Haptic Gesture Analysis and Recognition

Youssef Chahir, François Jouen, Michèle Molina, Bahjat Safadi

Haptic Gesture Analysis and Recognition

Youssef Chahir, Michèle Molina,.François Jouen and Bahjat Safadi

***Abstract*—Haptic perception properties is achieved by the execution of specific exploratory procedures "Eps". An "EP" is a stereotyped movement pattern which is dictated by the object properties that the haptic system chooses to process, both perceptually and cognitively. The aim of the present work is to devise an automatic manual testing procedure able to extract tactile information effectively. Our system has been conceived to recognize the texture and the hardness of an object through the video analysis of hand actions. Two objects properties have been tested: texture and consistency. For each property, two modalities were proposed. The texture of the object being explored could either be smooth or granular. Its consistency could either be hard or soft. In this paper, we propose an automatic approach for hand gesture analysis and recognition for understanding human action and manipulation. To enhance robustness, each hand sequence is characterized globally by a volume which is characterized by its 3D geometrical moments. Neural networks are then used using distance between vectors of features. We tested the proposed approach with different hand gestures and results showed that our method is effective, achieving a high recognition rate.**

I. Introduction

Klatzky and Lederman [1] give many evidence that in adults, haptic perception properties is achieved by the execution of specific exploratory procedures "Eps". An "EP" is a stereotyped movement pattern which is dictated by the object properties that the haptic system chooses to process, both perceptually and cognitively. Klatzky and Lederman [1] described six basic exploratory properties. Lateral motion is an EP aimed at perceiving the texture of an object. It is achieved by rubbing the fingers across a surface. Pressure is an EP associated with the hardness of an object. By pressing down on an object, we gain information about object hardness. This pressure is a force applied on the object while the object is stabilized. Static contact is an EP associated with object temperature. It corresponds to a contact in one spot by a large skin surface without effort to mold to contours. Unsupported holding, that is holding an object out away from a support, is an EP associated with object weight. Enclosure is defined wrapping the hand around an object and provides information about its global shape and volume. Contour following is is an EP defined as moving the fingers about the perimeter of an object and provides information about the exact shape of an object. The identification of these exploratory procedures is necessary to design robots but also, for human beings, to interact with machine. The next generation of computers is expected to interact and communicate with users in a cooperative and natural manner while users engage in everyday activities. By being situated in users' environments, intelligent computers should have basic perceptual abilities, such as understanding what people are talking about (speech recognition), what they are looking at (visual object recognition), and what they are doing (action recognition) [2]

Vision based hand gesture recognition methods are generally categorized in two groups: feature based [3] and appearance based methods [4]. In feature based methods, initially, we need to extract model or features from images. In appearance based methods, images can be used directly for hand gesture recognition.

Several methods have used motion history images as temporal templates for gesture recognition [5]. Appropriate features are extracted from this image and different methods are used for classification such as Neural Network [6] and Hidden Markov Model [7].

The aim of the present work is to devise an automatic manual testing procedure able to extract tactile information effectively. Our system has been conceived to recognize the texture and the hardness of an object through the video analysis of hand actions.

Our work, based on the seminal description of Lederman and Klatzky [1], is aimed at answering to the following question: what are the hands doing: what the action they are exerted is aimed at?

Two objects properties have been tested: texture and consistency. For each property, two modalities were proposed. The texture of the object being explored could either be smooth or granular. Its consistency could either be hard or soft. Consequently, EPs extracted by our system were:

Lateral motion: For the smooth object, lateral motion corresponds to a continuous rubbing of the hand all over the object surface. For the granular object, lateral motion corresponds to low amplitude movements of one or more

Manuscript received July 6, 2008.

Y. Chahir is with the GREYC - CNRS UMR 6072 Laboratory. Computer Science Department. University of Caen Lower-Normandy, Bd Maréchal Juin, BP 5186, 14032 Caen , France (phone: 33-2-31-56-73-75; fax: 33-2-31-56-73-30; email: youssef.chahir@info.unicaen.fr)

M. Molina, is with the PALM Laboratory, JE 2528, University of Caen Lower-Normandy, France. (e-mail: michele.molina@unicaen.fr).

F. Jouen is with the CHArt Laboratory, EA 4004, Practical School of High Studies (EPHE) Paris, France (e-mail: francois.jouen@ephe.sorbonne.fr).

B. Safadi is with the GREYC - CNRS UMR 6072 Laboratory. University of Caen Lower-Normandy, (e-mail: bhjat_s@hotmail.com).

fingers scratching the surface in a left-right or up-down direction.

