In Proc. of IEEE Conf. on CVPR'2000, Vol.II, pp.88-94, Hilton Head Island, SC, 2000

View-independent Recognition of Hand Postures

Ying Wu, Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{yingwu, huang}@ifp.uiuc.edu

Abstract

Since human hand is highly articulated and deformable, hand posture recognition is a challenging example in the research of view-independent object recognition. Due to the difficulties of the model-based approach, the appearance-based learning approach is promising to handle large variation in visual inputs. However, the generalization of many proposed supervised learning methods to this problem often suffers from the insufficiency of labeled training data. This paper describes an approach to alleviate this difficulty by adding a large unlabeled training set. Combining supervised and unsupervised learning paradigms, a novel and powerful learning approach, the Discriminant-EM (D-EM) algorithm, is proposed in this paper to handle the case of small labeled training set. Experiments show that D-EM outperforms many other learning methods. Based on this approach, we implement a gesture interface to recognize a set of predefined gesture commands, and it is also extended to hand detection. This algorithm can also apply to other object recognition tasks.

1 Introduction

In current VE applications, keyboards, mice, wands and joysticks are still the most popular devices. However, they are inconvenient and unnatural. In recent years, the use of human movements, especially hand gestures, serves as a motivating force for research in gesture modeling, analyzing and recognition.

Although hand gestures are complicated to model since the meanings of hand gestures depend on people and cultures, a set of specific hand gesture vocabulary can be always predefined in many applications, such as Virtual Environment (VE) applications, so that the ambiguity can be limited. Generally, these hand gestures can be either static hand postures or temporal hand gestures. Hand postures express some concepts by hand configurations and hand shapes, while temporal hand gestures represent some actions by hand movements. Sometimes, hand postures act as special transition states in temporal gestures, and supply a cue to segment and recognize temporal hand gestures. Some research results show that static hand signs and temporal hand gestures seldom present simultaneously, which suggests us study static hand gestures and temporal gestures separately.

Different from sign languages, the gesture vocabulary in VE applications is structured and disambiguated. In such scenarios, some simple controlling, commanding and manipulative gestures are defined to fulfill natural interaction such as pointing, navigating, moving, rotating, stopping, starting, selecting, etc. These gesture commands can be simple in the sense of motion; however, many different hand postures are used to differentiate and switch among those commanding modes. For example, only if we know a gesture is a pointing gesture, it makes sense to estimate its pointing direction. This problem is an empirical problem in most VE applications.

Although this problem can be formulated as a classification problem of different predefined static hand postures, there are still many difficulties. The first is view-independent hand posture recognition, which means hand postures must be recognized from any view direction. This is a natural requirement in many VE applications. In most cases, since users do not know where the cameras are, the naturalness and immersiveness will be ruined if users are obliged to issue commands to an unknown direction. Another difficulty is that human hand is highly articulated and deformable, the large variation in hand postures should be handled to make a user-independent system.

Since hand postures can express some concepts as well as act as special transition states in temporal gestures, recognizing or estimating hand postures or human postures is one of the main topics in gestures recognition. Some work has been done in this area.

One approach is the 3D model-based approach, in which the hand configuration is estimated by taking advantage of 3D hand models [7, 8, 10, 11, 13, 15, 17, 21]. Since hand configurations are independent to view directions, these methods could directly achieve view-independent recognition. Different models take different image features to construct feature-model correspondences. Joint angles can be estimated by minimizing a projected surface model and some image evidences such as silhouettes in the light of "analysis-by-synthesis" [10, 11, 7]. However, this approach needs good surface models and the process of projection-and-comparison is expensive. Alternatively, point and line features are employed in kinematical hand models to recover joint angles [15, 17, 21]. Hand postures could be estimated accurately if the correspondences between the 3D model and the observed image features are well established. Physical models and statistical models [8] were also employed to estimate hand configurations. However, the ill-posed problem of estimating hand configuration is not trivial. Many current methods require reliable feature detection which is plagued by self-occlusion. Another drawback is that it is not trivial to achieve user-independence, since 3D models should be calibrated for each user; otherwise the accuracy will be scarified.

