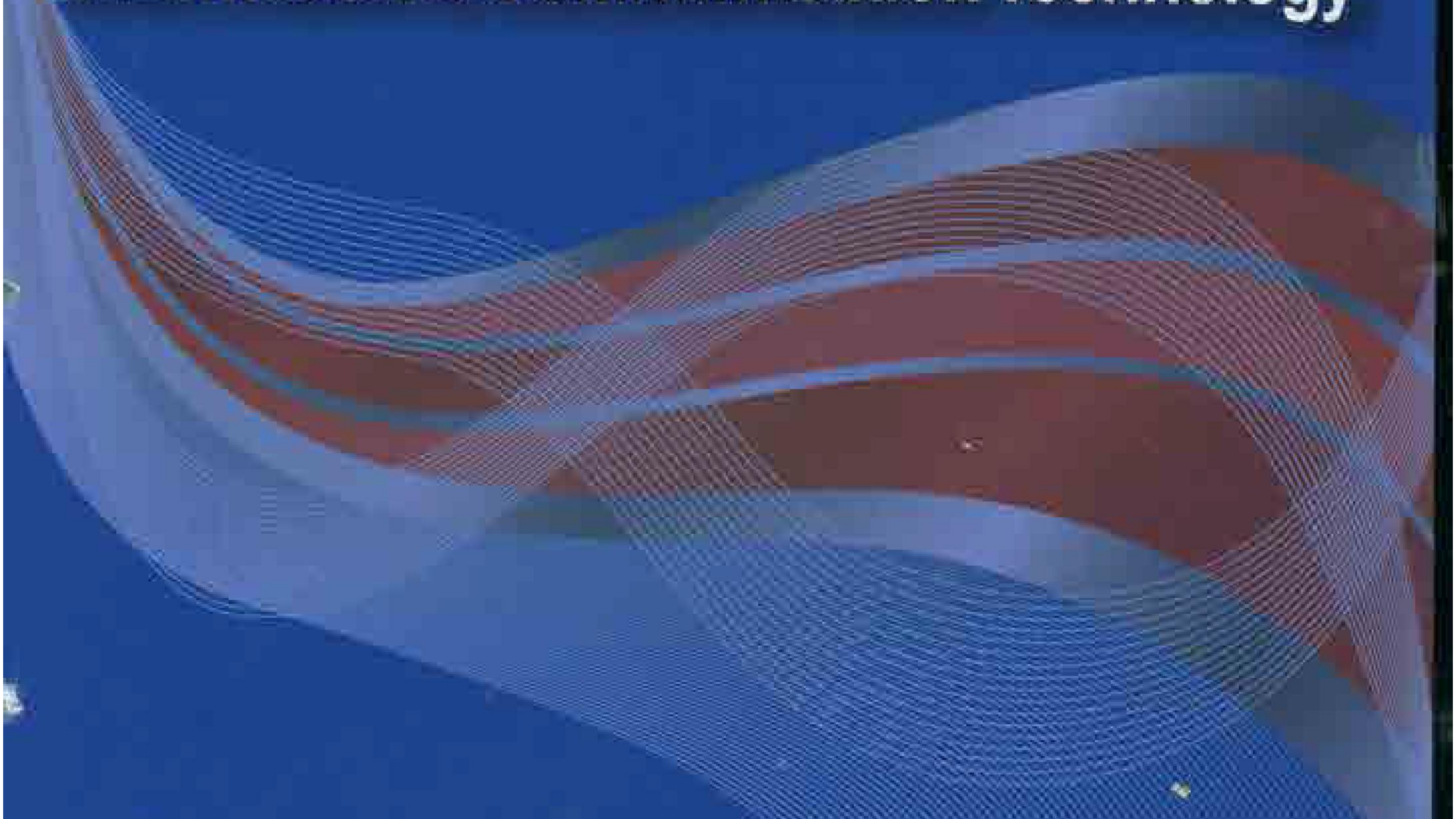
MINISTRY OF INFORMATION AND COMMUNICATIONS

Journal on INFORMATION TECHNOLOGIES & COMMUNICATIONS

NEW SERIES
ISSN 1859 - 3534

Volume E-1, No. 4(8)
10.2011

Research, Development and Application on Information & Communication Technology



CÁC CÔNG TRÌNH NGHIÊN CỨU, PHÁT TRIỂN VÀ ỨNG DỤNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Journal on INFORMATION TECHNOLOGIES & COMMUNICATIONS

NEW SERIES

Volume E-1,
No. 4(8)
[illegible]

JOURNAL OFFICE:
95E Ly Nam De, Hanoi, Vietnam
Tel: 04.37737136 - 04.37737137, Fax: 84.4.37737130
Email: chuyensanbcvt@mic.gov.vn; tapchibcvt@mic.gov.vn;
Website: http:/www.tapchibcvt.gov.vn

HOCHIMINH BRANCH OFFICE:
No 8A, D2 Street, 25 Ward, Binhthanh District, HoChiMinh
Branch Manager: Nguyen Van Nguyen *nvnguyen@mic.gov.vn*
Tel (Fax) : (84.8) 38992898; Mobile : 0913876194

License No. 433/GP-BTTTT dated 03/4/2009 Issued by Ministry of Infomation and Communications.

Printed by Army Printing Company.

Price: 20.000 VND

Contents

Vision-based Hand Gesture Recognition: a Survey

Tran Thi Thanh Hai
Int. Research Center in Multimedia Information, Communication and Applications
MICA – HUST - CNRS/UMI-2954 - GRENOBLE INP
Hanoi University of Science and Technology
1 rd. Dai Co Viet, Hanoi, Vietnam
Email: thanh-hai.tran@mica.edu.vn

Abstract: **Hand gesture is a natural mean of communication. The use of hand gesture as an efficient interface leads to a lot of applications in reality, which motivates more and more researches in modeling, analyzing and recognizing hand gestures. In the literature, there exists many works in this domain. Approaches range from using glove-based to vision-based. Each has some advantages and drawbacks. This paper gives a review on vision-based approaches for hand gesture recognition as well as its applications. This study helps to compare different techniques for hand representation and recognition in term of precision and computational time. By this way it allows identifying an appropriate method for a certain specific application using hand gesture.**

Keywords: hand gesture, feature extraction, machine learning, human computer interaction.

I. Introduction

Gesture is an intuitive and efficient mean of communication between human and human in order to express information or to interact with environment. In Human Computer Interaction (HCI), hand gesture can be an ideal way that a human control or interacts with a machine. In that case, machine must be able to recognize human hand gesture. Recently, hand gesture recognition becomes a hot research topic in the HCI and Computer Vision field due to its wide applications such sign hand language, lie detection, game, e-learning, human-robot interaction, etc. Approaches to recognize hand gestures can be divided into 2 categories: glove-based approach and vision-based approach.

