



**HAL**  
open science

## Reconnaissance et catégorisation de l'activité manuelle humaine

Youssef Chahir, Michèle Molina, François Jouen

► **To cite this version:**

Youssef Chahir, Michèle Molina, François Jouen. Reconnaissance et catégorisation de l'activité manuelle humaine. *Studia Informatica Universalis*, 2010, 8 (4), pp.31-57. hal-00808267

**HAL Id: hal-00808267**

**<https://hal.science/hal-00808267>**

Submitted on 5 Apr 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# Reconnaissance et catégorisation de l'activité manuelle humaine

## Human Hand Movement Recognition

Youssef Chahir \* — Michèle Molina \*\* — François Jouen \*\*\*

\* GREYC - CNRS UMR 6072

Université de Caen, France

youssef.chahir@info.unicaen.fr

\*\* Laboratoire PALM JE 2528

Université de Caen, France

michele.molina@unicaen.fr

\*\*\* Laboratoire CHArt EA 4004-CNRS UMS 2809

EPHE, Paris, France

francois.jouen@ephe.sorbonne.fr

---

**RÉSUMÉ.** Dans cet article, nous avons adopté une approche unifiée pour l'analyse structurale de bases de vidéos, basée sur l'analyse spectrale et les réseaux de neurones.

Cette approche offre un cadre unifié pour la réduction, la catégorisation et la reconnaissance de données complexes. Le but est de développer une méthode d'analyse vidéo capable d'interpréter des gestes de toucher de la main. Dans un premier temps, nous avons cherché à différencier les deux mains Gauche ou Droite. Ensuite, nous avons étudié les objets selon leurs consistance et leur texture. Pour chaque propriété, deux modalités ont été proposées. La texture de l'objet peut être lisse ou granuleuse. L'objet pouvait être dur ou mou. Une action humaine étant fortement liée au mouvement, nous proposons dans cet article de suivre les mains en mouvement et de former un volume dans l'espace 3D ( $2d+t$ ). Ce volume qui représente un geste 3D sera caractérisé par des moments géométriques 3D qui sont invariants à la translation et au changement d'échelle. Dans cet article, nous présentons une nouvelle approche de catégorisation des gestes basée sur la diffusion géométrique par marches aléatoires sur graphe, et une méthode de réseaux de neurones pour la reconnaissance de la nature des gestes et leur but. Nous décrivons l'implémentation de notre approche et nous montrons les résultats de validation très prometteurs sur un corpus vidéo de gestes manuels.

**ABSTRACT.** *In this paper, we adopted a unified framework for structural analysis of video databases using spectral graph techniques and neural networks. The framework provides an efficient approach for both clustering, data organization, dimension reduction and recognition. The aim is to develop vision-based approach for analysis of haptic gesture. In this study, we investigated how 3D hand gesture was analyzed when the objects are explored using touch. We focused on the extraction of perceptual similarity of haptic modality and investigated how changes in the type of hand movement used to explore the objects affected similarity space. We propose an automatic approach for hand recognition Left or Right, and hand gesture analysis and recognition for understanding human action and manipulation. Our system has been conceived to recognize the texture and the consistency of an object through the video analysis of hand actions, and to recognize which hand is presented in the video. The texture of the object being explored could either be smooth or granular. Its consistency could either be hard or soft. To enhance robustness, each given video sequence is characterized globally by a volume which is characterized by spatio-temporal features including 3D geometrical moments which are invariants with the translation and the scaling. We used two approaches as recognition methods. Firstly, the diffusion maps is used to categorize gestures and actions in the video, and the neural networks are used to recognize them. We tested the proposed approaches with different hand gestures. Results showed that our framework is effective, achieving a high recognition rate in both approaches.*

**MOTS-CLÉS :** *Analyse vidéo, gestes 3D, marches aléatoires, coupe de graphe, cartes de diffusion, flux optique, moment géométrique 3D*

**KEYWORDS:** *Video analysis , Hand Gestures, Graph-Cut, Diffusion Maps , Neural Networks ,Pattern Recognition , Random walk, Nyström , hand gesture ,3D geometrical moments*

---

## 1. Introduction

Ces dernières années, le problème de la reconnaissance et de la classification des activités humaines a suscité l'intérêt de communautés de recherche de plus en plus larges qui vont des neurosciences aux sciences de l'ingénierie en passant par la biomécanique, l'informatique et les sciences de la communication. Dans différentes applications, telles la vidéo surveillance, l'archivage ou l'indexation de vidéos, il est important de reconnaître les mouvements des personnes pour pouvoir interpréter leurs comportements. La reconnaissance d'activité nécessite l'extraction de données multiples, l'interprétation automa-

tique des séquences vidéos, et fait appel à des techniques d'analyse vidéo (perception visuelle, estimation de mouvement...) et des méthodes d'analyse et de classification de données. Ce problème d'identification devient crucial lorsque le nombre d'individus augmente, lorsque les points de vue de caméras sont différents, ou encore lorsque les sujets sont dans des environnements complexes.

Le problème de l'identification est également crucial lorsque l'on s'intéresse à des activités segmentaires, comme l'activité manuelle, et non plus seulement à des mouvements globaux de l'ensemble du corps. Cette difficulté est augmentée par la double fonction des mains. Les mains ont, en effet, deux fonctions imbriquées : une fonction motrice, dans laquelle les perceptions tactiles aident à la réussite des actions, et une fonction perceptive, dans laquelle la motricité est au service de la perception pour identifier et connaître les objets. Selon Klatzky et Lederman [LK87], l'identification et la reconnaissance des propriétés des objets sont obtenues par l'exécution de procédures exploratoires élémentaires ("PEs"). Selon ces auteurs, l'exploration tactile est une activité séquentielle réalisée via l'exécution de mouvements stéréotypés des doigts et de la paume des mains. Ces mouvements sont intentionnels et dépendent de la propriété d'objet que le système tactile choisit de traiter. Une PE se définit donc comme un mouvement stéréotypé de main dont la spécificité permet d'établir la relation la plus efficace entre une propriété de l'objet et les récepteurs sensoriels impliqués dans la perception haptique. Par exemple, la rugosité d'un objet sera évalué par des mouvements circulaires de la pulpe des doigts sur la surface de l'objet alors que la forme globale sera analysée en fermant la main autour de l'objet. Klatzky et Lederman ont ainsi montré que, chez l'adulte privé d'informations visuelles, une procédure exploratoire spécifique doit être appliquée pour chaque propriété des objets : frottement latéral de la surface pour la texture, pression pour la consistance, enveloppement et suivi des contours pour la forme et la taille, etc.[?]. Ces mouvements sont intentionnels et la propriété sera mal perçue et par conséquent mal identifiée si ces mouvements ne sont pas produits par le sujet. L'existence d'une multiplicité de procédures haptiques exploratoires pour traiter les différentes propriétés des objets de notre environnement rajoute encore à la complexité de l'analyse vidéo des mouvements manuels.