Although accurate estimation of hand configuration is important in some applications such as manipulating virtual objects or multi-DOF input devices, a classification of hand postures is often enough in many other applications such as commands switching. Since the appearances are much different among different hand postures and these differences are not large among different people, an alternative approach is appearance-based approach [5, 6, 14, 19], in which classifiers are learned for a set of image samples. Although it is easier for the appearance-based approach to achieve user-independence than model-based approach, there are two major difficulties of this approach: automatic feature selection and training data collection. Although there have been many discussions about feature extraction [19, 14, 13] and selection [5, 6], little has been addressed on the training data. The generalization of many current methods have to largely depend on their training data sets. In general, good generalization requires a large and representative labeled training data set. However, to manually label a large data set will be very time-consuming and tedious. Although unsupervised schemes has been proposed to clustering the appearances of 3D objects[1], it is hard for pure unsupervised approach to achieve accurate classification without supervision.

In this paper, we take an appearance-based approach and try to investigate this *training data problem*. As we observed, although it is expensive to manually label sample images, it is not difficult to collect a large set of unlabeled hand images, which motivates us to train a good classifier by a small set of labeled data but with a large unlabeled data set. We propose a novel and powerful learning approach, the Discriminant-EM (D-EM) algorithm, to hybrid supervised and unsupervised learning paradigms. Experiments show that D-EM outperforms many other learning methods. This algorithm can also be applied to other object recognition tasks.

A formulation of the problem is given in section 2. Section 3 and section 4 describes our proposed approach and the D-EM algorithm. Some of our experiment results can be found in section 5, and section 6 concludes the paper and give some future work.

2 An Inductive Problem

View-independent hand posture recognition is to identify a posture in any view direction. Some hand posture images are shown in Figure 1. Each row should be classified into the same posture class.

Traditionally, feature extraction and selection are independent to the designation of classifier. Since the raw image space is huge, some physical features, which can be extracted by some image processing techniques, can be employed as a compact representation of an object. Even though selecting such physical features needs quite a lot domain knowledge and experiences, this step in many applications is *ad hoc* and under large risk. For example, some statistics, such as the area of the bounding box, the compactness of the hand and some moments of the edge map, could be used to represent a hand posture. However, totally different hand posture might share the same set of such statistics. To avoid this problem, many researchers employ mathematical features. Although they may not have substantial physical meanings, mathematical features can preserve most of the image information of the object. The Principal Component Analysis (PCA) technique offers such a way by persevering some largest "energy" components. Generally, statistical methods to extract mathematical features need a large training data set.

Although either physical or mathematical features can largely reduce the raw image space and serve as compact representations of an object, different features play different roles in object recognition and they should be weighted differently. Some heuristics can be used to weight physical features. How-