Glove-based approach (e.g. [1], [2], [3], [4], [5]) requires the participant to wear a specific glove. This glove, built with several sensors, measures the location and bending of each finger as well as the global pose of the hand. The glove is connected to a server via wire or wireless connection for data transmission and processing. The type of data depends strictly of the type of sensors to be used as magnetic, lighting or acceleration sensor.

Glove-based approach for hand gesture recognition is very fast and accurate. However, this approach has some remarkable drawbacks. First, the fact of wearing glove is not comfortable because this is not a normal glove but a sensor-based glove. Therefore, the realized gesture will be no more natural and human reflection is slower. Secondly, the glove-based method can work only in a limited environment due to the connection between devices. Finally, such specific glove is very expensive for experiment.

To avoid these difficulties, **vision-based approach** (e.g. [6], [7], [8], [9], [10], [11]), studied since 20 years still interests a lot of researches. The vision-based approach uses only one or several cameras instead of gloves. The image of hand gesture captured by camera will be next processed by hand gesture recognition system using image processing/computer vision techniques.

Vision-based approach is not evident because of multiple following reasons. First, as hand posture has dynamic hand gesture must be represented in temporal-spatial domain.

The recognition of dynamic hand gesture requires analysis of hand motion. In many works for example sign hand language or automatic control, static hand postures are enough to express or to control something. However, for interactive game, moving objects requires to recognize dynamic hand gesture. Therefore, this paper will study both *hand posture classification* and *dynamic hand gesture recognition.*

C. *Main components of a hand gesture recognition system*

A vision-based hand gesture recognition system is generally composed of 4 main components: 1) hand detection; 2) hand modeling; 3) posture classification; 4) dynamic hand gesture recognition. In the following section, we will detail each component of the system.

III. HAND DETECTION AND REPRESENTATION

A. *Hand detection*

Hand detection aims to determine the presence of hand regions and locate them in the image. It allows removing non-interest regions that will be not considered in the next steps therefore reduces the computational time. Hand detection could be considered as hand region segmentation. The next step will be assignment of detected hand-regions to hand posture classes.

The segmentation of hand regions will be based on hand features (e.g. shape, color, texture, motion) that allow distinguishing hand from other objects ([23], [24], [25], [26], [27], [28]).

The typical technique for hand detection is based on active contour or snake [28]. The main idea is to represent each object by a set of points (e.g. points on the contour line) and to fit a contour detected in the image with a predefined hand contour using deformation technique by minimizing a predefined energy function. Hand detection based on active contour is not suitable in case of local occlusion or bad image quality. The first problem can be resolved using a hand tracker while the second issue can be carried out using only key-points on active contour.

Neural network technique is another solution for hand segmentation. A neural network will be trained to learn background images (negative examples) and hand images (positive examples). In [23], the authors used RGB vector as input of neural network. A drawback of this technique is the network configuration is static, it cannot predict or/and adapt to changes in data representation.

Color is important information to distinguish objects whenever the shape cannot do it. Color is invariant to transformations (i.e. rotation, translation) but quite sensitive to occlusion, color temperature, camera sensor. In addition, there are a lot of objects of same color that confuses the segmentation. To segment skin color pixel in image, many methods ([23], [24], [25], [29], [30], [31], [32], [33], [34], [35], [36]) have been proposed such as Probability Model, Mixture of Gaussians, Continuously Adaptive Mean-Shift.

Each method of hand detection has its own advantage and disadvantage. Apart from requirement to adapt to image acquisition condition, each method should be as simple as possible for real-time applications. Among above studied methods, probability model based hand segmentation seems to be simple and efficient. Some authors showed that the true positive rate of segmentation using this technique could attain to 94% ([25], [37]). Although the result is promising, these methods cannot distinguish a face region from hand regions because they are of the same skin color.

B. *Hand modeling*

Hand modeling approaches can be divided into 2 categories: 3D model and appearance-based model.

1) *3D hand modeling*

3D hand modeling approach represents explicitly hand posture through parameters as angles at junction point and hand pose. This approach allows reconstructing 3D hand posture in a very accurate manner. The methods of hand modeling in this category are classified into 2 groups: *volumetric model* and *skeletal model*.

Volumetric model has been widely used in computer graphic to describe 3D visual appearance of human body. In computer vision, researchers recognize a hand posture by synthesize 3D model of this one then varying its parameters until the deformed synthesized model and the real posture appear as the same visual image [37], [38]. To synthesize a 3D model, a lot of authors used simple geometrical structures like cylinders, super-quadrics to be rendered in near real-time. There are two problems in *volumetric modeling*. First, the dimensionality of the parameter space is very high. Secondly, the computation of such parameters via computer vision technique seems to be quite complex.

Skeletal model is a solution to avoid dealing with all parameters of volumetric model. Instead, skeletal model represents hand with a reduced set of parameters as joint angles or segment lengths.

In [39], the authors proposed a Digit Eyes system that treats hand tracking as a model-based sequential estimation problem: given a sequence of images and hand model, they estimate the 3D hand configuration at each frame. All possible hand configurations are represented by vectors in a state space. Each hand configuration generates a set of image features, 2D lines and points, by projection via a camera model. A feature measurement process extracts these hand features from gray-scale image by detecting the occluding boundaries of finger links and tips. The state estimate for each image is computed by finding the state vector that best fits the measured features. In the paper, the authors showed the good function of the system in a constrained condition: invariant lighting, uniform background. The Digit Eyes system can recover the state of 27DoFs hand model at 10Hz.

The authors in [40] proposed to model 3D hand using Point Distribution Model (PDM). The PDM is a deformable model built from statistical analysis of examples of the object being modeled. Each hand posture is represented by a vector of 3D coordinates of landmarks. To reduce the high dimensionality of vector space, Principal Component Analysis (PCA) technique is used. Real-time tracking is achieved by finding the closest possibly deformed model matching the image. The experiments have showed that this method can work at 10Hz with uniform background images on 134MHz workstation.

[41] presented a method for 3D hand tracking using hand model built from truncated quadrics. This allows for the generation of 2D profiles and for an efficient method to handle self-occlusion. The tracking works at 3Hz on a Celeron 433MHz.

