# Dynamic Gesture Recognition Using PCA with Multi-scale Theory and HMM

Hai Wu
Alistair Sutherland
School of Computer Applications
Dublin City University
Dublin 9, Ireland
wuhai@compapp.dcu.ie

## ABSTRACT

In this paper, a dynamic gesture recognition system is presented which requires no special hardware other than a Webcam. The system is based on a novel method combining Principal Component Analysis (PCA) with hierarchical multi-scale theory and Discrete Hidden Markov Models (DHMM). We use a hierarchical decision tree based on multi-scale theory. Firstly we convolve all members of the training data with a Gaussian kernel, which blurs differences between images and reduces their separation in feature space. This reduces the number of eigenvectors needed to describe the data. A principal component space is computed from the convolved data. We divide the data in this space into two clusters using the k-means algorithm. Then the level of blurring is reduced and PCA is applied to each of the clusters separately. A new principal component space is formed from each cluster. Each of these spaces is then divided into two and the process is repeated. We thus produce a binary tree of principal component spaces where each level of the tree represents a different degree of blurring. The search time is then proportional to the depth of the tree, which makes it possible to search hundreds of gestures in real time. The output of the decision tree is then input into DHMM to recognise temporal information.

**Keyword:** PCA, Multi-scale theory, Decision tree, DHMM

## 1. Introduction

In the past decade, the computational power of computers has doubled every eighteen months, while the human computer interface hasn't changed too much. When we work with a computer, we are constrained by intermediary devices: keyboards and mice. This has become a bottleneck in human-computer interaction, particularly in virtual reality. Gesture is a natural means for human beings to communicate with each other. In daily life, people use gestures to point, emphasise and navigate. To make the computer understand this, however, is not an easy job.

A lot of effort has been put into hand gesture recognition[1]. Some of the most popular methods are vision-based systems[1][2]. These take advantage of one or more cameras and allow users to implement gestures freely and naturally. Many stereomatching systems based on multiple cameras have been constructed, such as Akira Utsumi[10] and Hitoshi Hongo[11]. To achieve high accuracy, many researchers have employed a 3D-hand model. Tony Heap[13] presented a 3D deformable model of the human hand using a Simplex Mesh. Ying Wu and Thomas S. Huang proposed a kinematical hand model to handle articulated hand motion[14]. Both systems use a genetic algorithm to estimate model parameters. Another kind of system is "appearance-based". This type of system uses only one single camera and constructs a set of templates from training data beforehand. The real-time frame is then compared to the predefined templates and is classified according to the best fit. One example is Hermann Hienz [16], who introduced a single camera system to track the human hand and arm in real-time. Starner and Alex Pentland[15] also presented a HMM-based system using a single camera to recognise 40 different dynamic gestures in American Sign Language(ASL). Apart from the above, one thing needing to be noticed is that human gestures involve not only hand posture and motion but also facial expression and body movement. Gong at Queen Mary and Westfield College has done a lot of research on face recognition and body tracking. Lots of his ideas are very useful for hand gesture recognition as well [17]. In this paper, however, we consider only hands at the moment. We will look at the face and body in future research.


Image Extraction, Segmentation, and Recognition, Tianxu Zhang, Bir Bhanu, Ning Shu, Editors, Proceedings of SPIE Vol. 4550 (2001) © 2001 SPIE · 0277-786X/01/$15.00

Our interest is in developing a real-time hand gesture recognition system with a single camera that is able to run on a normal desktop computer. A PCA-based system has been developed. The key difference of our system from the others is that we adopt a hierarchical architecture to speed up the search process in real-time. Search is particularly time-consuming especially when dealing with large vocabulary. This makes our system very suitable for real-time purposes even on a cheap machine. And it is also easy to expand to a large vocabulary without large reduction in the performance. Principal Component Analysis (PCA) has been widely used in gesture recognition systems and has achieved remarkable success. Alex Pentland [12] has successfully used it for face recognition. PCA constructs a low dimensional space automatically. This PC space can represent most of the information in the gestures, thus significant dimensionality reduction is achieved.

