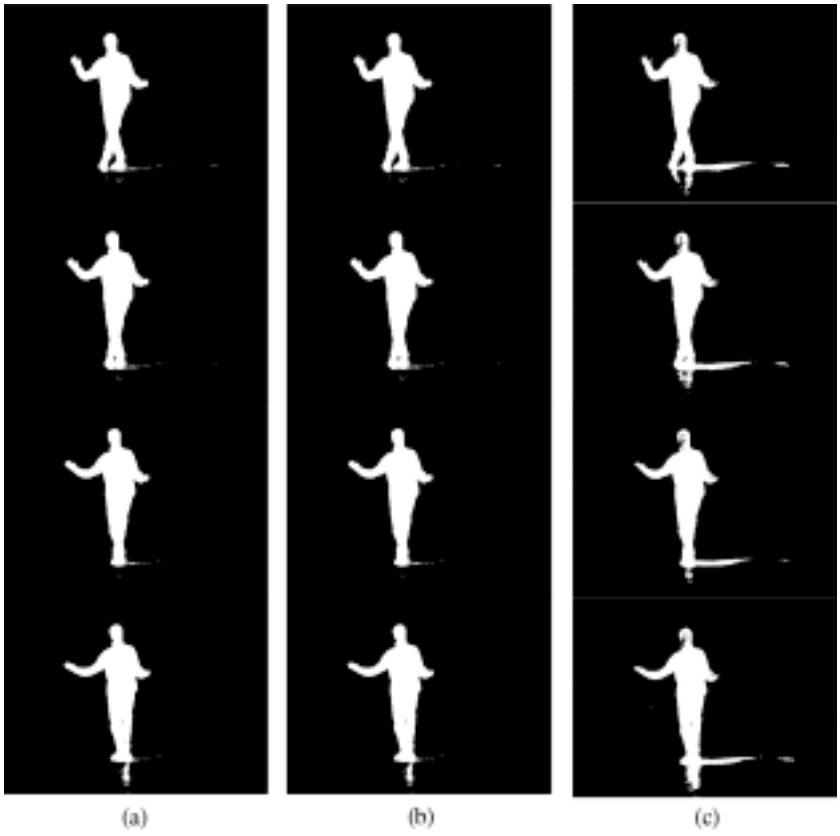

Figure 5. Comparaison des espaces couleur RGB (a), HS (b), YUV (c) dans le cas d'un algorithme de soustraction de fond.

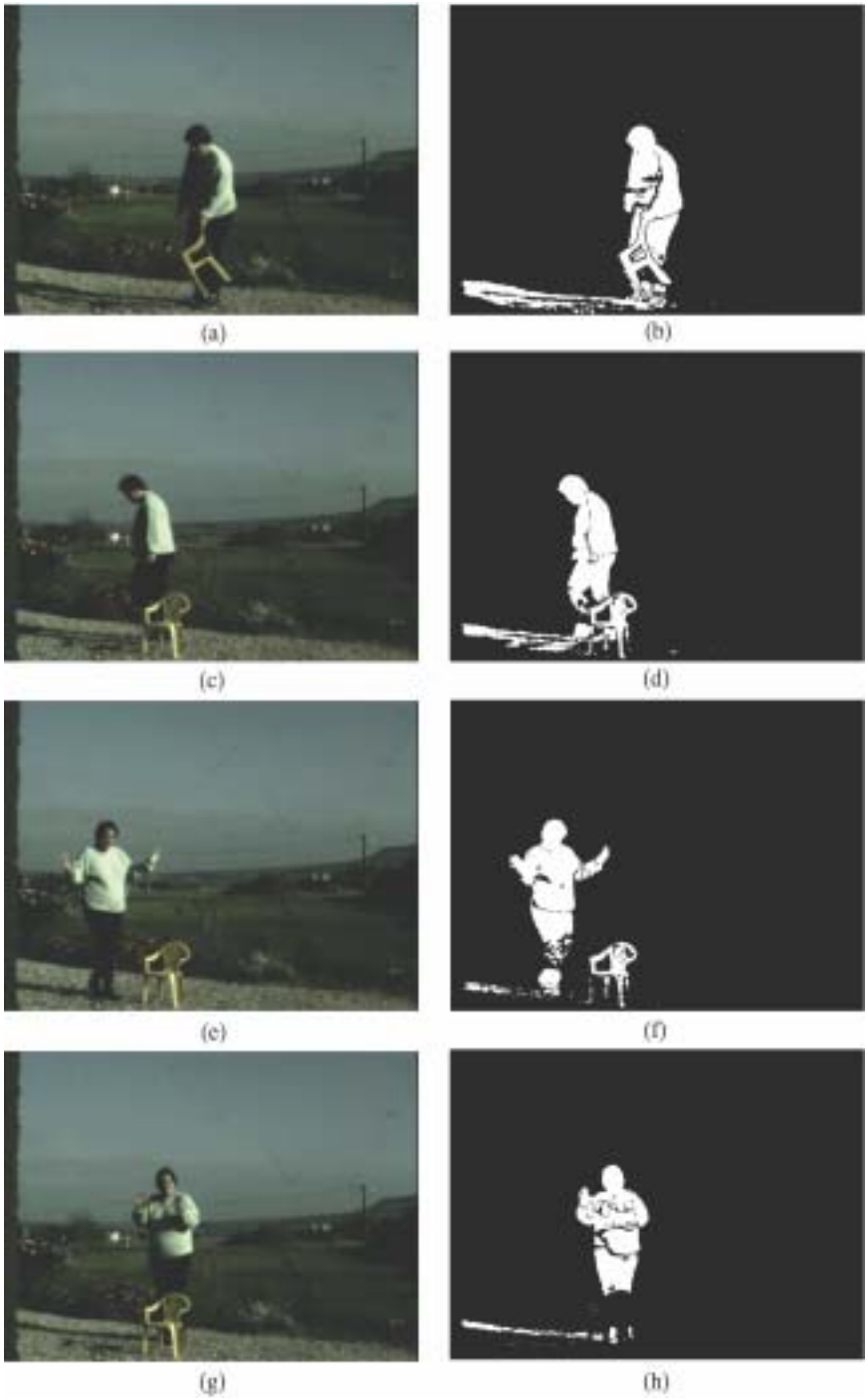

Figure 6. Illustration de la phase de suppression de fond sur une scène d'extérieur. Les images de gauche représentent les images originales. Les images de droite sont des images binarisées où le fond appraraît en noir et le premier plan apparaît en blanc.

