TWO-STAGE CLASSIFICATION USING SELECTIVE ATTENTION FOR FAST FACE DETECTION

*Steve R. Jones, David W. Capson**

Department of Electrical and Computer Engineering
McMaster University
Hamilton, Ontario, Canada, L8S 4K1
email: {jonessr, capson}@mcmaster.ca

ABSTRACT

A novel system for face detection in images and video sequences is presented. The system incorporates a two-stage linear discriminant and nonlinear support vector machine classifier coupled with a front-end biologically-inspired search scheme. Results based on the CMU test set demonstrate that by using such a classifier arrangement with a non-exhaustive searching scheme, a significant reduction in computational complexity is achieved while maintaining comparable accuracy to other leading face detection systems.

1. INTRODUCTION

Support vector machine (SVM) classification has recently been demonstrated as a valuable tool for face recognition and detection in computer vision. SVMs are capable of systematically learning the complex non-linear decision boundaries from a given sparse training set, and have been successfully applied in many face processing tasks such as face recognition, pose discrimination, and face detection.

Such encouraging empirical results obtained from these experiments can be, in part, attributed to the SVMs ability to learn a decision function under conditions where estimating the parameters of a probability density model for objects in high-dimensional image space would be otherwise difficult. Furthermore, SVMs perform structural risk minimization on sparse training data in order to maximize generalization to novel test examples, which, theoretically is superior to other non-parametric learning algorithms based on empirical risk minimization (such as bayes classifier, ANN's) [1].

1.1. Related Work

Over the past decade, a wide range of face detection systems have emerged from well founded algorithms including template matching and active shape models as well as statistical learning methods such as PCA based classification [2], and neural networks [3]. In recent years, image-based methods involving statistical learning approaches such as SVMs and ANNs have proven their effectiveness from the reported high percentage detection rates (most over 90%) and have absorbed much of the attention in face detection research. Rowley et al. [3] created a neural network based face detector which has become an apparent benchmark for other face detection systems. A number of SVM-based face detection systems have also been proposed. Osuna et al. [4] created a single SVM classifier trained on 19x19 face and non-face patches obtained from bootstrapping. A SVM ensemble-based classifier for face detection was used in in [5].

Many of these proposed techniques exhaustively search the image space in a brute-force manner by sliding a search window through the image at multiple scales. Most techniques that have emerged to speed up searching by focusing attention to "active" image regions use system available cues such as color and motion (when available) to direct attention. The more challenging problem of selective attention in grayscale images is considered here.

This paper overviews a method for improved frontal-view face detection over a single SVM system. A two-stage face detector is described which speeds up classification by using a simple, fast linear classifier for the majority of "easy" patterns, and invokes a more complex but accurate nonlinear classifier when required. Rather than using a sliding window approach, the classifier invokes a front-end selective attention server which predicts regions in an image attracting visual attention. Using this approach, a much faster classifier needs to examine only a small subset of the total number of possible image windows.

Section 3 gives an overview of the method and experimental results are shown in Section 4.

2. BACKGROUND

2.1. Support Vector Machines

Given a data set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ of m examples $\mathbf{x}_i \in \mathbb{R}^d$ where d denotes the dimensionality of the data points with labels $\mathrm{y}_i \in \{-1, 1\}$, and a kernel function $K(\mathbf{x}, \mathbf{x}')$, the SVM is formed by solving the convex quadratic programming problem:

$$\begin{aligned} \max \quad & w(\alpha) = \sum_{i=1}^{m} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C, i = 1, 2, ..., m \end{aligned} \tag{1}$$

Where C controls the weight of the classification errors ($C = \infty$ in the separable case). The points $\mathbf{x}_i$ which correspond to non-zero α_i are the support vectors which represent the significant test points making up the decision boundary. The classification func-

Financial support from Natural Sciences and Engineering Research Council of Canada (NSERC) is acknowledged.

tion is given by:

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \quad (2)$$

and the bias term

$$b = -\frac{1}{2}\sum_{i=1}^{m} \alpha_i y_i \left[K\left(\mathbf{x}_i, \mathbf{x}_r\right) + K(\mathbf{x}_i, \mathbf{x}_s)\right] \quad (3)$$

is computed using any support vectors $\mathbf{x}_r$ and $\mathbf{x}_s$.

2.2. Fisher Linear Discriminant Classifier

The classic fisher's linear discriminant function of the form $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ is defined for the data $\{\mathbf{x}_i, y_i\}_{i=1}^m$ such that the between class scatter matrix

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \quad (4)$$

of the projected data is maximized and the within class scatter matrix

$$S_W = \sum_{\mathbf{x}\in C_1} (\mathbf{x} - \mu_1)(\mathbf{x} - \mu_1)^T + \sum_{\mathbf{x}\in C_2} (\mathbf{x} - \mu_2)(\mathbf{x} - \mu_2)^T \quad (5)$$

is minimized, where μ_i is the mean vector of class C_i. The solution to the problem of maximizing the criterion function

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (6)$$

can be obtained using matrix inversion $\mathbf{w} = S_W^{-1}(\mu_1 - \mu_2)$[6].