We found that the 3D hand model offers a rich description that allows a wide class of hand gesture. This approach is ideal for realistic interactions in virtual environment. However, it has some disadvantages. First, at each frame, the initial parameters of the model have to be close to the solution; otherwise it is liable to find a suboptimal solution (i.e. local minima). Secondly, the fitting process is very sensitive to noise. Thirdly, as 3D hand models are articulated deformable objects with many DoFs, a very large number of images are required to cover all hand configurations under different views. Finally the 3D model approach cannot handle the inevitable self-occlusion of the hand when mapping it to 2D plan. For these reasons, most works based on 3D model are hand tracking because the hand configuration doesn't change a lot between two consecutive frames.

2) Appearance-based modeling

Appearance-based modeling represents hand by image itself or its extracted features. In this category, model parameters are not directly derived from 3D spatial description of the hand. The gestures are modeled by relating the appearance of any gesture to the appearance of the set of predefined template gestures. There is a variety of methods of appearance-based hand modeling.

a) Deformable 2D template

The deformable 2D template is a set of points on the outline of the object used as interpolation nodes for object outline approximation. The simplest interpolation function is a piecewise linear function. The templates consist of average point sets, point variety parameters.

In a HCI application, [42] proposed to use 2D rigid model to represent hand. Each model is represented by silhouette contours, which are modeled by splines. This model is simple, permits to reduce the complexity of the involved computations and remains discriminative enough to track a limited set of hand postures. To detect hand, the author used skin segmentation. The confusion between face and hand (of the same skin color) is resolved by some heuristics. The recognition rate attained to 95% even with cluttered background.

b) Elastic graph

In [7], the authors represented hand by 2D elastic graph. Each vertex of the graph is labeled by a vector of Gabor kernels. Each edge is assigned with a distance between two vertices. A hand posture is built from 35 vertices and 70 edges. For one posture class, an elastic graph model will be built from 6 images (3 subjects x 2 backgrounds).

The model based on elastic graph has been largely used for object representation and face recognition. The fact of building an average graph from multiple ones allows improving the recognition rate when the background changes. Experiments showed that the method obtained 86.2% as recognition rate on complex background.

Template models or elastic graphs are usually used for hand tracking. They can be used to classify simple postures. The trajectory of deformable parameters will be used to recognize dynamic hand gestures.

c) Image features for hand representation

Recently, there is an increasing trend in the research of *significant features* for object representation in the field of computer vision. There are two questions to be posed: What is a good feature? How many features are enough for representing an object?

In [43], the author has indicated some criteria for evaluating a feature. They are *expressivity, repeatability, invariance, efficiency*. Depending of application, the criteria will be classified in priority order. Concerning the number of features, there is no exact answer for this. A little feature number cannot represent all information about object. A much more number of features cause unnecessary computations. There are two types of feature we will consider for hand representation: *global feature* and *local feature.*

Global features:

Global feature is a feature built from all pixels in the image. This does not require feature extraction. The difference between methods for object representation is the measurement used to describe the entire image. Some authors proposed to use the original image, some others used histogram of gray-scale / color image or motion history image.

Image-based hand representation: The simplest method for hand modeling based on appearance is directly use images of this one. The direct use does not require any feature extraction. However, an important drawback of this method is the high dimensionality of the parameter space (i.e. image resolution). Therefore the searching and recognition in this space is very time-consuming. To reduce the number of dimensions, PCA is a popular solution (e.g. [9], [44]). Once the new space is built, each hand configuration will be a point in this space. The difference between two configurations is measured by a Euclidian or Mahanalobis distance.

Global orientation histogram: Orientation histogram has been used for hand representation in

[45]. This histogram is computed for all pixels in the image. Due to the use of orientation, not intensity, orientation histogram is quite consistent to illumination change. This approach is simple, fast to compute because no salient region detection is needed.

However, this approach has some drawbacks: two similar configurations can have different representations (variant to orientation); two different configurations have the same representation. So, the method is not invariant to change in position and orientation of the hand.

Motion feature: As dynamic gesture is performed by moving hand, the motion information can be used to represent a dynamic hand gesture. The motion can be detected by subtracting two consecutive frames.

In [30], the author used motion history image to describe a gesture. A motion history image is a 2D image built by accumulating motion at each separated pixel at different times. By this way, the intensity of each pixel in the motion history image will represent the movement of pixel belonging to the object in question.

To reduce computational time on the overall image, in [46] the authors detected only peaks on the motion history image. Position and moment's detected peaks are considered as description vector for the Hidden Markov Model (HMM). This method allows a very compact description of the moving hand therefore the computation is very fast.

Evaluation of global features: Global features have some advantages such as they are simply computed and they don't require feature extraction. However, they have several disadvantages. While modeling object pixels, they model also background pixels. In addition, as they use all pixels of the entire image, feature space is of high dimensionality. Some feature types cannot distinguish two different images (e.g. histogram). Finally, they are not robust to occlusion, scale and view point change.

Local features:

To deal with above problems of global features, local features have been studied. The local features encode information about object that will be very difficult to learn from raw data. There are three principal steps for representing an object (hand) using local features: 1) salient region detection; 2) region description; 3) feature space building for an efficient searching. The difference between methods depends of the choice of description vector.

There is a discussion about if salient regions detection is really necessary. In [47], the authors showed that the use of all dense interest points can give better recognition rate of object. However, there are still very much works that do the salient region detection before further processing. There are some reasons for this: although salient region detection is time-consuming and sensitive to noise; they allow keeping only saliency of the object. Therefore, in the following, we will present methods which require or not salient region detection and methodes to compute description vector of the surrounded region.

Haar like features: Haar like feature has been widely used for face detection and hand detection (e.g. [48], [49], [50]). Haar like feature is a rectangle composed of black and white rectangles of different orientation and size. Each Haar like feature is represented by a value that is the difference of sum of all pixels value in the white rectangle and the sum of all pixels value in the black rectangle. These features represent an edge or a centre line of the object.

Haar like feature is very good for representing object when its shape doesn't change. As hand is an articulated object with a lot possible configurations, the number of classifiers that must be learnt to consider all configurations of the hand is numerous.

Local orientation histogram based on Scale Invariant Feature Transform (SIFT): To avoid disadvantages of global orientation histogram for hand representation that takes all pixels of the image into account, some authors proposed to detect salient features as SIFT that have been shown very robust to transformations, illumination and scale change [51]. SIFT is used in [11] to detect and recognize hand postures.

Structured features (ridge, blob) for hand representation: Ridge and blob are features studied from the year of 90 to describe natural lines (ridge) or around regions (blob) in image [52], [53], [43]. [43] showed that ridge and blob are good features for representing in a semantic ways structured objects like text, human, face, hand.

