MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1688 June, 2000
C.B.C.L Paper No. 188

People Recognition in Image Sequences by Supervised Learning

Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, Tomaso Poggio

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu. The pathname for this publication is: ai-publications/1500-1999/AIM-1688.ps.Z

Abstract

We describe a system that learns from examples to recognize people in images taken indoors. Images of people are represented by color-based and shape-based features. Recognition is carried out through combinations of Support Vector Machine classifiers (SVMs). Different types of multiclass strategies based on SVMs are explored and compared to k-Nearest Neighbors classifiers (kNNs). The system works in real time and shows high performance rates for people recognition throughout one day.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from Office of Naval Research under Contract No. N00014-93-1-3085 and Office of Naval Research under Contract No. N00014-95-1-0600. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, Daimler-Benz AG, Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc.

1 Introduction

Digital video cameras and computers have come into wide use recently but visual surveillance for security is still mainly performed by human observers. Automatic visual surveillance systems could play an important role in supporting and eventually replacing human observers. To become practical these systems should be able to distinguish people from other objects and to recognize individual persons with a sufficient degree of reliability, depending on the specific application and security level.

The development of automatic visual surveillance systems can now leverage techniques for detecting and recognizing people that have been developed recently: pedestrian detection [1, 2], face detection [3, 4, 5], face recognition [6, 7], and motion detection [8, 9]. In general, the unconstrained task of people recognition still presents a number of difficult challenges due to the similarity of people images, pose variations, change of clothes, different illumination and background conditions.

In our experiments we defined a suitably restricted task of recognizing members of our Lab (up to about 20) while they were using a coffee machine located in the Lab's main office. The camera was located in front of the coffee machine at a distance of about 15 feet; background and lighting were almost invariant. Recognition was based on the assumption that the same person was going to have the same general appearance (clothes) during the day. The recorded images were distributed to multiple computers that performed the recognition in real-time.

In our group, SVMs have been successfully applied to various two-class problems, such as pedestrian and face detection [1, 10, 11]. Recently several methods have been proposed in order to expand the application field of SVMs to multi-class problems [12, 13, 14]. In this paper, we used these methods to recognize people with multi-class SVMs. The experiments show high recognition rates indicating the relevance of our system for the development of more sophisticated indoor surveillance applications.

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

2 System Outline

The system consists of two modules: Image pre-processing and people recognition. Figure 1 shows an outline of the system. Each image from the camera is forwarded to the pre-processing module. The results of the pre-processing and the recognition modules are visualized on a display. Each module works independently on several computers connected through a network.

2.1 Pre-Processing

The pre-processing module consists of two parts: detection of moving persons and extraction of image features.

2.1.1 Person Detection

The system uses two steps to detect a moving person in an image sequence. The first step, known as background subtraction, computes the difference between the current image and the

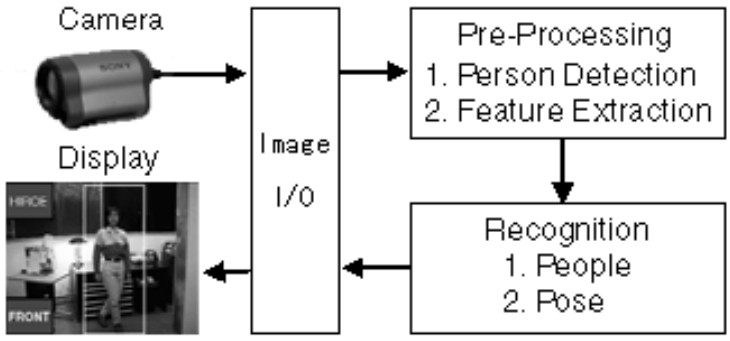


Figure 1: Outline of the system.

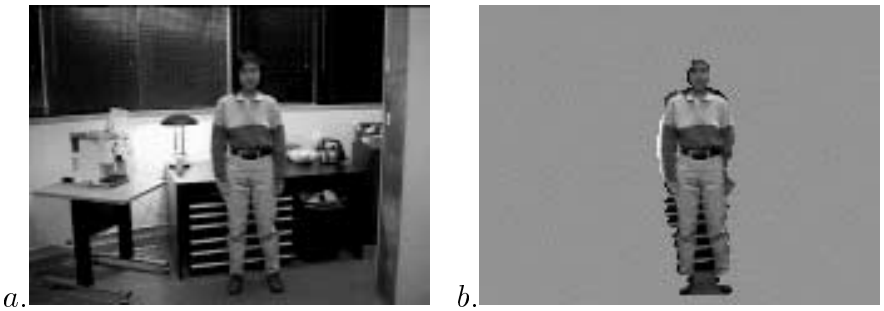


Figure 2: An example of moving person detection.

background image. The background image is calculated over the k most recent images[1] of the sequence. The result of background subtraction usually includes a lot of noise. For this reason we added a second step which extracts the silhouette of the moving person using edge detection methods. Figure 2-a shows an image from the sequence and Figure 2-b shows the combined result of the two steps.