Les méthodes d'identification des comportements sont généralement basées sur les modèles d'apparence 2D ou 3D. Une catégorie de travaux consiste à détecter les différentes parties du corps humain telles que la tête, les mains, les pieds ainsi que d'autres parties du corps telles que les articulations [PN94]. Haritaoglu et al. [BP97] proposent un système global de reconnaissance qui est fondé sur les projections horizontales et verticales de la silhouette de la personne, et de son orientation par rapport à la caméra (vue de face, vue de côté gauche, ...). Iwasawa et al [MK04] ont proposé une méthode qui consiste d'abord à déterminer le centre de gravité de la silhouette, à calculer ensuite l'orientation de la moitié supérieure du corps, et à estimer enfin les différentes parties significatives du corps en utilisant une analyse heuristique du contour de la silhouette. D'autres travaux, cherchent à suivre et interpréter le mouvement humain dans l'action.

Plusieurs méthodes utilisent l'historique du mouvement comme base pour la reconnaissance de gestes [YCS08, OTS04]. Des caractéristiques appropriées sont extraites de cette image et différentes méthodes de classification sont utilisées tels que les réseaux de neurones [YCS08, KSN04] et les modèles de Markov cachés. [MK04].

En ce qui concerne la reconnaissance des gestes la plupart des méthodes est fondée sur l'apprentissage supervisé de prototypes d'actions individuelles.

Pour ce qui est de la reconnaissance des comportements deux approches sont généralement proposées. La première concerne l'appariement de gabarits [PN94, CW97] et consiste à convertir une séquence vidéo des événements en une représentation statique, comme une silhouette image unique. Ensuite, ces gabarits sont comparés à une action type pour assurer la classification du comportement identifié. Cette méthode souffre néanmoins d'un défaut de généralisation, car elle rend difficilement compte de la variété spatio-temporelle de la réalisation motrice d'une action. Qui plus est, le modèle devrait idéalement représenter toute action indépendamment des différences interindividuelles qui existent entre différents sujets.

D'autres méthodes sont basés sur les automates [BW95, Bre97, BP97, JYI92, AL06] et fournissent généralement une solution au

problème de changement de phase avec l'utilisation de modèles probabilistes tels que les HMM. Une approche commune est de définir chaque caractéristique statique d'une action comme un état et d'apprendre la relation entre l'ensemble des caractéristiques. Pour classer une action, la probabilité conjointe à la valeur maximale est retenue comme critère de classification.

Bobick et Davis proposent d'utiliser les images d'énergie du mouvement et celles de l'historique du mouvement (MHI), et la distance de Mahalanobis entre les moments de Hue 2D pour comparer entre deux actions [AA01].

Chomat et Crowley [CC98] ont utilisé des filtres spatio-temporels pour générer des templates. Ensuite, une approche bayésienne a été utilisée pour la classification des actions. Wang et al. [WH03] ont proposé de classer des gestes par classification des descripteurs dans l'espace réduit (eigen space) par ACP.

Notre but n'est pas seulement de classer les gestes dans la scène, mais aussi d'extraire des descripteurs pertinents du comportement observé. Nous nous sommes intéressés non seulement à la représentation l'action, mais aussi aux caractéristiques pertinentes qui faciliteraient la classification et la reconnaissance. Cette phase d'extraction de caractéristiques est une tâche délicate et qui doit être soigneusement mis en oeuvre.

Efros et al. dans [BD01] comparent deux actions en se basant sur les caractéristiques extraites, dans l'espace spatio-temporel, à partir du flux optique. Blank et al [MB05] utilisent une pile de points de silhouettes qui sont extraites et évaluées en utilisant l'équation de Poisson pour chacun des points. La comparaison de deux actions se fait par distance euclidienne entre les vecteurs caractéristiques.

Une action humaine étant fortement liée au mouvement, nous proposons dans cet article de suivre l'objet en mouvement et de former un volume dans l'espace 3D (2d+t). Ce volume qui représente une action donnée sera caractérisé par des moments géométriques 3D qui sont invariants à la translation et au changement d'échelle. Nous présentons une nouvelle approche de catégorisation des actions basée sur la diffusion géométrique par marches aléatoires sur graphe. L'idée de base

est de considérer l'ensemble des actions (vidéos) comme un graphe pondéré, où les sommets du graphe sont représentés par les volumes 3D (séquences des actions), et les arêtes connectés représentent la similarité entre les noeuds. Cette mesure de similarité sera calculée par une distance euclidienne entre les vecteurs caractéristiques des actions.

Dans notre article, nous proposons une approche unifiée qui permet la classification et la reconnaissance de gestes complexes. L'approche présentée utilise les marches aléatoires sur graphe pour la catégorisation et des réseaux de neurones pour la reconnaissance. Nous commencerons par présenter notre approche de segmentation d'objets binaires 3D, fondée sur les coupes de graphes. Ensuite, nous décrirons dans la section 3 les caractéristiques spatio-temporelles extraites, notamment la caractérisation globale du volume binaire par des moments géométriques 3D. Après un rappel du principe de la diffusion géométrique par marches aléatoires sur graphe et des réseaux de neurones utilisés, dans la section 4, nous présentons dans la section 5 les résultats de validation sur des corpus vidéo d'analyse et de reconnaissance de gestes 3D. Enfin, nous concluons notre travail par quelques remarques et nous présentons la direction de nos futurs travaux dans ce sens.

