Face Components Detection using SURF Descriptors and SVMs

Donghoon Kim
Trinity College Dublin
Dublin, Ireland
donghook@cs.tcd.ie

Rozenn Dahyot
Trinity College Dublin
Dublin, Ireland
Rozenn.Dahyot@cs.tcd.ie

Abstract

We present a feature-based method to classify salient points as belonging to objects in the face or background classes. We use SURF local descriptors (Speeded Up Robust Features) to generate feature vectors and use SVMs (Support Vector Machines) as classifiers. Our system consists of a two-layer hierarchy of SVMs classifiers. On the first layer, a single classifier checks whether feature vectors are from face images or not. On the second layer, component labeling is operated using each component classifier of eye, mouth, and nose. This approach has the advantage about operating time because windows scanning procedure is not needed. Finally, this system performs the procedure to apply geometrical constraints to labeled descriptors. We show experimentally the efficiency of our approach.

1. Introduction

Thanks to the increasing computational power of computers, systems performing surveillance are getting more and more intelligent in combining and integrating dedicated computer vision based approaches and machine learning classification schemes. Automatic face detection and recognition are essential tasks in these surveillance systems, but also in a wide range of other applications (e.g. e-learning, teleconferencing, entertainment, indexing video libraries, etc.). For that reason, a lot of research efforts have aimed at detecting human faces in visual streams. Recent developments in object (e.g. face) detection or recognition involves the usage of local informative descriptors such as Haar wavelets [11], SIFT (Scale Invariant Feature Transform) [1] and SURF [5]. Using local descriptors versus global ones usually insures the system a certain natural robustness to partial occlusion. Moreover, adequate normalization of these descriptors allows them to be invariant to some transformations such as rotation, scale changes or illumination. These are interesting properties for detecting an object appearing with different scale or orientation in images.

Once the sets of representative descriptors are available for training both the target object class and its complement (non-object), detection is performed by classifying new observations between those two classes. Boosting and SVMs are classifiers that have been applied to face detection and have provided comparable results. However, there are still several challenges to deal with in order to get a reliable face detector. Low resolution images, partial occlusion, variation in lighting conditions or head-pose changes are all difficulties to overcome. As the environment becomes more complex, the procedure of reliable feature extraction becomes more important than the performance of classifiers. In particular, in the various and complex environment, it is necessary to extract salient features which are able to steadily discriminate each different class (e.g. face, non-face).

In this paper, we propose a feature-based method to classify salient points in between two classes: face or background (non face). We use SURF descriptors [5] to generate informative feature vectors and use SVMs as classifiers. Our system consists of a two-level hierarchy of SVMs classifiers. On the first level, a single classifier checks whether feature vectors are from face images or not. On the second level, component labeling is operated using component classifiers of eye, mouth, and nose. This approach is fast since no additional window scanning is needed. We show experimentally how our system performs with changes in the resolution of the images.

2. Related work

Yang et al. have proposed a survey of research on face detection up to early 2001 [12]. They classify face detection methods into four categories such as knowledge-based, template matching, appearance-

based methods, and feature invariant approaches. The first is knowledge-based methods to use simple rules to describe the features of face and their relations, but it is difficult to translate human knowledge into well-defined rules. The second method is feature invariant approaches to detect facial features including eyes, eyebrows, and nose by edge detectors and then infer the face presence. However, illumination and image noise problems have a great influence on the features. The third is template

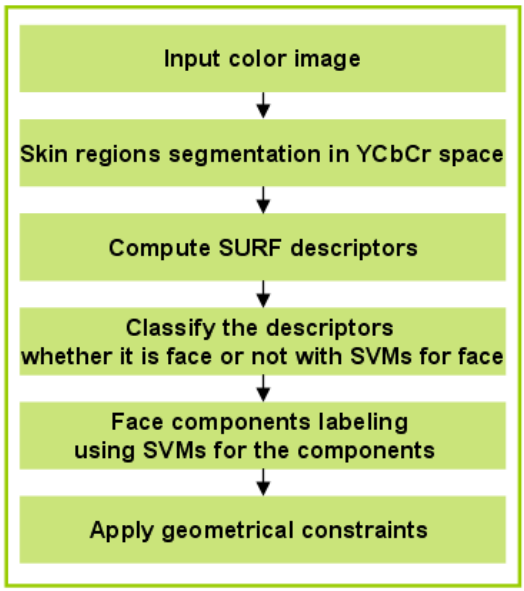

Figure 1. Overview of the proposed approach.