Pressure: For the soft object, the torque exerted on the object induces that the fingers gets closer some of the others. For the hard object, the torque exerted on the object did not lead to any change in the distance between fingers.

We would like to build a system which is able to recognize the action or the behavior of the hands. We categorize four Eps:

- Lateral Motion for Smooth Object (LMSO)
- Lateral Motion for Granular Object (LMGO)
- Pressure for Soft Object (PSO)
- Pressure for Hard Object (PHO)

The outline of the paper continues as follows. The proposed 3D hands segmentation is presented in section 2. Section 3 explains the global features used for shape descriptors. Then, in Section 4, we describe in detail our categorization and recognition approaches. Section 5 describes the experimental results while Section 6 present concluding remarks.

II. HAND GESTURE EXTRACTION

The hand gestures are captured by a cheap web camera and a standard Intel Pentium based personal computer. The skin color-based segmentation has been first applied. When initial color images are input, the first step is to convert RGB to both the HSV and the YUV color systems. Because the H and UV values of human skin color are in the invariant ranges, the probability distribution image is discerned using predefined rules obtained from data-mining techniques [8,9]. Pre-processing and morphological operations are also used to remove the noise. Then the mask of a region of interest (ROI) composed of hand regions is extracted, and the background is eliminated (cf. fig1).

We wish to partition hand video into three parts "left, right hand and the background". To segment the ROI (two hands) to both parts (left and right hands), and track each hand in each successive frame, we use an approach based on graph-cut technique. Motion-based estimation defines regions of foreground, background and boundary blocks. An automatic segmentation is realized by obtaining prior knowledge from foreground and background blocks. The result also can be improved by user adjustment.

Then, the segmentation problem is formulated as an energy minimization problem which is settled by using graph cut algorithm, according to the user-imposed constrains. As a pioneer work of object segmentation using a graph cut, the user-interactive segmentation technique was proposed by Boykov and Jolly [10]. They assumed that a given user imposes certain hard constraints for segmentation by indicating certain pixels (seeds) that belong to the object and certain pixels that belong to the background. *The main contribution of our approach is that the object and background seeds(regions) are estimated in every frame of sequences without user interaction.*

Basically each pixel in the image is viewed as a node in a graph, edges are formed between nodes with weights denotes how alike two pixels are, given some measure of similarity, as well as the distance between them. The edges for each pixel can be formed between the pixel with all the other pixels, In attempt to reduce the number of edges in the graph, we will predetermine neighborhood N that describes the neighbors of each pixel and we will be interested in the similarity "distance" between each pixel and its neighbors.

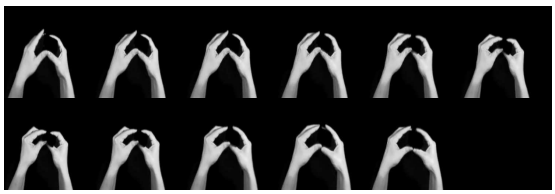


Fig. 1. Extraction of two hands sequences by skin-color model

There are two additional terminal nodes: an "object" terminal (a SOURCE) and a "background" terminal (a SINK) (cf. fig.2). These two terminal nodes don't correspond to any pixel in the image but instead they represent the object and the background respectively. The source is connected by edges to all nodes identified as object seeds and the sink is connected to all background seeds. Edges are formed between the source and sink and all other non-terminal nodes, where the corresponding weights are determined using models for the object and background. The min-cut of the resulting graph will then be the segmentation of the image. This segmentation should then be a partition such that, similar pixels close to each other will belong to the same partition. In addition, as a result of the terminal weights, pixels should also be segmented in such a manner so they end up in the same partition as the terminal node corresponding to the model *(object or background)* they are most similar to.

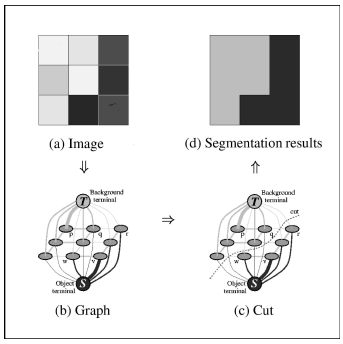


Fig.2. Example segmentation of a very simple 3-by-3 image. Edge thickness corresponds to the associated edge weight. (Image courtesy of Yuri Boykov.)

