Object Recognition and Detection by a Combination of Support Vector Machine and Rotation Invariant Phase Only Correlation

Chikahito Nakajima Norihiko Itoh
Central Research Institute of Electric
Power Industry, Tokyo 201-8511 JAPAN

Massimiliano Pontil Tomaso Poggio
Massachusetts Institute of Technology
45 Carleton, Cambridge, MA 02142 USA

Abstract

This paper proposes an object recognition and detection method by a combination of Support Vector Machine Classifier (SVM) and Rotation Invariant Phase Only Correlation (RIPOC). SVM is a learning technique that is well founded in statistical learning theory. RIPOC is a position and rotation invariant pattern matching technique. We combined these two techniques to develop an augmented reality system. This system can recognize and detect objects from image sequences without special image marks or sensors and show information about the objects through a head-mounted display. Performance is real time.

1 Introduction

We are developing a support system for maintenance training of electric power facilities using object recognition and detection techniques. To assist in the training, we are evaluating the use of augmented reality through a head mounted display, small cameras, and image processing.

Many systems in augmented reality have been developed in a variety of applications such as maintenance, repair, assistance of surgery, and guidance of navigation [1, 2]. They typically use special marks or sensors on the target objects to facilitate detection and classification of objects and determination of their poses. In this system, we intend to recognize and detect objects without special marks or sensors. This paper proposes an object recognition and detection method by a combination of Support Vector Machine (SVM) classifiers [3] and Rotation Invariant Phase Only Correlation (RIPOC) [4].

SVM is a technique for learning from examples that is well-founded in statistical learning theory. SVM has recently received a great deal of attention and has been applied to areas such as handwritten character recognition, 3D object recognition, text categorization and object detection [3]. If a large amount of sample data of the target objects is available, SVM is a very useful tool for summarizing the data in terms of the support vectors. A drawback of “pure” learning techniques from examples, like SVM, is that they need large data sets for effective training. Moreover object detection or recognition has to be performed at many locations and scales in a given image [6].

RIPOC is a pattern matching technique which measures rotation and translation between two images. RIPOC computes the correlation between the two images by means of the Fast Fourier Transformation (FFT). It uses the FFT amplitude for measuring the rotation and the FFT phase for measuring the translation. When applied to object recognition, RIPOC has to compute the correspondences among all the object templates and all the images in the sequence. This method is not sufficient to perform moving object recognition by its own because of its lack of robustness against changes in the background.

This paper describes a new method for object recognition and detection from image sequences. The system uses a hierarchy of SVMs for object recognition and RIPOC for pose detection. Our preliminary experimental results indicate the advantage of a combination of the two techniques and open the possibility of applying the system to maintenance training in an industrial domain. The paper is organized as follows. Section 2 presents a description of the system outline. Section 3 describes the results of the system. Section 4 summarizes our work and presents our future research.

2. System Outline

The system consists of three parts: Image I/O, Recognition and Detection. Figure 1 shows an outline of the system. Each image from a camera is distributed to the Recognition module and the Detection module through the Image I/O module. The results of recognition and detection are displayed on a Head Mounted Display (HMD). Each part is working independently on different computers.

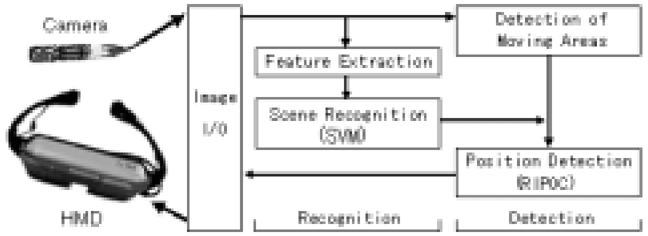

Figure 1. Outline of the system.

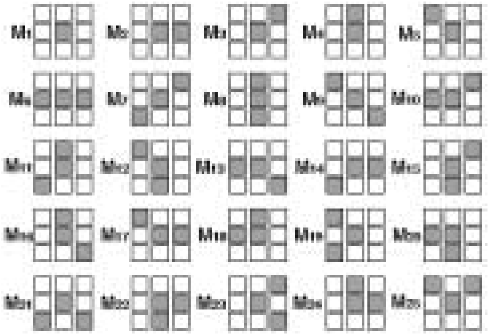

Figure 2. Mask patterns.