Blob and ridge have been used to hand tracking in [54]. The main idea is to represent hand at intrinsic scales such that ridges and blobs detected at sparse scale represent overall structure of the hand while ridges and blobs detected at smaller scales represent fingers.

The main advantages of the method is that the model represents semantically the structure of the hand. The approach works at 10Hz on a Pentium III Xeon 550 MHz on even complex background. However, this method has some drawbacks: it is assumed that no skin like object appeared on the image. In addition, the ridge and blob detection is quite time consuming and most importantly, the analysis of ridges at intrinsic scales is very difficult.

Multimodal features for hand representation: Each feature type for representing hand has its own drawback. Combining some types of feature improves recognition performance. In [55], authors used a set of following features for hand representation: original image, image filtered from Sobel vertical, horizontal, and magnitude, derivatives, motion energy, motion history image, skin color. The combination of several features allows a better recognition w.r.t [46]. The recognition rate attains to 92%.

To summarize, there are two approaches for hand modelling: 3D model and appearance-based model. 3D models give an accurate representation of hands and allow rebuilding hand in 3D space, so very suitable for interaction application. However, the fact of fitting model in the image is very sensitive to noise and experiments have been carried out only in ideal environment. The problem of local optima appears quite frequently when the initialization is far from the configuration of the hand in question. Therefore, the 3D hand model is suitable for hand tracking when the configurations do not change a lot between two consecutive frames.

The appearance-based modeling approach, like almost approaches based on image features, is an explicit representation because it doesn't express any information about a 3D hand. The feature extraction is not a simple problem. However, the use of image features allows representing a variety of hand configurations under illumination, scale change.

IV. Static Hand Posture Classification

A. *3D model posture classification*

When hand gesture is represented by 3D model, the 3D parameters of the model will be learnt. The recognition will be carried out by projecting the 3D model on the image that we need to detect hand. The recognition consists of looking for a transformation that minimizes the difference between points on contour lines of image in question and the projected model.

B. *Appearance based posture classification*

Instead of learning 3D model parameters, the approach of this kind learns features extracted from images of hands. In the following, we will study some methods for learning image features based hand models and hand posture classification.

1) Eigenspace-based classification

Eigenspace has been widely used in classification problem or object poses determination [56]. This approach has been used for face detection and recognition. The main idea is with a set of many examples, only eigenimages that represent approximately the variation in the set of examples will be used as basis vector of eigenspace

The eigenspace approach uses entire image as global feature, so does not require feature extraction. Each image is represented by an intensity vector. A posture is represented by a set of all images of the same configuration of the hand, observed in different conditions. The recognition is done by projecting the image of the hand in question to the eigenspace. The nearest configuration in the eigenspace will be the posture to be recognized. In reality, recognizing hand in an image requires some pre-processings as hand region segmentation to limit region of non interest; normalization of candidate regions following size, illumination before projecting in the eigenspace.

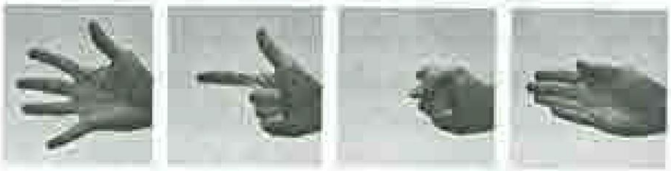


Figure 1. 4 posture classe used in [44]

In [44], 100 images belonging to 4 classes of posture (Figure 1) have been used to build the eigenspace of 25 D.

The authors have experimented the recognition system with 100 images rotated of 15 pixels. Euclidian distance has been used to determine the nearest candidate in the eigenspace. However, the authors demonstrated only the good working but not the quantitative evaluation of recognition rate of the system. Beside, learning and testing images are of uniform background. Hand size does not change between images. Constant illumination is considered.

In [9], eigenspace approach is used to recognize a set of 25 postures for hand sign language. Each posture represents a character in the alphabet. The Euclidian is used for measuring the distance in eigenspace. Training and testing images are both of uniform background, without arm occlusion.

Training data compose of 1000 images and testing data compose of 1500 images with resolution 256x248. A threshold is used for extracting hand region. The recognition rate is very significant: 99% (only 6 false recognitions). The frame rate is about 14Hz.

Eigenspace is a simple approach because it uses global features (i.e. entire image). However, this approach cannot recognize hand postures when they are partially occluded or there is a view point change. The papers reported only results when testing with uniform background image in ideal lighting condition and obtained very high recognition rate. If images are taken in an unconstrained condition, recognition rate will be significantly reduced.

2) Neural network for classification

Neural network is trained to do the task of classification by regulating weights of the network. The complexity of the network depends of the number of nodes, the transfer function and the connection between nodes of the network.

In [8] authors used neural network to learn a set of command gestures in a robot controlling application. The number of inputs and outputs equal to the sample resolution (20x20, 18x20 or 18x30). To recognize hand posture in the image, an active window will be determined. It is a region containing face. The skin detection will locate hand in the image based on the relative position between hand and face and hand posture will be classified using the trained neural network. The classification is carried out by computing the distance of the example to the posture set.

To evaluate the algorithm, the authors have built a database of 6 posture classes corresponding to A, B, C, 5, Point, V. Each posture is used to execute a command in a system for locating and tracking individual speaker (LISTEN). A thousand images have been taken in different conditions of lighting, background, view, scale.

When the images are of uniform background, the recognition rate obtains 94% while it is reduced significantly to only 75% in case of complex background.

The algorithm has been also evaluated with Jochen database of grayscale images with resolution 128x128, of 10 postures taken from 24 different subjects on white-black or complex background [57]. The recognition rate attains 93.7% with simple background images and 84.4% with complex background images (Figure 2).

Figure 2. Examples of hand postures image in database of Jochen used by Marcel [8]

In [58], Nguyen *et al.* proposed a method for gesture recognition by combining neural network and fuzzy ARTMAP. The neural network consists of 4 layers. Instead of using a vector composed from all pixels of the original image, the authors divided image into blocks to reduce the dimensionality of the input vector. Each input of the network is a binary value taken from binarized hand image.

The training and testing data are **binary** images of 36 postures. For each posture class, they took 200 samples. The total number of samples in the dataset is 7200 images. The authors evaluated the algorithm on 3438 images and obtained 92.19% as recognition rate.

The hand posture classification using neural network has following advantages: the neural network can model more complex distribution of data than traditional methods. However, it has some drawbacks. First, it is not able to describe the model of the data. This is why one cannot explain why in some cases it works well but not other cases. In addition, it is very hard to extract rules from neural network to help analyzer to explain the results. As all other methods, if the training data are not significant, the network cannot give good response.