This paper concentrates on the problem of hand-shape recognition from a single frame and dynamic gesture recognition from image sequences. It is divided into five parts: In the second section, we introduce the theoretical basis of our system, including PCA subspace, multi-scale theory and Hidden Markov Models. In the third part we explain the system in detail. In the fourth section, we show experimental results. Finally, we give some possible improvements and discuss the potential of this method and the perspective.

## 2 Theory

### 2.1 PCA Subspace

Given a set of N training images $f=\{f_1, f_2, f_3, .....f_N\}$, where each image has M pixels. We can use the PCA algorithm to construct a space of dimension much lower than M [5][8][9]. In this PC space, every image is projected onto a single point. The similarity between two images is equivalent to their Euclidean distance in the PC space. Different gestures form different clusters. A new image is categorised by finding the shortest distance between its projection point and the clusters. A problem arises, however, as the number of gestures increases, i.e. the number of clusters gets larger. In such a case, not only the dimensionality of the feature space becomes higher, but also the recognition error increases quickly. A solution is to build a PC space for every possible gesture. The distances from new images to all PC spaces are computed and the final result is the one with the shortest distance. This method improves recognition rate but it is unrealistic to search the whole vocabulary exhaustively. To overcome this problem, we employ a hierarchical binary search based on multi-scale theory.

### 2.2 Multi-scale theory

Given an image $I$, if we convolve it with a Gaussian kernel, a smoother version of it is obtained. Varying the blurring factor $\sigma$ of the Gaussian kernel, the image $I$ is then represented by a family of smoother versions: $I(\sigma)$. $\sigma=0$ corresponds to the original image and as $\sigma$ increases, more and more details of the image are eroded and no spurious structures will be created. Formally, this can be expressed by the following equation:

$$I(\sigma) = I * G(\sigma)$$

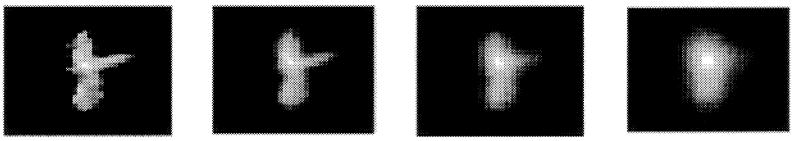

Figure 1. Images under different blurring factors: the leftmost is the original image, i.e. $\sigma=0$ . From left to right, the blurring factor $\sigma$ is increasing.

Where $G$ is the Gaussian kernel, $\sigma$ is the blurring factor or scale parameter and * stands for convolution. Figure 1 gives an example.

Given a set of gestures, if we blur them at different levels, different details will appear so that the same training set can be divided into different groups. To utilise this in practice, firstly we convolve all members of the training data with

a Gaussian kernel that blurs differences between images and reduces their separation in feature space. This reduces the number of eigenvectors needed to describe the data. A Principal Component (PC) space is computed from the convolved data. We divide the data in this space into two clusters using the k-means algorithm. Then the level of blurring is reduced and PCA is applied to each of the clusters separately. A new PC space is formed from each cluster. Each of these spaces is then divided into two and the process is repeated. We thus produce a binary tree of PC spaces where each level of the tree represents a different degree of blurring. The search time is then proportional to the depth of the tree. This makes it possible to search hundreds of gestures in real time.

### 2.3 Discrete Hidden Markov Model (DHMM)

We assume the readers have a basic knowledge about Hidden Markov Models (HMM) so that we only give a very brief introduction here, and because we adopt DHMM in the system, the next subsection is focused on the concepts of DHMM. People with great interest about it are referred to [19][20]. HMM is a doubly stochastic process for producing a sequence of observed symbols. This means two stochastic processes are underlying simultaneously, one is not observable (hidden), while the other can produce a sequence of observations, hence is observable. Discrete, continuous or semi-continuous HMM is used depending on different applications. A widely used notation of DHMM is

$$\lambda = (A, B, \pi)$$

This is a compact format. In fact, it contains five parameters: $N$, $M$, $A$, $B$ *and* $\pi$. The meaning of each parameter is listed below:

$N$ : The number of states in the model;
$M$: The number of distinct observation symbols;
$A$: The state-transition probability distribution matrix;
$B$ : The observation symbol probability distribution matrix;
$\pi$: The initial state distribution.

