Ensemble SVM Regression Based Multi-View Face Detection System

Jie Yan, Stan Li , Shujie Zhu, Hongjiang Zhang

Jan. 18. 20001

Technical Report
MSR-TR-2001-09

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

In this paper, we present a novel learning method for SVM regression ensemble used in multi-pose face detection. Firstly, several view-specific SVM classifiers are trained using corresponding positive and negative examples. And then, an ensemble mechanism (SVM regression) is trained to combine the results from the view-specific SVCs. Experimental results show that the detection accuracy of the ensemble is better than the view-specific SVCs. Moreover, the SVR ensemble does not need extra pose estimation process prior to the classification; it generates pose information in addition to its detection result.

Keywords

Multi-View Face Detection, Support Vector Classification/Regression, Wavelet Transform, Ensemble

1. Introduction

The problem of face detection has been studied for many years. To date, current systems are quite limited in that detection is only possible in regions associated with a frontal view of a person's face. It is very clear that the pure frontal face detection technique is quite limiting because approximately 75% faces in one picture is non-frontal. More recently there have been attempts to build a face detection and recognition system that works with faces rotated out of plane.

This paper is directed toward a face detection system that overcomes the aforementioned limitations in prior face detection systems. A system, termed a pose-adaptive face detection system, is developed to detect non-frontal faces as well as frontal ones, regardless of the scale or illumination conditions associated with the face. In addition, the techniques described in this paper can also be used for pose information at the same time.

However, the task of detecting faces of various poses from a single image has been remained a major challenge because large changes in orientation significantly changes the overall appearances of face. Attempts have been made to view-based appearance models using a set of view-labeled appearances[1-5]. It should be mentioned here that Gong et al. investigated multi-view face pose distribution[6], and further extended SVMs to model the appearance of human faces which undergo nonlinear change across multiple views. Their approach uses inherent factors in the nature of the input images and the SVM classification algorithm to perform both multi-view face detection and pose estimation [7]. The SVM is based on Structural Risk Minimization theory, what should be mentioned here is Osuna et al, who has introduced a support vector machine(SVM) based approach for frontal view face detection which is the first application of SVM on face detection[8].

Gong et al. implemented a multi-view face detection and recognition system under a support vector machine framework and achieved better performance on video sequences[9]. However, the problem will become even more challenge when dealing with multi-scale, complexity background and illuminating conditions in the real static image.

Schneiderman and Kanade[10] use a statistical model to represent the object's appearance over a small range of pose variation, to capture variation in the appearance of the object that cannot be modeled explicitly. This includes variation in the object itself, variation due to lighting, and small variations in pose. Another statistical model is used to describe non-objects-of-interest. Since each detector is designed for a specific view of the object, multiple detectors that span a range of the object's orientation are used. The results of these individual detectors are then combined.

In this paper, we propose a new architecture of SVM[10] which is called SVR ensemble to do the face detection and pose estimation work in a single image. In order to simplify the complexity of detection, we only consider face rotating in depth, other degrees of freedom such as image plane translation should be removed.

The paper is organized as follows: section 2 will explain our approaches, including the theoretic introduction of support vector classification. In section 3, we will introduce the whole architecture of our system. Section 4 will give some experimental results we have achieved in multi-view face detection and pose estimation. The conclusions are discussed in Section 5.

2. Support Vector Classification and Regression

2.1 Support Vector Classification

Consider a two classes classification problem. Given a set of training examples $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in R^n$ is a feature vector, the examples are labeled by $y_i \in \{+1, -1\}$. The support vector machine[11,12] separate this two classes by an optimal hyper-plane $W \bullet X + b = 0$, Then the optimal classification function can be found by solving the following constrained minimization.

$$E(W) = \frac{1}{2}\|W\|^2 \text{ , subject to}$$

$$y_i[W \bullet X_i + b] \geq 1, i = 1, \ldots m$$

The SVM classification decision function has the form as follows:

$$D(x) = \sum_{i=1}^{n} y_i \alpha_i k(X, X_i)$$

To the linear SVM classifier, the decision function of the optimal hyperplane is thus:

$$D(x) = sign(\sum_{i=1}^{m} y_i \alpha_i X \bullet X_i + b)$$

To the nonlinear SVM classifier, there are a number of kernel functions which have been found to provide good generalisation capabilities, e.g. polynomials, Multi-Layer Perception, Sigmod, Gaussian Radial Basis Function. Here we explore the use of a Gaussian kernel function(analogous to RBF networks) as follows:

Gaussian Kernel $k(x, y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right)$

The corresponding decision function given by:

$$D(x) = sign(\sum_{i=1}^{m} \alpha_i \exp\left(-\frac{|X - X_i|^2}{2\sigma^2}\right) + b)$$

Note that the number l of RBFs, the kernel centers, which correspond to the support vectors, and the coefficients α_i are all automatically determined as a result of quadratic optimization.