3) Boosting approach

Adaboost is an approach that reduces the number of non-significant features for representing an object. This is the case of using Haar like feature because th number of Haar like features extracted from an imag is very big.

To speed up Adaboost, some authors proposed t build Cascaded Adaboost. The main idea is as ther are a lot of regions in the image which do not contai hand, some classifiers will be used to remove non interest regions rapidly in order to concentrate more o difficult examples. The principle of Cascade Adaboost is to put several Adaboost classifiers i consecutive layer.

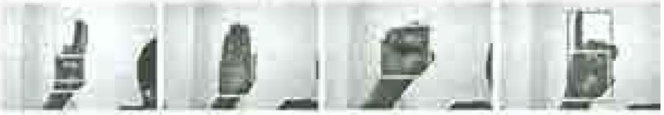


Figure 3. 4 posture classed are considered and recognition results [50]

In [50], the authors used Cascaded Adaboost fc hand posture classification using Haar like features. 4 postures are considered (Figure 3).

The dataset are taken with webcam at resolutio 320x240. Each posture class has 450 samples taken a different scale, angle view on simple background. 50 non-hand images are taken as negative images to trai the Cascaded Adaboost classifiers.

Each Cascaded Adaboost classifier is trained t recognize a posture class. To classify postures, th parallel structure of Cascaded Adaboost will be built The recognition rate is about 98% for each postur class when testing with 100 images. The algorithn works well even there is a rotation of 15 degree o the hand.

In [59], Haar like features are inputs of Cascade Adaboost. 6 classes of postures have been learnt closed, sidepoint, victory, open, Lpalm, Lback. Thes postures are taken in varying conditions.

To detect hand in image, image will be scanned a several scales. To evaluate the algorithm, the author have taken 2300 images of right hand of 10 men an 10 women with 2 different cameras in indirect har lighting condition. Images are then normalized to rotation and size. The authors showed that the Cascaded Adaboost built from 100 weak Adaboost classifiers gives recognition rate 92.23% with false recognition rate $1.01*10^{-8}$.

As Haar like features are not invariant to rotation, strong change in illumination and scale, [11] computed SIFT features and used SIFT descriptor as input vector for Adaboost. 3 hand postures are considered: Palm, Fist and Six, which are acquisitioned in different conditions. 642 training images are taken from Massey dataset [60]. 450 images of the *Fist* posture and 531 images of the *Six* posture are taken at different lighting conditions. Background image are taken from 830 images from internet and 149 images taken by the authors. To evaluate the algorithm, 275 images have been taken from webcam 320x240. The recognition rate obtained with Adaboost is 95.4% in average.

The authors showed that the combination of SIFT-Adaboost is better than Haar-Adaboost, mostly when there is noise and the background is complex. The SIFT-Adaboost works still well when the hand rotates an angle of 40 degree. We found that Adaboost is a good classification method for hand posture classification. SIFT-Adaboost is still better than Haar-Adaboost due to some interesting properties of SIFT features. However, **SIFT** is very time consuming therefore not suitable for real-time application.

4) Support vector machine (SVM)

SVM has been used for object classification and recognition. In [61], the authors used active contour algorithm to determine the human body and hand contour lines. Some heuristics are used to determine hand on this active contour. The feature vector is then built from the coordinates of the 4 points: top of head and fingertip, feet position and the shoulder. 2 subjects participated into experimentation. 5 videos are captured. The SVM is used to recognize "Pointing" gesture and obtained recognition rates 71% when hand is observed from side camera and 94.4% in case of frontal camera.

V. TEMPORAL GESTURE MODELING AND RECOGNITION

Section IV presented methods for hand posture classification. In this section, we will study methods to recognize dynamic hand gestures.

A. Hidden Markov Model (HMM)

HMM is a statistical model to represent random signal. In [62] the authors resolved the problem of 3D object control in virtual environment using HMM. 3 dynamic hand gestures "Greate", "Quote", "Trigger" have been considered. For preparing data, the authors use CyberGlove. Each posture is represented by a state vector of 20 elements representing joint angles.

Each dynamic hand gesture is represented by a HMM and each state of HMM is a posture represented by a vector 20 elements. At the end of the training process, 3 HMM corresponding to 3 gestures are obtained. Recognizing a gesture returns to finding a HMM having the biggest probability given the observation.

[62] showed that HMM is suitable for dynamic hand gestures. However, the authors have experimented with data taken by glove; it will be more difficult to obtain these observations with camera.

Based on HMM, [63] used data obtained from the Cosine transformation of images taken from camera. At each time, image is segmented and hand is detected as blob in the image. Kalman filter is used to track hand in the video.

The authors have experimented with 36 hand gestures on uniform and complex background. The recognition rate is about 98% but it is very difficult to evaluate quantitatively and explain this result.

In [46], the authors have used HMM combining with analysis result of motion image. To reduce the number of dimensions of state space, they used only mountain regions on difference image between frames. Each mountain region is characterized by a vector of 7 elements corresponding to coordinates and deviations, distance between mountain centres in horizontal/vertical orientation and motion index.

In [64], the author proposed also HMM to recognize alphabets (A-Z) and numbers (0-9) from stereo image sequences. There are 3 stages in the system: segmentation and pre-processing for hand regions; feature extraction and classification. In the first stage, color and 3D depth map (due to the use of stereo images) are used to detect hand, then Mean-shift and Kalman filter are used to track the trajectory of the hand movement. In the second step, feature extraction combines 3 types of features: location, orientation and velocity with respected to Cartesian systems, then K-means is employed for HMMs codeword. To train all parameters of HMM, Baum-Welch algorithm is used. The method is tested with 240×320 sequences, 30 sequences for training and 20 sequences for testing for each gesture. This approach works well in real-time even in case of cluttered background and partial overlapping.

B. *Hybrid recognizer*

In [65], a hybrid approach has been proposed to improve the recognition rate of HMM approach. This is the combination of HMM with Artificial Neural Network (ANN) coming from the idea that ANN represents the non-linear local dependence of postures while HMM allows processing the change in time.

In the paper, the authors used state vector of 15 elements composed of second order moments of the region representing hand, angle between *x* axis and the straight line connecting right hand center and left hand center, the length of this straight line, normalized vector of movement of the hand. Following this principle, each hand gesture is represented by a HMM. ANN is trained to classify state vectors which are randomly distributed.

90 videos for each gesture at resolution 120x90 have been used for training and testing. Recognition rate is about 91%, which is better than using only HMM presented in their previous works.