3. IMPLEMENTATION

3.1. 3.1 Two Stage Classifier

The nonlinear SVM classification algorithm is limited by the computational burden incurred while calculating the decision function, making real-time SVM based vision systems hard to achieve.

A two-stage combined FLD/SVM classifier model is presented here which improves overall performance over a single SVM-based detection system in terms of combined speed and accuracy. Under most circumstances, the FLD classifier is computationally more efficient but less accurate than the nonlinear SVM classifier. The advantages of this speed-accuracy tradeoff are combined in the following manner. Given an unknown input $\mathbf{x}$, the FLD classifier produces a value for the decision function $f(\mathbf{x}) \lessgtr 0$. A more strict set of thresholds T_a and T_b are defined so that if classification of the unknown pattern $\mathbf{x}$ falls into the region $f(\mathbf{x}) \in [-T_a, T_b]$, the pattern is labelled ambiguous and is then passed into the second stage SVM classifier. The values of T_a, T_b are application dependent and reflect the speed/accuracy tradeoff of the system. The classifier rule is therefore:

$$f(\mathbf{x}) = \begin{cases} sgn\left(\mathbf{w}^T\mathbf{x} + b\right) & \text{if } \mathbf{w}^T\mathbf{x} + b \notin [-T_a, T_b] \\ sgn\left(\sum_{i\in SV} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right) & \text{otherwise} \end{cases} \quad (7)$$

This process consists of an initial quick detection phase by the FLD classifier followed by a more precise SVM phase. Using the FLD, the linear classifier always has a fixed number of computations O(d) (where d is the dimension of the training space) making it fast and ideal for classification of unambiguous patterns.

With this arrangement, the initial FLD classifier performs as a linear pre-filter for the SVM stage. Therefore, training the non-linear SVM classifier is simplified since it is only trained on the examples in the ambiguous region as well as some additional bootstrapped images, which simplifies the learning stage and produces a smaller set of support vectors. This in turn speeds up training, since the complexity of non-linear SVM classification is a direct function of the number of support vectors.

3.2. Training

The training data set used in these experiments was built from the MIT-CBCL frontal face database [7] consisting of over 400 male and female subjects cropped and centered in the image while posing under varying lighting conditions and facial expressions. Figure 1 shows a few sample images from the dataset. To improve generalization of the classifier for handling minor variations in frontal and upright pose, each face was transformed via small reflections, scaling, rotations and translations, expanding the final training set to 4858 19x19 grayscale face images.

Each training image was first convolved with a quasi-elliptical binary mask removing some of the pixels lying close to the boundary of the window pattern which introduce noise into the training process. Histogram equalization was then applied to each image compensating for variations in illumination brightness. These preprocessing steps are similar to those used in [3]. Finally, the face image space was reduced from 339 to 30 dimensions via PCA.

Fig. 1. Sample images from the MIT-CBCL training data set

An RBF kernel function was used for SVM learning. The value for parameters C and γ were found by using a coarse grid search in the (C,γ) space to find

$$\max_{C,\gamma} \text{ CVA} \quad (8)$$

where CVA is the cross-validation accuracy defined as the percentage of data correctly classified (true positives and true negatives) using 3-fold cross-validation. The training set is randomly split into 3 subsets of equal size. Successively, one subset is tested using the classifier trained on the other 2 subsets giving a total of $\binom{3}{1}$ training instances. Each part of the whole training set is predicted once by the classifier and the results are aggregated. The number of retained eigenvalues k can also be chosen using eq. (8) by adding k to the list of optimization parameters.

To further improve generalization of the non-face class, non-face samples were collected though a bootstrapping procedure similar to [8], and [4]. A set of random image patches from a natural scene were used to generate the non-face samples. 10000 non-face samples were first applied to the linear classifier. The resulting false positives and ambiguous patterns were appended to the intial training set and then passed to the second stage SVM classifier. Any false positives collected from the second stage were appended

to the SVM training set. The classifiers were then retrained using their newly appended non-face images. This procedure was iterated a total of three times.

From the final initial training set, 342 faces and 547 non-faces fell into the ambiguous region of the linear classifier and were used to train the non-linear SVM giving a total of 269 support vectors. Note that when the SVM classifier was trained and used independently on the training data, between $800 - 2500$ support vectors were present depending on system parameters.