In gesture recognition systems, a temporal sequence of motion can be represented as the evolution from one frame to another, hence it can be modelled as transitions between states in HMM. There are different topologies available. While in gesture recognition systems, the most frequently used is left-right HMM. A first-order left-right DHMM with five states is used in our system, as illustrated in Fig. 2:

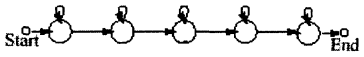

Fig.2: The five state HMM used.

## 3. Outline of the system

The whole system is composed of two parts: off-line and online part. Off-line part is used to acquire training data, calculate Principal Components (PC), construct binary decision tree and train the DHMM recogniser. On-line part is responsible for real-time recognition. The diagram is shown in Fig. 3.

### 3.1 Hand Segmentation and Normalization

The video camera is set up in front of the user. To extract the hand from the rest of the image the user wears a coloured glove whose colour is not likely to appear in the background of the streamed video images. A standard colour segmentation method is applied. Because the area of the hand within the image varies greatly when the hand moves with respect to the camera, we scale the extracted hand by area to a 32x32 greylevel image centred on the centroid and normalize it by energy.

### 3.2 Decision Tree Construction

It is well known that PCA can reduce the dimensionality of data by constructing a low-dimensional feature space

from the top few eigenvectors of the covariance matrix of the training data. Each hand shape corresponds to a single point in this space. Given an unknown gesture, one can recognise it by finding the nearest neighbourhood in this feature space. However, as the number of gestures in the vocabulary increases, the number of eigenvectors needed increases as well. Furthermore, it becomes impossible to search all gestures templates exhaustively.

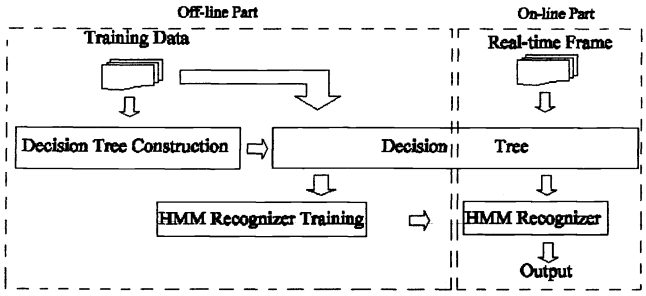


Fig. 3: Diagram of our dynamic gesture recognition system.

To overcome these problems we use a hierarchical decision tree based on multi-scale theory. Firstly we convolve all members of the training data with a Gaussian kernel which blurs difference between images and reduces their separation in feature space. This reduces the number of eigenvectors needed to describe the data. We divide the data in this space into two clusters using K-means algorithm. Then the level of blurring is reduced and PCA is applied to each of the clusters separately. A new feature space is formed from each cluster. Each of these spaces is then divided into two and the process is repeated. We thus produce a binary decision tree of principal component spaces where each level of the tree represents a different degree of blurring. The search time overall vocabulary is thus only proportional to the depth of the tree which makes it possible to search many more gestures than linear search in real time. The flowchart in Figure 4 explains the procedure of Decision Tree construction.

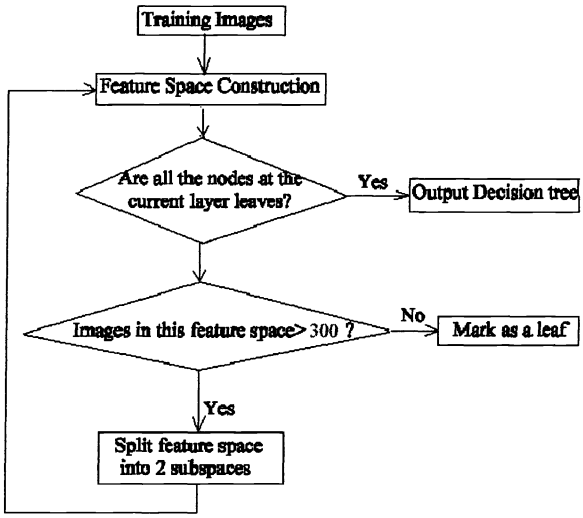


Fig. 4: Flowchart of Decision Tree Construction.

