Scale and Orientation Invariance in Human Face Detection

Kin Choong Yow and Roberto Cipolla
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, England

Abstract

Human face detection has always been an important problem for face, expression and gesture recognition. Though numerous attempts have been made to detect and localize faces, these approaches have made assumptions that restrict their extension to more general cases. In this research, we propose a feature-based face detection algorithm that can be easily extended to detect faces under different scale and orientation. Feature points are detected from the image using spatial filters and grouped into face candidates using geometric and gray level constraints. A probabilistic framework is then used to evaluate the likelihood of the candidate as a face. We provide results to support the validity of the approach, and show that the algorithm can indeed cope efficiently with faces at different scale and orientation.

1 Introduction

Human face recognition is a field that has important applications in our daily activities, such as for security verification, criminal identification, and human-computer interactions. However, a pre-requisite for automatic face recognition is to detect and localize faces in a scene. This task is certainly not trivial when the background is complex, the illumination is varied, and the pose of the face not fixed. Though many approaches have been attempted towards face detection and localization, the assumptions and the constraints made in these approaches are too restrictive, making the algorithm incapable of extension to more general cases.

2 Related Work

There are a variety of approaches to face detection. The model-based approach assumes a different face model at different coarse-to-fine scales. For efficiency, the image is searched at the coarsest scale first. Once a match is found, the image is searched at the next finer scale until the finest scale is reached. Some of the work using this approach were reported by Yang and Huang [13], and Lanitis *et.al.* [4]. In general, only one model is assumed in each scale (usually in the fronto-parallel view) and thus it is difficult to extend this approach to multiple views.

The feature-based approach searches the image for a set of facial features and groups them into face candidates based on their geometrical relationship. Leung *et.al.* [5], Sumi and Ohta [11], and Yow and Cipolla [15] reported work using this approach. Though this approach can be easily extended to multiple views, it is unable to work well under different imaging conditions because the image structure of the facial features vary too much to be robustly detected by the feature detectors.

The approach based on neural networks detects faces by subsampling different regions of the image to a standard-sized image and then passing it through a neural network filter. Recent work was reported by Sung and Poggio [12], and Rowley *et.al.* [8]. The algorithm performed very well for fronto-parallel faces but is difficult to be extended to different views of the face.

Lastly, the colour-based approach labels each pixel according to its similarity to skin colour, and subsequently labels each subregion as a face if it contains a large blob of skin colour pixels (Chen *et.al.* [1], Dai and Nakano [2]). It can cope with different viewpoint of faces but it is sensitive to skin colour and the face shape.

3 About Image Evidence

So what can we learn from the attempts of these various researchers ? We find that Lanitis *et.al.* 's approach is able to localize the human face very well. It performs well because they make use of gray-level information in addition to edge information. Sung and Poggio's neural network approach works very well also because almost every pixel in the 19x19 subimage is used to evaluate the output, and many of these pixels encode spatial and gray-level information. This emphasized the importance of using different types of image evidence to support the face detection process.

On the other hand, we can see why Leung *et.al.* 's, Sumi and Ohta's method did not perform as well. The system is dependent on too few image features, which cannot be extracted robustly due to image noise or noise in the feature detector. Leung *et.al.* use the response from a set of steerable-scalable filters to find facial features, and Sumi and Ohta use template matching to identify eyes. In both these cases the only evidence for a feature to be present is the response of the filter or the correlation output. As a result, there is a lack of evidence to support the hypothesis of a face and therefore the performance of the algorithm is affected.

Apart from the usual image evidence of edges, gray level, etc., a form of evidence that is less well-exploited is that of contextual evidence, i.e. the knowledge that certain features occur in the vicinity of other features. For example, we know that eyes occur in pairs. So, when we find an eye in the image, the existence of this eye is evidence for the existence of the other eye.

In this paper, we present a framework that makes use of the specified structure between various components of a face to propagate and update evidence in each of the components. Evidence for the hypothesis of a face thus comes from two sources, from detected image features and from model knowledge, resulting in a high detection confidence in the true instance of a face.

4 A Feature-Based Face Detection System

In this section, we will describe a feature-based face detection algorithm that works in a bottom-up fashion from low-level image features to high-level model features. The algorithm extracts feature points using spatial filtering techniques, applies perceptual grouping principles to group feature points into face candidates, and makes use of a probabilistic framework for the selection of candidates. Invariance to scale and orientation is achieved to some extent and will be described in the next section.