Given an image, we try to find the labeling X that minimizes the energy E:

$$E(X)=\lambda\sum_{p\in P} D_p(x_p)+\sum_{pq} B_{pq}\delta(x_p,x_q)\ \ p,q\in E$$

In the above equation, coefficient λ specifies the relative importance of the data term $D(.)$ and the smoothness term $B(.)$.

$X:=(x_1,x_2....x_p...x_{|P|})$ is a binary vector whose component x_p specifies labels to pixels p. Each x_p value can be either 1 or 0 where 1 represents an object and 0 represents a background area. The vector X defines segmentation.

$\delta(x_p,x_q)$ denotes the delta function defined by 1 if $x_p\neq x_q$, and 0 otherwise. The B_{pq} are defined by:

$$B_{pq}=K.\exp\left(-\frac{\|I_p-I_q\|^2}{\sigma_2}\right)$$

I_p and I_q are intensities/colors at pixels p and q. K is a constant. The data term D_p measures how well label x_p fits pixel p given the observed data. We modeled the object and background color likelihood's of $P(./"Obj.")\equiv P(\cdot/1)$. and $P(./"Back.")\equiv P(\cdot/0)$ using Gaussian mixtures in the RGB color space, using the data taken from the previous frame according to the labels given by the output of the segmentation process.

In every frame, our object and background estimation processes determine O and B which are the sets of pixels belonging to the estimated object and background regions respectively.

We compute the edge weights between pixels as the following. The edge weight between pixels p and q will denoted as $W(p,q)$ and the terminal weights (source and sink) between pixel p are given by:

$$W(p,S)=-\lambda\ln(P(I_p/"Background").$$

$$W(p,T)=-\lambda\ln(P(I_p/"object").$$

$$W(p,q)=B_{pq}.$$

$W(p,q)$ contains the inter-pixel similarity, that ensures that the segmentation more coherent. $W(p,S)$ and $W(p,T)$ describe how likely a pixel is to being background and foreground respectively.

In a video we construct a 3D graph that is obtained from a series of images that describes the video. Each node from the graph is connected to 26 (Pixels) neighbors, that means it has a 26 edges with weights calculated as described in the 2D graph. We applied the same ideas as above with slightly changes in 3D.

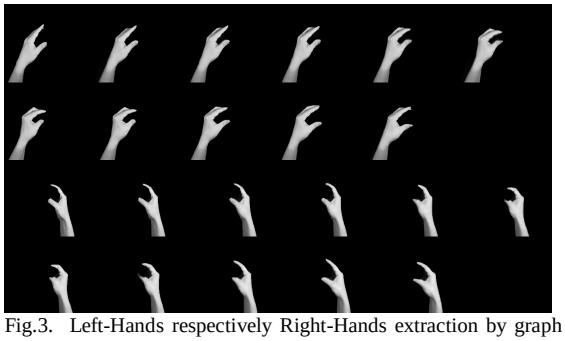


Fig.3. Left-Hands respectively Right-Hands extraction by graph cut approach

III. GLOBAL FEATURES

A. *Motion History Image Density*

Motion History Image Density uses the same technique as MHI [11] with one major change that it takes into account how many times the pixel is belonging to the object in a video. It takes a video as an input and returns 2D image which represents the historical information about this video that contains the projection of all the images in a video into one image 2d.

$$H_\tau(x,y,t)=\begin{cases}\tau & if(D(x,y,t)=1\\ \max(0,H(x,y,t-1)-1) & otherwise\end{cases}$$

where H_τ is Motion History Image and D is the binary difference between successively images. x,y and t are pixels coordinates. τ is a threshold for extraction of moving patterns in video image sequence. Thus, MHI is a scalar-valued image where more recently moving pixels are brighter. We extend the MHI filter by giving the pixel a value equal to how many times it is white in the video (it belongs to the object), except pixels that had never changed.

Fig.4. Some results of MHID filter.

B. *3D-Geometrical moments*

In general terms, shape descriptors are a set of numbers that are found to describe a shape in compact form. A shape descriptor should ideally be a simplification of its representative region but still hold enough information so that different shapes are discriminated. Usually it either describes the shape boundary or the image region.

In our approach we use the features based region description. Moments are a measure of the spatial distribution of 'mass' of the shape of an object. Objects in a binary image are represented as a set of white pixels (in 2D) and voxels (in 3D), videos are a 2D+t (time) images, so we can represent it as a 3D image with Depth equal to t.

A set of 14 moments derived by HU gives information about region-based shape descriptor that are rotation, scaling and translation invariant in a 3D dimensions.