To summarize, the using of ANN helps to improv recognition rate of hand gestures. However, drawback of this method is it is time consuming due t the computations performed on the image. A solutio for that is to detect only candidate regions (han detection) before hand recognition.

C. *Conditional Random Field (CRF) and Finite Stat Machine (FSM)*

HMM is a powerful generative model that include hidden state structure, this generative model assume that observations are conditionally independen This restriction makes it difficult or impossible t accommodate long-range dependencies amon observations.

CRF, a discriminative model, was first introduce in natural language processing for tasks such as nou co-reference resolution, name entity recognition an information extraction. Recently, there has bee increasing interest in using CRF in vision communit CRF uses an exponential distribution to model th entire sequence given the observation sequenc allowing to avoid the independence assumptio between observations, and allows non-loca dependencies between state and observations.

In [66], the authors compared gesture recognitio using HMM, CRF and Hidden CRF (HCRF). Hidde CRF is CRF with hidden variables. A HCRF allow modeling the conditional probability of a class lab given a set of observations. In the case of gestur recognition, users were asked to perform these gesture in front of a stereo camera. From each image fram a 3D cylindrical body model, consisting of a heac torso, arms and forearms was estimated. The joir angles and relative coordinates of the joint and the arr are used as observation. 6 arm gesture classes (labels are considered as in Figure 4. The experimental result showed that the HCRF outperforms HMM and CR for certain head gesture recognition tasks (72% (HCRF) against 65.33% (HMM) and 66.53% (CRF in term of recognition accuracy). In case of han gesture recognition, HCRF's recognition accuracy obtains to 97.55%.

Finite State Machine (FSM), a behavior model composed of a finite number of states, transitions between those states, and actions has been used to model dynamic hand gesture [67]. Features, computed from the input video images, used for input to gesture modeling and recognition are the 2D positions of the centers of the user's face and hands, which are determined by skin color based segmentation. A gesture is defined as an ordered sequence of states in the spatial-temporal space. Each state has 5 parameters. The FSM is trained for each hand gesture via 2 phases: 1) spatial clustering phase determines the number of states for each hand gesture; 2) temporal alignment phase builds the state sequence corresponding to the data sequence. Each state sequence is a FSM recognizer for a gesture. When a new feature vector arrives, each gesture recognizer decides whether to stay at the current state or to jump to the next state based on the spatial parameters and the time variable. When a recognizer reaches it final state, the gesture is recognized. This method is quite similar to HMM-based method. The difference is that with HMM, the number of states and the structure of the HMM must be predefined. The FSM does not define the number of states because the gesture model is produced thanks to the segmentation and alignment of training data. Unfortunately, in the paper, the authors did not mention the performance of hand gesture using FSM.

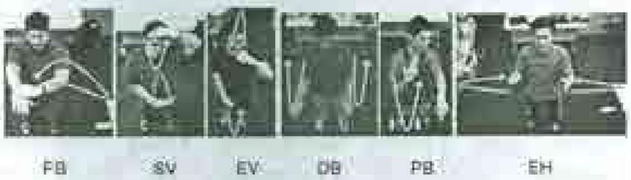


Figure 4. Illustration of 6 gestures classes [66]

VI. Conclusions

In this paper, we reviewed techniques for hand gestures recognition methods used in different applications ranging from robotic, control devices to hand sign language. We started our studies with definition of hand gestures and its classification. The framework of an overall system for hand gestures recognition is proposed. Following this framework, we presented each component: hand detection, hand representation; posture classification and dynamic hand gesture recognition.

Hand detection is an optional step that helps to remove all non interest regions in the image to be processed later therefore reduce the computational time. The technique most frequently used is skin segmentation. However this technique requires more data for training and depends a lot of lighting condition and human origin. Even obtaining a segmentation rate to 94%, skin segmentation remains a big challenge.

Concerning hand representation, there are 2 categories of approach: 3D-based modeling and appearance-based modeling. 3D based-model based on 3D parameters of a hand (e.g. joint angles, length finger) so it is good to recover the 3D hand for interactive applications (e.g. moving object in a virtual environment). However, the fact of determining model parameters and the searching for the most similar model in the image is an optimization process and can easily return to a local extreme. Therefore, 3D model is generally applied in hand tracking because the hand configuration does not change a lot between two consecutive frames. Appearance-based model seems to be more convenient because it used features computed from images. Appearance models do not give an explicit representation about hand structures, but lead a lot of methods for hand recognition. Features range from global (e.g. entire image, PCA technique) to local (e.g. Haar like feature), static to motion features, numeric (e.g. oriented histogram) to structured features (e.g. elastic graphic, ridge and blob). Global features give a simple learning and recognition but they are not robust to occlusions. Local features as Haar like give the good results of recognition. In addition, the computation is very fast. However, Haar like features require more examples for training and depend strongly of hand subjects to be learnt. Ridge and blob are good features for representing structured objects but blob and ridge extraction is quite sensitive to noise (when connecting ridge at intrinsic scale) and time - consuming.

Techniques for hand gesture recognition depend of features and the method used for hand representing. PCA is suitable for global features to reduce the high dimensionality. Cascaded Adaboost is a good accompany of Haar like features, first to reduce the number of features to be considered, secondly, to reject rapidly non-candidate hand gestures. The recognition rate is about 92%. The combination of SIFT and Cascaded Adaboost gives better recognition rate than Haar like features and Cascaded Adaboost. However, SIFT is time-consuming and not widely used for real-time applications. SVM is a typical classifier for all types of numeric features. However, currently the using of SVM for hand gesture recognition does not obtain good result. Neural network is used in some cases and give quite good results.

For dynamic hand gestures, HMM is commonly used in signal processing and also in hand gesture recognition. Generally, each state of HMM represents a configuration of hand (hand posture) and the number of nodes in HMM represents key configurations representing dynamic hand gesture. The using of ANN in the choosing of the best state that fits the observation data is a good solution in the framework of HMM. It gives a better recognition rate (90%) than using only HMM. However ANN is time consuming.

In conclusion, there is no exact answer for the question: which method is the best for hand gesture recognition. The recognition rate reported in the cited papers is not evaluated on the same database. Therefore, this information is just for reference. However, we have some qualitative conclusions as follows: For posture classification, Haar like features are good accompanied to Cascaded Adaboost. To avoid miss detection, it would be possible to regular some parameters of Cascaded Adaboost then remove false detections by verifying if the candidate regions satisfy criteria on some properties of Hu moments or skin color. For dynamic hand gesture recognition, HMM is a typical model. The difference between methods is the features used to model states of HMM. The contextual information could be integrated into the process to improve the recognition rate.

