To appear in the Proceedings of the 11th Scandinavian Conference on Image Analysis (SCIA'99), Kangerlussuaq, Greenland, June 7th – 11th, 1999

Facial Features detection by Saccadic Exploration of the Gabor decomposition and Support Vector Machines

F. Smeraldi, N. Capdevielle and J. Bigün
Microprocessor and Interface Laboratory
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne

Abstract

Facial features detection is of primary importance to face authentication and recognition. In this paper, we present an attention–driven approach to eyes and mouth detection inspired by the human saccadic system. The algorithm is centred around a log–polar retinotopic grid that is used to sample the Gabor decomposition of the image. Detection is achieved by displacing the grid according to a saccadic pattern. Saccade planning is performed using eye and mouth models implemented by means of Support Vector Machine classifiers.

1 Introduction

Facial features detection is of primary importance to face recognition and authentication. The problem is particularly challenging because of the high variability in feature appearance due to changes in pose and expression as well as to the presence of artifacts, such as eyeglasses. In this paper, we present a bio–inspired approach to eyes and mouth detection based on a model of saccadic eye movements. Central to the algorithm is a log–polar retinotopic grid, which is used to sample the Gabor decomposition of the image being processed [2]. The grid is displaced in "jumps" (saccades) to the facial features of interest, emulating the way human subjects explore a scene [3, 9]. Saccade planning relies on coarse appearance–based models of the search targets, consisting in Support Vector Machine classifiers trained to recognise a specific facial feature encoded as a vector of Gabor responses. This attentional mechanism allows the system to collect a number of candidates for the looked–for features, thus emulating the human saccadic system in the selection of task relevant information [7, 8].

Further processing relies on a more detailed model of each facial feature implemented by means of another SVM classifier. This is used to improve the precision of the localisation as well as to rate the different candidates. Expected probabilistic geometrical relations between the eyes and the mouth are also used in the final assessment.

2 Modelling the facial features

Our saccadic search strategy requires the use of two levels of modelling for each of the targets [10]. The coarser, or pointwise, model is obtained for each facial feature from the vector of Gabor responses extracted at the location of that feature in the images of the training set. Each vector consists of thirty filter responses arranged in five frequency channels. The finer, or global model is obtained by placing the whole retina at the location of the facial feature on the training images (figure 1) and collecting the set of Gabor filter responses from all of the retinal points.These Gabor features are then arranged into a single vector.

In both cases, the example vectors for each facial feature are complemented with negative examples and used to train a Support Vector Machine classifier [1, 13]. In our experiments we have been using a training set consisting in 202 frontal images from the M2VTS database. This allowed us to obtain 202 positive examples for each facial feature. Negative examples were collected by extracting Gabor vectors (or, re-

Figure 1: The retinotopic sampling grid placed on a person's eye for model creation.

spectively, positioning the retina) at 10 random locations in each training image. Also, right eyes and mouths have been included as negative examples in the left eye models, and so on. The SVM engine we have used was developed following the ideas in [4] and [6].

The training set is also used to estimate the parameters of the probability distributions of the position of each facial feature relative to one of the other two (we assumed the distributions to be Gaussians).

3 The saccadic search strategy

Facial feature detection begins with the retinotopic sampling grid randomly placed on an image. Search is initially aimed at one randomly selected feature among the three possible targets. A Gabor feature vector $\vec{v}_i$ is extracted at each retinal point i and the SVM classifier encoding a pointwise model of the selected target is used to rate the retinal points with the classifier output

$$f_{pwise}(\vec{v}_i) = \sum_j \alpha_j y_j K(\vec{s}_j, \vec{v}_i) + b, \qquad (1)$$

where K is a (generally nonlinear) kernel function and $\vec{s}_j$ are the support vectors. The retina is subsequently centred at the position that maximises the score. This procedure is iterated until the centre of the retina obtains the best score, meaning that the sampling grid is centred on a local maximum. Note that one advantage of this search strategy is that, due to the topology of the retina, the search automatically becomes finer as a local maximum is approached.

After saccades have converged to a local maximum, the retinotopic grid is displaced in a pixel-by-pixel fashion to maximise the output of the more accurate global SVM model for the detected facial feature candidate. The maximum score output by the global model is checked against a rather relaxed threshold to avoid candidate proliferation, and the feature candidate along with its score is pushed onto a stack. After that, a saccade is performed to the most probable location of the next facial feature (in a cyclic fashion), and an attempt at detection is made directly with the corresponding global model. If this succeeds, another saccade is performed based on the probability distribution to reach the third facial feature; otherwise, saccadic search is restarted at random to look for the feature for which the fewest candidates have been detected.

Saccadic search is continued until a complete set of facial features is found which has a very high score (see section 4) or until the retina has been centred on the maximum allowed number of different locations.