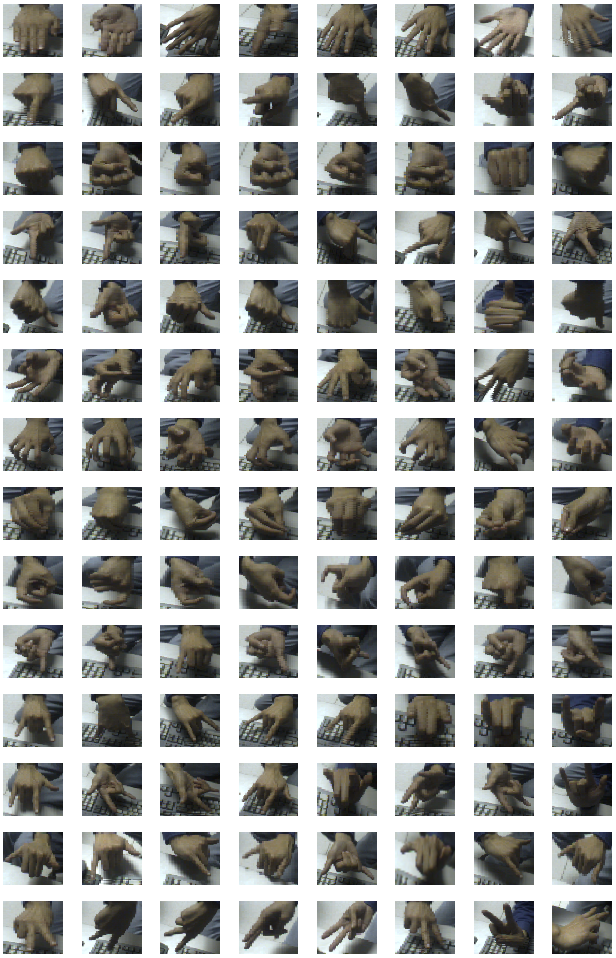


Figure 1: Posture data

ever, such heuristics are domain-dependent and are not always plausible. As of mathematical features, it is even harder to find such heuristics, since physical meanings for mathematical features are not available. Fortunately, the discriminant analysis technique offers a means to automatically select and weight classification-relevant features. In the area of face and gesture recognition, there have been some successful methods based on it [5, 6, 3]. However, the discriminant analysis technique puts a harsh requirement to the training data set: a large labeled data set. We do not expect discriminant analysis to output a good result, unless enough labeled data are available.

In fact, it seems that it might not be necessary to have every sample labeled in supervised learning. A very interesting result given by the theory of the support vector machine (SVM)[20] is that the classification boundary is related only to some support vectors, rather than the whole data set. Although the identification of these support vectors is not trivial, it motivates us to think about the roles of non-support vectors.

Fortunately, it may be easier to collect a large number of unlabeled data, which may be used to help supervised learning, since unlabeled data contain information about the joint distribution over features. If the probabilistic structure of data distribution is known, parameters of probabilistic models can be estimated by unsupervised learning alone, but it is still impossible to assign class labels without labeled data [2]. This fact suggests that labeled data (if enough) can be used to label the class and unlabeled data can be used to estimate the parameters of generative models.

In such circumstance, the hybrid training data set $\mathcal{D}$ consists of a labeled data set $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$, where $\mathbf{x}_i$ is feature vector, y_i is label and N is the size of the set, and an unlabeled data set $\mathcal{U} = \{\mathbf{x}_i, i = 1, \ldots, M\}$, where M is the size of the set. We make an assumption here that $\mathcal{L}$ and $\mathcal{U}$ are from the same distribution. This assumption is reasonable, because labeled data are selected from the unlabeled data set. We also assume that the unlabeled set $\mathcal{U}$ is much larger than the labeled set $\mathcal{L}$. Essentially, the classification problem can be represented as:

$$y_i = arg \max_{j=1,\ldots,C} p(y_j|\mathbf{x}_i, \mathcal{L}, \mathcal{U} : \forall \mathbf{x}_i \in \Omega) \qquad (1)$$

where Ω is the whole data space and C is the number of classes. The goal of the inductive learning is to learn a classifier which can be generalized to the whole data space Ω by using a small set of labeled date and a large set of unlabeled data.

3 In the EM Framework

The Expectation-Maximization (EM) approach can be applied to this learning problem, since the labels of unlabeled data can be treated as missing values. We employ a generative model which assumes that the hybrid data set is drawn from a mixture density distribution of C components $\{c_j, j = 1, \ldots, C\}$, which are parameterized by $\mathbf{\Theta} = \{\theta_j, j = 1, \ldots, C\}$. The mixture model can be represented as:

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{C} p(\mathbf{x}|c_j; \theta_j) p(c_j|\theta_j) \qquad (2)$$

where $\mathbf{x}$ is a sample drawn from the hybrid data set $\mathcal{D} = \mathcal{L} \bigcup \mathcal{U}$. We make another assumption that each component in the mixture density corresponds to one class, i.e. $\{y_j = c_j, j = 1, \ldots, C\}$.