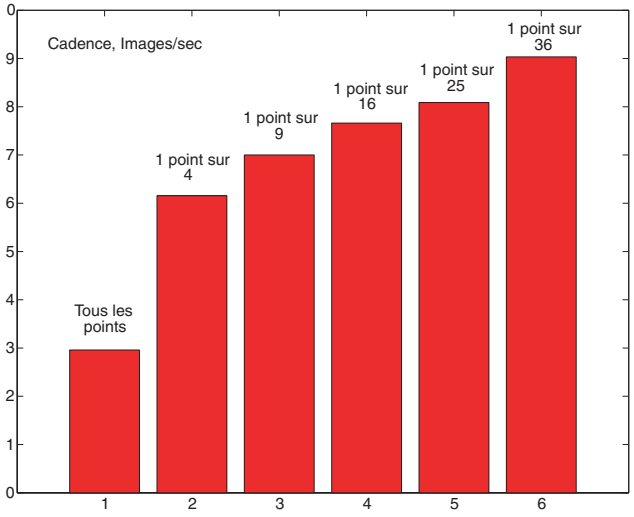

Figure 7. Nombre d'images par seconde moyen en fonction du nombre de pixels traités.

2. Détection de zones colorées

Le suivi d'objets dans des séquences d'images est un axe de recherche très largement abordé en traitement d'images. Les approches proposées peuvent se classer en deux catégories : les approches basées sur un apprentissage (*learning based approach*), et celles basées sur l'utilisation d'un modèle (*model based approach*).

Dans le premier cas, il s'agit de modéliser l'objet à suivre par une collection de vues, et d'utiliser cette base pour extrapoler, à chaque itération, la configuration de l'image courante en la comparant avec les images de la base. Dans [1], les auteurs utilisent une collection de silhouettes et recherchent la pose courante d'une personne par une méthode de régression dans cette collection de silhouettes.

Dans le cas de l'approche basée modèle, certains travaux comme [7] utilisent un modèle géométrique *a priori* de l'objet. Ce dernier peut être basé sur les contours [15], la texture [12], ou sur la couleur [5, 16]. Dans certains cas, le suivi doit être particulièrement robuste aux variations d'éclairement ou aux occultations partielles [6]. La méthode proposée ici appartient à la catégorie des approches basées modèle.

Dans le cas du suivi de la tête et des deux mains, il s'agit d'objets non rigides colorés. Nous considèrerons que le modèle de chair utilisé est le même pour les trois zones à suivre. Dans [22] les auteurs comparent les performances d'un détecteur de pixels de couleur chair pour cinq plans couleurs différents (*CIELAB*, Fleck *HS* [9], *HSV*, *RGB* normalisé et YC_rC_b). Les conclusions de ces travaux montrent que malgré des performances assez proches, les espaces couleur Fleck *HS* et *HSV* donnent les meilleurs taux de classification. Un état de l'art plus complet concernant les espaces couleur et la détection de couleur chair peut être consulté dans [21].

Dans [3], nous présentons une méthode de suivi des deux mains et de la tête d'une personne équipée de marqueurs colorés (un casque et deux gants de couleurs différentes). Nous proposons une méthode originale de détection d'une zone colorée dans une image. Dans le cas du suivi sans marqueur, il s'agit de détecter trois zones de couleur chair dans l'image. Nous étendons donc la méthode au cas de la détection de zones multiples.

2.1. La détection d'un amas coloré dans l'image

La méthode présentée permet de détecter la position, dans une image, d'un amas de points colorés dont la dimension est approximativement connue. Elle permet également d'extraire la couleur la plus représentative d'un ensemble de points.

2.1.1. L'apprentissage de la couleur

Le but de l'apprentissage de la couleur est de rechercher, dans une région d'intérêt W^i (définie de façon supervisée) de l'image I, la couleur la plus représentative (dominante). Cette dernière est codée dans un repère HSV (Hue, Saturation, Value) car il permet de séparer l'intensité des deux autres composantes. Soit $\mathbf{y}(\mathbf{u})$ le vecteur $(h,s,v)^T$, $h,s,v \in [0,1]$ (teinte, saturation, intensité) associé au pixel de coordonnées $\mathbf{u}$.

L'apprentissage de la couleur dominante est obtenu par un système de vote. Pour chaque pixel, on calcule une distance entre sa couleur et la couleur de chacun des autres pixels de la région d'intérêt. La somme de toutes les distances ainsi calculées forme alors un poids associé à la couleur du pixel courant. La fonction de poids $g_c^i(u)$, associée à la zone i, est alors définie par :

$$\forall \mathbf{u} \in W^i, g_c^i(\mathbf{u}) = \sum_{\mathbf{x}\in W^i,\ \mathbf{x}\neq\mathbf{u}} \exp\left[-\lambda_c \cdot d_c(\mathbf{y}(\mathbf{x}),\mathbf{y}(\mathbf{u}))\right] \qquad (14)$$

Le paramètre λ_c permet d'ajuster le calcul du poids. La distance d_c utilisée est une distance de Mahalanobis définie par :

$$d_c(\mathbf{y}_1,\mathbf{y}_2) = [\mathbf{y}_1 - \mathbf{y}_2]^T \cdot C_c^{-1} \cdot [\mathbf{y}_1 - \mathbf{y}_2] \qquad (15)$$

avec C_c, matrice de covariance associée aux trois composantes du vecteur couleur :

$$C_c = \begin{pmatrix} \sigma_h^2 & 0 & 0 \\ 0 & \sigma_s^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{pmatrix} \qquad (16)$$

Dans le calcul du modèle, l'intensité n'est pas prise en compte : on choisira donc $\sigma_v \to \infty$.