## 2. Segmentation Vidéo par Graph Cuts

Nombreux sont les problèmes en Vision par Ordinateur qui peuvent être vus comme un simple problème d'étiquetage. C'est le cas, par exemple, pour la segmentation. Un tel problème peut être représenté en termes de minimisation d'énergie, qui peut être résolue par les graph cuts sous certaines conditions [VK]. Cette approche de minimisation d'énergie par les graph cuts est à la base de plusieurs algorithmes tel que alpha-expansion et alpha-beta-swap [BJ01]. De nombreuses approches ont été proposées pour la segmentation interactive. Dans ce travail, nous utilisons les graphcuts [BJ01, YB03] pour segmenter nos régions d'intérêt "ROIs" (personne et mains) et pour suivre leurs mouvements dans une séquence vidéo. Dans les deux cas, on suppose que le fond est statique. La segmentation d'image peut être reformulée en termes de minimisation d'énergie qui peut être résolue par graph cuts. Toute l'idée des Graph Cuts est de ramener le problème de minimisation d'énergie à

un problème de coupe minimale dans un graphe. Greig et al. ont montré que cette minimisation (de type estimation du maximum à posteriori d'un champ aléatoire de Markov) peut être réalisée par la coupe minimale d'un graphe avec deux nœuds spécifiques "source" et "puits" pour la restauration d'images binaires [?]. L'approche a d'ailleurs été étendue aux problèmes non-binaires connus sous le nom "S-T Graph Cut". V. Kolmogorov [VK] a donné une condition nécessaire et suffisante, dite condition de sous modularité, pour qu'une fonction puisse être minimisable par graph cuts. Chaque pixel de l'image correspond à un nœud du graphe. Deux autres nœuds supplémentaires forment la source et le puits, représentant respectivement l'objet et le fond. Chaque nœud (pixel) est relié à ses voisins par des arrêtes n-liens, avec une connectivité choisie, et dont les capacités dépendent des différences de l'intensité. Chaque nœud (pixel) est aussi relié par des arrêtes, t-liens, aux terminaux (source et puits). Dans ce travail, nous proposons l'extraction d'objets d'intérêt par graph cuts en utilisant des contraintes "dures" sur le fond et sur les objets en s'inspirant de l'approche interactive présentée par Boykov and Jolly.

Le but étant de partitionner un graphe en un ensemble de classes disjointes, pour une image donnée, on construit un graphe pondéré  $G = (V, E, W)$ .  $V$  étant les nœuds du graphe,  $E$  les arcs reliant les nœuds avec des poids positifs (coûts)  $W$ . Les nœuds sont les pixels  $p$  de l'image  $P$  et les arcs sont représentés par les relations d'adjacence avec le voisinage  $q$  dans  $N$ . Le but de l'étiquetage est d'assigner un label unique  $A$  à chaque nœud.  $A = (A_1, A_2, \dots, A_p, \dots, A_{|P|})$  Cela peut être obtenu par minimisation de l'énergie  $E(A)$ .  $A$  est un vecteur binaire i.e.  $A_p \in \text{"Object"}, \text{"BackG"}$ .

$$E(A) = \lambda \cdot R(A) + B(A)$$

Le coefficient  $\lambda (\geq 0)$  représente l'importance relative des propriétés de régions  $R(A)$ , face aux propriétés de frontière  $B(A)$ .

$$R(A) = \sum_{p \in P} R_p(A_p)$$

$$B(A) = \sum_{p, q \in N} B_{p, q} \cdot \delta(A_p, A_q)$$



Le terme  $R(\cdot)$  indique comment le pixel  $p$  doit réagir en fonction des à priori sur les objets et le fond. Le terme  $B(A)$  décrit les propriétés de contour de la segmentation.  $B(A)$  doit être interprété comme une pénalité pour la discontinuité entre les pixels  $p$  et  $q$ .  $B_{p,q}$  est grand quand les pixels  $p$  et  $q$  sont proches.  $\delta(A_p, A_q) = 1$  si  $A_p \neq A_q$ , sinon  $\delta(A_p, A_q) = 0$ . Le but est d'encourager la segmentation (la coupe) qui passe par des régions où le gradient de l'image est assez fort.

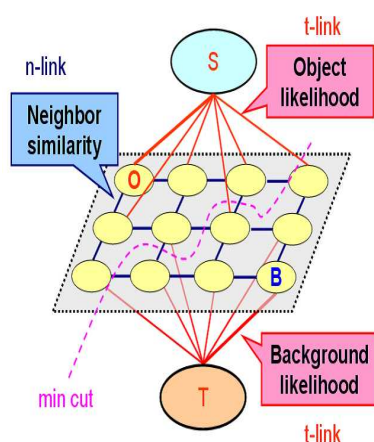


Figure 1 – Exemple de graphe d'une image 3x4.

Les contraintes dures sont spécifiées par l'utilisateur. L'utilisateur spécifie des régions dans l'image qu'il veut absolument voir appartenir à la classe "objet" et à la classe "arrière-plan".

Arcs	Coût	pour	
n-link	$p, q : W_{pq}^I$	$B_{p,q}$	
t-link	$p, s : W_p^s$	$\lambda \cdot R_p(\text{"BackG"})$	$p \in P, p \notin O \cup B$
		$K$	$p \in O$
		$0$	$p \in B$
	$p, t : W_p^t$	$\lambda \cdot R_p(\text{"Object"})$	$p \in P, p \notin O \cup B$
		$0$	$p \in O$
		$K$	$p \in B$

Tableau 1 – Construction du graphe

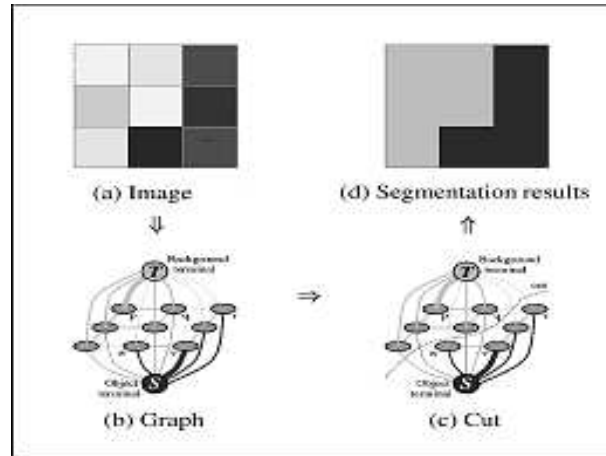


Figure 2 – Exemple de segmentation par graphcut (Yuri Boykov.)

La Table 1 montre comment on peut construire le graphe en tenant compte des contraintes sur chaque noeud. Les termes région et contour sont calculés de la manière suivante :

$$\begin{cases} R_p(\text{"Object"}) = -\ln Pr(C_p | O) \\ R_p(\text{"BackG"}) = -\ln Pr(C_p | B) \end{cases}$$

$$B_{p,q} = \exp\left(\frac{(d_{pq})^2}{2\sigma^2}\right) \cdot \frac{1}{\text{dist}(p,q)}$$

$$K = 1 + \max_{p,q \in N} B_{p,q}$$

Soient  $O$  et  $B$  les germes qui représentent respectivement l'objet et l'arrière plan.  $I_p$  et  $C_p$  représentent resp. le niveau de gris et la couleur du pixel  $p$ . La détection du contour des objets par rapport au fond se fait par minimisation de l'énergie du graphe  $G$ . où  $d_{pq} = I_p - I_q$  est la distance entre les pixels  $i$  et  $j$ ,  $K$  est une constante.  $\sigma$  est un paramètre de pondération entre les différentes caractéristiques.  $B_{p,q}$  reflète la similarité entre les pixels et donc permet de contrôler la cohérence de la segmentation.  $W_p^s$  et  $W_q^t$  décrivent l'appartenance d'un pixel aux deux classes. Les contraintes dures ne sont pas, en général, requises. La segmentation peut être faite sur la base de deux seuils déduits automatiquement à partir d'a priori sur le fond/objet, tels que l'histogramme par exemple.