matching methods using manually predefined templates or deformable templates which are parameterized by specific functions, but the approach can not effectively deal with the variation in scale, pose and shape. The fourth is appearance based methods to learn the relevant characteristics with training face and non-face images. Some popular approaches for face detection include neural network [8], support vector machines (SVMs) classifiers [9], a network of linear units [10], or the adaptive boosting approach [11]. Two important recent approaches, boosting [11] and SVMs [13], belong to the appearance based methods. In the appearance based methods, the task of face detection can be divided into two steps: feature extraction from images and classification of the extracted features.

Concerning feature extraction, there are many papers in the field of visually salient regions and obtaining descriptors for such regions according to the tutorial introduction [14] to salient point detectors. Since the first corner detectors, the Moravec corner detection algorithm, were developed in the late 1970's, dozens of interest point detectors have been proposed such as a Hessian detector, Harris detector[4], Hessian/Harris-Laplacian/Affine detector [3] based on affine normalization around Harris and Hessian points, MSER (Maximally Stable Extremal Regions) [7] detector, SIFT (Scale Invariant Feature Transform) [1] detector, and SURF [5] detector. Here, we aim at detecting face in real time with the best accuracy. We assess both SIFT and SURF features for classification of faces and non-faces. We found that that SURF was performing better in particular for dealing with low resolution images. These are also faster to compute.

For classification, the boosting algorithm has two drawbacks: training times is long and a huge number of training images are required. On the other hand, SVMs have faster training times and also generalize well on smaller training sets. Hence, we use SVMs to train SURF descriptors.

3. Proposed system steps

An outline of the proposed algorithm portrayed in Figure 1, contains the following major modules: (1) Skin region segmentation, (2) Compute SURF descriptors, (3) Classify the descriptors whether it is face or not with SVMs trained by face descriptors, (4) Face components labeling using SVMs trained by the descriptors of face components, (5) Apply geometrical constraints. The following sections present a brief summary of each step.

3.1. Skin region segmentation

In order to detect interest regions quickly, our approach starts with the segmentation of skin areas in the images using the YCbCr (Luminance, chrominance-blue, chrominance-red) color space and a set of experimentally defined thresholds. A luminance element largely depend on the variation of illumination, and therefore we only defined thresholds about elements of Cb and Cr which are more robust to the variation of illumination. A result of the skin region segmentation is shown in Figure2.

3.2. SURF descriptor

One of the main advantages of SURF is to be able to compute distinctive descriptors quickly. In addition, SURF descriptor is invariant to common image transformations including image rotation, scale changes, illumination changes, and small change in viewpoint. This section shows a brief summary of its construction process.

Figure 2. Skin color segmentation.

3.2.1. Interest point localization

The SURF detector is based on the Hessian matrix. Given a point $X = (x, y)$ in an image I, the Hessian matrix $\mathrm{H}(X, \sigma)$ at X at scale σ is defined as follows:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \quad (1)$$

where $L_{xx}(X,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I at point X, and similarly for $L_{xy}(X,\sigma)$ and $L_{yy}(X,\sigma)$. In contrast to SIFT, which approximates Laplacian of Gaussian (LoG) with Difference of Gaussians (DoG), SURF approximates second order Gaussian derivatives with box filters (mean or average filter) shown in Figure 3. These can be calculated rapidly through integral images [11]. The location and scale of interest points are selected by relying on the determinant of the Hessian matrix. Interest points are localized in scale and image space by applying non-maximum suppression in a $3 \times 3 \times 3$ neighborhood.

3.2.2. Interest point descriptor

SURF constructs a circular region around the detected interest points in order to assign a unique orientation and thus gain invariance to image rotations. The orientation is computed using Haar wavelet response in both x and y directions. The Haar wavelets can be quickly computed by integral images. When the dominant orientation is estimated and included in the interest point information, SURF descriptors are constructed by extracting square regions around the interest points. The windows are split up in 4×4 sub-regions. The underlying intensity pattern (first derivatives) of each sub-region is described by a vector $V = \left[\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right]$.

