People Recognition and Pose Estimation in Image Sequences

Chikahito Nakajima
Central Research Institute of
Electric Power Industry, 2-11-1,
Iwado Kita, Komae, Tokyo Japan.
nakajima@criepi.denken.or.jp

Massimiliano Pontil and Tomaso Poggio
Center for Biological and Computational
Learning, MIT 45 Carleton Steet,
Cambridge, MA, 02142 USA.
{pontil,tp}@ai.mit.edu

Abstract

This paper presents a system which learns from examples to automatically recognize people and estimate their poses in image sequences with the potential application to daily surveillance in indoor environments. The person in the image is represented by a set of features based on color and shape information. Recognition is carried out through a hierarchy of biclass SVM classifiers that are separately trained to recognize people and estimate their poses. The system shows a very high accuracy in people recognition and about 85% level of performance in pose estimation, outperforming in both cases k-Nearest Neighbors classifiers. The system works in real time.

1 Introduction

This paper presents a system which learns from examples to automatically recognize people and to estimate their poses in image sequences.

The problem of people recognition presents a number of difficulties due to the similarity of people images and the high variability of conditions under which such images are recorded. As a first attempt to attack the problem, we have made two assumptions about the context in which people are observed: (a) each person wears the same clothes throughout the day, and (b) images are taken from a fixed camera with a static background. The second assumption is introduced for practical reasons (images are taken in a small indoor environment) and facilitates the detection of a person in the image. The first assumption simplifies the recognition task as it allows us to extract features based on color information. In the paper, we discuss both features based on color and pixel histograms and local shape information [5].

People recognition and pose estimation is formulated as a multiclass classification problem. In this work, classification is based on a hierarchy of biclass Support Vector Machine (SVM) classifiers [1, 11]. SVM is a technique for learning from examples motivated in the framework of statistical learning theory [11]. This technique has received a great deal of attention in the last few years due to its successful application to different problems (see [3] and references therein). We have used two types of hierarchies of biclass SVM classifiers: a bottom-up decision tree [7] and a top-down decision graph [6].

We have performed a preliminary set of experiments, where the system is trained to recognize four different people and estimate as many possible poses (front, back, left and right side). The system has been trained with 640 examples, 40 for each person at a given pose. The system works in real time and shows very high performance in people recognition and about 85% level of performance in pose estimation, outperforming in both cases k-Nearest Neighbors classifiers. The experimental results also indicate a potential application of the system to daily surveillance in indoor environments.

The paper is organized as follows. Section 2 presents a description of the system outline. Section 3 describes the experimental results. Section 4 summarizes our work and presents our future research.

2 System Outline

The system consists of three modules: Image I/O, Pre-Processing and Recognition. Figure 1 shows an outline of the system. Each image from a camera is distributed to the Pre-Processing module through

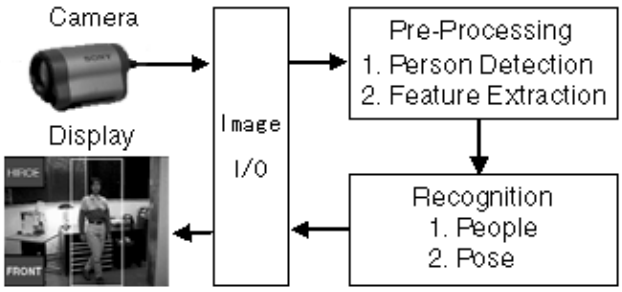


Figure 1: Outline of the system.

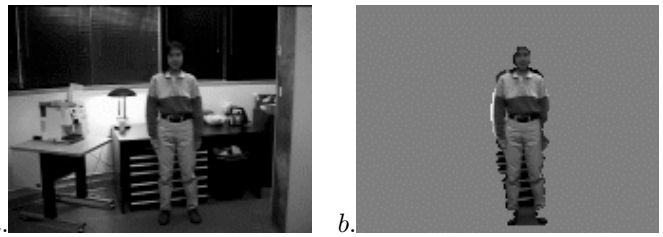


Figure 2: An example of moving person detection.

the Image I/O. The results of the Pre-Processing and the Recognition modules are shown on a display. Each module works independently on three separated computers connected through a network.

2.1 Pre-Processing

The Pre-Processing module consists of two parts: moving person detection and feature extraction.

2.1.1 Moving Person Detection

The system uses two filters to detect a moving person in an image sequence. The first filter computes the difference between a given image and an average image. The average image is calculated over the k previous images in the sequence. In our experiments we choose $k = 3$. Generally, the result of this filter has a lot of noise. To reduce the noise, we use another filter which extracts moving edges from the image sequence and fills the interior part with the original pixel values. Figure 2-a shows an image from the sequence and Figure 2-b is the combined result of the two filters.

2.1.2 Feature Extraction

Once the person has been detected in the image, a set of features is extracted. Typically color or shape features are used for surveillance systems [4]. In this paper, we used features based on color and pixel histograms and local shape filters.