2.2 Support Vector Regression

To support vector regression, the output of a SVM can be defined as any real value[13], assuming a set of training examples $(x_1, y_1), \ldots, (x_m, y_m)$ is given, where $x_i \in R^n$ is a feature vector. The linear SVM regression function is $D(x) = W \bullet X + b$. Then the optimal regression function can be found by solving the following constrained minimization.

$E(W) = \frac{1}{2}\|W\|^2$, subject to $|D(X_i) - (W \bullet X_i + b)| \le \varepsilon, i = 1, \ldots m$

Solving it results in the optimal linear regression function

$$D(x) = \sum_{i=1}^{m} \alpha_i (X \bullet X_i) + b$$

Additionally, the above formula can be generalizaed to nonlinear regression as follows:

$$D(x) = \sum_{i=1}^{m} \alpha_i k(X \bullet X_i) + b$$

3. Multi-view Face Detection and Pose Estimation in SVM Framework

Our face detection system is depicted in Figure 1. In Figure 1, the subwindow is a small rectangle area of input image which may contain face. We use wavelet transform to extract features of the subwindow under different viewpoint. Then the features are input in 3 view-based SVM classifiers. The raw output of 3 set of SVM classifiers then input in an ensemble SVR to give the final result.

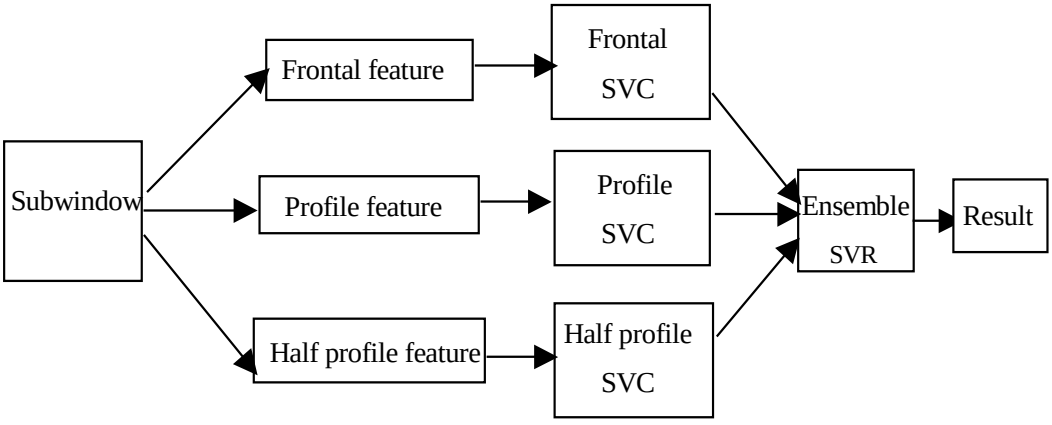

Figure 1. System diagram of the Multi-view Face Detection System

The architecture of each view-specific SVC set is consisted of two layer SVM classifiers. The first layer is several linear SVM classifier. The input feature vector is selected from the different combination of the wavelet transform coefficient. The second layer is a nonlinear SVC which use RBF kernal. The first layer of the SVC give a rough estimation of whether the given region of an input image depict an face or not. Generally speaking, we need train several such linear SVCs using different feature vector. All the detected subwindow from the first layer SVC is inputted into the second layer SVC and made the final decisions. As the second layer SVC is a nonlinear SVC, it will use much more time to decide whether the inputted subwindow of the given image depicted a face or not, by using the first layer SVC, we may expect that the second layer SVC need not search all the image for the candidate subwindow. But only make decisions on the output of the first layer SVC. Thus increase the overall efficiency of the system. Once all the view-specific SVM classifiers have been trained, an ensemble mechanism is brought on-line and the set of feature vectors associated with each respective training image is, in turn, simultaneously input into the appropriate view-specific SVC. The output of each SVM classifier is used to train the ensemble SVM regression. It is noted that the training examples set of ensemble SVR should be different from the training examples set of View-specific SVCs.