2.1.2 Feature Extraction

Once the person has been detected and extracted from the background, we calculate different types of image features:

1. RGB Color Histogram

   The system calculates one dimensional color histogram with 32 bins for each color channel. The total number of extracted features is 96 (32×3) for a single image.

2. Normalized Color Histograms

   The system calculates two dimensional normalized color histograms; $r = R/(R+G+B)$, $g = G/(R+G+B)$. Again, we chose 32 bins for each color channel. Overall, the system extracts 1024 (32×32) features from a single image.

[1] In our experiments we chose $k = 3$.

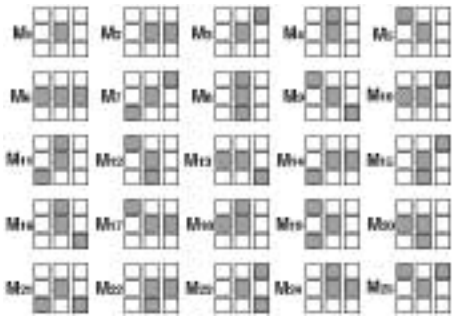


Figure 3: Shape Patterns.

3. RGB Color Histogram + Shape Histogram

   We calculate simple shape features of people by counting pixels along rows and columns of the extracted body images. We chose a resolution of 10 bins for column histograms and 30 bins for row histograms. The total number of extracted features is 136, 32×3 for the RGB histograms and $10 + 30$ for the shape histograms.

4. Local Shape Features

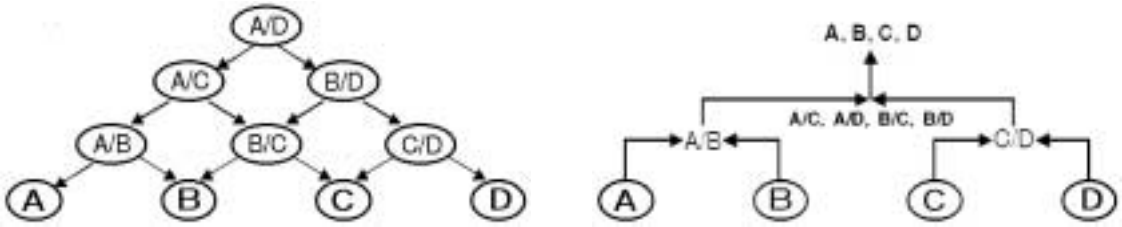
   Local features of an image are obtained by convolving the local shape patterns shown in Figure 3. These patterns have been introduced in [15] for position invariant person detection. Let M^i, $i = 1, \ldots, 25$, be the patterns in Figure 3 and V_k the 3×3 patch at pixel k in an image. We consider two 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

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

   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 mainly extracts edges and has been inspired by recent work in the field of brain models [16]. The shape features are extracted for each of the following color channels separately : $R+G-B$, $R-G$ and $R+G$. This color model has been suggested by physiological studies [17]. The system extracts 75 (25×3) features from the three color channels.

2.2 Recognition

Our recognition system is based on linear SVMs. Briefly, a linear SVM [13, 18] 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 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 and can play a very important role in designing the multi-class classifier.



a. A Top − down decision graph. b. A Bottom − up decision tree.

c. One-vs-all type.

Figure 4: Multi-class of SVMs.

To deal with multi-class problems, we used three types of combinations of linear SVMs[2]: the top-down decision graph recently proposed in [12], the bottom-up decision tree described in [13] and the one-vs-all type classifiers [14]. These methods are illustrated in Figure 4-a,b,c for the case of four classes. Each node in the decision graph in Figure 4-a represents an SVM classifier. Classification of an input vector starts from the root node of the graph and follows the decision path along the graph. For example, if the A/D SVM in the root node of Figure 4-a classifies the input as belonging to class A, the node is exited via the left edge. Notice that the classification result depends on the initial position of each class in the graph, as each node can be associated with different pairs of classes. A possible heuristic to achieve high classification performance consists of 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. To classify an input vector, the A/B and C/D SVM classifiers in the bottom nodes are evaluated. In the top node, the SVM trained to distinguish between the two winning classes is evaluated. For example, if A and D win in the bottom layer, the A/D SVM is evaluated in the top node. In the one-vs-all type technique, there is one classifier associated to each class. This classifier is trained to separate this class from all remaining classes. A new input vector is classified in the class whose classifier has the highest score among all classifiers. Figure 4-c shows an example for four classes. Note that at run time all three strategies require the evaluation of $n - 1$ SVMs (n being the number of classes).