L'approche cherche à minimiser la fonction d'énergie globale, avec des termes de vraisemblance et d'a priori. L'image à segmenter est convertie en graphe où chaque pixel est représenté par un nœud intermédiaire. Il y a également deux nœuds terminaux, un pour chaque classe. Les liens possèdent tous un poids. Le terme de vraisemblance lie entre les nœuds intermédiaires et les nœuds terminaux. Tandis que le terme d'a priori lie entre les nœuds intermédiaires. Le principe consiste à séparer le graphe en deux partitions, chacune d'elle devant contenir un nœud terminal. Cela consiste à modifier le poids des liens entre les nœuds intermédiaires et les nœuds terminaux dans le graphe.

### 2.1. Résultats de la segmentation

Nous avons appliqué l'algorithme de minimisation par graphcut sur des images  $2D$  et  $2D + t$ . Les résultats sont encourageants



Figure 3 – Segmentation 2D Person/Arrière-Plan par l’algorithme basée sur Graph cut

Dans la figure 3, les contraintes dures sont présentées en rouge et bleu. Sur l’image gauche, on observe les contraintes sur l’objet d’intérêt en rouge, ainsi que celles sur le fond. Le résultat de la segmentation (image de droite) montre l’extraction de la personne en mouvement. Dans le cas de la segmentation des gestes de la main, la phase d’initialisation est très importante afin de distinguer chacune des deux mains de l’arrière-plan. On note d’assez bons résultats. Comme on peut le voir dans les figures (3-4-a), le processus d’extraction est très robuste et ne nécessite pas beaucoup de contraintes. Pourtant, la difficulté de la séparation des deux mains, le chevauchement des mains et des problèmes d’occultation, nous ont obligés à adapter le choix de ces contraintes.

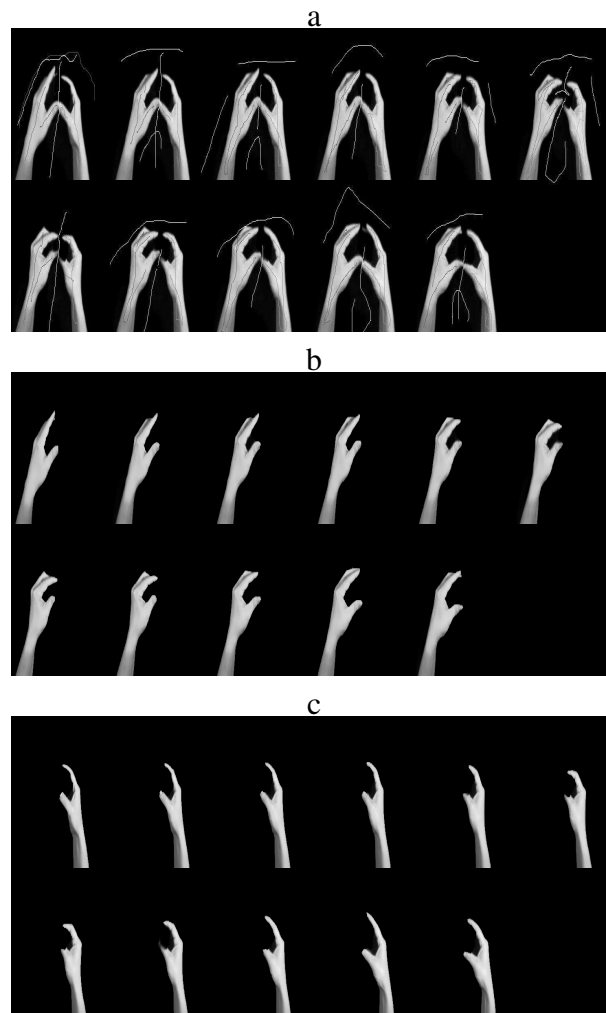


Figure 4 – Extraction des gestes des deux mains à partir d'une vidéo par l'approche graph-cut : Une série de frames contenant les deux mains. Les figures b et c montrent l'extraction de la main droite et de la main gauche

### 3. Caractéristiques spatio-temporelles

Dans cette section, nous présentons quelques descripteurs visuels utilisés en analyse vidéo, qui permettent de donner des informations vi-

suelles et compactes sur les objets en mouvement .

### **Energie du mouvement**

L'énergie du mouvement (MEI) et l'historique d'un mouvement (MHI) ont été introduits pour capturer l'information  $2D$  du mouvement dans les images. L'énergie du mouvement est essentiellement une image du mouvement cumulé. Elle indique l'emplacement spatial du mouvement. L'historique du mouvement décrit quant à lui les caractéristiques temporelles du mouvement.

Soit  $D(x; y; t)$  le volume binaire de l'objet segmenté,  $D = 1$  indique qu'il y a eu mouvement de pixel à cette position  $(x; y)$ , à l'instant  $t$ . MEI est calculée comme suit :  $H_\tau(x, y, t) = \begin{cases} \tau & D(x, y, t) = 1 \\ \max((H_\tau(x, y, t - 1), 0) & \text{Otherwise} \end{cases}$  où  $\tau$  est l'intervalle temporelle de capture entre deux frames consécutives.



Figure 5 – Résultats de MHI sur la séquence "pression sur un objet dur" de la fig.4-b .

### **Historique de la densité de mouvement**

L'historique de la densité de mouvement dans le temps (MHIDT) capture la fréquence du mouvement dans les images. Il est calculé à partir des images consécutives. Soit  $D(x, y, t)$  un volume (vidéo) binaire, on calcule le MHIDT par :

$$H_\tau = D(x, y, t) - D(x, y, t - 1)$$

où  $(t = 1 \dots n)$  est l'intervalle temporelle de capture entre deux images consécutives.



Figure 6 – Résultats de MHIDT sur la séquence ”pression sur un objet dur” de la fig.4-b .

### 3.1. Descripteurs 3D de forme

Un descripteur de forme propose de décrire une forme en caractérisant son contour et en exploitant certaines propriétés topologiques ou géométriques. Cela conduit à une variabilité assez importante en termes de perception intuitive de la forme. Une opération de reconnaissance de formes se déroule en général suivant deux phases : une phase d’apprentissage et une phase de décision. Au cours de chacune de ces deux phases, on retrouve une phase d’extraction des paramètres représentatifs de l’image. Ces valeurs doivent présenter la particularité d’être invariantes à certaines transformations telles que la rotation et/ou la translation.