The system is now ready to accept prepared input image regions, and to indicate if the region depicts a face, as well as indicating the pose range exhibited by the face. Each input image region extracted in the foregoing manner from any scale version of the input image is abstracted in a way similar to the training images to produce a set of feature vectors, which are, in turn, simultaneously input into the first level SVCs and then to the second level ensemble SVR. For each feature vector set input the system, an output is produced from the ensemble having one active node. The active node will indicate first whether the region under consideration depicts a face, and secondly, if a face is present, into what pose range the pose of the face falls.

3.1 Feature Extraction

Because wavelets are a type of multi-resolution function approximation that allow for the hierarchical decomposition of signal.

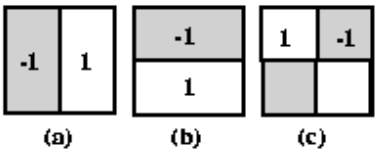

Figure 2. Three types of 2-dimensional Haar wavelets (a)"vertical" (b) "horizontal", (c)" diagonal"

When applied at different scales, wavelets encode information about an image from the coarse approximation all the way down to the fine details. The Haar wavelet are a natural and the simplest set basis functions which encode differences in average intensities between different regions. Harr transform shifts each wavelets by n, the quadruple density transform shifts the wavelet by $\frac{1}{4}2^n$ in each direction, shown in Figure 2. So Haar wavelets of two different scales(4*4 and 2*2) are used to generate a multi-scale representation of the images. At each scale, three different orientations of Haar wavelets are used. In this manner, information about how gray varies in the horizontal, vertical, and diagonal direction are obtained.

The Harr wavelet transform formula is described as follows:

$$\psi_H(x) = N_1(2x) - N_1(2x-1)$$

$$\psi_H(x) = N'_2(2x)$$

The subwindow is clipped to the dimensions 20*20 such that different viewpoint face can be centered and approxiamtely of the same size. The selection of wavelets feature coefficients under different viewpoints is different. This is due to the fact that the important wavelet coefficients that are consistent along the ensemble of face images are comprised of strong response in the coefficients corresponding to the sides of the organs and weak response in the coefficients along the face cheek areas and forehead areas.

3.2 Classifier design

Because faces in images could have different viewpoints. The difference between them is significant. So we have to design view-based SVM classifiers to deal with them separately, and then, the problem is how to combine the classifiers together. Thus the classifier of the system consists of view-based SVM classifiers and ensemble SVM Regression.

3.2.1. View-specific SVC. Generally Speaking, the standard method for N-class SVCs is to construct N SVCs. The *i*th SVC will be trained with all of the examples in the *i*th class with positive labels, and all other examples with negative labels. The output of each SVM classifier is the distance of the test point from the

corresponding decision hyperplane. This distance is a rough measure of how "well" a test point fits into its designated class. SVCs trained in this way can be refer as *1-v-r* SVCs(one-versus-rest).The final output of the N *1-v-r* SVCs is the class that corresponds to the SVC with the highest output value. Unfortunately, there is no bound on the generalization error for the *1-v-r* SVCs.

In our system, the SVCs for each view have been trained respectively. The positive examples are the normalized face images from the corresponding view. The negative examples are selected from images that do not contain faces. It is note that the negative examples of each view-specific SVC do not contain any faces of other view. For example, in order to train the half profile SVC, we use a set of half profile face images as positive examples, and the negative examples contains only non-face patterns, but not face images of frontal view and profile view.

It is due to the face patterns of each view are similar with each other. Generally speaking, they share some common features. While the face patterns and non-face patterns are much different from each other. So for the view-specific SVC, the main task is to distinguish faces of corresponding view and non-face as well as reduce the missing face rate, but not necessary distinguish faces of different view too clearly. That's to say, each view-specific SVM classifier can be trained using the same negative examples set.

3.2.2. Linear SVC to speed-up. From the decision function listed in section 2, it is easy to see that the computational cost of nonlinear SVC is very high. To increase the speed of the detection, a linear SVC is added to each view-specific classifier as a fast face/nonface filter. Therefore, the actual view-specific SVC consists of two layer SVCs, a linear one and a kernel one. The filtering using the linear can discard most of the false alarms with a small computational cost, and significantly reduce the number of subwindows passed to the much more costly nonlinear kernel SVC. This will be illustrated using experiments.