3 Experiments

In this section we report on two different sets of experiments. In our experimental setup a color camera recorded Lab members in our main office while they were using a coffee machine. The camera was located in front of the coffee machine at a distance of about 15 feet. Images were recorded at a fixed focus, background and lighting were almost invariant.

[2]The same techniques can be applied to non-linear SVMs [18].

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.

In the first experiment, we evaluated the use of different sets of image features and different types of classifiers (multi-class SVMs and kNNs). The task in the first experiment was to distinguish four different people and to recognize their poses using recordings of one day. In the second experiment, we chose the best features as determined in the first experiment and increased the number of people to be recognized to eight and the time span of our recordings to 16 days.

3.1 People Recognition and Pose Estimation

In this experiment the system was trained to recognize four people and to estimate their poses (front, back, left and right). Training and test images were recorded during one day. Example images of a person in four poses are shown in Figure 5; example images of the four people are shown in Figure 6. We used 640 images to train the system, 40 for each person at each pose. First, we trained a multi-class classifier to recognize people. The training set contained 160 images per person, 40 per pose. Then, multi-class pose classifiers were trained for each person separately. To summarize, five multi-class classifiers were trained, one for people recognition and four for pose estimation. The system first recognized the person and then selected the proper multi-class classifier to determine the pose of the person.

Figure 7 shows an example of the output of the system for the four people in frontal pose. The upper left corner shows the name of the recognized person, the lower left corner shows the estimated pose. The white boxes in the center of the images are the results of the detection module. Figure 8 shows a similar example for different poses of the same person. Table 1 reports the test classification rates for different types of features and different types of classifiers

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

	Features (*dimension*)	*SVM*			$k-NearestNeighbor$		
		TopDown	*Bot.,Up*	*OneAll*	$k=1$	$k=3$	$k=5$
People Recognition	*RGB* (96)	99.5	99.2	99.5	99.0	98.7	98.5
	NormalizedRG (1024)	100	100	100	100	100	100
	RGB+ GlobalShape(136)	91.4	91.6	96.2	94.7	94.4	94.1
	LocalShape (75)	99.5	99.5	97.5	88.3	85.0	84.8
Pose Estimation	*RGB* (96)	74.9	75.9	83.8	70.1	70.6	72.3
	NormalizedRG (1024)	86.5	86.3	87.8	85.5	85.8	86.0
	RGB+ GlobalShape(136)	68.0	68.2	70.1	67.8	66.8	65.7
	LocalShape (75)	84.5	84.3	84.0	82.0	82.7	82.0

including three versions of multi-class SVMs and a kNN classifier[3]. The test set consisted of 418 images of the four people at all possible poses. For both tasks, people recognition and pose estimation, the best results were obtained with normalized color features (1024 dimension). The three types of SVM classifiers showed similar recognition rates, which were slightly better than the recognition rates achieved by kNN classifiers. Note that recognition rates for poses are lower than that for people. People can be easily distinguished based on their clothes. Pose estimation is more difficult because of the similarity between right/left poses and front/back poses. We expected global shape features based on row and column histograms to be helpful for pose estimation. However, the performance decreased when adding row and column histograms to the input features. This is because of arm movements of people and varying distances between people and camera that lead to significant changes in the shape and size of the body silhouettes. On the other hand, local shape features performed well for both: person recognition and pose estimation.

3.2 Increasing the Data Set

In the second experiment we repeated the previous experiment on a data set containing images of eight people recorded over several days. About 3500 images were recorded during 16 days. From the 3500 images we selected 1127 images belonging to the eight most frequent users of the coffee machine. Some example images[4] of these eight people are shown in Figure 9. The images were represented by their normalized color histograms. We chose these features because they showed the best performance in the first experiment.

We performed five different sets of experiments where the system was trained to recognize the eight people. In the first four experiments we used about 90%, 80%, 50%, and 20% of the image database for training. The remaining images were used for testing. In the fifth

[3]In case of a tie, the system chose the class whose nearest neighbors had minimum average distance from the input.

[4]More details about this data set can be found in Appendix A.

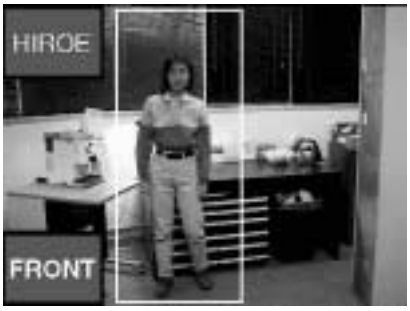
a. Cpapa

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 2: Recognition rates for eight people. The classifier was applied to 1024 normalized color features.