1. Color Histogram and Pixel Histogram

   These features consist of two sections. The first section is the RGB color histograms. The system computes a 1-D color histogram of 32 bins per color channel of an image. The second section is the pixel histograms along the horizontal and vertical lines in the image. To compute the pixel histograms first a 50×120 window is centered at the detected person and then the horizontal and vertical (HV) histograms are calculated inside the window. We have chosen a resolution of 10 bins for the vertical histogram and 30 bins for the horizontal histogram. The total number of extracted features is 136, 32×3 for the RGB histograms and $10 + 30$ for the HV pixel histograms.

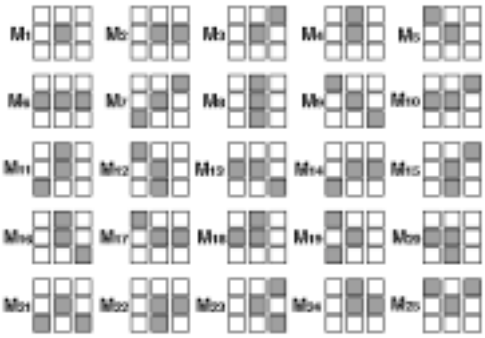
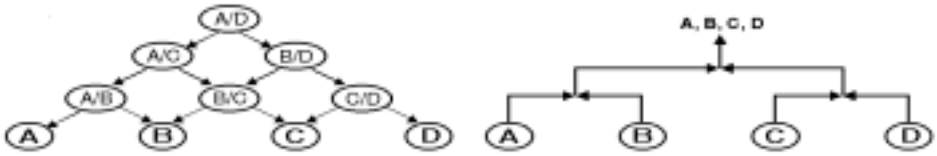


Figure 3: Shape Patterns.



a. A top − down decision graph. b. A bottom − up decision tree.

Figure 4: Hierarchy of SVMs.

2. Local Shape Features

   A second set of features is obtained by convolving the local shape patterns shown in Figure 3 with a given image. These patterns have been introduced in [5] for position invariant person detection. Let M^i, $i = 1, \ldots, 25$, be the pattern in Figure 3 and V_k the 3×3 patch at pixel k in an image. We consider two different types of convolution operations. The first is the linear convolution given by $\sum_k M^i \cdot V_k$, where the sum is on the image pixels. The second is a non-linear convolution given by $F_i = \sum_k C_{(k,i)}$, where

$$C_{(k,i)} = \begin{cases} V_k \cdot M^i & : \quad if \ V_k \cdot M^i = \max_j (V_k \cdot M^j) \\ 0 & : \quad otherwise. \end{cases}$$

   The system uses the simple convolution from the pattern 1 to 5 and the non-linear convolution from the pattern 6 to 25. The non-linear convolution works mainly on edge areas in the image and has been inspired by recent work in [8]. This paper uses a simple combination model, such as "R+G-B", "R-G" and "R+G", suggested by physiological study [10]. The system extracts 75 (25×3) features from the three types of the above RGB combinations.

2.2 Recognition

We have used two similar types of multiclass classifiers based on the combination of biclass SVM: the top-down decision graph recently proposed in [6] and the bottom-up decision tree described in [7]. These methods are illustrated in Figure 4-a and Figure 4-b in the case four classes. They both require the computation of all the possible biclass SVM classifiers, each trained on a pair of classes (i.e. two different people or two different poses).

Each class is represented by a set of input vectors, each vector consisting of the features extracted above. Briefly, a linear SVM [11, 7] finds the hyperplane $\mathbf{w} \cdot \mathbf{x} + b$ which best separates two classes. The $\mathbf{w}$ is the weight vector, the $\mathbf{x}$ is the vector of features, and b is a constant. This hyperplane is the one which maximizes the distance or *margin* between the two classes. The margin, equal to $2\|\mathbf{w}\|^{-1}$, is an important geometrical quantity because it provides an estimate of the similarity of the two classes

a.Front b.Back c.Left d.Right

Figure 5: Examples of the four poses for one person.

a.Cpapa b.Chika c.Hiroe d.Bernd

Figure 6: Examples of the four people in the frontal pose.

and can play a very important role in designing the multiclass classifier. This idea can be extended to non-linear SVM; see [11] for more information.

Each node in the decision graph in Figure 4-a represents a biclass SVM and has two children (except the bottom nodes). Classification of an input vector starts from the root node of the graph and follow the decision path along the graph. For example, looking at Figure 4-a, if the root note classifies the input in class A (resp. D), the node is exited via the left (resp. right) edge and so on. Notice that the classification result depends on the initial position of each class in the graph as each node can be associated with different class pairs. A possible heuristic to achieve high classification performance consists in selecting the SVMs with the largest margin in the top nodes of the graph.