The parameters $\mathbf{\Theta}$ can be estimated by maximizing *a posteriori* probability $p(\mathbf{\Theta}|\mathcal{D})$. Equivalently, this can be done by maximizing $\lg(p(\mathbf{\Theta}|\mathcal{D}))$. Let $l(\mathbf{\Theta}|\mathcal{D}) = \lg(p(\mathbf{\Theta})p(\mathcal{D}|\mathbf{\Theta}))$. When assuming that each sample is independent to the others, and introducing a binary indicator $\mathbf{z}_i = (z_{i1}, \ldots, z_{iC})$, where $z_{ij} = 1$ iff $y_i = c_j$ and $z_{ij} = 0$ otherwise, we have:

$$l(\mathbf{\Theta}|\mathcal{D}, \mathcal{Z}) = \lg(p(\mathbf{\Theta})) + \sum_{\mathbf{x}_i \in \mathcal{D}} \sum_{j=1}^{C} z_{ij} \lg(p(O_j|\mathbf{\Theta})p(\mathbf{x}_i|O_j; \mathbf{\Theta}))$$

In the EM framework, probability parameters $\mathbf{\Theta}$ can be estimated by an iterative hill climbing procedure, which alternatively calculates $E(\mathcal{Z})$, the expected values of all unlabeled data, and estimates the parameters $\mathbf{\Theta}$ given $E(\mathcal{Z})$. The EM algorithm generally reaches a local maximum of $l(\mathbf{\Theta}|\mathcal{D})$. It consists of two iterative steps:

- E-step: set $\hat{\mathcal{Z}}^{(k+1)} = E[\mathcal{Z}|\mathcal{D}; \hat{\Theta}^{(k)}]$
- M-step: set $\hat{\Theta}^{(k+1)} = arg\ max_{\theta}\, p(\Theta|\mathcal{D}; \hat{\mathcal{Z}}^{(k+1)})$

where $\hat{\mathcal{Z}}^{(k)}$ and $\hat{\Theta}^{(k)}$ denote the estimation for $\mathcal{Z}$ and $\mathbf{\Theta}$ at the k-th iteration respectively.

When the size of the labeled set is small, EM basically performs an unsupervised learning, except that labeled data are used to identify the components. Although the EM algorithm can be applied straightforwardly, one of the difficulties is that the probabilistic structure of data distribution must be determined in advance. When the assumed probabilistic structure in the generative model does not align to the ground truth structure, EM hardly gives a good estimation, which is partly the reason that unlabeled data hurt the classifier. In many cases, the data dimension is high so that the size of training data should be accordingly large, otherwise, the parameter estimation of the generative model will be highly biased.

4 Inductive Learning by D-EM

Since we generally do not know the probabilistic structure of data distribution, EM often fails when structure assumption does not hold. One approach to this problem is to try every possible structure and select the best one. However, it needs more computational resources. An alternative is to find a mapping such that the data are clustered in the mapped data space, in which the probabilistic structure could be simplified and captured by simpler Gaussian mixtures. The Multiple Discriminant Analysis (MDA) technique offers a way to relax the assumption of probabilistic structure, and EM supplies MDA a large labeled data set to select most discriminating features. At the mean time, MDA also reduces the data dimension, which makes the task of statistical estimation easier.

MDA is a natural generalization of Fisher's linear discrimination (LDA) in the case of multiple classes[2]. The basic idea behind MDA is to find a linear transformation $\mathbf{W}$ to map the original d_1 dimensional data space to a new d_2 space such that the ratio of the between-class scatter and the within-class scatter is maximized in some sense. Details can be found in [2]. MDA offers a means to catch major differences between classes and discount factors that are not related to classification. Some features most relevant to classification are automatically selected or combined by the linear mapping $\mathbf{W}$ in MDA, although these features may not have substantial physical meanings any more. Another advantage of MDA is that the data are clustered to some extent in the projected space, which makes it easier to select the structure of Gaussian mixture models.