Let (x,y,t) be a binary video, that means its voxels values equal to 1 for the voxels belonging to the object (hand) and zero for the background. We can define the moment as:

$$A_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^{p} y^{q} t^{r}\, dxdydt$$

A_{000} represent the area of the object and $(A_{100}, A_{010}, A_{001})$ the center of the object.

$$M_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-A_{100}}{A_{200}^{1/4} * A_{020}^{1/4}}\right)^{p}\left(\frac{y-A_{010}}{A_{200}^{1/4} * A_{020}^{1/4}}\right)^{q}\left(\frac{t-A_{001}}{A_{002}^{1/2}}\right)^{r} dxdydt$$

In discrete, the integration is changed to summation. The 3D geometrical feature descriptors are calculated from our videos by applying the 3D geometrical moments directly on the videos. 14 moments are extracted as a feature vector:

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

IV. RECOGNITION USING NEURAL NETWORKS

Pattern recognition is the study of how machines can observe the environment, learn how to distinguish patterns of interest and make decisions about categories of the pattern.

In our approach the patterns are videos of one moving hand, the movement of the hand gives information about what is the gesture. So we would like to give the system the ability to recognize the action done by the hand. Several approaches are supposed here, the recognition part of our approach is based on neural networks. Neural networks are efficient techniques in machine learning: they learn from a set of examples (training set) and have the ability to give an prediction answer when presenting a new data. We used in our approach the NN based back-propagation Algorithm.

We built two neural networks based on the back-propagation algorithm. The first network returns whether the hand in a video is left or right and the second net answers what is the action or the gesture. We arranged our videos (after the segmentation and separation process) into two sets, the training and test set, training set contains 120 videos of both hands and both actions (20 videos of Pressure for Soft Object, 29 videos of Lateral Motion for Granular Object, 37 videos of Pressure for Hard Object and 34 videos of Lateral Motion for Smooth Object) the test set contains 34 new videos (7 videos of PSO, 7 videos of LMGO, 10 videos of PHO and 10 videos of LMSO), each video we have contains 100 frames. Both networks were trained using matlab neural networks tool, using the sigmoid activation function and Levenberg-Marquardt learning rule. The figure 5 shows our flow processing diagram.

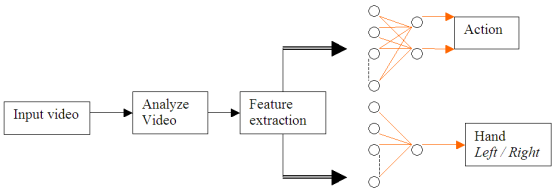


Fig.5. Flow processing diagram

The network was trained successfully with the 3D-geometrical moments as the feature descriptor of our videos composed of all hand gestures. We used the 14 geometrical moments as input of the two networks and we selected manually the output of each example in the training set.

The first network contains 3 layers the first layer is the input layer and consist of 14 nodes (the moments) the hidden layer contains 30 neurons where each neuron have a weight matrix of 14 weights, one base and one output it uses the sigmoid activation function finally the output layer consists only one layer which takes as input the outputs of the hidden layer neurons (30) and gives as output the result of which hand is moving in the video (*left or right hand*).
The network was trained successfully on the training set. Figure 6 show various hand videos used for training set. It gave 100 % right answers.

TABLE II: Recognition Using Neural Networks

NN BASED BACK-PROPAGATION ALGORITHM
Initialize the weights in the network (often randomly) *Do* *For each example x in the training set* *O = neural-net-output(network, x) ; forward pass* *T = teacher output for x* *Calculate error (T - O) at the output units* *Compute delta_wi for all weights from hidden layer to output layer ; backward pass* *Compute delta_wi for all weights from input layer to hidden layer ; backward pass continued* *Update the weights in the network* *Until all examples classified correctly or stopping criterion satisfied* *Return the network*

We presented the new examples which are in the test set (*34 videos "16 Left and 16 Right"*); it recognized well the hands in these videos, 97% right answers.