4.1 The Face Model

The face is modelled as a plane with 6 oriented facial features. Due to occlusion, or missing features (eyebrows, in general), there is a need to decompose the face model into components consisting of 4 features, which are common occurrences of faces under different viewpoints, and of different identity. These groups are called Partial Face Groups or PFGs (Yow and Cipolla [15]). These PFGs are further subdivided into components consisting of 2 features (horizontal and vertical pairs - Hpair and Vpair) for the purpose of perceptual grouping and evidence propagation. The different component groups are shown in fig. 1.

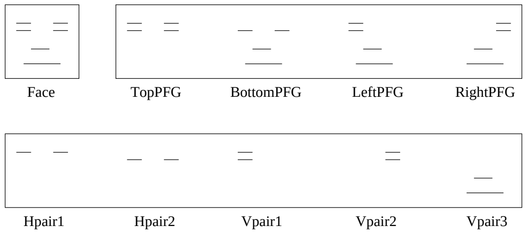


Figure 1: The face model and the component face groups.

It is observed that at low resolutions, all the 6 facial features will appear only as dark elongated blobs against the light background of the face. Since edges are illumination invariant to a large extent, we model the 6 facial features as pairs of oriented edges shown in fig. 2. The image should be smoothed before the feature detection process so that any high-resolution features will take the form of the lower resolution ones. The vertical edges in the eye and nose model may or may not be present after smoothing but they should not be an important criteria in the detection of the feature.

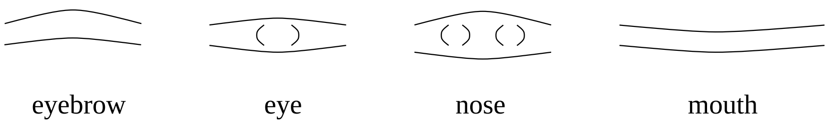


Figure 2: The facial feature models.

4.2 Perceptual Grouping

With such a low resolution model, there will be lots of false positive feature candidates. We therefore propose a perceptual grouping framework to group these feature candidates into faces using geometrical, gray-level and spatial information. Feature candidates that cannot be grouped will be discarded.

The grouping process is modelled as a two stage model of perception. The first stage, known as pre-attentive perception, extracts image information into points and regions of interest, which directs the attention of the processing efforts of the next stage. The second stage of perception, the attentive stage, will perform grouping, comparison, evaluation, and reasoning activities based on the detection and identification of meaningful object groups in the image.

4.2.1 Preattentive Feature Selection

The preattentive feature selection stage is performed in two steps. First, a list of interest points is found from the image. This can be achieved by filtering the image using a preattentive filter (a matched bandpass filter used in Yow and Cipolla [15]) and then searching for local maxima. Next, the edges around each interest point are examined. Similar edges are linked using a boundary following algorithm. If we find the existence of two roughly parallel edge segments with opposite polarity on both sides of the interest point, then this point is flagged as a feature point. The extent of the feature region is then defined by finding a boundary box around the two edges. Fig. 3 illustrates this process.

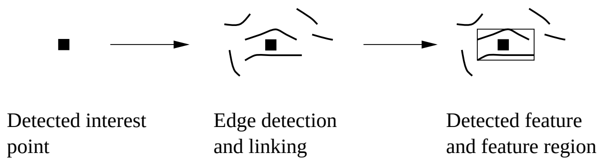


Figure 3: Preattentive feature selection process.

Measurements of the region's image characteristics (such as edge length, edge strength, gray-level variance, etc.) are then made and stored into a feature vector $\mathbf{x}$. From the training data of the facial features, e.g."eyebrow", a mean vector μ_{brow} and covariance matrix $\mathbf{\Sigma}_{brow}$ are also obtained which define the class of valid "eyebrow" feature vectors in a n-dimensional space, where n is the number of components defining the feature vector $\mathbf{x}$.

A facial feature candidate i is a valid facial feature j if the Mahalanobis distance $\mathcal{M}_{ij}$ of the feature vector $\mathbf{x}_i$ is within an admission threshold τ_j from the class mean μ_j, i.e.

$$\mathcal{M}_{ij} < \tau_j \text{ , where } \mathcal{M}_{ij} = (\mathbf{x}_i - \mu_j)^T \mathbf{\Sigma}_j{}^{-1} (\mathbf{x}_i - \mu_j)$$

This is repeated for all the 4 classes of facial features, namely, eyebrow, eye, nose, and mouth. If the facial feature does not belong to any of the 4 classes, it is discarded from the list.

4.2.2 Attentive Feature Grouping

After obtaining a set of feature points and the associated feature regions, these feature regions are then actively grouped using our model knowledge of the face. Single features are grouped into vertical and horizontal pairs, pairs are grouped into face groups, and face groups are grouped into face candidates (fig. 4).