3.3. Image Search Approach

Instead of using a global multiscale sliding window technique to search for face candidate regions, a system developed by Itti et al. [9] is used to predict potential face regions in a serial fashion. The system linearly combines a set of adjustable weighted channels (color, intensity, orientation, motion) and performs a center-surround operation between several pairs of scales in order to construct a master saliency map of the image. A winner-take-all neural network then uses the map to serially select output targets for the image. Figure 2 shows the overall architecture.

Fig. 2. Selective attention model architecture (based on [9])

Note that saliency in an image is the degree of conspicuousness (or a measure of visual attractiveness) that a feature exhibits in the presence of its surroundings. This notion is used to make the assumption here that face objects are generally distinguishable from their surrounding background. Furthermore, since this model uses simple cues in the visual cortex such as intensity, color and orientation, it does not make any inference about high level object targets that it is focusing on. To this extent, it applies as a general predictor of regions of visual attention. In the experiments presented here, grayscale images were used. Therefore, the color and motion channels were given a weighting of zero.

The CMU test set [3] was used to measure the selective attention model's ability to identify correct face regions in images. Table 1 shows results for the test set A-C containing 130 images with 507 total unoccluded faces. The top two rows use 10 attention shifts per image while the bottom two rows use 30. The best results are obtained from images containing less than 5 faces per image. From only the first 10 targets, nearly 78% of the targets land on the faces in the images (within 25 pixels horizontally and vertically), with an average computation time of 2425ms for the tenth target. Using 30 targets, approximately 84% of the targets landed on all faces in the images with an average computation time of 3132ms for the 30th target.

Two example images taken from the CMU face test set and processed by the selective attention model are shown in figure 3. Only the first few target coordinates are shown. The resulting targets are overlaid on the images as large circles, with arrows indicating sequence of targets acquired. Note that in the right image the set of face targets is not exhaustive and does not pick out all faces in the image in the first 7 targets. In contrast to the right image, clearly as more faces are present in an image, the less salient each individual face becomes since they are competing for visual attention.

	on face			within 25 pixels		
	faces found	total faces	%	faces found	total faces	%
images with $1-4$ faces	83	127	65.3	99	127	77.9
all images	148	507	29.2	192	507	37.8
images with $1-4$ faces	91	127	71.6	107	127	84.2
all images	157	507	30.9	204	507	40.2

Table 1. Selective attention results for CMU test set [3] using 10 and 30 attention shifts per image.

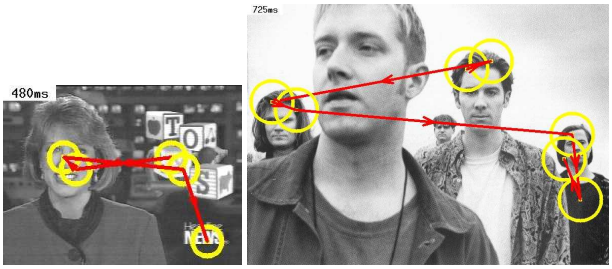

Fig. 3. sample images from CMU test set after being processed by the front-end selective attention system.

The face detection system is organized in a client-server framework over TCP/IP using wireless 802.11b in which the client (two-stage classifier) requests potential target locations from the selective attention server after transmitting the image. The client responds with a set of points (x, y) indicting spots of potential face regions for the classifier to inspect. Since the classifier can work independently of the selective attention model once it has been given at least one target point, the model can operate in a parallel fashion.

A limited range sliding window (25 pixels horizontal and vertical) at multiple scales is then applied to each received target point. A pyramid of images is created by repeatedly subsampling the window creating 10 levels from 19x19 up to 190x190 in size. A similar technique was done in [3] for example. As described, usually true faces will give correct classification in 2 or 3 consecutive scales and locations. This notion is used to help drastically reduce false positives occurring in the image and confirm correct face regions. A 3x3x3 neighborhood filter is used to remove isolated detections by only retaining detections with multiple hits within a 3x3 region and within 3 scales in the image pyramid. A visual overview of the face detection system is shown in figure 4.

4. EXPERIMENTAL RESULTS

The CMU test set [3] was used to evaluate the performance of this system. For each image, the selective attention server found 30 targets. For each target, the two-stage classifier in eq. (7) examined $187,500$ 19x19 windows using 10 scales and a region of 25x25 pixels around each target.