2.1 Recognition Module

The Recognition module has two processes: a feature extraction process and a scene recognition process.

2.1.1 Feature Extraction

The system uses a set of features which are obtained by convolving the local mask patterns shown in Figure 2 with a given image. These masks have been introduced in [7] for position invariant object detection. Let M^i, $i = 1, \ldots, 25$, be the mask pattern in Figure 2 and V_k the 3×3 patch at pixel k in an image. In our system, the *i-th* feature, F_i, is given by $\sum_k M^i \cdot V_k$, where the sum is on the image pixels. We have used not only the simple convolution but also a non-linear convolution, $F_i = \sum_k C_{(k,i)}$ 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 mask 1 to 5 and the non-linear convolution from the mask 6 to 25. The non-linear convolution works mainly on edge areas in an image. The non-linear operation has been inspired by recent work in [8] and has shown good performance for people recognition in our previous work [9].

To detect color, the human visual system uses responses of three types of the receptors, known as red, green and blue, and combines them in the retina. 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.

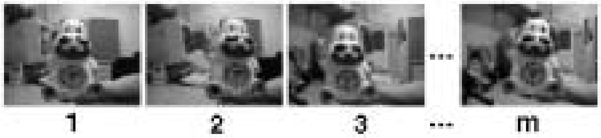

Figure 3. An example of a training data set for SVM.

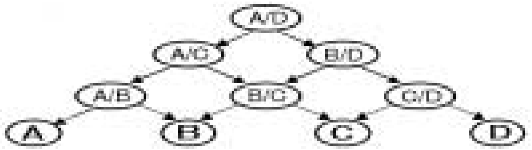

Figure 4. A decision graph of SVM.

2.1.2 Recognition of a Scene

Figure 3 shows an example of training images for one class. The m in Figure 3 is the number of training images for the class and is set to 60 in this experiment. Each class is represented by a set of m vectors, each vector consists of the 75 features extracted above. The system uses a linear SVM [3, 5] that determines 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 estimator of the similarity of the two classes.

The SVM is computed for each pair of n classes. Recognition of the object in a new image is based on the decision graph of SVMs [11]. The graph for four classes (A, B, C, D) is shown in Figure 4. Each node in the graph is associated with a pair of classes. Classification of an input vector starts from the root node (A/D) of the graph. Notice that the classification result depends on the initial position of each class in the graph. A possible heuristic to improve classification performance consists of selecting the SVMs with the largest margin in the top node (A/D) of the graph; we use this strategy. A similar method based on a binary decision graph of SVMs is also discussed in [5]. In both cases, classification of an input vector requires the evaluation of $n - 1$ SVMs.

2.2 Detection Module

The object detection module has two parts: detection of moving areas and position detection of the object. In the position detection part, the system selects a template image from a set of object templates by using the result of the Recognition module.

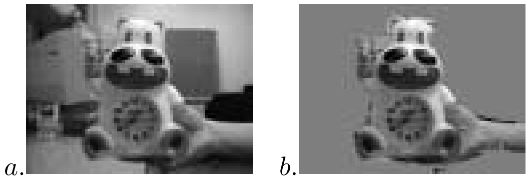

Figure 5. An example of moving area extraction.

2.2.1 Detection of Moving Area

In this system, we are using two filters to detect moving areas from images. One is the extraction of differences between the new input image and the average image. The average image is automatically calculated over k frames without moving edge extraction. In this experiment k is 3. Generally, the result of this filter has a lot of noise. To reduce the noise, the system uses another filter which extracts moving edges from the image sequence and fills the interior of the edges with the original pixel values. Figure 5-a is an image in the image sequences and Figure 5-b is the combination of the two filters. In Figure 5-b, the moving area is extracted.