For each level k of grouping, a set of n_k measurements is made of the component features and stored into a n_k-dimensional vector. This vector is then projected into its n_k-dimensional class space, which was determined from measurements obtained from faces in training data. The Mahalanobis distance $\mathcal{M}_{ij}$ of this feature vector is then evaluated and used to determine its membership in the class.

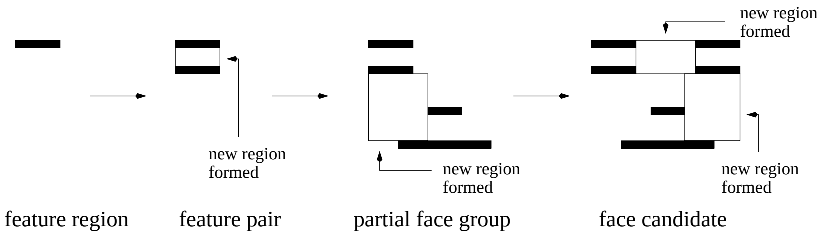


Figure 4: Attentive feature grouping process.

This grouping process is effective in removing false positives because a lot of geometric and gray-level measurements are used to determine its validity, in particular the edge and spatial information about the new region formed that is not part of the components themselves (see fig. 4).

The measurements we choose include :

1. the ratio of feature lengths (obtained from edge linking) to image size.
2. the ratio of feature lengths to other feature lengths.
3. the aspect ratio of a feature region.
4. the ratio of inter-feature distances.
5. the difference in orientation between features.
6. the number of directional edgels in a region (normalized to region size).
7. the ratio of edge strengths in a region to edge strengths of facial features.
8. the mean gray level of a region (normalized to intensity distribution).
9. the variance in the gray-level distribution of a region.

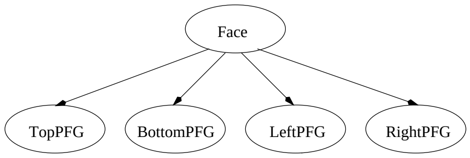
One important advantage of this process is that though the spatial region to be analyzed gets larger at higher levels, there are fewer of these regions to process. Therefore, the processing time is kept small throughout the whole algorithm.

4.3 Evidence Propagation

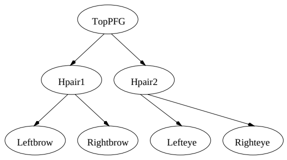
The perceptual grouping framework enables us to reject grossly incorrect groupings of face candidates. However, there are still a number of false positives which cannot be discarded this way. As such, we examine the use of belief networks for this task.

Bayesian networks, which are also known as belief networks, are directed acyclic graphs, with nodes representing random variables and arcs signifying conditional dependencies specified in terms of conditional probabilities (Sarkar and Boyer [10]). Each node has a conditional probability table (CPT) associated with it, describing the conditional probability of each value of the variable, given each possible combination of the values of the parent nodes.

In Yow and Cipolla [14], the use of belief networks was proposed for defining the probability for each face candidate. The belief network used is shown in fig. 5(a). A uniform prior is assumed, and the CPT entries estimated directly using the statistics of the set of examples (Russell *et.al.* [9]).



(a)



(b)

Figure 5: (a) Main belief network. (b) Example of component belief network.

To obtain higher confidence in true instances of the child nodes, we build component belief networks for each of the 4 child nodes. An example is shown in fig. 5(b). When evidence for a component network node becomes available, a corresponding virtual node is created and instantiated, propagating the evidence through the network and updating the probability of all the other nodes. These probabilities are in turn used as evidence in the main belief network. A propagation algorithm for trees given by Pearl [6] is used to update the probabilities.

The evidence for each facial feature or face group i is related to its Mahalanobis distance, $\mathcal{M}_{ij}$, and the admission threshold for the jth feature class, τ_j, by :

$$P_i = \left(1 - \frac{\mathcal{M}_{ij}}{\tau_j}\right), \text{ where } \mathcal{M}_{ij} < \tau_j$$

Each facial feature that is detected is assigned 4 probability values, P_{brow}, P_{eye}, P_{nose} and P_{mouth}, using the above equation. When a higher level group is formed, only the probability of the corresponding feature is propagated. For example, if a vertical brow-eye pair (Vpair1) is formed, only P_{brow} of the upper facial feature and P_{eye} of the lower feature is propagated. Likewise, only these values are updated in the propagation process. As a result, only true positive faces are updated to a high confidence level.