On distingue deux grandes approches. L’approche contour qui caractérise la forme à partir de son contour sans tenir compte de la texture et l’approche globale qui étudie la forme dans son ensemble. Dans ce qui suit, nous décrivons quelques types de représentations globales fondées sur les moments : les moments statistiques 2D tels que les moments de hu et Les moments géométriques 3D.

A partir du volume des images binaires, il faut extraire des caractéristiques représentatives de l’action. Nous avons choisi de caractériser le volume 3D par les moments géométriques 3D. Soit  $x, y, t$  l’ensemble des points appartenant au volume où  $x, y$  et  $t$  représentent les coordonnées spatio-temporelles. Le moment d’ordre  $(p+q+r)$  de ce volume est représenté par :

$$A_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q t^r dx dy dt$$

$A_{000}$  représente le volume de l'objet et  $(A_{100}, A_{010}, A_{001})$  sont les coordonnées du centre de l'objet.

$$M_{pqr} = \int \int \int \left[ \frac{x - A_{100}}{A_{200}^{1/4} * A_{020}^{1/4}} \right]^p \left[ \frac{y - A_{010}}{A_{200}^{1/4} * A_{020}^{1/4}} \right]^q \left[ \frac{t - A_{001}}{A_{200}^{1/2}} \right]^r dx dy dt$$

Pour caractériser le volume  $(2d + t)$ , nous avons utilisé un vecteur de caractéristiques composé d'une dizaine de moments d'ordre 2 et 3. Un vecteur caractéristique composé de 14 moments est ensuite construit pour représenter le volume binaire (la vidéo) :

$$M_{3d} = \begin{cases} M_{200}, M_{011}, M_{101}, M_{110}, M_{300}, M_{030}, M_{003}, \\ M_{210}, M_{201}, M_{120}, M_{021}, M_{102}, M_{012}, M_{111} \end{cases}$$

## 4. Categorisation par diffusion sur graphe

### 4.1. Marches aléatoires sur graphe

Dans cette section, nous discutons de la méthode de représentation d'un ensemble de données  $X = \{x_1, \dots, x_n\}$  où  $n \in \mathbb{R}^n$  en termes de coordonnées de diffusion, et nous montrons le rapport entre le processus de diffusion sur graphe et les marches aléatoires sur l'ensemble de données. La diffusion géométrique [18, 24] sur graphe offre un cadre unifié pour la réduction, la visualisation, la catégorisation et la fusion de données de grande dimension.

L'idée de base, issue de la théorie des graphes, est de représenter cette variété de données comme un graphe  $G = (V, E)$  qui consiste en un ensemble fini de sommets  $V = v_1, \dots, v_n$  et un ensemble fini d'arêtes  $E \subset V \times V$ . Deux sommets  $v_i$  et  $v_j$  sont adjacents si l'arête  $(v_i, v_j) \in E$ . Nous considérons que le graphe  $G$  est un graphe pondéré. On lui associe une fonction de poids  $G$  qui reflète le degré de similarité entre les deux sommets du graphe et décrit ainsi l'interaction du premier ordre entre les sommets du graphe. Son choix dépend généralement de l'application considérée.

$$W_{ij} = e^{-\frac{d_{ij}}{2\sigma^2}}$$



où  $\sigma$  est la variance de la Gaussienne et  $d_{ij}$  est la distance entre les vecteurs caractéristiques de  $v_i$  et de  $v_j$ . Le degré d'un sommet  $v_i \in V$  est défini par :

$$D_i = \sum_{j=1}^n w_{ij}$$

On définit la matrice diagonale des degrés des sommets  $D$  par :  $D_{ii} = D(v_i, v_i) = d_i$ , et  $D_{ij} = 0$  pour  $i \neq j$ . On définit la matrice  $L$  telle que :

$$L_{ij} = \begin{cases} d_i - w_{ii} & \text{si } v_i = v_j \\ -w_{ij} & \text{si } v_i \text{ et } v_j \text{ sont adjacents} \\ 0 & \text{sinon} \end{cases}$$

Ainsi, le Laplacien du graphe  $G$  peut être défini par :

$$\mathfrak{L} = D^{-1/2} L D^{-1/2} \text{ où } D_{ii}^{-1} \equiv 0 \text{ if } d_i = 0$$

La probabilité de transition du  $v_i$  à  $v_j$  en chaque étape est :  $P_{ij} = w_{ij}/d_i$ . Ceci permet de définir la matrice de transition  $P$  de la chaîne de Markov.

$$\forall v_i, v_j, 0 \leq p_{ij} \leq 1 \text{ et } \sum_{j \in V} p_{ij} = 1$$

$$\text{On peut aussi écrire } P = D^{-1} W$$

Nous présentons, dans le tableau suivant, notre approche de catégorisation qui est basée sur la diffusion par marches aléatoires sur graphe.

#### 4.2. Extension de Nyström

Soit  $\Omega = \{x_1, x_2, \dots, x_n\} \subset \mathfrak{R}^d$  l'ensemble des données de l'apprentissage. Le noyau est une fonction  $k : \Omega \times \Omega \rightarrow \mathfrak{R}$  telle qu'il existe  $\varphi : \Omega \rightarrow H$ , où  $H$  est un espace d'Hilbert et  $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ ,  $i, j = 1, \dots, n$

---

**Algorithm 1** Algorithme de cartes de diffusion
 

---

- 1: **Input** : Ensemble de données  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d, t, m, \varepsilon$
- 2: **Output** : Classes  $A_1, \dots, A_k$  avec  $A_i = \{j | y_j \in C_i\}$
- 3: Construction de la matrice de similarité. Soit  $W$  sa matrice d'affinité.