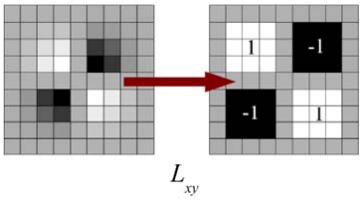

Figure 3. Gaussian second order partial derivatives and corresponding box filter.

3.3. The first layer classifier

Generally, SVMs [2] perform classification for two-class problems by determining the separating hyperplane with maximum distance to the closest points of the training set. These points are called support vectors. In this paper, linear SVMs classifiers are used for recognizing feature vectors from face images. To train SVMs, our system used SURF descriptors that have a dimension of 128. These descriptors are computed from face images (65x70) which are manually cropped and background images (65x70) which are randomly cropped. Some images are shown in Figure 4.

Figure 4. Example of training images (top: faces, bottom: background).

The number of training descriptors is 251 from face images and 340 from background images. The classifier for recognizing the descriptors extracted from face images is computed by total training descriptors (591 descriptors). Figure 5 is the result of this step, and each point in Figure 5 means the descriptor is classified as a face. We can notice that most of the 'face features' are located on the face region of the image, however some false alarm also appears on the window and the corners on the wall. This first result is

promising but shows that we need a second step to discard better false alarms.

Figure 5. The result of face descriptors classifier.

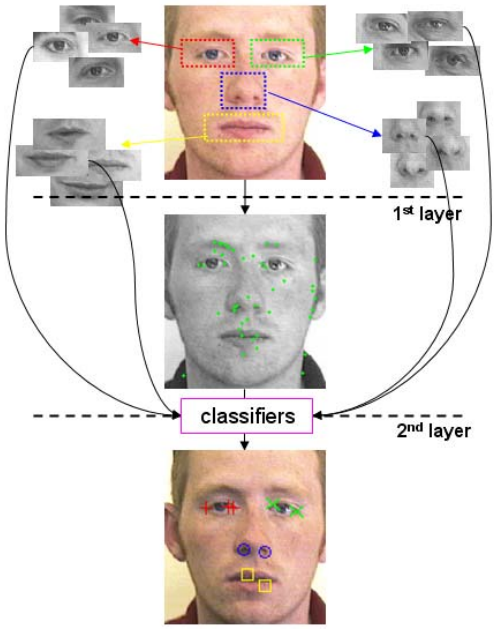

Figure 6. System overview of the facial components classifier (Red color plus: left eye, Green color cross: right eye, Blue color circle: nose, Yellow color rectangle: mouth).

3.4. The second layer classifiers

This step is to assign component labels to each object in a face image. In other words, for each feature classified as face in the previous step, we add another label corresponding to the subclasses left eye, right eye, nose, or mouth. We train a SVM classifier for each of these subclasses (therefore the number of classifiers in this step is four). The classifiers are shown in Figure 6. For example, to compute a left eye classifier, the classifier is trained using the descriptors extracted from the left eye versus mouth and nose images (as negative examples). The right eye subclass is excluded from the negative samples in the training since it is too similar to the target left eye subclass. When the mouth classifier is trained, all the other subclasses (right and left eyes, and nose) are used as negative examples.

Training data images are manually cropped at high resolution 130x140 to have a maximum number of selected features. The number of left eye descriptors is 100, the number of right eye descriptors is 126, the number of mouth descriptors is 149, and the number of nose descriptors is 61. The result is shown in Figure 9, 10. Left eye descriptors are classified with red plus marks, and the green cross marks mean right eye descriptors.

3.5. Apply geometrical constraints

This step is to eliminate the wrongly classified descriptors and also to estimate position and scale of the face components using both their label (subclass) and geometrical information. For using geometrical information, we start with eye pairing process. This choice is motivated by the fact that the most robust classification results are obtained for the left and right eye descriptors in face components. Once eyes features have been localized, using the difference between the x and y coordinated of the two eyes, the coordinates of the other facial descriptors are rotated until obtaining a frontal view face position where both eyes have the same y value. It allows to easily interpret each facial component using geometrical constraints. When eye pairing procedure is finished, we can estimate the position of nose and mouth through both label information and geometrical constraints. Falsely classified features can also be discarded thanks to this process. For instance, in Figure 7, the condition of nose position is that the descriptors of labeled nose have to exist in triangle region. The condition can eliminate some labeled nose descriptors.