3.2.3. Ensemble mechanism. When we use multiple learning systems to learn to solve a certain problem, the information present to those learning systems are uniform (or near uniform). So, all the learning systems have equal right to "express" their "opinion" on the problem. In his situation, voting could work. On the contrary, when the information present to the learning systems are different to some extent, they only have right to "express" their "opinion" to their special problem, that is, a sub-problem. Thus, we should add the ensemble mechanism to get the final decisions. Accordingly, when we have trained the view-specific SVCs, the other thing we should concern about is to train the ensemble mechanism to fusing the decisions we got from the view-specific SVM classifiers and give the combination result.

As the face examples of other view is not served as the negative examples of current view, so actually each SVC is not trained exclusively, so to some degree, the decision space is overlapped. That is to say, each SVC has the redundancy decision ability, it is possible that the frontal view SVC can give the positive answer when the probe image is a half profile image, so as to other view-specific SVMs. In addition, this redundancy is difficult to formulation, as it is very difficult to lump several feature together due to their diversified forms. In order to fusing all the information from the view-specific SVMs, an ensemble mechanism should be introduced.

In this work, SVM regression is used to do the ensemble task. The raw outputs of the SVM classifiers of each view are fed into the ensemble SVM regression to get the final result. We define 4 distinct values of the SVR output, namely {-1,1,2,3} to represent the non-face, frontal, half profile, and profile faces, respectively.

For example, when a person's frontal face fed into the classifier, the raw output of the view-based SVCs may be [0.5 -0.3 0.3], then the raw vector input in the ensemble SVR, we should get 1, it denotes the subwindow may contains a frontal face. If raw vector is [-0.1 -2 -0.5], the output Ensemble SVR should be -1, here it denotes the subwindow contains no face.

4. Experiments

4.1 Data preparation

A total of approximately 12000 multi-view face examples and 84000 nonface examples are collected. This is by cropping multi-view faces from VCD frames. We select 1/3 of these examples to train the first level view-specific SVM classifiers respectively, the other 1/3 to train the second level ensemble SVM regression, and the last 1/3 to test the whole system. Each of the 3 SVM classifiers is trained based on the face samples of the corresponding view.

4.2 View-specific SVC

Here we trained three view-specific SVM classifiers (corresponding to frontal, half-profile, profile), each SVM classifier is trained based on the face samples of the corresponding view. Each view-specific SVC include a linear SVC and a kernel SVC, the linear SVC will sever as a pre-filter of the face patterns and the nonface patterns to increase the overall efficients of the whole system.

In order to test the filtrating ability of the linear SVC, We run the linear SVC detector on 210 images to get the

filter rate. The experimental result rare described in Figure 3.

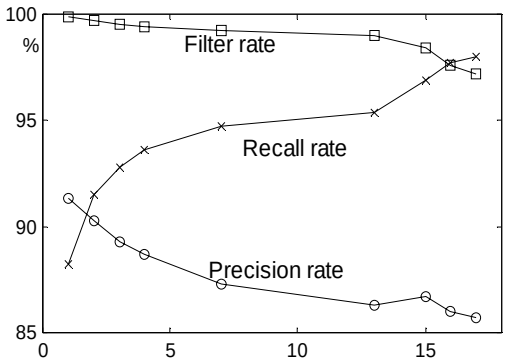

Figure 3. The Classification Ability of Linear SVC

Here the recall rate is the precision for face patterns while the precision rate is for all face and nonface patterns. The filter rate is the rate of sub-windows discarded as nonface pattern by linear SVC to all subwindows to be examined by the combined classified. We run the detector on 210 images to get this filter rate.

From Figure 3 we can see that the recall rate could be improved 11.1% while the precision rate loses 6.1%. The recall rate is even greater than 98.1%. The filter rate is between 96% and 99.5% because most of the nonface patterns could be easily classified from the face patterns. Thus the times to compute the decision function will be decreased greatly.

4.3 Ensemble SVR

In order to train the ensemble SVM regression , we should first extract the corresponding features and feed them to the first level view-specific SVCs and collect all the output from each view-specific SVC, the real values from the first level SVCs would be considered as the input of the second level ensemble SVR.