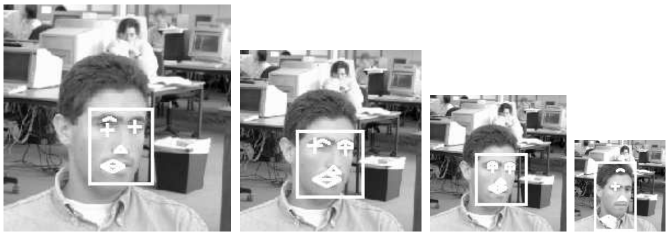
5 Scale and Orientation Invariance

In this section, we will look at the effects of varying scale and orientation on the detection of faces. We first vary the scale of the preattentive filter from $\sigma = 3.0$ to $\sigma = 1.0$ while keeping the scale of the edge detection filter fixed at $\sigma = 3.0$. The results in fig. 6 show that although 3 different values of σ are used, the face detected for all the 3 cases is the same. We also observe that at a scale ($\sigma = 1.0$) smaller than the one required for matched filtering, the correct facial features are still detected, though there is a much larger number of false detected features.

(a) (b) (c) (d)

Figure 6: Varying the scale of the preattentive filter. (a) $\sigma = 3.0$ (81 points). (b) $\sigma = 2.0$ (177 points). (c) $\sigma = 1.0$ (332 points). (d) Face detected (same for all 3 cases of σ).

We subsequently vary the size of the image while keeping the scale of the preattentive filter constant at $\sigma = 1.0$. Fig. 7 shows the result for the image being reduced to 80%, 60%, 40% and 20% of the original size. We fail to detect the facial features when the face is too small because the image structure of these facial features are corrupted by quantization noise. However, we are successful in detecting the facial features of large faces. This is because the size of the facial features are actually determined by the edge detection and edge linking process.



(a) (b) (c) (d)

Figure 7: Varying the size of the face in the image. (a) percentage size = 80%. (b) percentage size = 60%. (c) percentage size = 40%. (d) percentage size = 20%.

We do a further test by varying the aspect ratio of the preattentive filter. Fig. 8 shows the result of varying the aspect ratio from 3:1 to 1:1. We observe that the facial features are still detected even though the aspect ratio is 1:1. The significance of this is that we can steer a 1:1 second derivative Gaussian exactly by using only 3 basis filters (Freeman and Adelson [3]), instead of 16 basis filters for a 1% error approximation for 3:1 filter (Perona [7]) - a huge saving in computation.

(a) (b) (c) (d)

Figure 8: Varying the aspect ratio of the preattentive filter. (a) aspect ratio = 3:1 (81 points). (b) aspect ratio = 2:1 (110 points). (c) aspect ratio = 1:1 (201 points). (d) Face detected (same for all 3 values of aspect ratio).

The edge detection filter is a standard Canny edge detector. Choosing a small σ will result in very noisy edges that are difficult to link. A large σ, however, may cause 2 edges to be smoothed into one. For an application of, say, detecting the face of a person sitting in front of a computer, a $\sigma = 1.0$ was found to be sufficient.

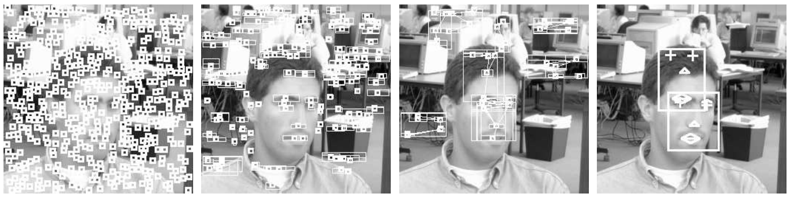
(a) (b) (c) (d)

Figure 9: Varying the orientation with aspect ratio = 1.1. (a) orientation = $-30°$ (205 points). (b) orientation = $0°$ (283 points). (c) orientation = $30°$ (223 points). (d) Face detected (same for all 3 values of orientation).

We now keep the preattentive filter at the 1:1 aspect ratio and rotate the filter in $30°$ increments. The results are shown in fig. 9. We observe that the facial features are detected in all the different orientations of the filter. The significance of this is that we can make do without steerable filters completely, and just use a single orientation of the preattentive filter to obtain the interest points.

6 Results

We implement the above face detection algorithm using only one single scale and orientation of the preattentive filter. The scale of the filter is chosen to be the same as the edge detection filter ($\sigma = 1.0$) so that we only need to smooth the image once. A set of 40 images is used as a training set to obtain the necessary parameters. The algorithm takes about 10 seconds to run on 256x256 images on a SUNSparc20 workstation. The intermediate results are shown in fig. 10.