La couleur dominante associée à la zone, notée $\mathbf{y}^{i*}$ est donnée par celle du pixel qui posséde le poids maximum :

$$\mathbf{y}^{i*} = \mathbf{y}(\text{argmax}(g_c^i)) \qquad (17)$$

2.1.2. Localisation de l'amas

Une fois le modèle de couleur associé à la zone chair ($\mathbf{y}^{i*}$) connu, il convient de rechercher, dans chaque nouvelle image, la position des trois zones de couleur chair. La plupart des algorithmes de suivi sont basés sur la définition d'une région d'intérêt remise à jour à chaque image. Nous avons choisi une approche différente, qui consiste à balayer toute l'image afin de rechercher le plus gros amas de points dont la couleur est proche de celle du modèle. Ainsi, une liste de points candidats L^i est construite :

$$L^i = \{\mathbf{u} \in I_s / \ |\mathbf{y}(\mathbf{u}) - \mathbf{y}^{i*}| < \sigma\}\ ,\ \sigma = (\sigma_h, \sigma_s, \sigma_v)^T \qquad (18)$$

I_s est constitué des points de l'image de premier plan I_p dont la valeur de l'intensité est supérieure à un seuil v_l. En effet, un point dont l'intensité est proche de zéro possède une teinte mal définie :

$$I_s = \{\mathbf{u} \in I_p / \ \mathbf{y}(\mathbf{u}) > (0,0,v_l)^T\} \qquad (19)$$

Dans les équations 18 et 19, l'opérateur d'inégalité $<$ (resp. $>$) utilisé entre 2 vecteurs est vérifié si les inégalités concernant chaque élément des 2 vecteurs sont vérifiées.
La localisation du centre de l'amas de points le plus représentatif est effectué par une méthode similaire à celle utilisée pour la détermination du modèle de couleur. La fonction de poids $g_s^i(u)$, est alors définie par :

$$\forall \mathbf{u} \in L^i, g_s^i(\mathbf{u}) = \sum_{\mathbf{x} \in L^i,\ \mathbf{x} \neq \mathbf{u}} \exp[-\lambda_s . d_s(\mathbf{x}, \mathbf{u})] \qquad (20)$$

Le paramètre λ_s permet d'ajuster le calcul du poids. De plus, la distance d_s est une distance de Mahalanobis définie par :

$$d_s(\mathbf{u}_1, \mathbf{u}_2) = [\mathbf{u}_1 - \mathbf{u}_2]^T . C_s^{-1} . [\mathbf{u}_1 - \mathbf{u}_2] \qquad (21)$$

avec C_s, matrice de covariance associée aux 2 composantes spatiales de l'image :

$$C_s = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \qquad (22)$$

Le choix de σ_x et σ_y permet d'ajuster la taille des amas que l'on veut détecter. La position du centre de gravité de chaque zone $\hat{\mathbf{u}}_i$ est alors estimée par :

$$\hat{\mathbf{u}}_i = \text{argmax}(g_s^i) \qquad (23)$$

2.2. Localisation de plusieurs amas

Dans le cas d'une image dans laquelle plusieurs zones sont présentes, la relation (23) fournit la position du centre de gravité de l'amas le plus représentatif. En considérant que la taille de cet amas est grossièrement connue (σ), une nouvelle liste de points candidats L_{k+1}^i est constituée à partir de L_k^i (en considérant que l'indice k est associé au numéro de l'amas courant détecté) en supprimant les points situés dans un voisinage (de taille $K.\sigma$) de $\hat{\mathbf{u}}_i^k$:

$$L_{k+1}^i = \{\mathbf{u} \in L_k^i | (\mathbf{u} - \hat{\mathbf{u}}_i^k)^T (\mathbf{u} - \hat{\mathbf{u}}_i^k) < (K.\sigma)^2\} \qquad (24)$$

K est un coefficient qui vaut habituellement 2 ou 3. Les relations (20) et (23) peuvent alors être appliquées pour estimer la position de l'amas k dans l'image.

2.3. Résultats obtenus

Les images exposées dans la figure 8 montrent les pixels détectés comme couleur peau pour plusieurs images test.
D'une manière générale, l'extraction fonctionne correctement. Certains points appartenant essentiellement au bord de la silhouette sont parfois classés comme peau.

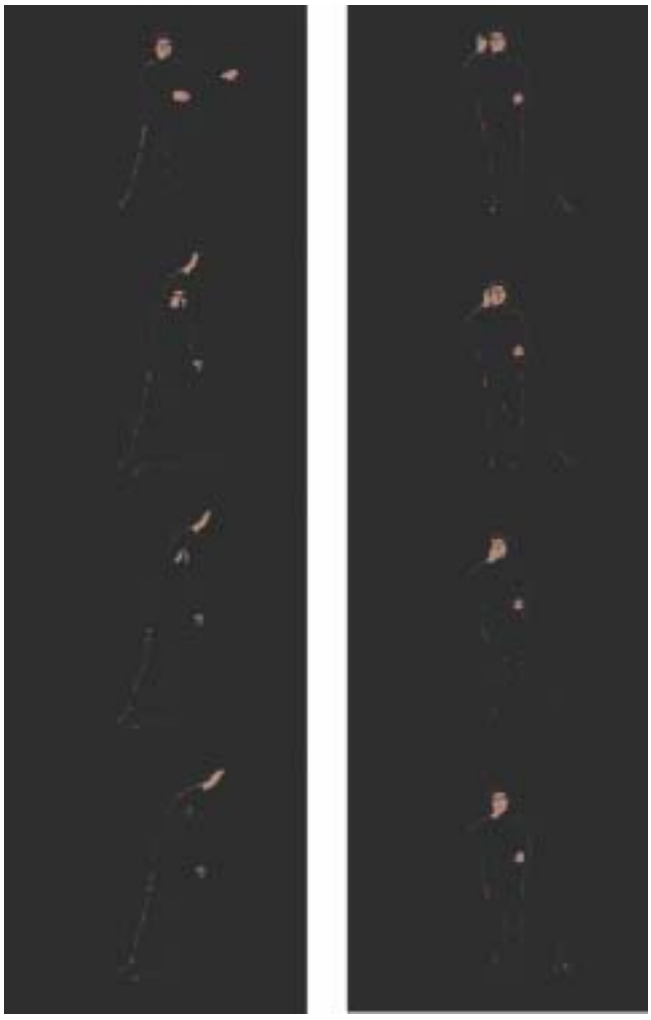

Figure 8. Exemples d'extraction des pixels de couleur peau.