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\varepsilon}}$$

- 4: Normalisation par la méthode de Laplace-Beltrami :

$$\widetilde{w}_{ij} = w_{ij} / (\sqrt{d_i d_j})$$

- 5: Calcul de la matrice de transition :

$$p_{ij} = \widetilde{W}_{ij} / (\sqrt{\widetilde{d}_i \widetilde{d}_j}) \text{ avec } d_i = \sum_{i=0}^{n-1} w_{ij}$$

- 6: Puissance de la matrice de transition  $P^t$  avec  $t = 1..2$  :
  - 7: Diagonalisation de la matrice  $P^t$
  - 8: **Espace de diffusion :**
  - 9: Calcul des  $k$  premiers vecteurs propres  $v_1, \dots, v_k$  de  $P^t$
  - 10: Normalisation des vecteurs propres par le premier vecteur propre
  - 11:  $Y =$  tri des vecteurs par ordre croissant :  $\lambda_1, \lambda_2, \lambda_3$
  - 12: Classifier les points  $(y_i)_{i=1, \dots, n}$  dans  $R_k$  avec l'algorithme des k-means en  $C_1, \dots, C_k$  classes
-

Soit  $K$  la matrice qui contient les valeurs du noyau,  $K_{ij} = k(x_i, x_j)$ . Si la matrice est définie semi positive, alors  $k$  est un noyau de l'ensemble  $\Omega$ . Une correspondance satisfaisante du produit scalaire peut être la décomposition en vecteurs propres de la matrice du noyau  $K$  :  $K = U \Lambda U^T = U \Lambda^{1/2} (U \Lambda^{1/2})^T$  où  $U$  est la matrice dont les colonnes sont les vecteurs propres  $\phi_i$ ,  $i = 1, \dots, n$ , et  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  est la matrice diagonale des valeurs propres dans l'ordre décroissant.

Si on définit  $\varphi(x_i)$  comme la  $i^{\text{th}}$  rangée de  $U \Lambda^{1/2}$ , et comme les vecteurs propres sont non-négatifs, on obtient :

$$\varphi(x_i) = [\sqrt{\lambda_1} \phi_1(x_i), \sqrt{\lambda_2} \phi_2(x_i), \dots, \sqrt{\lambda_n} \phi_n(x_i)]$$

La fonction du noyau peut alors être considérée comme une généralisation du produit scalaire et, par conséquent, agir comme une mesure de similarité entre l'ensemble des données

Si l'on veut calculer tenir compte d'un nouveau vecteur  $x$ , on utilise la formule de Nyström [19], pour évaluer l'extension des vecteurs propres.

Soit  $x \in \mathfrak{R}^d$  une nouvelle donnée à classer. La méthode de Nyström indique que la  $j^{\text{th}}$  coordonnée du noyau  $\phi$  de ce point peut être approximé par :

$$\varphi_j(x) = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^n k(x, x_i) \phi_j(x_i) \quad j = 1, 2, \dots, n$$

or in vector form :

$$\varphi(x) = \frac{1}{\sqrt{\Lambda}} U^T k_x$$

où  $k_x = [k(x, x_1), k(x, x_2), \dots, k(x, x_n)]$ , et  $\frac{1}{\sqrt{\Lambda}}$  signifie  $(\sqrt{\Lambda})^{-1} = \text{diag}(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}})$ .

## 5. Applications

Dans cette section, nous présentons les résultats de notre expérimentation sur deux catégories de comportements : la reconnais-

sance des gestes des mains et la catégorisation d'actions (activités humaines).

### 5.1. *Reconnaissance de gestes tactiles*

L'objectif de ce travail est de développer une méthode d'analyse vidéo capable d'interpréter efficacement des gestes de toucher. L'application a été conçue pour la classification et la reconnaissance des objets selon leurs consistance et leur texture. En se basant sur la description donnée par Lederman et Klatzky [1], nous avons tenté de répondre par analyse vidéo aux interrogations suivantes :

- Le geste est effectué par quelle(s) main(s) ?
- Nature du geste ?
- Nature de l'objet manipulé ?

Deux objets de propriétés différentes (texture et consistance) ont été testés. Pour chaque propriété, deux modalités ont été proposées. La texture de l'objet pourrait être lisse ou granuleuse. Sa consistance pourrait être dure ou molle. Par conséquent, nous avons essayé de définir les procédures exploratoires élémentaires ("PEs") suivantes :

**Mouvement Latéral :** Le mouvement latéral sur un objet *lisse* correspond à un frottement de la main le long de sa surface. Tandis que sur un objet *rugueux*, un mouvement latéral correspond à un faible mouvement d'un ou de plusieurs doigts tout en grattant la surface dans un sens (gauche-droite ou de haut-bas).

**Pression :** Quand on exerce une pression sur un objet doux, les doigts s'enfoncent et donc parfois s'approchent entre eux. Tandis que sur objet dur, la distance entre les doigts ne varie pas beaucoup ou très peu.

Nous avons mis en place un système capable de reconnaître la nature du geste et de l'objet manipulé par les deux mains. Nous avons défini à cette fin quatre classes de "PEs" qui sont :

- 1) Mouvement Latéral pour Objets Lisse (*LMSO*)
- 2) Mouvement Latéral pour Objets Rugueux (*LMGO*)
- 3) Pression sur Objet Doux (*PSO*)

#### 4) Pression sur Objet Dur (*PHO*)

Les gestes de la main sont capturés par une caméra web avec un ordinateur personnel Intel Pentium. D'abord, nous avons exploité la couleur de la peau pour la segmentation des régions d'intérêt (ROI) en utilisant des règles de décision dans les espaces HSV et YUV [MH03, CE06]. Ensuite, des pré-traitements par opérations morphologiques sont utilisés pour éliminer le bruit. Nous avons proposé une méthode de segmentation semi-interactive basé sur les graphcuts pour segmenter les images vidéo en trois parties : main gauche, main droite et l'arrière-plan.

Nous avons utilisé une base d'apprentissage composé de 120 vidéos qui représentent les quatre actions (PEs) présentés ci-dessus : 20 vidéos de pression de PSO, 29 vidéos de LMGO, 37 vidéos de PHO et 34 vidéos de LMSO. Nous avons testé notre approche sur une base de test composée de 34 nouvelles vidéos : 7 PSO, 7 LMGO, 10 PHO et 10 vidéos LMSO. Chaque vidéo est composée de 111 frames. Pour la classification et la reconnaissance, nous avons utilisé deux méthodes de classification : notre approche basée sur la diffusion sur graphe par marches aléatoires et des réseaux de neurones par rétro-propagation.