The comparative table shows that the ensemble SVR can achieve higher detection accuracy and lower false alarm rate than simple voting of the output of the view-specific SVCs, from Table 1, it is very clear that face detection using ensemble SVR can fix more than 4.47% detection errors and gives the right pose information at the same time.

Table 1 Experimental results of Ensemble SVR

POSE	Frontal face	Half Profile	Profile	Nonface	Detection Rate	False Alarm Rate
Method\Test Set	1815	1176	1176	29910	/	/
Vote	1721	983	936	29583	87.35%	1.09%
SVR	1774	1063	989	29709	91.82%	0.67%

4.4 Pose invariant face detection using view-specific SVC + ensemble SVR

At the test stage, when one input image is given, we scan all the image in different scale and get the current sub-window and calculate the feature coefficients of the sub-window, and then feed these feature coefficients to the first level view-specific SVCs for classification and then to the second level ensemble SVR for fusing. At last, we can get whether the sub-window depict a face or not, if the answer is yes, the pose information can also be provided.

We Collected a test set for benchmarking multi-scale face detection performance for out of plane rotation. This test set consists of 50 images with 146 multi-view faces that vary from frontal to side view. We gathered these images from VCD. Of these faces 129 faces are detected with correct pose estimation report, 17 false alarms.

Figure 4 shows some results of multi-view face detection and pose estimation.

5. Conclusion

The main strength of the present method is the ability to detect the location and estimate the pose of the face at the same time. To do this, we first train several view-specific SVM classifiers according to each view range, and then, an ensemble mechanism (SVM Regression) is introduced to combine the decisions we got from the view-specific SVM classifiers and made the final decisions.

To increase the detecting speed, we designed a two layer SVM classifier as the view-specific SVM classifier. The first layer of linear SVCs serves as a filter that can discard many background subwindows in the search procedure.

The experimental results also show that our ensemble architecture is effective. It is note that the ensemble mechanism can achieve high detection accuracy, more than 4.47% better than the results of the separate SVM classifiers. In addition, it can also provide the pose information of the corresponding faces.

Figure 4. Multi-scale face detection and pose estimation result

6. References

[1] S. Baker, S. Nayar, and H. Murase. Parametric feature detection. IJCV, 27(1): 27-50, March 1998.

[2] H. Murase and S. K. Nayar. "Visual learing and recognition of 3-D objects from appearance". International Journal of Computer Vision, 14: 5-24, 1995.

[3] S. Nayar, S. Nene, and H. Murase. Subspace methods for robot vision. RA, 12(5): 750-758, Oct. 1996.

[4] A. Pentland, B. Moghaddam, and T. Starner. "View-based and modular eigenspaces for face recognition". In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 84-91, 1994.

[5] A. R. Pope and D. G. Lowe. "learning appearance models for object recognition". In Object Representation in Computer Vision, 201-219. Springer-Verlag, 1996.

[6] S.Gong, S. McKenna, and J. Collins. An investigation into face pose distributions. In IEEE int. Conf. On Face& Gesture Recognition, pages 265-270, Vermont, 1996.

[7] J. Ng and S. Gong. Performing multi-view face detection and pose estimation using a composite support vector machine across the view sphere. In *Proc. IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 14-21, Corfu, Greece, 26-27 September 1999

[8] E. Osuna, R. Freund, F. Girosi, "Training Support Vector Machines: an Application to Face Detection." Proc. CVPR'97, June17-19, Puerto Rico.

[9] Y.M. Li, S.G. Gong, and H. Liddell, Support Vector Regression and Classification Based Multi-view Face Detection and Recognition. In IEEE int. Conf. On Face & Gesture Recognition, pages 300-305, France, 2000.

[10] H. Schneiderman. A Statistical Approach to 3D Object Detection Applied to Faces and Cars, CMU-RI-TR-00-06.

[11] C. Cortes and V. Vapnik. Support vector networks. Machine Learining, 20:273-297, 1995.

[12] V. N. Vapnik. Statistical learing theory. John Wiley & Sons, New York, 1998.

[13] A. Smola, B. Scholkopf, and K. R. Muller. General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, Proc. Of the Ninth Australian Conf. On Neural Networks, pages 79-83, Brishane, Australia, 1998.