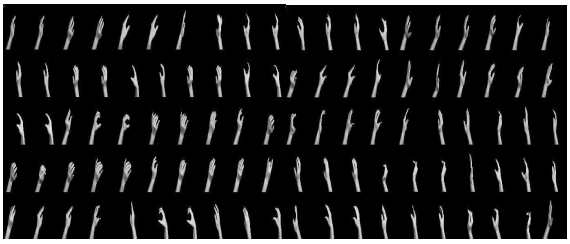


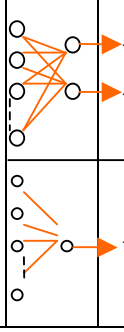
Fig.6 . Various hand videos used for training set

The second network, which is more important for us, recognizes the action done by the hands. This network also contains 3 layers. The first layer is the input layer and consists of 14 nodes (the moments). The hidden layer contains 30 neurons where each neuron have a weight matrix of 14 weights, one base and one output that uses the sigmoid activation function. Lastly the output layer consists of two neurons which take as input the outputs of the hidden layer neurons (30) and giveas output the result of what is the gesture or the action done by the moving hand in a video. Here we represent the output values as function in the following transformation:

1	1	PSO
-1	1	LMGO
1	-1	PHO
-1	-1	LMSO

The network was trained successfully on the training set. it gave 100% right answers. We presented the new examples that are in the test set (34 videos), it successfully recognized the gestures in these videos: 82.4% right answers. The result for two categories (left and right) respectively for the four actions of the test examples is summarized in the following table (cf. Figure 7).

Example:

Input video	features	ANN	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 -0.0003755 -0.000164941 0.000227694		-1 -1 -1	LMSO Left Hand

V. Conclusion

A new approach based on graphcut and 3D geometrical moments for hand gesture recognition is presented. The proposed framework classifies haptic properties through the video analysis of hand actions. Two objects properties have been tested: texture and consistency. For each property, two modalities were proposed. The texture of the object being explored could either be smooth or granular. Its consistency could either be hard or soft.

In this algorithm, after extracting efficient dynamic hand (in 3D) and applying necessary processing on theses videos of hand gestures, robust global features are extracted, based on 3D geometrical moments. These vectors are then used for training of neural network and hand gesture recognition.

We tested the proposed algorithm with the collection data set and the results showed the correct haptic gesture recognition rate of 82.4 percent. On the other hand, the results also showed a high recognition rate of 97 percent for left/right hand recognition. In addition, the proposed automatic approach is robust to traditional problems of gesture extraction and recognition. The framework can be used for interaction handicapped persons with computer and increase their abilities.

(a) First Neural Network for L/R HAND	100 %	for Left Hand
	93.8%	for Right Hand
(b) 2nd Neural Network for Gesture recognition	71.4%	for PSO
	85.7%	for LMGO
	80%	for PHO
	90%	for LMSO

Fig.7. Recognition rate of Left/Right hand recognition (a) , and of hand gesture recognition (b)

References

[1] S.J. Lederman and R.L. Klatzky, R.L. Hand movements: A window into haptic object recognition. Cognitive Psychology, 19(3), 342-368 (1987)

[2] C Yu and D Ballard, A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions ACM Transactions on Applied Perceptions, Vol. 1, No. 1, July 2004, Pages 57–80.

[3] S. Mu-Chun : A Neural Network-Based Approach to Recognizing 3D Arm Movement, Biomedical Engineering Application. Basis and Communication 15(1), 17-26 (2003)

[4] X. Hou et al. Direct Appearance Models. IEEE Conf. On Computer Vision and Pattern Recognition, vol. 1, pp. 828-833 (2001)

[5] T. Ogata, J.K. Tan, S. Ishikawa. High-Speed Human Motion Recognition Based on Motion History Image and an Eigenspace. IEICE Trans. On Information and Systems E89-D, 281-289 (2006)

[6] S. Kumar et al. Classification of Hands Movements Using Motion Templates and Geometrical based Moments, 3, pp. 299-304 (2004)

[7] M. Kenny, S.J. McKenna. An Experimental Comparison of Trajectory-Based Representation for Gesture. LNCS LNAI, vol. 2915, pp. 152-163. Springer, Heidelberg (2004)

[8] M. Hammami, Y. Chahir, L.Chen et D. Zighed . Détection des régions de couleur de peau dans l'image. EGC03 . RSTI, Vol.17, N°. 1-2-3, pp. 219-231, (2003)
[9] Y. Chahir and A. Elmoataz. Skin-color detection using fuzzy clustering, IEEE-EURASIP ISCCSP . ISBN 2-808848-17-8, Marrakech, Morocco , March 13-15, (2006.)
[10] Y. Boykov, and M. Jolly. Iterative graph cuts for optimal boundary and region segmentation of objects in N-D Images. Proc. IEEE 8th Int. Conf. on Computer Vision, Canada, CD-ROM, 2001.
[11] G. Bradski and J. Davis, Motion Segmentation and Pose Recognition with Motion History Gradients, IEEE Workshop on Applications of Computer Vision, December 2000.