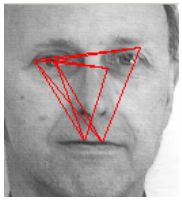

Figure 7. Geometrical constraint for nose position.

4. Experimental results

Test data set of faces consists of 3 subsets such as high, lower, and lowest resolution images. The purpose of the subsets is to evaluate the performance with

respect to the scaling factor. The high resolution (130x140) subset of face comprises of 100 face images randomly selected and cropped from the AR face database [6] and Caltech face database [16]. The other subsets, the lower resolution (65x70) and the lowest resolution (43x46) subsets of face, are made using resized high resolution images. The example is shown in Figure 8. All images in the test set are different from the training set. To calculate error rate, we also made the coordinate data file of facial components. Test data set of non-face comprises of 241 background images (130x140) collected from the web. In all our experiments, we used SURF descriptor as implemented in [15]. For testing the performance of the proposed approach, we performed two sets of experiments.

1. Calculation detection rate and true positive and false positive of each classifier in the face data set.

2. Calculation false positive of each classifier in the non-face data set.

In the first experiment, we estimate true positive and false positive by the condition whether labeled descriptor exists in correct or wrong region taken from the coordinate data file. For example, if a descriptor labeled as left eye is in the region of left eye, the case is true positive, but if not, the case is false positive. In Table1, error rate is calculated using the false positive.

In the second experiment, we estimate false positive. All detected descriptors become false positive. The error rate is calculated by the false positive as well. The result is shown in Table 1.

Table1. Result of the subclass classifiers

Classifier / DB		Left eye	Right eye	Mouth	Nose	Number of images	Total number of Descriptors
Face high resolution	Detection rate	**97%**	**97%**	**93%**	**72%**	**100**	**9095**
	Error rate	**6.4%**	**8.2%**	**4.7%**	**0.57%**		
Face lower resolution	Detection rate	**88%**	**93%**	**56%**	**28%**		**2914**
	Error rate	**3.0%**	**3.7%**	**2.4%**	**0.68%**		
Face lowest resolution	Detection rate	**43%**	**49%**	**5%**	**1%**		**1218**
	Error rate	**0.65%**	**0.98%**	**0%**	**0.16%**		
nonFace	Error rate	**6.7%**	**6.7%**	**6.7%**	**0.87%**	**242**	**18012**

(second layer classifiers).

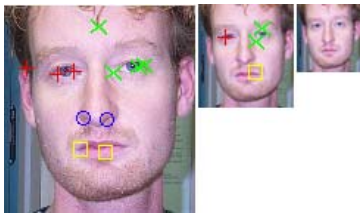

Figure 8. Example of face test data set (Left: high resolution image, Center: lower resolution image, Right: lowest resolution image)

In the high and lower resolution test data, classification results are extremely good for the eyes. Similarly mouth detection is also giving good result whereas there is more missed for the nose. However, at the lowest resolution, detection results are not very good for all components (below 50%). The method depends on the salient point descriptors. Although SURF descriptor is invariant to scaling, the system does not work that well in small faces. This is because, in small faces, there are a few feature vectors computed by SURF descriptor and therefore the performance worsens. For comparison, at that (lowest) resolution, the detection rate of the OpenCV face detector [11, 17] is just 25%.Our proposed method is then better. All the processes (skin detection, first layer classifier, second layer classifiers and geometrical constraints) have been applied for these results in Table 1.

Some results of detection are shown in Figures 9 and 10. All the remaining features after the different processes belong to the face in the image. Figure 5 (detection after the first classifier) can be compared with Figure 9. We see that the only face features have been kept and all false alarms have been discarded.