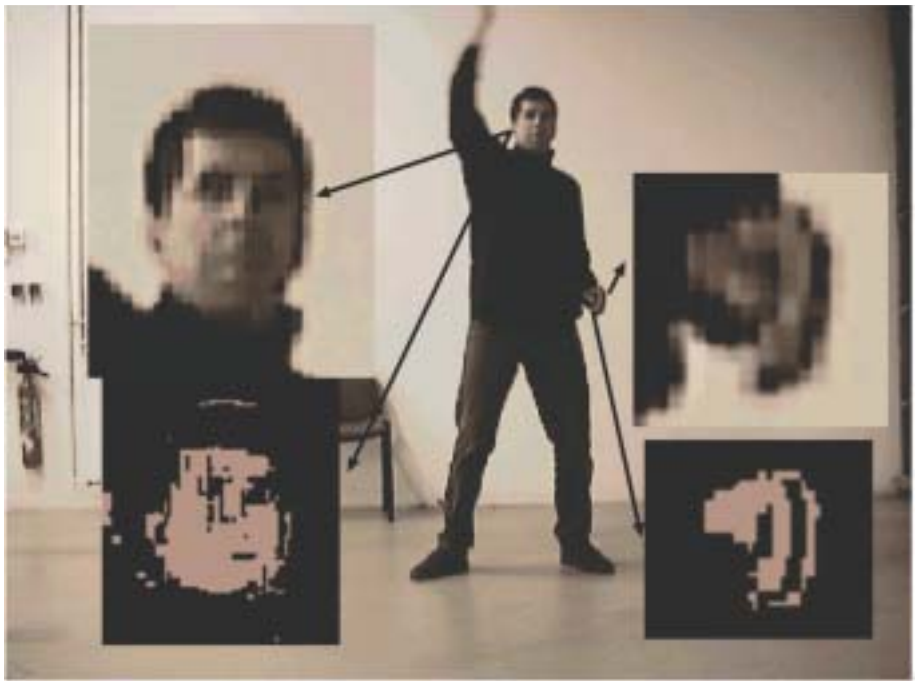

Figure 9. Illustration de la robustesse de l'extracteur de peau pour des conditions d'éclairement défavorables.

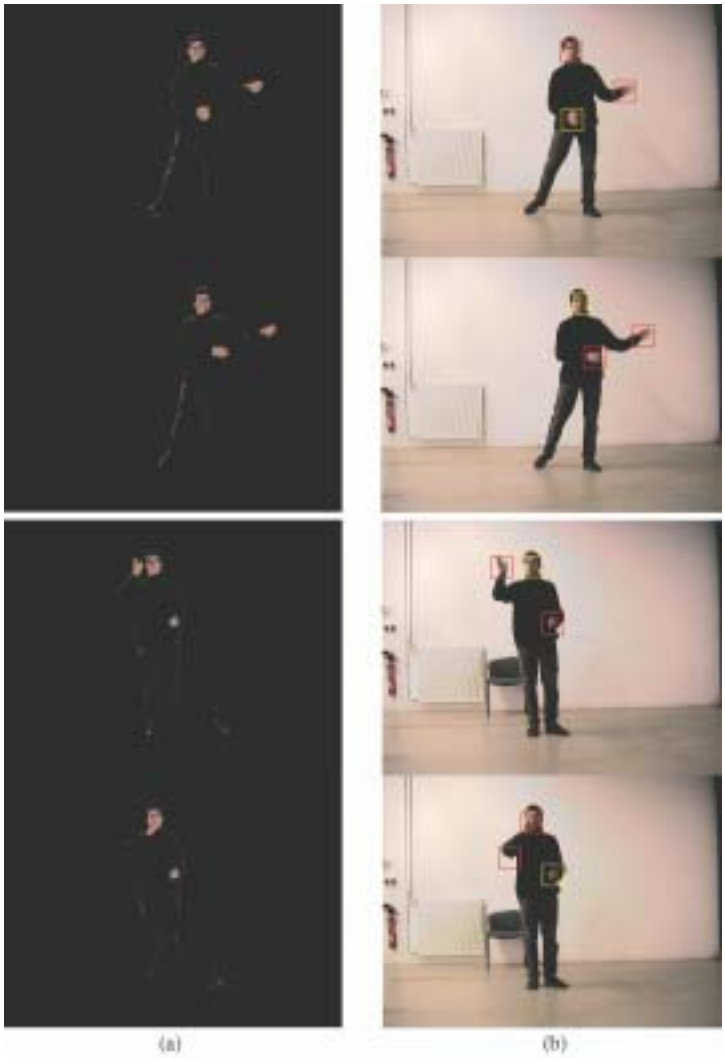

Figure 10. Exemples de localisation des trois zones de couleur peau.

La méthode proposée est particulièrement robuste face aux variations d'intensité lumineuse. La figure 9 montre le résultat de l'extraction, pour une image dons laquelle l'éclairage n'est pas constant au niveau des mains et de la tête. Les performances de l'extracteur sont peu sensibles à ce phénomène d'ombrage. La figure 10 montre les positions des trois zones détectées. Les résultats obtenus sont satisfaisants et, malgré la présence de points parasites, la localisation est correcte pour les trois premières images. Pour la quatrième image, la main droite vient masquer le visage. Il n'apparaît plus que deux zones significatives dans l'image. La détection de la troisième zone s'effectue alors sur des points parasites.