It is apparent that MDA is a supervised statistical method, which requires enough labeled samples to estimate some statistics such as mean and covariance. By combining MDA with the EM framework, our proposed method, Discriminant-EM algorithm (D-EM), is such a way to combine supervised and unsupervised paradigms. The basic idea of D-EM is to enlarge the labeled data set by identifying some "similar" samples in the unlabeled data set, so that supervised techniques are made possible in such an enlarged labeled set. D-EM employs a generative model in the lower dimensional space mapped by the transformation $\mathbf{W}$ from MDA.

$$p(\mathbf{y}|\mathbf{\Theta}) = \sum_{j=1}^{C} p(\mathbf{y}|c_j; \theta_j)p(c_j|\theta_j) \tag{3}$$

where $\mathbf{y} = \mathbf{W}^T\mathbf{x}$.

D-EM begins with a weak classifier learned from the labeled set. Certainly, we do not expect much from this weak classifier. However, for each unlabeled sample $\mathbf{x}_j$, the classification confidence $\mathbf{w}_j = \{w_{jk}, k = 1, \ldots, C\}$ can be given based on the probabilistic label $\mathbf{l}_j = \{l_{jk}, k = 1, \ldots, C\}$ assigned by this weak classifier.

$$l_{jk} = \frac{p(\mathbf{W}^T\mathbf{x}_j|c_k)p(c_k)}{\sum_{k=1}^{C} p(\mathbf{W}^T\mathbf{x}_j|c_k)p(c_k)} \tag{4}$$

$$w_{jk} = \lg(p(\mathbf{W}^T\mathbf{x}_j|c_k))\ k = 1, \ldots, C \tag{5}$$

Euqation(5) is just a heuristic to weight unlabeled data $\mathbf{x}_j \in \mathcal{U}$, although there may be many other choices. This E-step outputs a probabilistic label and a weight for each unlabeled sample, given a fixed transformation $\mathbf{W}$ and a generative model.

After that, the D-step is to perform MDA on the new weighted data set $\mathcal{D}' = \mathcal{L} \bigcup \{\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$, to find a linear transformation $\mathbf{W}$, by which the data set $\mathcal{D}'$ is linearly projected to a new space of dimension $C - 1$, but unchanging the labels and weights, $\hat{\mathcal{D}} = \{\mathbf{W}^T\mathbf{x}_j, y_j : \forall \mathbf{x}_j \in \mathcal{L}\} \bigcup \{\mathbf{W}^T\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$. Then the M-step estimates parameters $\mathbf{\Theta}$ of the probabilistic models on $\hat{\mathcal{D}}$, so that the probabilistic labels are given by the Bayesian classifier according to Equation(4). The algorithm iterates over these three steps, "Expectation-Discrimination-Maximization". The algorithm can be terminated by several methods such as presetting the iteration times, comparing a threshold and the difference of the parameters between consecutive two iterations, and using cross-validation.

It should be noted that the simplification of probabilistic structures is not guaranteed in MDA. If the components of data distribution are mixed up, it is very unlikely to find such a linear mapping. In this case, nonlinear mapping should be found so that simple probabilistic structure could be used to approximate the data distribution in the mapped data space. Generally, we use Gaussian or 2-order Gaussian mixtures. Our experiments show that D-EM works better than pure EM.

5 Experiments

In this section, we describe the collection of training and testing data set, data preprocessing, extraction of physical and mathematical features, investigation of the effect of using unlabeled data in training, comparison among different classification schemes, and an application to hand detection.