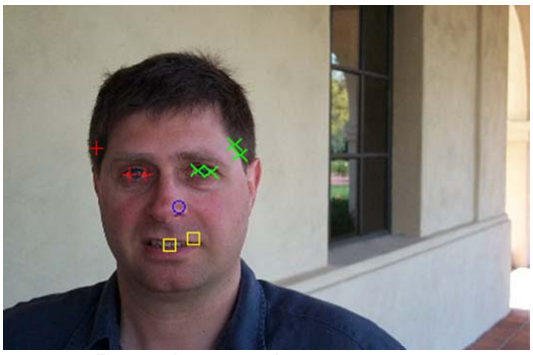

Figure 9. Detection results (Red color plus: left eye, Green color cross: right eye, Blue color circle: nose, Yellow color rectangle: mouth).

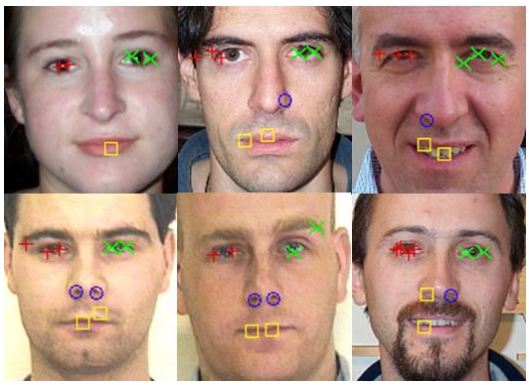

Figure 10. Experiment results

5. Conclusion and Future work

In this paper, we presented a method to detect face and face components based on SVMs classifier trained with SURF descriptors. We have shown that the method has high detection rate. The method is also able to localize face components and can be applied to both the recognition whether there are faces or not in video sequence and other object detection tasks in computer vision.

This system is not yet fully optimized and need further development in particular we will need to remodel better geometrical priors to pair different face components together, in particular for dealing with several faces in a same image, and different head poses.

Also, as SURF descriptors are detected in fewer numbers in low resolution images, decreasing the overall performance of the detection, Additional work is required for dealing with that problem.

ACKNOWLEDGEMENTS

This research is supported by the "Irish Research Council for Science, Engineering and Technology in collaboration with INTEL Ireland Ltd: funded by the National Development Plan" and Enterprise Ireland (project IP-2006-0412).

6. References

[1] David G.Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, January 2004, pp. 91-110.

[2] V.Vapnik, "Statistical learning theory", *John Wiley and Sons*, New York, 1998.

[3] Mikolajczyk, K., and Schmid, C", "An affine invariant interest point detector", *In European Conference on Computer Vision(ECCV)*, Copenhagen, Denmark, 2002, pp. 128-142".

[4] Harris, C. and Stephens, M, "A combined corner and edge detector", *In Fourth Alvey Vision Conference*, Manchester, UK, 1998, pp. 147-151.

[5] Bay, H., Tuytelaars, T., and Van Gool, L, "SURF: Speeded Up Robust Features", *In Proceedings of the Ninth European Conference on Computer Vision*, May, 2006.

[6] A.M. Martinez and R. Benavente, The AR Face Database, CVC Technical Report #24, June 1998.

[7] J.Matas, O. Chum, M. Urban, and T. Pajdla, " Robust wide baseline stereo from maximally stable extremal regions ", *In BMVC*, 2002, pp. 384-393.

[8] H. A. Rowley, S. Baluja and T.Kanade, Neural network-based face detection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, January 1998, pp. 23-29.

[9] E. Osuna, R. Freund and F. Girosi, "Training support vector machines: An application to face detection", *Computer Vision Pattern Recognition(CVPR)*, June 1997, pp. 130-136.

[10] D. Roth, M.-H. Yang and N.Ahuja, "A SNoW-based face detector", *Advances in Neural Information Processing Systems*, November 2000, pp. 855-861.

[11] P. Viola and M. Jones, "Robust real time face detection", *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, July 2001,pp. 747.

[12] M.-H. Yang, D. Kriegman, and N. Ahuja. "Detecting faces in images: A survey.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[13] S. Romdhani, P. Torr, B. Scholkopf, and A. Blake, Computationally efficient face detection. *In ICCV*, 2001.

[14] T. Tuyttellaars. Local Invariant Features: What? Why? When? How? How?, In *ECCV Tutorial*, 2006.

[15] "SURF Source", http://www.vision.ee.ethz.ch/~surf/download.html.

[16] Caltech face database, http://www.vision.caltech.edu/html-files/archive.html.

[17] OpenCV face detector, http://www.intel.com/technology/computing/opencv/index.htm