3. Suivi temporel par filtrage particulaire

Le module de détection de zones colorées fournit, à l'image k, un ensemble de trois vecteurs $\mathcal{U}_k = \{\widehat{\mathbf{u}}_1^k, \widehat{\mathbf{u}}_2^k, \widehat{\mathbf{u}}_3^k\}$ représentant la position du centre de gravité de la tête et des deux mains dans l'image courante. À ce stage, aucune labellisation n'est effectuée ; on ne sait pas à quel élément de l'ensemble correspond la position de la tête. D'autre part, d'une image sur l'autre, cet élément peut changer. Ces deux remarques font apparaître deux problèmes distincts. D'une part, il faut être capable de localiser, parmis les trois zones, celle qui correspond à la tête. D'autre part, il faut connaître l'évolution temporelle de chaque zone. Pour ce faire, nous proposons d'utiliser un filtre à particules, particulièrement adapté aux systèmes dont les modèles d'évolution et de mesure ne sont pas gaussiens.

3.1. Principe du suivi d'objets par filtrage particulaire

Le problème du suivi d'un objet dans une image peut être décrit par la recherche de la densité de probabilité *a posteriori* $p(\mathbf{x}_t|\mathbf{z}_{1:t})$, à partir de la densité de probabilité $p(\mathbf{z}_{1:t}|\mathbf{x}_t)$; où $\mathbf{x}_t$ et le vecteur d'état composé des paramètres du modèle à suivre et $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$ est le vecteur d'observation qui regroupe l'historique des mesures. Le suivi s'effectue alors en deux étapes. La première phase consiste en une prédiction de $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ à partir de $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ et d'une densité de transition $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. La deuxième phase utilise la règle de Bayes afin de mettre à jour les $p(\mathbf{x}_t|\mathbf{z}_t)$ en fonction des nouvelles mesures.
Dans le cas où les densités de probabilités sont gaussiennes, le filtre de Kalman fournit une solution optimale au problème d'estimation. Malheureusement, la plupart du temps, les distributions ne sont pas gaussiennes. C'est notamment le cas de cette application. L'estimation de la solution peut passer par l'utilisation de méthodes de Monte-Carlo.
La simulation de Monte-Carlo effectuée à l'aide d'un filtre à particules se base sur une densité de probabilité *a posteriori* estimée par un ensemble de particules pondérées $\{(\mathbf{s}_k^{(i)}, \pi_k^{(i)})\ i = 1, \ldots, N_F\}$. Une présentation détaillée du filtrage particulaire est donnée dans [2]. L'application la plus connue de cette méthode en traitement d'images est l'algorithme CONDENSATION proposé par Isard et Black [15], et repris dans de nombreux travaux [13, 8, 14, 16]. Il se décompose en cinq étapes :

1. Durant l'initialisation, un jeu de particules est généré, en fonction de la valeur initiale du vecteur d'état $\mathbf{x}_0$. La valeur du poids associé à chaque particule est fixé à l'inverse du nombre de particules.
2. Une prédiction de l'état de chaque particule est alors calculée, en fonction d'un modèle d'évolution. Dans le cas du suivi d'objets dans une séquence d'images, soit on considère un modèle simple (position constante, vitesse constante, accélération constante, etc.), soit le modèle d'évolution est construit à partir d'un apprentissage.
3. Le poids associé à chaque particule est alors mis à jour en fonction des mesures.
4. Le calcul de l'espérance du jeu de particules fournit une estimation de l'état du système.
5. Un nouveau jeu de particules est choisi. Il s'agit d'un tirage au sort avec remise, de N_F particules parmi le jeu courant. Plus une particule du jeu courant a un poids élevé, plus elle est susceptible d'être choisie plusieurs fois ; de même, une particule de poids faible a de fortes chances d'être éliminée. Une fois le nouveau jeu de particules construit, le système boucle à l'étape 2.

Dans le cas présent, le vecteur d'état $\mathbf{x}_k$, à l'instant k, est constitué des coordonnées, dans l'image, des centres de gravité des trois zones à suivre :

$$\mathbf{x}_k = (\mathbf{c}_k^1, \mathbf{c}_k^2, \mathbf{c}_k^3)^T \tag{25}$$

Pour chaque particule, il convient alors de définir une fonction de mesure permettant de mettre à jour le poids $\pi_k^{(i)}$ associé. Cette dernière est basée sur la distance de Haussdorf, définie par :

$$\mathrm{H}(\mathcal{U}_k, \mathcal{C}_k) = \max(\mathrm{h}(\mathcal{U}_k, \mathcal{C}_k), \mathrm{h}(\mathcal{C}_k, \mathcal{U}_k)) \tag{26}$$

$$\mathrm{h}(\mathcal{U}_k, \mathcal{C}_k) = \max_i(\min_j \mathrm{D}(\mathbf{u}_i^k, \mathbf{c}_k^j)) \tag{27}$$

où $\mathcal{C}_k = \{\mathbf{c}_k^1, \mathbf{c}_k^2, \mathbf{c}_k^3\}$. La distance D utilisée est une distance Euclidienne.
Le poids $\pi_k^{(i)}$ est alors défini par une relation exponentielle par rapport à la mesure de distance :

$$\pi_k^{(i)} = \exp[-\lambda H(\mathcal{U}_k, \mathcal{C}_k)] \tag{28}$$

où λ est un paramètre permettant d'ajuster la vitesse de décroissement du poids en fonction de la distance.

Figure 11. Les positions détectées de la tête et des mains avant (images de gauche) et après utilisation d'un filtre à particules (images de droite).