ACKNOWLEDGMENTS

The research leading to this paper was supported by the National Project B2011-01-052 "*Study and develop an object detection and recognition system in smart and perceptive* environments". We would like to thank the project and people involved in this project.

REFERENCES

[1] D. J. Sturman, D. Zeltzer. A Survey of Glove-based Input. IEEE Computer Graphics 1994;14(1).

[2] http://www.idemployee.id.tue.nl/g.w.m.rauterberg/presentations/HCI-history/tsld065.htm.

[3] http://www.5dt.com/products/pdataglove5u.html.

[4] http://www.vrealities.com/pinch.html.

[5] http://www.vrealities.com/cyber.html.

[6] F. Dadgostar, A. L. C. Barczak, A. Sarrafzadeh. A Color Hand Gesture Database for Evaluating and Improving Algorithms on Hand Gesture and Posture Recognition. Research Letters in the Information and Mathematical Sciences 2005;7:127-34.

[7] J. Triesch, C. Malsburg. Robust Classification of Hand Postures against Complex Backgrounds. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition 1996; Killington, VT , USA 1996. p. 170 - 5

[8] S. Marcel. Hand Posture Recognition in a Body-Face centered space. CHI'99 Conference on Human Factors in Computer Systems; 1999; Pittsburgh, PA USA; 1999. p. 15-20.

[9] H. Birk, T. B. Moeslund, C. B. Madsen. Real-Time Recognition of Hand Alphabet Gestures Using Principal Component Analysis. In Proc of 10th Scandinavian Conference on Image Analysis; 1997; Lappeeranta, Finlande: Pattern Recognition Society; 1997. p. 261-8.

[10] Q. Chen, N. D. Georganas, E.M Petriu. Real-time Vision based Hand Gesture Recognition Using Haar-like features. Conference Proceedings of IEEE on Instrumentation and Measurement Technology IMTC'07 2007: 1 - 6

[11] C. C. Wang, K. C. Wang. Hand Posture recognition using Adaboost with SIFT for human robot interaction. Proceedings of the International Conference on Advanced Robotics (ICAR'07); 2008; Jeju, Korea; 2008.

[12] Y. Wu, J. Lin, T.S. Huang. Analyzing and capturing articulated hand motion in image sequences. IEEE

Transactions on Pattern Analysis and Machine Intelligence 2005;7(12):1910–22.

[13] M. Baldauf, P. Fröhlich. Supporting Hand Gesture Manipulation of Projected Content with Mobile Phones. Proceedings of the Workshop on Mobile Interaction with the Real World (MIRW), 2009.

[14] T. Schlomer, B. Poppinga, N. Henze, S. Boll. Gesture recognition with a Wii controller. Proceedings of the 2nd international conference on Tangible and embedded interaction. Bonn, Germany, 2008.

[15] G. Bernstein, N. Lotocky, D. Gallagher. Robot Recognition of Military Gestures. CS 4758 Term project: Cornell University; 2009.

[16] V. I. Pavlovic, R. Sharma, T. S. Huang. Visual interpretation of hand gesture for human-computer interaction: a review. IEEE Trans on Pattern Analysis and machine Intelligence 1997;19(7):677-95.

[17] M. Kohler, S. Schroter. A survey of video-based gesture recognition - stereo and mono systems: University of Dortmund, August 1998; 1998.

[18] Y. Wu, T. S. Huang. Vision-Based Gesture Recognition: A Review. Lecture Notes in Computer Science 1999:103-15.

[19] Y. Wu, J. Y. Lin, T. S. Huang. Capturing natural hand Articulation. In Proc 8th Int Conf on Computer Vision,. Vancouver, Canada, 2001: 426–32.

[20] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, X. Twombly. Vision-based hand pose estimation: a review. Computer Vision and Image Understanding 2007;108:52-73.

[21] P. Garg, N. Aggarwal, S. Sofat. Vision Based Hand Gesture Recognition. World Academy of Science, Engineering and Technology 2009;49:972-7.

[22] S. Marcel, O. Bernier, J. Viallet, D. Collobert. Hand Gesture recognition using Input/Ouput Hidden Markov Models. FG '00 Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition; 2000; Washington, DC, USA: IEEE Computer Society; 2000. p. 456-.

[23] K. A. McCrae, D. W. Ruck, S. K. Rogers, M. E. Oxley. Color Image Segmentation. Proc of SPIE, Applications of Artificial Neural Networks; 1994; 1994. p. 306-15.

[24] L. Sigal, S. Sclaroff, V. Athitsos. Skin Color-Based Video Segmentation under Time-Varying Illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 26:862–77.

[25] B. D.Zarit, B. J. Super, F. K. H. Quek. Comparison of five color models in skin pixel classification. ICCV'99 Int'l Workshop on recognition, analysis and tracking of faces and gestures in Real-Time systems; 1999; 1999. p. 58-63.

[26] A. Birdal, R. Hassanpour. Region Based Hand Gesture Recognition. 16th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. Bohemia, Plzen, Czech Republic, 2008: 1-7.

[27] M. C. Vlaardingen. Hand Models and Systems for Hand Detection, Shape Recognition and Pose Estimation in Video; 2006.

[28] D.H. Cooper, T.F. Cootes, C.J. Taylor, J. Graham. Active shape models their training and application. Computer Vision and Image Understanding 1995;61:38–59.

[29] V. V. Vassili, V.Sazonov, A. Andreeva. A Survey on Pixel-Based Skin Color Detection Techniques. Proc Graphicon, 2003: 85-92.

[30] J. H. Shin, J.S. Lee, S. K. Kil, D. F. Shen, J. G. Ryu, E. H. Lee, et al. Hand Region Extraction and Gesture Recognition using entropy analysis. IJCSNS International Journal of Computer Science and 216 Network Security 2006;6(2A).

[31] P. Peer, J. Kovac, F. Solina. Human skin colour clustering for face detection. International Conference on Computer as a Tool - EUROCON 2003.

[32] S. L. Phung, A. Bouzerdoum, D. Chai. A novel skin color model in YCbCr Space and its application to human face detection. In Proc of ICIP, 2002.

[33] J. Y. Lee, S. I. Yoo. An elliptical boundary model for skin color detection. In Proc of the International Conference on Imaging Science, Systems, and Technology, 2002.

[34] D. Gokalp. Learning Skin Pixels in Color Images Using Gaussian Mixture. Workshop in Computer Science. Bilkent University, 2005.

[35] P. Gejgus, J. Placek, M. Sperka. Skin color segmentation method based on mixture of Gaussians and its application in Learning System for Finger Alphabet. International Conference on Computer Systems and Technologies - CompSysTech, 2004.