	$(test : training)$	$1:9$ $(113:1014)$	$1:5$ $(188:939)$	$1:1$ $(564:563)$	$5:1$ $(939:188))$	$NewDay$ $(122:1005)$
SVM	$Top-Down$	92.3	91.2	90.5	73.3	45.9
	$Bottom-Up$	90.6	91.7	90.6	66.1	45.9
	$One-All$	87.2	90.6	85.9	84.6	49.2
	$One-All(Poly)$	98.3	96.4	94.7	88.1	45.9
kNN	$k=1$	92.9	92.0	92.7	85.1	53.3
	$k=3$	92.9	92.0	92.2	81.3	50.0
	$k=5$	94.7	91.0	90.1	76.0	50.8

experiment the training set consisted of all images recorded during the first 15 days, the test set included all images recorded during the last day.

Recognition rates are shown in Table 2. The system performed well when the training set contained images from all 16 days (first four experiments). The recognition rate decreased to about 50% when the system was tested on images recorded during a new day (last experiment). This is because people wore different clothes every day, so that the system was not able to recognize them based on the color of their clothes only. Notice that this rate is still considerably better than chance (12.5%). Overall kNN was slightly better than linear SVMs. Preliminary tests with SVMs with second degree polynomial kernel showed a significantly better performance than kNN.

4 Conclusion and Future Work

We have presented a system that recognizes people in a constrained environment. People recognition was performed by multi-class SVMs that were trained on color images of people. The images were represented by different sets of color and shape-based features. The recognition rate of the system was about 90% for linear SVMs trained on normalized color histograms of peoples' clothes. The high recognition result indicates the relevance of the presented method for automatic daily surveillance in indoor environments. However, since the clothes of people usually change every day we need to expand the capabilities of the system in order to recognize people over multiple days. For this reason we plan to combine the system described here with a face recognition system.

Figure 9: Images of eight people recorded at different days.

Figure 10: Results examples of the sub-system.

A Data Sets of People Images

We recorded members of our Lab (up to about 20) while they were using a coffee machine located in the Lab's main office. The system recorded the foreground and background images. The system recorded about 3500 images during 16 days (May/1,2,3,4, April/10,12,13,19,24, Mar/6,7,8,21,24,27,29 2000) and selected the images of the eight most frequent users of the coffee machine using clustering techniques. Example images of these eight people are shown in Figure 9. Figure 10 shows a sequence of images which was recorded while one of the lab members was using the coffee machine. The complete database can be downloaded from the CBCL Homepage.

References

[1] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. *Proc. of ICIP*, 25-28, 1999.

[2] C. Papageorgiou and T. Poggio. A trainable object detection system: Car detection in static images. *MIT AI Memo*, 1673 (CBCL Memo 180), 1999.

[3] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *MIT AI Memo*, 1521 (CBCL Memo 112). 1994.

[4] H. Schneiderman. A statistical approach to 3D object detection applied to faces and cars. *CMU-RI-TR-00-06*, 2000.

[5] V. P. Kumar and T. Poggio. Learning-based approach to real time tracking and analysis of faces. *Proc. of AFGR*, 2000.

[6] R. Brunelli and T. Poggio. Face Recognition: Features versus templates. *IEEE on PAMI*, Vol.15, No. 10. 1993.

[7] A. Pentland and T. Choudhury, Face recognition for smart environments, *Computer*, Vol. 32, 2000.

[8] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking, *Proc. of CVPR*, 1999.

[9] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring: VSAM final report, *Technical CMU-RI-TR-00-12*, 2000.

[10] B. Heisele, T. Poggio and M. Pontil. Face detection in still gray images. *MIT AI Memo No. 1687*, 2000.

[11] A. Mohan. Object detection in images by components. *MIT AI Memo No. 1664*, 1999.

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

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

[14] O. Chapelle, P. Haffner and V. Vapnik. Support Vector Machines for histogram-based image classification. *IEEE Trans. Neural Networks*, 1055–1064, 1999.

[15] 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.

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

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

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

[19] C. Nakajima, M. Pontil and T. Poggio. People recognition and pose estimation in image sequences. *Proc. of IJCNN*, 2000.

[20] C. Nakajima, N. Itoh, M. Pontil and T. Poggio. Object recognition and detection by a combination of Support Vector Machine and Rotation Invariant Phase Only Correlation. *Proc. of ICPR*, 2000.