3.2. Résultats obtenus

3.2.1 Filtrage de la trajectoire

La figure 11 montre les résultats obtenus avant filtrage (figure 11 (a)), puis ceux obtenus après filtrage (figure 11 (b)).
Losqu'une main est proche ou devant la tête, les deux amas de pixels de couleur peau se superposent et il n'y a plus que deux zones cohérentes dans l'image. La troisième zone détectée correspond alors aux bruits résiduels. C'est le cas des images (c), (e) (g) et (i). L'utilisation du filtre à particules permet de gérer la cohérence temporelle de la trajectoire des mains. On constate que sur les images (d), (f), (h) et (i), la localisation est correcte, malgré le phénomène d'occultation.
La figure 12 montre les trajectoires images obtenues avant et après filtrage.
Les mauvais résultats obtenus sans l'utilisation du filtre à particule sont dus à deux phénomènes différents : d'une part, les occultations produisent des erreurs de localisation lors de la détection des trois zones de couleur peau. D'autre part, l'ordre de détection des trois amas n'est pas toujours le même, ainsi, sur certaines images, le bras droit est inversé avec le bras gauche. Ce problème de labélisation est résolu grâce au filtre à particules. Les trajectoires obtenues à la sortie de ce dernier sont satisfaisantes.
La figure 13 montre les résultats obtenus dans le cas d'une séquence d'extérieur. Le système est capable de gérer des superpositions temporaires de zones (images (g) (h) (i)). Dans le cas de cette image (g), la main droite passe devant le visage. Deux zones sont alors visibles dans l'image. Sans filtrage, la troisième zone se localise sur des pixels correspondant à du bruit (h). Lorsque le filtrage est actif (image (i)), les trois zones restent correctement localisées. Les particules (points verts sur l'image) se diffusent alors dans toutes les directions.

3.2.2. Traitement total du système

La figure 14 illustre les résultats obtenus pour les différentes étapes du système.

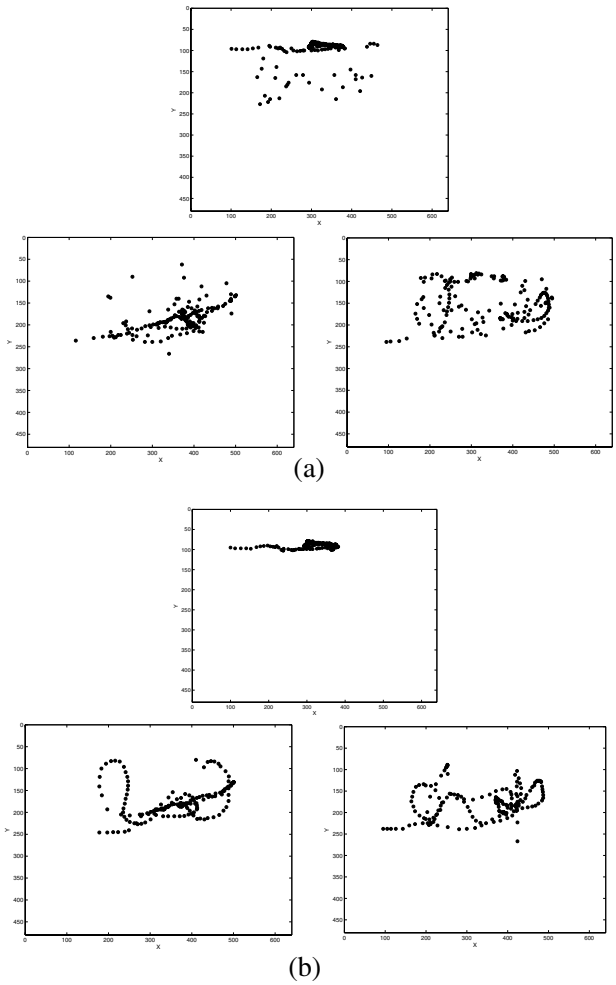

Figure 12. Trajectoires parcourues avant (a) et après utilisation du filtre à particule (b). Pour chaque sous-figure, la trajectoire de la tête correspond au graphe du haut et la trajectoire des deux mains correspond aux deux graphes du bas.