5.1 Data Collection and Setting

As shown in Figure 1, the gesture vocabulary in our gesture interface is 14, each of which represents a gesture command mode, such as navigating, pointing, stopping, grasping, hooking, cutting, etc. A hand localization system [23] has been developed to automatically collect hand images which serve as the unlabeled data, since the localization system only outputs bounding boxes of hand regions, regardless of hand postures. A large unlabeled database can be easily constructed. Currently, there are 14,000 unlabeled hand images in our database. It should be noted that the bounding boxes of some images are not tight, which introduce noise to the training data set. However, including such noise can make the recognition algorithm more robust. For each posture class, some samples are manually labeled. To investigate the effect of using unlabeled data and to compare different classification algorithms, we construct a testing data set, which consists of 560 labeled images.

The step of data preprocessing includes color-based background subtraction [23], histogram equalization and lighting correction [16]. This step largely bypasses the influence of the backgrounds.

Physical and mathematical features are both used as hand representation in our experiments. Taking advantage of the texture and edge information, we extract and normalize 28 physical features, ($d_2 = 28$). Gabor wavelet filters with 3 levels and 4 orientations are used to extract 12 texture features, each of which is the standard deviation of the wavelet coefficients from one filter. 10 coefficients from the Fourier descriptor are used to represent hand shapes. We also use some statistics such as the hand area, contour length, total edge length, density, and 2-order moments of edge distribution. Therefore, we have 28 low-level image features in total.

To extract mathematical features, hand images are resized to 20×20, which gives a raw image space of dimension ($d_1 = 400$). PCA is employed to find a lower-dimensional feature space $\mathcal{R}^{d_2}$. We experiment with different d_2 and a cross-validation approach is taken to find a good d_2.

5.2 The Role of Unlabeled Data

To investigate the effect of the unlabeled data used in D-EM, we feed the algorithm a different number of labeled and unlabeled samples. In this experiment, we use the mathematic features extracted by PCA with 22 principal components, and the dimension for MDA is set to 10. In this experiment, we use 500, 1000, 2500, 5000, 7500, 10000, 12500 unlabeled samples and 42, 56, 84, 112, 140 labeled data, respectively. From Figure 2, in general, combining some unlabeled data reduce the classification error by 20% to 30%. It is not surprising to see this result, because D-EM is able to automatically label some samples by its embedded unsupervised mechanism.

In Figure 3, we study the effect of the dimension parameters in PCA and MDA. If less principal components of PCA are used, some minor but important discriminating features may be neglected so that those principal components may be insufficient to discriminate different classes. On the other hand, if more principal components of PCA are used, it would include

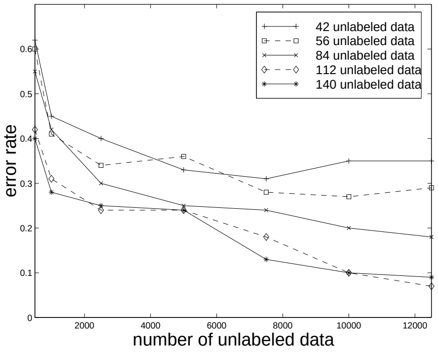


Figure 2: The effect of labeled and unlabeled data in D-EM

more noise. Therefore, the number of principal components of PCA is an important parameter for PCA. The dimension of MDA ranges between 1 to $C-1$, where C is the number of classes. We are interested in a lower dimensional space in which different classes can be classified. In this experiment, we use 112 labeled data and 10000 unlabeled data, and we find that a good dimension parameter of PCA is around 20 to 24, and 8 to 13 for MDA.

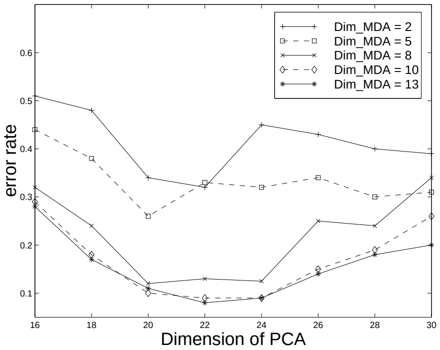


Figure 3: The effect of the dimension of PCA and MDA in D-EM

5.3 Comparison

Four classification algorithms are compared in this experiment. We test both physical (P-Features) and mathematical features (M-Features). For M-Features, the number of principal components of PCA is set to 22, and a set of 560 labeled data is used to perform MDA with dimension of 10.