2.2.2 Position Detection

Let $f_1(m,n)$ be a template image, and $f_2(m,n)$ an image in the image sequence. Both image sizes are $M \times N$ pixels, $m = 0,...,M-1$, $n = 0,...,N-1$. The discrete Fourier transformation for f_1 and f_2 are $F_1(u,v)$ $= A(u,v)e^{j\theta(u,v)}$ and $F_2(u,v)=$ $B(u,v)e^{j\phi(u,v)}$, $u = 0,...,M-1$, $v = 0,...,N-1$. $A(u,v)$ and $B(u,v)$ are amplitude spectral functions, θ and ϕ are phase spectral functions. Now we define new discrete Fourier images $F_1^{'}(u,v) = e^{j\theta(u,v)}$ and $F_2^{'}(u,v) = e^{j\phi(u,v)}$ where the amplitude spectral functions A and B are set to 1. The correlation of F_1' and F_2' is $H_{12}(u,v) = e^{j(\theta-\phi)}$. The inverse Fourier transform of H_{12} is :

$$G_{12}(r,s) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} e^{j(\theta-\phi)} e^{-j2\pi(\frac{ur}{M}+\frac{vs}{N})}$$

where $r = 0,...,M-1$, $s = 0,...,N-1$. The transform value is calculated by G_{12}.

If the image f_2 is translated along the horizontal direction by τ, $f_3(m,n) = f_2(m+\tau,n)$, the inverse Fourier transform becomes $G_{13}(r,s) = G_{12}(r+\tau,s)$. The peak of G_{12} moves in the same direction by τ. It is possible to detect image translation (direction and distance) from the peak shift. This method is called Phase Only Correlation (POC) [12]. However POC is not adequate for image rotation.

To treat image rotation, the system uses Rotation Invariant Phase Only Correlation (RIPOC) [4]. At first, it considers only the amplitude spectral functions A and B of F_1, F_2. The A and B are transformed to polar-space. Next it extracts the shift transformation value on the polar-space by using the above POC. This shift value indicates the rotation of the two images. Then the input image is rotated based on the measured rotation result. Finally, it measures the transformation from the template image to the input image by POC. Thus RIPOC is able to measure rotation and transformation of the images.

3 Experimentation

In this experiment, we used four classes for the system. Sixty images for each class, taken under a moving camera as in Figure 3, were used for training the SVM. One template image, as in Figure 5-b, for each class was provided to the RIPOC. This system runs in real time by distributing processing on three computers.

Figure 6 shows one result from the system. The lower left corner of each image shows the result of the SVM classification and the upper left corner shows the result of the RIPOC for each frame. The center cross in each image is the average position of RIPOC results over 5 frames. Figure 7 shows the SVM results for the image sequence in Figure 6 which contains 170 frames. The misclassifications in Figure 7 were caused by blurred images which occurred in the sequence from time to time.

Figure 8 shows the results for a rotated object and a object sitting on a table. The rotated object as in Figure 8-a was recognized and detected by the system. The object sitting on a table was recognized by the SVM, as in the lower left corner of Figure 8-b, but the object area wasn't detected, because moving areas weren't included on the image. Figure 9 shows the result when the camera moves from Figure 9-a to Figure 9-b. The detection result of moving areas, as in the upper left corner of Figure 9-b, included not only the moving object but also a part of background. However it found the correct position by the RIPOC, because the system used the true template which was selected by the SVM.

Figure 10-c,d are results where we replaced the object detection point with images of a brain, which are selected automatically, by the result of the SVM.

4 Conclusion

This paper presents an object recognition and detection method by a combination of SVM and RIPOC to develop an augmented reality system. The system can recognize and detect objects in real time from image sequences without special image marks or sensors and can show information about the objects through a head-mounted display.

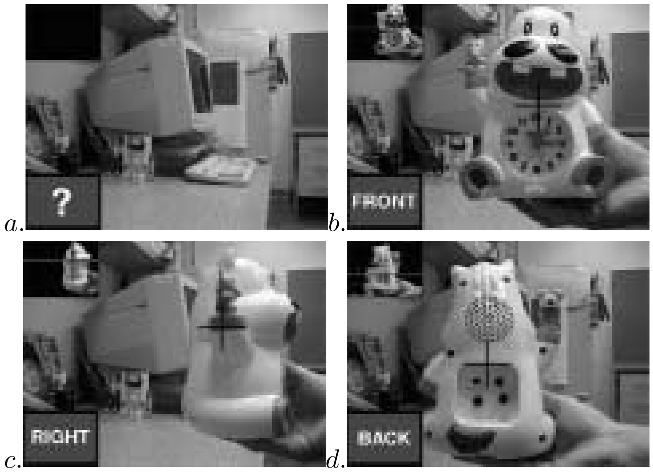

Figure 6. Results for four classes.