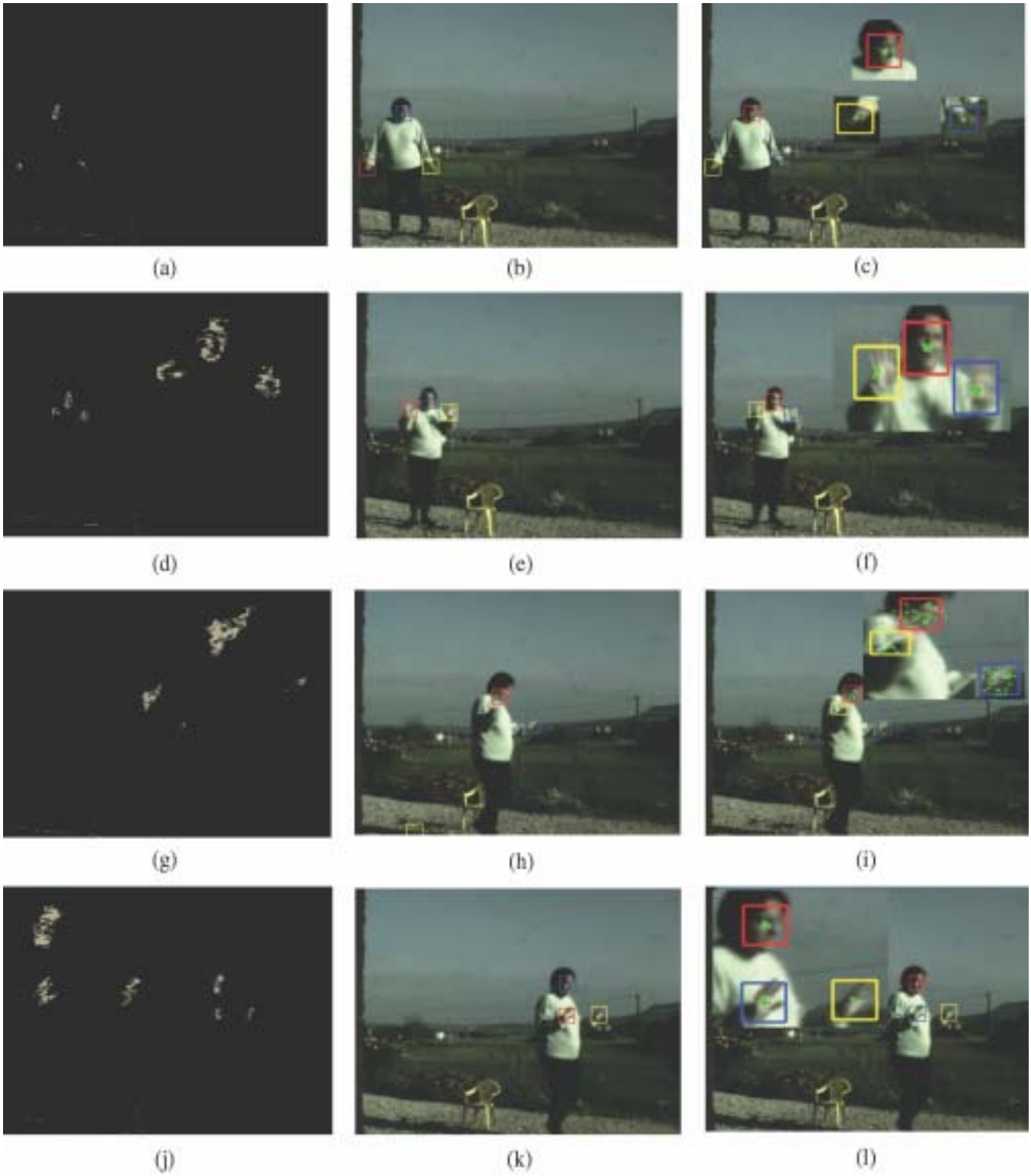

Figure 13. Illustration de la phase de filtrage, sur une séquence d'extérieur. Lorsque les mesures ne permettent pas de localiser les 3 zones pendant quelques itérations (images (g) (h) (i)), le filtre à particules permet de prédire une position estimée. Les points verts représentent les positions des particules. Les coins hauts droits des images (d) et (g), ainsi que le coin haut gauche de l'image (j) sont des zooms de la zone proche de la tête de la personne.

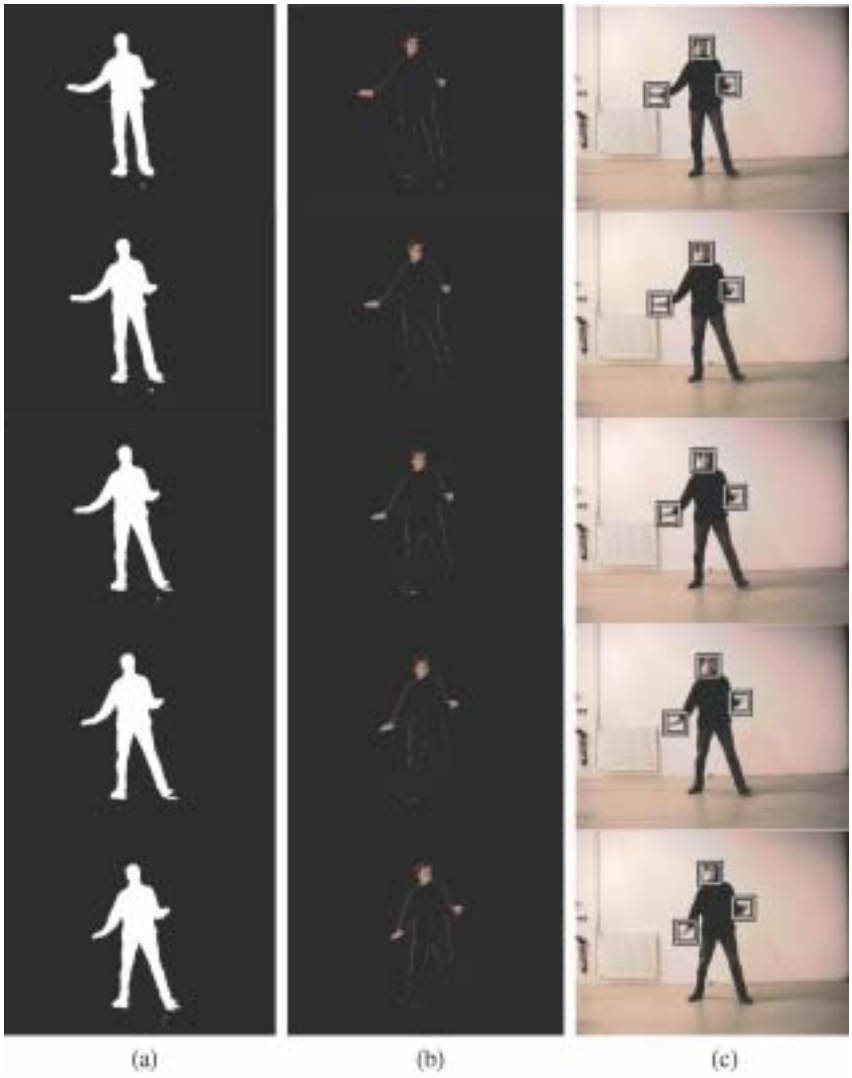

Figure 14. Le traitement total du suivi. En (a), l'extraction du fond, en (b), la détection de la peau et en (c), le filtrage et le renvoi des coordonnées de la tête et des mains.

4. Conclusion

Ce article propose une solution au problème de suivi de la tête et des deux mains, sans marqueurs, d'une personne par vision monoculaire couleur.

Une pré-selection de points appartenant à la personne à suivre est donnée par l'utilisation d'une méthode de soustraction de fond. Cette dernière, adaptative, permet de prendre en compte l'arrivée de nouveaux objets dans la scène, ou les variations lentes de luminance au cours du temps. Nous avons comparé les performances de cette approche, dans le cas de trois plans couleurs.