4 Selecting the best feature set

The final decision on which feature candidates constitute the best facial feature set is based both on the individual score and on the relative position of the candidates. The left eye, right eye and mouth candidates present in the stack are grouped in all the possible ways and are assigned a score according to the expression

$$\text{Score}(i,j,k) = \mathcal{S}(i)\mathcal{S}(j)\mathcal{S}(k)\mathcal{P}(i,j,k), \qquad (2)$$

where $\mathcal{S}$ is a function of the score of the single facial feature and $\mathcal{P}$ depends on their relative location on the image. Function $\mathcal{S}(i)$ is computed, for each candidate feature i, by transforming the output $f_{global}(i)$ of the corresponding global SVM model through a sigmoid function

$$\sigma_f(x) = \frac{1}{2}\left(\tanh\left(\gamma_f(x - c_f)\right) + 1\right). \qquad (3)$$

The position term $\mathcal{P}(i,j,k)$ is obtained by multiplication of three factors, each one of the form $\sigma_P(P_{ij}(\vec{x_{ij}}))$, where $P_{ij}(\vec{x_{ij}})$ is the probability that facial feature j is found at location $\vec{x}_{ij}$ relative to facial feature i (that should precede it in clockwise order on the image) and σ_P is a sigmoid function analogous to (3). The parameters of σ_P are chosen so that candidate features that deviate more than twice the standard deviation from their expected relative positions are penalised.

If none of the possible feature sets gives a satisfactory score, saccadic search is restarted from scratch for a second attempt. Sets of fewer than three facial features are considered if no candidates have been detected for one or more of the search targets.

5 Testing the models

We assessed the performance of the pointwise and global facial feature models by feeding into each classifier the appropriate set of Gabor responses extracted at the location of the facial feature it is supposed to recognise, together with negative examples it is supposed to reject. The negative examples included, for each classifier, the remaining two facial features plus ten random points per image. The test set consisted in the remaining 349 images from the M2VTS database; none of the subjects represented in the training set appeared in the test set. The main purpose of this preliminary test phase was to allow the choice of an optimal kernel function for the SVM classifiers.

Due to the wider availability of negative examples during the training process, all classifiers showed a tendency towards false rejection. In order to compare the performance obtained with different kernels, we modified the standard SVM decision function, given by the sign of (1), with the introduction of a threshold τ:

$$\sum_j \alpha_j y_j K(\vec{s}_j, \vec{v}_i) + b \gtrless \tau \qquad (4)$$

We then adjusted the threshold to determine the equal error rate, the corresponding value of the threshold and the value of τ at which the false rejection goes to zero. The dependence of these quantities on the kernel type is displayed in tables 1, 2 and 3 respectively. It must be emphasised that, since our algorithm doesn't make use of a hard threshold, but all the candidates are allowed to compete, the behaviour of the τ constant is of greater interest to us than the precise value of the EER. A small value of τ both at the EER and at the point where the false rejection becomes zero indicates that the outputs of the SVM classifier are "nicely grouped". Also, it is of interest to our algorithm that the classifiers trained on the three facial features have approximately the same behaviour. As evidenced by tables 2 and 3, these requirements are best met by classifiers based on a fourth order polynomial kernel. We therefore employed this type of kernel for the feature detection experiments described in the following section.

6 Experimental results

The full saccadic search feature detection algorithm was tested on 349 images from the M2VTS database, portraying different subjects from those present in the training set. Several subjects in the database wear eyeglasses, even with strong reflections, or have their eyes shut. Images were down-sampled to a 175×143 pixel resolution (256 grey levels). The retina employed consisted of 50 points arranged in 5 circles. We allowed the system to centre the retina at a maximum of 1000 different locations and to iterate the search 4 times at most. In a typical situation the system can perform eye and mouth detection correctly after a single trial consisting of about 100 displacements of the retina. Figure 6 shows successful detections. The right eye is not detected in one of the test images (349 total), while in 4% of them the location of one or more features is imprecise (figure 6). In 91% of the images both the eyes and the mouth are localised within a 3 pixel tolerance, i.e. one tenth of the typical inter–eye distance. Results on facial features detection on the M2VTS database by another research group have been published in [5], where an 86.5% success rate for simultaneous detection of eyebrows/eyes, nostrils/nose and mouth over 37 frontal images from the database is reported.

Although the results of our experiments seem to fall somewhat short of the optimal model performance reported in the last column of table 1, one must keep in mind that the difficulty of the problem is now considerably increased. This difficulty seems to be confirmed by others using the same database [5]. On one hand, the system is now required to find its own facial feature candidates without exploring the whole image; on the other hand, no particular threshold is adjusted to bring each classifier to its optimal working point; finally, the number of negative examples is now potentially enlarged to all of the image pixels that do not belong to the looked–for features.

EER	Kernel type			
Pointwise Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	11%	11%	11%	**10%**
Right eye	13%	11%	11%	**11%**
Mouth	6%	7.5%	7.5%	**7%**
Global Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	0.5%	0.3%	0.3%	**0.3%**
Right eye	0.8%	0.5%	0.4%	**0.3%**
Mouth	2%	1.8%	1.7%	**0.3%**

Table 1: The table displays the value of the EER of the pointwise and global models as a function of the kernel type employed. Notice how especially the global mouth model benefits from a higher order kernel.