In the bottom-up decision tree of Figure 4-b, there are 2 nodes in the bottom layer and one node in the second layer, each representing a biclass SVM. To classify an input, first the biclass SVM classifiers in the bottom nodes are evaluated and, depending on their result, a last biclass SVM is evaluated in the top node. For example, if A and D win at the bottom node classes, the A/D biclass SVM is evaluated as the final classification at the top node. This method can easily be extended to any number of classes [7].

Both the bottom-up decision tree and the decision graph requires the evaluation of $n-1$ SVMs (n being the number of classes) and are very fast if compared to other classification strategy like in [9, 2].

3 Experiments

We have performed a preliminary set of experiments where the system is trained to recognize within four different people, and estimate as many poses (front, back, left and right side). These poses are shown in Figure 5 for one person. The frontal image of the four people are shown in Figure 6. The system has been trained with 640 examples, 40 for each person at a given pose. First, we have trained a multiclass classifier to recognize people. In this case, each class contained 40 images of one person, 10 per each different pose. For each person another classifier was trained to classify the pose. In this case, each class contained 40 images of the same person at approximately the same pose.

To summarize, five multiclass classifiers have been trained, one for people recognition and four for pose estimation. The system uses first the multiclass classifier to recognize the person in the image. Once the person has been recognized, the pose multiclass classifier relative to that person is used to classify

a. Cpapa b. Chika c. Hiroe d. Bernd

Figure 7: Examples of people recognition results.

a. Front b. Back c. Left d. Right

Figure 8: Examples of pose estimation results.

Table 1: People recognition and pose estimation rates from the test set.

	$Features$	SVM	SVM	$k-NearestNeighbor$			
		$(Top-down)$	$(Bottom-up)$	$k=1,2$	$k=3$	$k=4$	$k=5$
$People$	$RGB+HV$	91.4	91.6	94.7	94.4	94.9	94.1
$Recognition$	$LocalShape$	99.5	99.5	88.3	85.0	86.3	84.8
$Pose$	$RGB+HV$	68.0	68.2	67.8	66.8	66.8	65.7
$Estimation$	$LocalShape$	84.5	84.3	82.0	82.7	82.7	82.0

the pose.

Figure 7 shows an example of the output display of the system for different people in the front position. The upper left corner of the display shows the name of recognized person and the lower left corner shows the estimated pose. The center square is the result of the detection module. Figure 8 shows a similar example for different poses of the same person.

Table 1 reports the test classification rates of the two types of multiclass SVM discussed above (see Figure 4-a,b) and k-Nearest Neighbor classifiers[1]. We have always used linear SVM. The test set consisted of 418 images of the four people at all possible poses. Notice that the two type of SVM classifiers have very similar performance and mostly outperform the k-Nearest Neighbor. The best results are achieved when the local shape features are used. Notice also that poses have a lower classification rate than people. People are easy to recognize as they wear the same clothes. On the other hand, pose estimation is a harder task because of the similarity between right and left poses or front and back. Finally, notice that the best results are achieved when the local shape features are used. This fact is interesting as features based on shape might be discriminative also when our assumption about people wearing same clothes is relaxed.

[1]In case of a tie, the system chooses the class whose nearest neighbors have minimum average distance from the input.

4 Conclusions

We have presented a system which is able to recognize people and estimate their poses in an image sequence. The core of the system is a multiclass classification problem which we have approached using two types of hierarchical SVM classifiers.

The experimental results indicate the effectiveness of the system to solve the task and a better performance of the two hierarchical SVM classifiers with respect to $k-$Nearest Neighbors.

This system is part of the surveillance system currently under development. In the future research we plan to combine this system and a face recognition system to develop a robust surveillance system.

References

[1] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273-297, 1995.

[2] J. Friedman. Another approach to pilychotomous classification. *Stanford University, Dept. of Statistics*, Technical Report, 1996.

[3] I. Guyon. SVM Application List. *http://www.clopinet.com/isabelle/Projects/SVM/applist.html*.

[4] I. Haritaoglu, D. Harwood and L. Davis. Hydra: multiple people detection and tracking using silhouettes. *Proc. of International Workshop on Visual Surveillance*, 6-13,1999.

[5] T. Kurita, K. Hotta, and T. Mishima. Scale and rotation invariant recognition method using higher-order local autocorrelation features of log-polar image. *Proc. of ACCV*, 1998.

[6] J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. *Advances in Neural Information Processing Systems (to appear).*

[7] M. Pontil and A. Verri. Support vector machines for 3-d object recognition. *IEEE Trans. PAMI*, 637–646, 1998.

[8] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1037, 1999.

[9] B. Scholkpof and C. Burges and V. Vapnik. Extracting support data for a given task. *Proc. of the first Int. Conf. on Knowledge Discovery and Data Mining*, AAAI Press, 1995.

[10] K. Uchikawa. Mechanism of color perception. *Asakura syoten*, (Japanese), 1998.

[11] V. Vapnik. Statistical learning theory. *John wiley & sons, inc.*, 1998.