Nous avons proposé une méthode de détection de plusieurs amas colorés, permettant de fournir les centres de gravité de trois zones correspondant à la tête et aux deux mains de la personne filmée. L'utilisation d'un codage teinte, saturation permet d'obtenir des résultats satisfaisants lorsque l'illuminant n'est pas constant suivant la position de la zone détectée.

La cohérence temporelle des trois zones détectées a été prise en compte par l'utilisation d'un filtre à particules, qui permet de gérer efficacement les occultations partielles et le chevauchement de deux zones.

Le système proposé fonctionne actuellement à une cadence de 6 images par seconde, ce qui est encore insuffisant dans le cadre des applications de suivi de geste courantes. Des optimisations des modules d'extraction de fond et de reconnaissance sont en cours afin d'améliorer ces performances.

Références

[1] A. AGARWAL, B. TRIGGS, 3D Human pose from Silhouettes by Relivance Vector Regression, In *To appear in IEEE Internationnal Conference on Computer Vision and Pattern Recognition*, 2004.

[2] S. ARULAMPALAM, S. MASKELL, N. GORDON, T. CLAPP, A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking, *IEEE Transactions on Signal Processing*, 50(2):174-188, February 2002.

[3] F. BARDET, T. CHATEAU, F. JURIE, M. NARANJO, Interactions geste-musique par vision artificielle, In *Workshop Acquisition du geste humain par vision artificielle, dans RFIA04, Congrès sur la Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, January 2004, Actes sur CDROM.

[4] T. CHATEAU, F. JURIE, R. MARC, Reconnaissance de gestes par vision monoculaire en temps réel : application la formation des chargés de manoeuvres pour la conduite des ponts polaires, In *Workshop Acquisition du geste humain par vision artificielle, dans RFIA04, Congrès sur la Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, January 2004, Actes sur CDROM.

[5] D. COMANICIU, V. RAMESH, P. MEER, Real-time tracking of non-rigid objects using mean shift, *Conference on Computer Vision and Pattern Recognition*, 2:142-149, 2000.

[6] D. CREMERS, T. KOHLBERGER, C. SCHNÖRR, Nonlinear Shape Statistics in Mumford-Shah Based Segmentation, In *7th European Conference on Computer Vision*, volume 2, 93-108, Copenhagen, Denmark, May 2002.

[7] D. DEMIRDJIAN, T. DARELL, 3-D Articulated Pose Tracking for Untethered Deictic Reference, In *ICMI 2002*, Pittsburgh, Pennsylvania, USA, October 2002.

[8] E. MEIER, F. ADE, Using the condensation algorithm to implement tracking for mobile robots, In *Third European Workshop on Advanced Mobile Robots, Eurobot99, IEEE*, 73-80, 1999.

[9] M. M. FLECK, D. A. FORSYTH, C. BREGLER, Finding naked people, In *European Conference on Computer Vision*, volume 2, 592-602, 1996.

[10] D.M. GAVRILA, The Visual Analysis of Human Movement: A Survey, *Computer Vision and Image Understanding*, 73(1):82-98, January 1999.

[11] E. HAYMAN, Jan-Olof Eklundh, Statistical background subtraction for a mobile observer, In *Int. Conf. Computer Vision*, 67-74, Nice, France, 2003.

[12] F. JURIE, M. DHOME, Real time template matching, In *Proc. IEEE International Conference on Computer vision*, 544-549, Vancouver, Canada, July 2001.

[13] K. NUMMIARO, E. KOLLER-MEIER, L. VAN GOOL, Object Tracking with an Adaptive Color-Based Particle Filter, In Symposium for Pattern Recognition of the DAGM, September 2002.

[14] E. B. KOLLER-MEIER, L. VAN GOOL, Modeling, Recognition of Human Actions Using a Stochastic Approach, In *2nd European Workshop on Advanced Video-based Surveillance Systems AVBS'01*, London, 4 September 2001.

[15] M. ISARD, A. BLAKE, Condensation – conditional density propagation for visual tracking, *IJCV : International Journal of Computer Vision*, 29(1):5-28, 1998.

[16] P. PEREZ, C. HUE, J. VERMAAK, M. GANGNET, Color-Based Probabilistic Tracking, In *Computer Vision ECCV 2002*, volume 1, 661-675, May 2002.

[17] J. SOBOTTKA, I. PITTAS, Segmentation and tracking of faces in color images, In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 236-241, 1996.

[18] C. STAUFFER, W. E. L. GRIMSON, Adaptative background mixture models for a real-time tracking, In *Conference on Computer Vision and Patern Recognition*, volume II, 246-252, 1999.

[19] A. TRÉMEAU, C. MALOIGNE-FERNANDEZ, and P. BONTON, *Image numérique couleur*, Dunod, 2004.

[20] M.H. YANG, N. AHUJA, Mark. Tabb, Extraction of 2D Motion trajectories and Its Application to Hand gesture Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061-1074, August 2002.

[21] B. D. ZARIT, Skin detection in video images, Technical report, University of Illinois, Chicagi, 1998.

[22] B. D. ZARIT, B. J. SUPER, K.H. QUEK, Comparison of Five Color Models in Skin Pixel Classification, In *ICCV'99 International Workshop on Recognition, Analysing, and Tracking of Faces and Gestures in Real-Time Systems, RATFG-RTS'99*, 58-63, Corfu, Greece, September 1999.

Thierry **Chateau**

Thierry Chateau est ingénieur du CUST (1995) et docteur de l'Université Blaise Pascal (1998). Il est maitre de conférences au CUST depuis 2001, et effectue ses travaux de recherches au LASMEA, UMR6602, CNRS / Université Blaise Pascal. Ses recherches portent sur le suivi d'objets et la reconnaissance des formes dans des séquences d'images.

Antoine **Vacavant**

Antoine Vacavant est titulaire d'une maitrise d'informatique de l'Université Blaise Pascal, Clermont Ferrand (2004). Il effectue actuellement un Master Recherche à l'Université Claude Bernard de Lyon en option Informatique, Graphique et Image.