### 3.3 DHMM Recognizer Training

After the construction of the decision tree, we can start to train the DHMM recognizer. Firstly the same set of training data are fed into the decision tree, the output is a set of 2 dimensional vector:

$$I = \{\{I_{11}, I_{12}\}, \{I_{21}, I_{22}\}, \ldots\{I_{i1}, I_{i2}\}, \ldots\{I_{N1}, I_{N2}\}\}$$

Where $I_{i1}$ is the depth number of $i$th training sample after the decision tree, and $I_{i2}$ is the leaf number of $i$th training sample at this layer; $N$ represents the size of the training set. $I$ is then used as inputs to train the DHMM network. This is done by Cambridge's Hidden Markov Model Toolkit (HTK). To avoid the boot strapping segmentation, that is, the manually labelling procedure, we use the global probabilities to initialize the original models, then embeded training is employed to optimize the parameters. The reason why we choose discrete model rather than continuous is because of the way that we construct the decision tree. Since similarity is the standard by which we split the data clusters, the variance inside the clusters will become progressively smaller. As it reaches the leaf level, the variances inside one leaf node are so small, typically much less than 0.1 on every dimension, that it becomes more like a point instead of a cloud of points and will make Gaussian mixture model meaningless and even invalid, i.e. it will cause the likelihood of a Gaussian larger than 1! On the other hand, DHMM has some other advantages, such as the training procedure is much easier than continuous model and it runs much faster in real-time. For either reason, DHMM is more appropriate than continuous model in our system for the time being.

## 4. Experimental Results

Although the system has the potential on very large vocabulary, in the current paper, we only show the recognition result of 10 dynamic gestures in Irish Sign Language (ISL). ISL is widely used in Ireland by deaf people and contains thousands of gestures. It is well-defined so it is suitable for evaluation of system performance. Ten dynamic gestures we used to test our system performance are: right, desire, elect, fail, doubt, TV, cruel, x, no and j. The details of these gestures can be found in [18]. This experiment was based on DELL OptiPlex Gx1 P2 350MHz and Creative Webcam 3. No other special hardware was used. The training images were grabbed under normal office illumination conditions. First, for each of the ten gestures, we acquired 60 samples, i.e., 600 in total. Using these samples, a binary decision tree was constructed with the above procedure. The selection of blurring factors were all determined by trail and error and decreased in logarithm order at different layers of the tree. The equation is given below [4]:

$$\sigma = \varepsilon * \exp(k / c)$$

Where $\sigma$ is the blurring factor, $\varepsilon$ and $c$ are constants. Leung et al [4] found the following values to be the best $c$=1/log (1.05) and $\varepsilon$=0.01. The reason why the blurring factor has to be adjusted logarithmically is that this reflects the changing structure in the images. Readers who are interested in this are referred to [3][4].

Once the decision tree was constructed, the same set of training samples were fed into the decision tree and its output was then input into HTK to train HMM recognizer. As stated above, discrete HMM model was adopted. After this, the whole system can be used to recognise real-time sequences.

The decision tree will analyse every real-time frame acquired by the camera and eventually classify it into one of the leaves. Figure below shows the classification procedure of a frame in the decision tree. The coarse-to-fine procedure of the decision tree can be clearly seen. Firstly, the image is blurred very much so that can only be classified roughly, then as the blurring factor decreases, more and more details appear and eventually the true class is found out which leaf it belongs to.