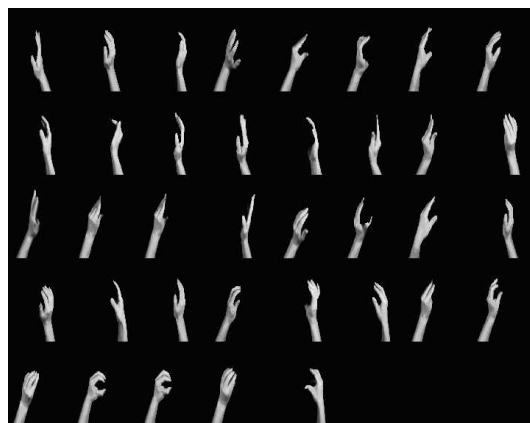


Figure 7 – Une partie du corpus vidéos

##### 1) *Diffusion sur graphe*

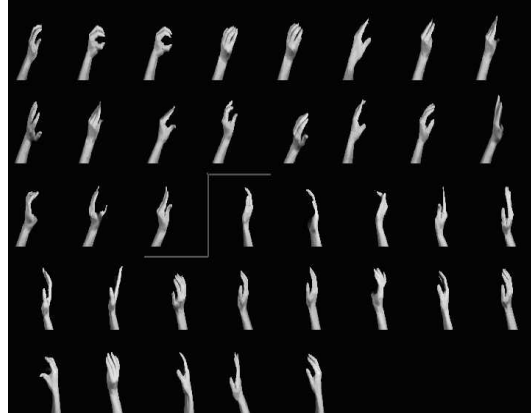


Figure 8 – Catégorisation des mains séparément par diffusion par marches aléatoires

Nous avons appliqué notre approche de la diffusion sur graphes (DG) en utilisant les moments géométriques 3D pour catégoriser les mains droites et gauches. Nous avons appliqué cette méthode sans connaissance à priori et sans apprentissage. Les résultats ont été très probants comme le montre la figure 8. Comme on peut le constater, la base de mains 3D est maintenant bien ordonnée. On distingue bien deux classes : une classe regroupant les mains gauches et une autre regroupant les mains droites. Pour la nouvelle vidéo (main en mouvement), on calcule les moments géométriques 3D, et on calcule la mesure de similarité avec le reste de la base. Ensuite nous utilisons la méthode de Nyström suivante :

$$\varphi_j(x) = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^n k(x, x_i) \phi_j(x_i) \quad j = 1, 2, \dots, n$$

Ici,  $k$  est le vecteur de transition de la similarité entre la nouvelle vidéo et l'ensemble d'apprentissage. Pour réduire le calcul, nous n'avons pris que les trois premières valeurs propres  $(\lambda_1, \lambda_2, \lambda_3)$ , et les trois premiers vecteurs propres  $(\phi_1, \phi_2, \phi_3)$ . Le résultat de l'extension de Nyström est un vecteur de trois éléments qui représente la nouvelle vidéo. Nous calculons la distance entre ce vecteur et le centre des deux classes, le centre le plus proche détermine la classe de cette nouvelle main (gauche ou droite).

Le résultat de la reconnaissance des mains dans ces nouvelles vidéos est de 100% .

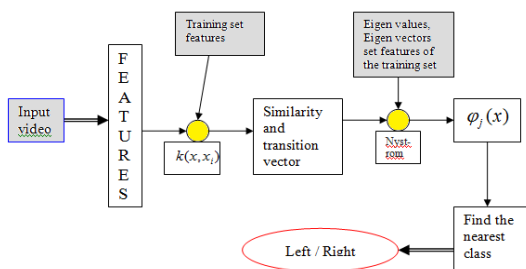


Figure 9 – Processus de catégorisation des mains G/D par diffusion sur graphe par marches aléatoires

2) **Réseaux de neurones** Nous avons construit un réseau de neurones (NN) basé sur l’algorithme de rétro-propagation pour répondre à la question principale : *Quel est le geste effectué par les deux mains ?*. L’apprentissage a été effectué sous Matlab, en utilisant la fonction d’activation sigmoïde et la règle d’apprentissage de Levenberg-Marquardt. Notre réseau de neurones reconnaît l’action effectuée par les mains. Il contient trois couches : la première couche est la couche d’entrée et se compose de 14 nœuds (les moments $3D$ ), la couche cachée contient 30 neurones, chaque neurone a une matrice de 14 poids, une base et une sortie. La couche de sortie est constituée de deux neurones qui prennent en entrée les sorties de neurones de la couche cachée (30) et donne en sortie le résultat la nature du geste des mains. Les valeurs de sortie sont représentées dans la transformation suivante :

1	1	<b>PSO</b>
-1	1	<b>LMGO</b>
1	-1	<b>PHO</b>
-1	-1	<b>LMSO</b>

L’apprentissage par NN a été réalisé avec succès sur l’ensemble des *120 videos*. Sur les 34 vidéos test, le réseau de neurones reconnaît bien 82.4% des gestes dans les vidéos.

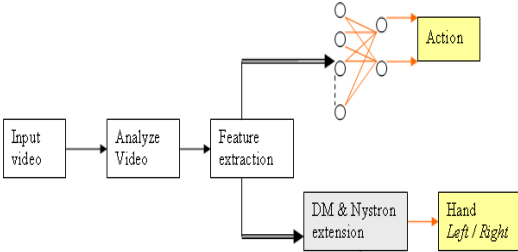
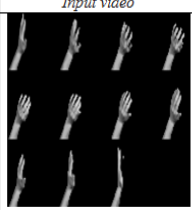
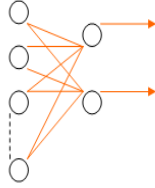


Figure 10 – Processus de reconnaissance des mains et du geste

Le résultat de test pour les deux catégories (main droite/gauche et les quatre actions) est résumée dans le tableau suivant :

Reconnaissance de la main par DG	100%	<i>Main Gauche</i>
	100%	<i>Main Droite</i>
Reconnaissance du geste par NN	71.4%	<i>PSO</i>
	85.7%	<i>LMGO</i>
	80%	<i>PHO</i>
	90%	<i>LMSO</i>



Input video	features	NN	Output	Result
	0.439233 -0.0554236 -0.143021 -0.668137 1.32221e-05 -0.00154504 0.000157242 0.000612796 -0.000125942 -0.00154664 0.00109113 -0.0003735 -0.000164941 0.000227694		-1 -1	The action is LMSO


Input video	features	Kernel Method	$\varphi_j(x)$	Result
	0.439233 -0.0554236 -0.143021 -0.668137 1.32221e-05 -0.00154504 0.000157242 0.000612796 -0.000125942 -0.00154664 0.00109113 -0.0003735 -0.000164941 0.000227694	Similarity and transition vector then calculate the Nystrom extension	$\varphi_1(x) = -0.6378$ $\varphi_2(x) = -0.1059$ $\varphi_3(x) = -1.3213$	Left hand

Figure 11 – Exemple de reconnaissance de gestes par NN et catégorisation de la main par DG