τ at EER	Kernel type			
Pointwise Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-5.0	-2.0	-2.0	$\mathbf{-1.9}$
Right eye	-7.0	-2.3	-2.0	$\mathbf{-1.9}$
Mouth	-3.5	-1.8	-1.7	$\mathbf{-1.6}$
Global Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-0.30	-0.15	-0.05	$\mathbf{-0.03}$
Right eye	-0.43	-0.30	-0.25	$\mathbf{-0.22}$
Mouth	-0.90	-0.82	-0.80	$\mathbf{-0.25}$

Table 2: The table shows the value of τ at the EER as a function of the kernel type, for both the pointwise and the global models. The value of τ steadily increases with kernels of higher order, meaning that the tendency towards false rejection is reduced.

τ at FR $= 0$	Kernel type			
Pointwise Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-14.0	-5.5	-5.0	$\mathbf{-5.0}$
Right eye	-19.0	-8.0	-7.0	$\mathbf{-6.5}$
Mouth	-8.0	-9.0	-8.0	$\mathbf{-6.0}$
Global Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-0.5	-0.2	-0.1	$\mathbf{-0.2}$
Right eye	-0.7	-0.6	-0.5	$\mathbf{-0.5}$
Mouth	-1.4	-1.3	-1.2	$\mathbf{-0.4}$

Table 3: The table shows the maximum value of τ for which FR $= 0$ as a function of the kernel type, for both the pointwise and the global models. Although this measure is sensitive to the presence of outliers, its marked tendency to increase with the kernel order indicates that the fluctuations of the classifiers' outputs become smaller with an higher order kernel.

Figure 2: Examples of successful eyes and mouth detection.

Figure 3: The subject's right eye was not detected in the upper frame left. The other images show the worst examples of imprecise detections.

7 Conclusions

We have described a facial features detection algorithm based on saccadic sampling of the Gabor decomposition. Support Vector Machines have been used to drive the saccades and to accept or reject the detected features. The system shows good robustness to the presence of eyeglasses, even with strong reflections, as well as to variation in the aspect of the features. The experimental results presented are however largely preliminary, and further tests will be performed on the extended version of the M2VTS database, featuring images[1] from 300 subjects. Although we believe our saccadic search strategy to be of wider applicability, as has been demonstrated in [12], we focused on facial features detection because of its relevance to the face authentication and recognition problems. Future work will include integration with the face verification algorithm described in [11], which uses the Gabor responses extracted in the region of the eyes to authenticate faces.

Acknowledgement

This research has been supported by the European Community IT-LTR Project VIRSBS.

References

[1] C. Cortes and V. Vapnik. Support–Vector Networks. *Machine Learning*, 20:273–297, 1995.

[2] F.Smeraldi, O. Carmona, and J. Bigun. Saccadic search with Gabor features applied to eye detection and real-time head tracking. *Accepted to the journal: Image and Vision Computing*, 1998.

[3] D. Hubel. *Eye, brain and vision.* Scientific American Library, 1988.

[4] T. Joachims. *Making Large-scale SVM Learning Practical*, chapter 11 of Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998. Eds. B. Schölkopf, C. J. C. Burges, A. J. Smola.

[5] C. Kotropoulos and I. Pitas. Rule–based face detection in frontal views. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume IV, pages 2537–2540, April 1997.

[6] E. Osuna, R. Freund, and F. Girosi. Improved training algorithm for Support Vector Machines. In *Proceedings of IEEE NNSP'97*, pages 276–285, September 1997.

[7] J. B. Pelz. *Visual representations in a natural visuo-motor task.* PhD thesis, Carlson Center for Imaging Science, Rochester Institute of Technology, 1995.

[8] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in visual cognition: a computational study. Technical Report 97.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, 1997.

[9] J. D. Schall, D. P. Hanes, K. G. Thompson, and D. J. King. Saccade target selection in frontal eye field of macaque. I. Visual and premovement activation. *The Journal of Neuroscience*, 15(10):6905–6918, 1995.

[10] F. Smeraldi and J. Bigun. Facial features detection by saccadic exploration of the Gabor decomposition. In *International Conference on Image Processing, ICIP-98, Chicago, October 4-7*, volume 3, pages 163–167, 1998.

[11] F. Smeraldi, N. Capdevielle, and J. Bigun. Face authentication by retinotopic sampling of the gabor decomposition and support vector machines. In *Audio and Video based Person Authentication - AVBPA99 (Accepted)*, 1999.

[12] F. Smeraldi, O. Carmona, and J. Bigun. Real–time head tracking by saccadic exploration and gabor decomposition. In A. T. Almeida and H. Araujo, editors, *Proceedings of the 5th International Workshop on Advanced Motion Control*, volume IEEE Cat. Num. 98TH8354, pages 684–687. IEEE Service Center, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331, USA, 1998.

[13] V. N. Vapnik. *The nature of statistical learning theory.* Springer–Verlag, 1995.

[1] Approximately 2400 frontal images taken at four different periods.