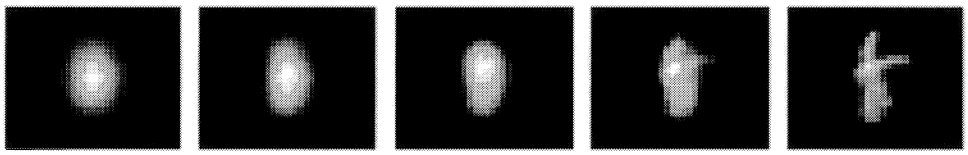

Figure 5: The classification procedure of a real time frame "F". From left to right corresponds to the top-down layers in the classification.

The depth and leaf numbers were sent over to DHMM recognizer. In theory, the DHMM recognizer should be able to run in real-time. However, we are taking advantage of HTK, which is a command-line based application. We haven't

connected it with our vision system yet. So the evaluation of DHMM recognizer in this experiment was done off-line. We will bring it online in future research. At the moment, all the outputs from the decision tree are buffered in a file, then sent to DHMM to get the final result.

For a fair test, we grabbed another similar group of samples, i.e. 60 for each of the ten gestures, which were never used for any portion of the training. The recognition rate and some other results are all shown in Table 1. No grammar was used. There are three types of error: substitution (*S*), deletion (*D*) and insertion (*I*). For example, for a sentence " I am a student", the substitution error is " I am a teacher", the deletion error is " I a student", and the insertion error is " I am a student student". The absolute values of these three types of error are shown in Table 1 as well. Basically we should treat those errors in different ways. Substitution errors and deletion errors are more severe than insertion errors because insertion error can be easily got rid of by grammar constraints, while there is no obvious way to improve the other two types of error. In this case, the recognition rate is also worth to be divided into two types: Correct percentage (Corr) and Accuracy percentage (Acc):

$$Corr = \frac{N - D - S}{N} * 100\%$$

$$Acc = \frac{N - D - S - I}{N} * 100\%$$

Where $N$ is the total number of each test set, here 60. Correct percentage excludes the insertion error and reflects the potential recognition rate of the system with grammar constraints, while Accuracy percentage includes insertion error and is a more strict evaluation standard.

Table 1. Recognition Results of ten dynamic gestures in ISL using Hierarchical PCA and DHMM

| Gestures | Test on Independent Samples | | | | | |
|---|---|---|---|---|---|---|
| | Corr(%) | Acc(%) | *D* | *S* | *I* | *N* |
| Right | 100.00 | 98.33 | 0 | 0 | 1 | 60 |
| Desire | 88.33 | 88.33 | 4 | 3 | 0 | 60 |
| Elect | 100.00 | 96.67 | 0 | 0 | 2 | 60 |
| Fail | 100.00 | 96.67 | 0 | 0 | 2 | 60 |
| Doubt | 100.00 | 100.00 | 0 | 0 | 0 | 60 |
| TV | 85.00 | 81.67 | 0 | 9 | 2 | 60 |
| Cruel | 100.00 | 98.33 | 0 | 0 | 1 | 60 |
| X | 100.00 | 100.00 | 0 | 0 | 0 | 60 |
| Born | 100.00 | 100.00 | 0 | 0 | 0 | 60 |
| J | 100.00 | 98.33 | 0 | 0 | 1 | 60 |

We notice the recognition rates are high in terms of both Correct percentage and Accuracy percentage. Although partly it is because of the small vocabulary, this preliminary result does show our approach's potential.

Since we use a discrete HMM with very small input dimension, the speed of DHMM recognizer is really fast. When evaluating, it spent less than 2 seconds on 5000 inputs. Under this situation, most of the processing time was spent on the decision tree classification. The speed of this part varies from 3 to 4 fps. However, it would not slow down too much even we expand our vocabulary to several hundreds, and this is the most important reason to stimulate us in the first place.