Using 1000 labeled training data, the multi-layer perceptron used in this experiment has one hidden layer of 25 nodes. We experiment with two schemes of the nearest neighbor classifier. One is just of 140 labeled samples, and the other uses 140 labeled samples to bootstrap the classifier by a growing scheme, in which newly labeled samples will be added to the classifier according to their labels. The labeled and unlabeled data for both EM and D-EM are 140 and 10000, respectively. Table 1 shows the comparison.

Algorithm	P-Features	M-Features
Multi-layer Perceptron	33.3%	39.6%
Nearest Neighbor	30.2%	35.7%
Nearest Neighbor(growing)	15.8%	20.3%
EM	21.4%	20.8%
D-EM	9.2%	7.6%

Table 1: Comparison among different algorithms

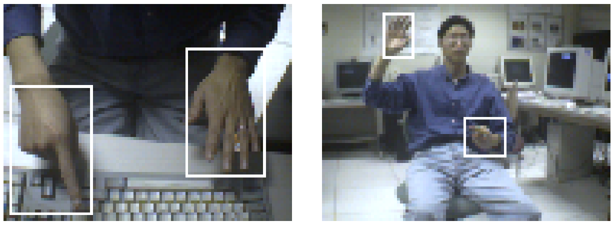
From Table 1, the D-EM algorithm outperforms the other three methods. The multi-layer perceptron is often trapped in local minima in this experiment. The poor performance of the nearest neighbor classifier is partly due to the insufficient labeled data. When the growing scheme is used, it reduces the error by 15%, since it automatically expends the stored templates. The problem of this scheme is that it is affected by the order of inputs, because there is no confidence measurement in growing so that the error of labeling will be accumulated. Pure EM algorithm hardly converges to a satisfactory classification in our experiments. However, D-EM ends up with a pretty good result.

5.4 Hand Detection

Combining with skin color segmentation[23], view-independent posture recognition can be used to detect hands. Since skin color segmentation has already limited the searching range, hand detection can be very efficient. Figure 4 shows two examples, in which the skin color regions from color-based segmentation often contain the arm. Hand detection gives a more accurate bounding box of hand region.

6 Conclusion

View-independent hand posture recognition is important to achieve natural and immersive interaction in many gesture-based virtual environment applications. Although many supervised learning approaches has been proposed to this problem, the generalization of these methods often suffers from the training data set, because collecting a large labeled training set is time-consuming. However, manually labeling all samples is not necessary. In this paper, we propose a novel



(a) (b)

Figure 4: Hand detection

and powerful learning approach, the Discriminant-EM algorithm, which uses an unlabeled data set to help supervised learning to reduce the number of labeled samples. By combining MDA and the EM algorithm, MDA makes EM perform more supervised learning, and EM supplies MDA enough labeled data to perform discriminant analysis. Experiments show that the D-EM outperforms some learning methods such as multi-layer perceptron and nearest neighbor. This algorithm can also be applied to other object recognition tasks.

Since current D-EM uses linear MDA and the simplification of probabilistic structure cannot be guaranteed in some cases, the non-linear case of MDA will be investigated in the future. The convergence and stability analysis of the D-EM algorithm will also be studied. More physical features for hand images will be studied to make a more extensive comparison to mathematical features. The work of hand detection will be extended in our current hand localization system. The applications of the D-EM algorithm to other object recognition tasks are worth pursuing.

7 Acknowledgments

This work was supported in part by National Science Foundation Grant IRI-9634618 and Grant CDA-9624396. The authors would like to appreciate the anonymous reviewers for their comments.