## 6. Conclusion

La gestion satisfaisante d'une grande collection de vidéos est toujours un défi. Nous avons présenté dans cet article, une nouvelle approche de catégorisation basée sur les marches aléatoires sur graphe. Nous avons représenté chaque vidéo de gestes de mains par son volume qui a été caractérisé par ses moments géométriques 3D. La segmentation vidéo a été réalisée par l'approche basée sur les graphcuts. Ces vecteurs caractéristiques sont utilisés pour la catégorisation des mains gauches/droites par l'approche de diffusion et pour la reconnaissance de gestes par réseaux de neurones. Nous avons testé ces méthodes sur un corpus vidéo varié et nous avons obtenu un taux de reconnaissance de 82,4% de gestes manuels. D'autre part, nous avons validé l'approche de diffusion sur les mains avec un taux de satisfaction de 100%. Cette approche unifiée de classification et de reconnaissance est très robuste aux problèmes classiques qui sont très sensibles à la phase de segmentation. Nous avons mené des expériences sur les mains et les gestes tactiles en rapport avec la nature des objets. Deux propriétés des objets ont été testés : la texture et la consistance. Pour chaque propriété, deux moda-

lités ont été proposées. La texture peut être lisse ou rugueuse et l'objet peut être mou ou dure. Nous avons montré l'utilité de la décomposition spectrale de la matrice de transition et des réseaux de neurones pour la catégorisation et la reconnaissance des gestes. Ce travail sera utilisé comme outil d'expérimentation et d'interaction avec des personnes à capacités motrices réduites.

### 6.1. Travaux futurs

Des travaux en cours sont menés pour prendre en compte d'autres informations telles que le rythme et la forme 3D à priori. Aussi, nous souhaitons aborder le problème des actions dans des environnements complexes, mais avec un apprentissage semi-supervisé. Une amélioration de la segmentation des gestes ne peut qu'améliorer le système et nous permettre une reconstruction virtuelle fidèle aux données réelles. D'ailleurs, ce travail fera partie de l'outil d'expérimentation que nous sommes en train de développer.

### Références

- [AA01] A. Ali and J. K. Aggarwal. Segmentation and recognition of continuous human activity. *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.
- [AL06] M. Ahmad and S.-W. Lee. Hmm-based human action recognition using multiview image sequences. *International Conference on Pattern Recognition*, 1 :263,À266, 2006.
- [BD01] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence*, 23(3) :257,À267, 2001.
- [BJ01] Y. Boykov and M. Jolly. Iterative graph cuts for optimal boundary and region segmentation of objects in n-d images. *Int. Conf. on Computer Vision*, 2001.
- [BP97] Oliver N. Brand, M. and A. Pentland. Coupled hidden markov models for complex action recognition. *Computer Vision and Pattern Recognition*, page 994, 1997.

- [Bre97] C. Bregler. Learning and recognizing human dynamics in video sequences. *Computer Vision and Pattern Recognition*, page 568, 1997.
- [BW95] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. *International Conference on Computer Vision*, page 382, 1995.
- [CC98] O. Chomat and J. Crowley. Recognizing motion using local appearance. *International Symposium on Intelligent Robotic Systems*, page 271, 279, 1998.
- [CE06] Y. Chahir and A. Elmoataz. Skin-color detection using fuzzy clusterin. *IEEE-EURASIP ISCCSP*, 2006.
- [CW97] Y. Cui and J. Weng. Hand segmentation using learning based prediction and verification for hand sign recognition. *Computer Vision and Pattern Recognition*, page 88, 93, 1997.
- [JYI92] J. O. J. Yamato and K. Ishii. Recognizing human action in time sequential images using hidden markov model. *IEEE International Conference on Computer Vision and Pattern Recognition*, page 379, 385, 1992.
- [KSN04] Sharma A. Kumar S., Kumar D. and McLachlan N. Classification of hands movements using motion templates and geometrical based moments. pages 299–304, 2004.
- [LK87] S.J. Lederman and R.L. Klatzky. Hand movements : A window into haptic object recognition. *Cognitive Psychology*, 19(3) :342–368, 1987.
- [MB05] E. Sechtman M. Irani M. Blank, L. Gorelick and R. Basri. Actions as space-time shapes. *ICCV*, 2005.
- [MH03] L.Chen et D. Zighed M. Hammami, Y. Chahir. Détection des régions de couleur de peau dans l'image. *Extraction et Gestion de Connaissances*, 17(1-2-3) :219–231, 2003.
- [MK04] S.J. McKenna M. Kenny. An experimental comparison of trajectory-based representation for gesture. *LNAI*, 2915 :152–163., 2004.
- [OTS04] Tan J.K. Ogata T. and Ishikawa S. An experimental comparison of trajectory-based representation for gesture. *LNAI*, 2915 :152–163., 2004.

- [PN94] R. Polana and R. Nelson. Exploiting human actions and object context for recognition tasks. *IEEE CS Workshop on Motion of Non-Rigid and Articulated Objects*, page 77, 82, 1994.
- [VK] Youri Boykov Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence*, 26 :1124–1137.
- [WH03] Tan T. Ning H. Wang, L. and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12) :1505, 1518, 2003.
- [YB03] Vladimir Kolmogorov Youri Boykov. Geodesics and minimal surfaces via graph cuts. *International Conference on Computer Vision*, 1 :26–33, 2003.
- [YCS08] F. Jouen Y. Chahir, M. Molina and B. Safadi. Haptic gesture analysis and recognition. *IEEE/RSJ, Intl. Conf. on Intelligent Robots and Systems, Workshop on Grasp and Task Learning by Imitation*, pages 65–70, 2008.

**ANNEXE POUR LA FABRICATION**  
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE  
PAPIER  
DE LEUR ARTICLE

1. ARTICLE POUR LA REVUE :  
*Studia Informatica Universalis.*
2. AUTEURS :  
*Youssef Chahir* \* — *Michèle Molina* \*\* — *François Jouen*  
\*\*\*
3. TITRE DE L'ARTICLE :  
*Reconnaissance et catégorisation de l'activité manuelle humaine*
4. TITRE ABRG POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :  
*Hand Movement*
5. DATE DE CETTE VERSION :  
*10 octobre 2010*
6. COORDONNES DES AUTEURS :
  - adresse postale :
    - \* GREYC - CNRS UMR 6072
    - Université de Caen, France
    - youssef.chahir@info.unicaen.fr
    - \*\* Laboratoire PALM JE 2528
    - Université de Caen, France
    - michele.molina@unicaen.fr
    - \*\*\* Laboratoire CHArt EA 4004-CNRS UMS 2809
    - EPHE, Paris, France
    - francois.jouen@ephe.sorbonne.fr
  - tlphone : 02 31 56 73 26
  - tlcopie : 00 00 00 00
  - e-mail : youssef.chahir@info.unicaen.fr
7. LOGICIEL UTILIS POUR LA PRPARATION DE CET ARTICLE :  
L<sup>A</sup>T<sub>E</sub>X, avec le fichier de style `studia-Hermann.cls`,  
version 1.2 du 03/12/2007.