(a) (b) (c) (d)

Figure 10: (a) Interest points (466 points). (b) Feature regions (180 points). (c) Face groups (28 top, 10 bottom, 5 left, 1 right). (d) Face candidates (2 faces). The probability associated with the upper and lower face candidates are 0.6124 and 0.9578 respectively.

We achieve a successful face detection rate of 89% on a database of 90 images. Some of the test results are shown in fig. 11. We can see from the results that the algorithm is able to cope well with variations in scale as well as orientation. Presence of glasses, occlusion and absence of facial features are also tolerated to some extent.

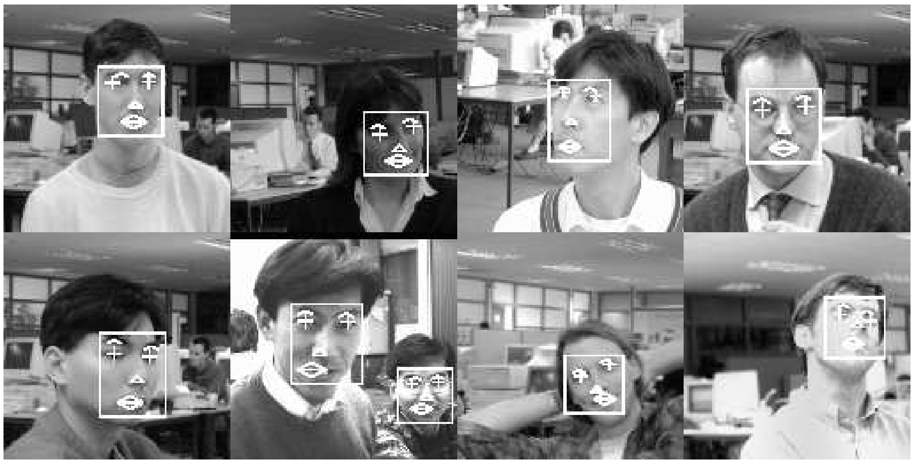


Figure 11: Result of face detection on various face images at different scale and orientation.

7 Conclusions

We have proposed a feature based face detection algorithm which detects interest points using spatial filters, groups them into face candidates using geometric and gray-level information, and selects true faces using a probabilistic framework. We have studied the effects of varying the scale and orientation parameters of the algorithm and showed that the algorithm is able to work well under different scale and orientation of the face.

References

[1] Q. Chen, H. Wu, and M. Yachida. Face detection by fuzzy pattern matching. In *Proc. 5th Int. Conf. on Comp. Vision*, pages 591–596, MIT, Boston, 1995.

[2] Y. Dai and Y. Nakano. Extraction of facial images from complex background using color information and SGLD matrices. In *Proc. Int. Workshop on Auto. Face and Gesture Recog.*, pages 238–242, Zurich, 1995.

[3] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Patt. Analy. and Machine Intell.*, 13(9):891–906, 1991.

[4] A. Lanitis, C. J. Taylor, and T. F. Cootes. An automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, 1995.

[5] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using labelled random graph matching. In *Proc. 5th Int. Conf. on Comp. Vision*, pages 637–644, MIT, Boston, 1995.

[6] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, 1988.

[7] P. Perona. Steerable-scalable kernels for edge detection and junction analysis. In G. Sandini, editor, *Proc. 2nd European Conf. on Comp. Vision*, pages 3–18, Italy, 1992. Springer–Verlag.

[8] H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158, CMU, July 1995.

[9] S. Russell, J. Binder, D. Koller, and K. Kanazawa. Local learning in probabilistic networks with hidden variables. In *Proc. Int. Joint Conf. on Artif. Intell.*, 1995.

[10] S. Sarkar and K. L. Boyer. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Trans. Patt. Analy. and Machine Intell.*, 15(3):256–273, 1993.

[11] Y. Sumi and Y. Ohta. Detection of face orientation and facial components using distributed appreence modeling. In *Proc. Int. Workshop on Auto. Face and Gesture Recog.*, pages 254–259, Zurich, 1995.

[12] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, CBLC Paper 112, MIT, Dec. 1994.

[13] G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[14] K. C. Yow and R. Cipolla. Finding initial estimates of human face location. In *Proc. 2nd Asian Conf. on Comp. Vision*, volume 3, pages 514–518, Singapore, 1995.

[15] K. C. Yow and R. Cipolla. Towards an automatic human face localization system. In *Proc. 6th British Machine Vision Conference*, volume 2, pages 701–710, 1995.