[36] X. Zhu, J. Yang, A. Waibel. Segmenting Hands of Arbitrary Color. Proc in Face and Gesture recognition Conference, 2000.

[37] A. C. Downton, H. Drouet. Image Analysis for Model-Based Sign Language Coding. Progress in Image Analysis and Processing II: Proc Sixth Int'l Conf Image Analysis and Processing, 1991: pp. 637-44.

[38] N. Magnenat-Thalmann, D. Thalman, editors. Computer Animation:Theory and Practice: New York: Springer-Verlag, 1990.

[39] J. M. Rehg, T. Kanade. DigitEyes: vision-based hand tracking for human-computer interaction. Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, TX , USA 1994: 16 – 22.

[40] A.J. Heap, D.C. Hogg. Towards 3-D hand tracking using a deformable model. In 2nd International Face and Gesture Recognition Conference. Killington, USA, 1996: 140 –5.

[41] B. Stenger, P. R. S. Mendonca, R. Cipolla. Model-Based 3D Tracking of an Articulated Hand. In proc British Machine Vision Conference. Manchester, UK, 2001: 63-72,.

[42] L. Brethes, P. Menezes, F. Lerasle, J. Hayet. Face tracking and hand gesture recognition for human-robot interaction. Proceedings ICRA '04: 1901 - 6.

[43] H. Tran. Etude de lignes d'interet naturelles pour la representation d'objets en vision par ordinateur: Institut National Polytechnique de Grenoble; 2006.

[44] M. J. Black, A. D. Jepson. Eigen tracking: Robust matching and tracking of articulated objects using a view-based representation. International Journal of Computer Vision 1998:329-42.

[45] W. T. Freeman, M. Roth. Orientation Histograms for Hand Gesture Recognition. IEEE In International Workshop on Automatic Face and Gesture Recognition; 1994; Zurich; 1994. p. 296-301.

[46] G. Rigoll, A. Kosmala, S. Eickeler. High Performance Real-Time Gesture Recognition using Hidden Markov Models. In International Gesture Workshop Bielefeld; 1998; Bielefeld, Germany: Springer-Verlag; 1998. p. 69–80.

[47] T. Tuytelaars. Dense interest points. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2010; San Francisco, CA 2010. p. 2281 - 8

[48] R. Lienhart, J. Maydt. An extended set of Haar-like features for rapid object detection. IEEE Int Conf Image Processing, 2002: 900-3.

[49] A. L. C. Barczak, F. Dadgostar. Real-time hand tracking using a set of co-operative classifiers based on Haar-like features. Res Lett Inf Math Sci 2005;7:29-42.

[50] Q. Chen, N. D. Georganas, E. M. Petriu. Hand Gesture Recognition Using Haar-Like Features and a Stochastic Context-Free Grammar. IEEE Transaction on Instrumentation and Measurement 2008;57(8):1562-71.

[51] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004;60(2):91-110.

[52] J. Crowley. A Representation for Visual Information: Ph.D Thesis, CMU-RI-TR-82-07, Robotics Institute, Carnegie Mellon University; 1981.

[53] T. Linderberg. Automatic scale selection as a pre-processing stage for interpreting the visual world. Proc Fundamental Structural Properties in Image and Pattern Analysis FSPIPA'99. Budapest, Hungary, 1999: 9-23.

[54] L. Bretzner, I. Laptev, T. Lindeberg. Hand gesture recognition using multiscale color features, hieracrchichal models and particle filtering. In Proceedings of Int Conf on Automatic face and Gesture recognition. Washington D.C., 2002: 63-74.

[55] P. Dreuw, T. Deselaers, D. Keysers, H. Ney. Modeling Image Variability in Appearance-Based Gesture Recognition. ECCV Workshop on Statistical Methods in Multi-Image and Video Processing (ECCV-SMVP). Graz, Austria, 2006: 7-18.

[56] H. Murase, S. K. Nayarb. Detection of 3D objects in cluttered scenes using hierarchical eigenspace Pattern Recognition Letters 1997;18(4):375-84

[57] http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/gestures/main.html.

[58] D. B. Nguyen, T. Ejima. A Fuzzy Neural Network for Gesture Recognition. ICGST International Journal on Graphics, Vision and Image Processing 2006;6:23-9.

[59] M. Kolsch, M. Turk. Robust Hand Detection. International Conference on Automatic Face and Gesture Recognition. Seoul, Korea, 2004: 614-9.

[60] http://www.massey.ac.nz/fdadgost/xview.php?page=farhad.

[61] Z. Cernekova, N. Nikolaidis, I. Pitas. Single Camer pointing gesture recognition using spatial features anc support vector machines. European Signal Processin Conference (EUSIPCO-2007). Poznan, Poland, 2007 130-4.

[62] Q. Chen, A. El-Sawah, C. Joslin, N. D. Georganas. A dynamic gesture interface for virtual environments basec on hidden Markov models. IEEE International Worksho on Haptic Audio Visual Environments and thei Applications, 2005.

[63] D. B. Nguyen, E. Shuchi, T. Ejima. Real-Time Hanc Tracking and Gesture Recognition System. Proceedings of International Conference on Graphics, Vision and Image Processing (GVIP-05), 2005: 362-8.

[64] M. Elmezain, A. Al-Hamadi, B. Michaelis. Hand Gesture Recognition Based on Combined Features Extraction. World Academy of Science, Engineering and Technology 2009;60.

[65] A. Corradini. Real-Time Gesture Recognition by Mean of Hybrid Recognizers. GW '01 Revised Papers from the International Gesture Workshop on Gesture and Sigr Languages in Human-Computer Interaction. London, UK 2001: 34-46

[66] S. B. Wang, A. Quattoni, L. P. Morency, D. Demirdjian T. Darrell. Hidden Conditional Random Fields for Gesture Recognition. Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006: pp.1521-.

[67] P. Hong, M. Turk, Huang TS. Gesture Modeling anc Recognition Using Finite State Machines. In Proc of IEEE Conference on Face and Gesture Recognition, 2000.

AUTHORS' BIOGRAPHIES

Tran Thi Thanh Hai graduated in Information Technology from Hanoi University of Science and Technology in 2001. She got her MS degree in Imagery Vision and Robotic at Grenoble Institute of Technology in 2002. She received her PhD degree from Grenoble Institute of Technology, France in 2006. She is currently lecturer/researcher at Compute Vision group, International centre MICA, Hanoi Universit of Science and Technology. Her main research interests are visual object recognition, video understanding, and human-robot interaction. E-mail: thanh-hai.tran@mica.edu.vn