## 5. Discussion and perspective

In this paper, we presented a novel appearance-based system that is able to recognize dynamic gestures using PCA, decision tree and DHMM. It runs fast even on a cheap machine without the help of any other special hardware except a single camera. Although it is not a real-time system at the moment because of the DHMM recognizer, we are expecting to use it in real-time in very near future. The key difference between our system and other appearance-based systems is that we employed a hierarchical search decision tree. On one hand, the decision tree reduces the search time significantly given an unknown frame: from O(n) to $O(\log_2 n)$. As the size of vocabulary gets larger, the reduction gets

more significant; On the other hand, it also reduces the dimension of input feature vector of DHMM recognizer significantly, thus speeds up DHMM processing. This doubles the time-saving effect. The construction of bigger vocabulary is now in progress. No obvious problem lies ahead as far as we can see.

However, some improvements will possibly enhance the performance. First of all, the decision tree only extracts the local shape information of hands due to the scaling, while no global information is available after it. Encoding global movements into the input feature vector of DHMM recognizer will help it to "see" the hand motion better and should improve the recognition rate of some special dynamics gestures that are described by the traces of a fingertip without great shape variance during the motion. Secondly, many dynamic gestures in sign languages involve not only the motion of hands but also facial expression and body movement. Extending our system to handle these information will make it more useful in many situations.

## References:


1. Vladimir I. Pavlovic, Rajeev Sharma, Thomas S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 19, No. 7, July 1997

2. Ying Wu, Thomas S. Huang, "View-Independent Recognition of Hand Postures". *Proc. of IEEE Conf. On CVPR'2000, Vol.2, pp.88-94, Hilton Head Island, SC,2000*

3. Lindeberg Tony, *Scale Space Theory in Computer Vision.* Kluwer Academic Publishers,1994

4. Yee L., Jiang-She Zhang, Zong-Ben Xu, "Clustering by Scale-Space Filtering". *IEEE Trans. on PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* Vol.22, No.12, DECEMBER 2000, pp.1396-1410

5. K.I.Diamantaras, S.Y.Kung, *Principal Component Neural Networks: Theory and Applications.* John Wiley & Sons,Inc. 1996

6. Wey-Shiuan Hwang, Juyang Wen, "Design and Evaluation of Classifiers: Hierarchical Discriminant Regression" *IEEE Trans. on PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* Vol.22,No.11,November 2000, pp.1277-1293

7. Aleix Martinez, "Face Image Retrieval Using HMMs". *Proc. of IEEE Workshop on Content-Based Access of Images and Video Libraries,* 1999

8. Kailash Jha, B.Gurumoorthy, "Multiple Feature Interpretation Across Domains". *Computers in Industry 42(2000)* pp.13-32

9. Vincent Colin de Verdiere, James L.Crowley, "Local Appearance Space for Recognition of Navigation Landmarks". *Robotics and Autonomous Systems 31 (2000)* pp.61-69

10. Akira Utsumi, Jun Ohya, "Direct Manipulation Interface using Multiple Cameras for Hand Gesture Recognition". *IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA COMPUTING AND SYSTEMS,* 1998

11. Hitoshi Hongo *et al,* "Focus of Attention for Face and Hand Gesture Recognition Using Multiple Cameras". *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition* 2000

12. Turk, M., and Pentland, A., "Eigenfaces for Recognition". *Journal of Cognitive Neuroscience 3(1):* 71–86. 1991

13. Tony Heap, David Hogg, "3D Deformable Hand Models". *Gesture Workshop 1996:* 131-139

14. Ying Wu, Thomas S. Huang, "Capturing Articulated Human Hand Motion: A Divide-and-Conquer Approach." *Proceedings of International Conference on Computer Vision,* 1999

15. Thad Starner, Alex Pentland, "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models". *IEEE International Symposium on Computer Vision,* 1995.

16. Hermann Hienz, Kirsti Grobel, and Georg Offner, "Real-Time Hand-Arm Motion Analysis using a single Video Camera". *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*

17. Gong's research website: http://www.dcs.qmw.ac.uk/research/vision/

18. Stanislaus J.Foran, *The Irish Sign Language*, Revised Edition, Elo Press Ltd. 1996

19. Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition,* Prentice Hall, 1996

20. Claudio Becchetti, Lucio Prina Ricotti, *Speech Recognition, Theory and C++ Implementation,* John Wiley & Sons, LTD