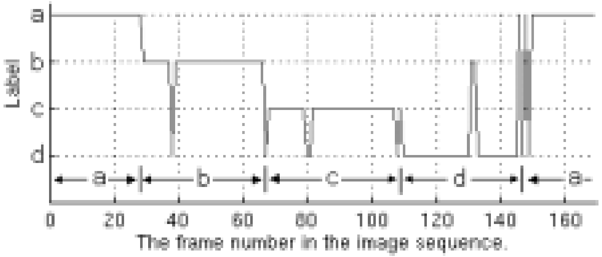

Figure 7. Recognition results for the image sequence. The bottom labels (a, b, c and d) show the true classes. The spikes in the graph are misclassifications.

This system is one part of the augmented reality system that we have been developing. We are now planning to combine the system and an image retrieval system to introduce the augmented reality system into maintenance training of electric power facilities.

References

[1] J. Rekimoto and K. Nagao. The world through the computer: Computer augmented interaction with real world environments. *Proc. of UIST*, 1995.

[2] S. Feiner, B. MacIntyre, T. Hollerer, and A. Webster. A touring machine: prototyping 3d mobile augmented reality system for exploring the urban environment. *Proc. of ISWC*, 1997.

[3] V. Vapnik. Statistical learning theory. *John Wiley & Sons inc.*, 1998.

[4] G. X. Ritter and J. N. Wilson. Pattern matching and shape detection. *Computer Vision Algorithms in Image Algebra*,CRC Press,1996.

[5] M. Pontil and A. Verri. Support vector machines for 3-D object recognition. *IEEE Trans. PAMI*, 1998.

[6] C. Papageorgiou and T. Poggio. Pattern classification approach to dynamical object detection. *Proc. of ICCV*, 1999.

[7] 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.

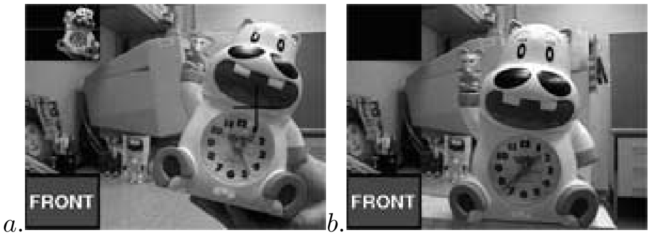

Figure 8. Results for a rotated object and a object sitting on a table.

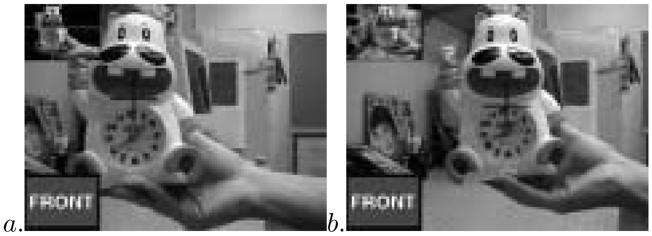

Figure 9. Results for a moving camera.

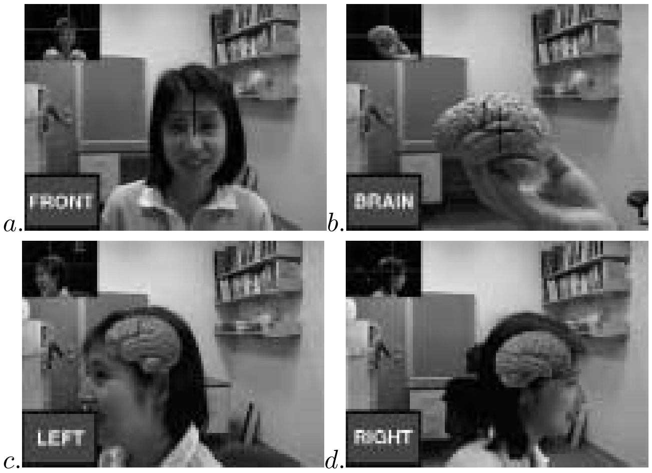

Figure 10. Results for four classes and overlaid pictures.

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

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

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

[11] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, 2000 (to appear).

[12] K. Kobayashi, H. Nakajima, T. Aoki, K. Kawamata and T. Higuchi. Filtering on phase only correlation domain and its applications. *ITE Technical Report*, 21(42),1997.