References

[1] R.Basri, D.Roth, D.Jacobs, “Clustering Appearances of 3D Objects”, *IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 1998

[2] R.Duda and P.Hart, “Pattern Classification and Scene Analysis”, New York:Wiley, 1973 (The 2nd Version with D.Stork unpublished)

[3] P. Belhumeur, J. Hespanha, and D. Kriegman, ”Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, *European Conference on Computer Vision*, April 1996

[4] P.Belhumeur, D.Kriegman, “What is the Set of Images of an Object Under All Possible Lighting Conditions?”, *IEEE CVPR'96*, 1996

[5] Y.Cui, D.Swets, J.Weng, “Learning-based Hand Sign Recognition Using SHOSLF-M”, *Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, pp.201-206, 1995

[6] Y. Cui, J. Weng, “Hand Sign Recognition from Intensity Image Sequences with Complex Background”, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.8893. 1996

[7] J.Davis, M.Shah, “Visual Gesture Recognition”, *Vision, Image, and Signal Processing*, 141(2), pp.101-106, 1994

[8] T.Heap, D.Hogg, “Towards 3D Hand Tracking Using a Deformable Model”, *IEEE Int'l Conf. Automatic Face and Gesture Recognition*, Killington, VT, 1996.

[9] C.Hummels, P.Stappers, “Meaningful Gestures for Human Computer Interaction: Beyond Hand Postures”, *IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 1998

[10] J.Kuch, “Vision-Based Hand Modeling and Gesture Recognition for Human Computer Interaction”, *M.S. Thesis, Dept. of ECE, Univ. of Illinois at Urbana-Champaign*, 1994

[11] J.Lee, T.Kunii, “Model-based Analysis of Hand Posture”, *IEEE Computer Graphics and Applications*, Sept., pp.77-86, 1995

[12] K.Nigam, A.Mccallum, S.Thrun, T.Mitchell, ”Text Classification from Labeled and Unlabeled Documents Using EM”, *Machine Learning*, 1999

[13] C. Nolker, H. Ritter, “Illumination Independent Recognition of Deictic Arm Postures”, *Proc. 24th Annual Conf. of the IEEE Industrial Electronics Society*, Germany, pp. 2006 2011. 1998

[14] F. Quek, M. Zhao, “Inductive Learning in Hand Pose Recognition”, *IEEE Automatic Face and Gesture Recognition*, 1996

[15] J.Rehg, T.Kanade, “Model-Based Tracking of Self-Occluding Articulated Objects”, *IEEE Int'l Conf. Computer Vision*, pp.612-617, 1995

[16] H.Rowley,S.Baluja,T.Kanade, “Neural Network-Based Face Detection”, *IEEE PAMI*, Jan, 1998

[17] N.Shimada, et al., “Hand Gesture Estimation and Model Refinement Using Monocular Camera - Ambiguity Limitation by Inequality Constraints”, *Proc. of the the 3rd Conf. on Face and Gesture Recognition*, 1998

[18] J.Segen, S.Kumar, “Fast and Accurate 3D Gesture Recognition Interface”, *IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 1998

[19] J. Triesch, C. Malsburg, “Robust Classification of Hand Postures Against Complex Background”, *Int'l Conf. On Automatic Face and Gesture Recognition*, 1996

[20] V.Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, 1995

[21] Ying Wu, Thomas S. Huang, “Capturing Human Hand Motion: A Divide-and-Conquer Approach”, *IEEE Int'l Conf. Computer Vision*, Greece, 1999

[22] Ying Wu, Thomas S. Huang, “Vision-Based Gesture Recognition: A Review”, *The 3rd Gesture Workshop*, Gif-sur-Yvette, France

[23] Ying Wu, Qiong Liu, Thomas S. Huang, “An Adaptive Self-Organizing Color Segmentation Algorithm with Application to Robust Real-time Human Hand”, *Proc. Asian Conference on Computer Vision*, Taiwan, 2000