A detection rate between 74.2% and 88.7% was achieved based on a small number of false detections ($< 10^{-4}$) depending on system parameters. On average, using $T_a = T_b = 0.00035$ (see

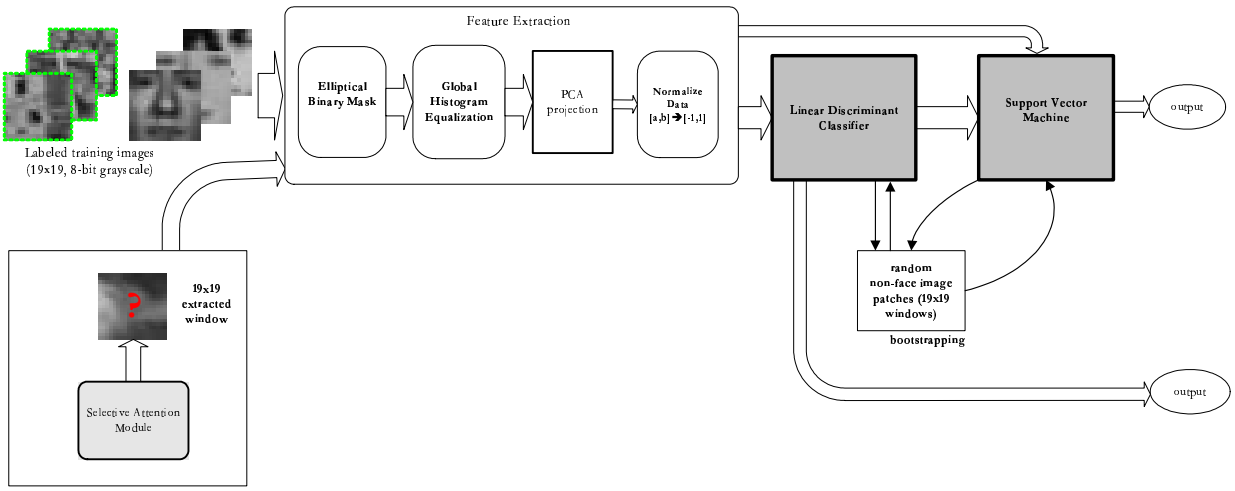

Fig. 4. System overview of face detection system.

eq. (7), k=30, $C = 2000$,$\gamma = 0.35$, the SVM classifier was invoked $\sim 58\%$ of the time. Figure 5 shows an ROC curve based on these parameters. The results show how overall accuracy of the two-stage classifier compares to a single SVM based classifier while providing a reduced complexity advantage.

While the detection results are comparable to other published face detection systems such as [3], the added gain lies in the reduced complexity in searching and classification. Table 2 shows that whereas Rowley et al.'s [3] system examined $83,099,211$ 20x20 pixel windows for the CMU test set, only $24,375,000$ 19x19 windows are examined in this system.

An extension to video sequences is currently being examined by integrating a motion channel into the selective attention process to further improve the searching strategy.

	# Windows Examined	Detection Rate ($< 10^{-4}$ FP%)
Proposed system	$24,375,000$	$74.2 - 88.7\%$
Rowley et al. [3]	$83,099,211$	$77.9 - 90.3\%$

Table 2. Performance of face detection system based on CMU test set.

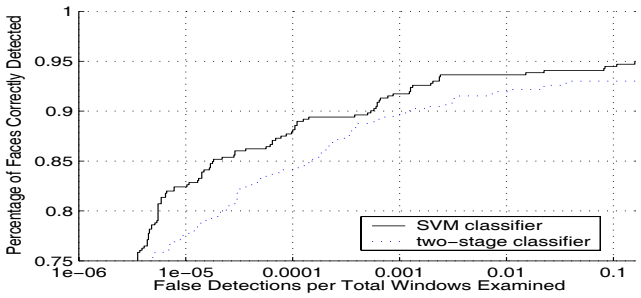

Fig. 5. ROC curve comparing two-stage and SVM classifers based on CMU test set A,B and C.

5. CONCLUSION

A combined two-stage classifier coupled with a front-end selective attention system has been shown to reduce the computational complexity of face detection in images. And, by using the selective attention module to select quality targets, a non-exhaustive search of the image can be performed yielding overall detection results comparable to other face detection systems.

6. REFERENCES

[1] V Vapnik, *The nature of statistical learning theory*, Springer, New York, 1995.

[2] M. Turk and A.P. Pentland, "Eigenfaces for recognition," *CogNeuro*, vol. 3, no. 1, pp. 71–96, 1991.

[3] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *PAMI*, vol. 20, no. 1, pp. 23–38, January 1998.

[4] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *CVPR97*, 1997, pp. 130–136.

[5] Hong-Mo Je, Daijin Kim, and Sung Yang Bang, "Human face detection in digital video using svmensemble," *Neural Process. Lett.*, vol. 17, no. 3, pp. 239–252, 2003.

[6] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New-York, 1973.

[7] "Cbcl face database #1," *MIT Center For Biological and Computation Learning http://www.ai.mit.edu/projects/cbcl*.

[8] Kah-Kay Sung and Tomaso Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.