Towards an Automatic Human Face Localization System

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

Abstract

This paper describes a method to detect and locate human faces in an image given no prior information about the size, orientation, and viewpoint of the faces in the image. This method uses a family of Gaussian derivative filters to search and extract human facial features from the image and then group them together into a set of partial faces using their geometric relationship. A belief network is then constructed for each possible face candidate and the belief values updated by evidences propagating through the network. Different instances of detected faces are then compared using their belief values and improbable face candidates discarded. The algorithm is tested on different instances of faces with varying sizes, orientation and viewpoint and the results indicate a 91% success rate in detection under viewpoint variation.

1 Introduction

Recognizing a human face in a scene is becoming an area of immense interest in the computer vision community. Clinical evidence suggests that the human brain has specific neural hardware to perform face recognition (Tranel *et.al.* [11]), indicating that face recognition is an important task for humans. Moreover, the fact that human can robustly recognize faces in a large variety of conditions (different illumination, viewpoint, expression, etc.) poses an interesting and challenging task for us to build an efficient model for computational face recognition.

The possible applications of automatic face-recognition systems are in criminal identification, security monitoring and man-machine interfacing. In all of these applications, face detection and localization is the first step of a solution. One major difficulty of a face detection algorithm is the lack of *a priori* information about the scale, orientation, viewpoint, etc. of the face in the image. A small difference in, say, the viewpoint of the subject leads to a great difference in the image structure and thus makes the problem extremely hard. Previous work on face detection (Govindaraju [4], Yang and Huang [12], Leung *et.al.* [6]) have been unable to cope with significant changes in viewpoint. In this paper, we will attempt to address the issue of detecting a human face in an image given no prior information about the size, orientation and viewpoint of the face.

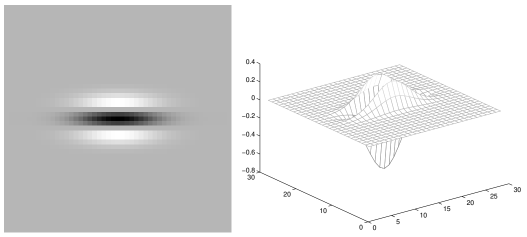
2 Feature Detection

A natural first step in detecting a human face in an image is to detect features that are unique to the structure of the human face. Methods have been proposed to detect features such as the eyes and mouth (e.g. Yuille *et.al.* [13]) as they have a very rich and unique image structure. However, the image structure of these features changes very rapidly even with a small change in viewpoint. Hence, we must seek to look for coarser features that will remain invariant under different viewpoints and orientations.

2.1 Gaussian Derivative Filters

Filters built from Gaussian and derivatives of Gaussian have been a popular choice in many applications (Canny [2], Leung *et.al.* [6]). This family of filters have high signal-to-noise response and good localization capabilities.

We observe that one of the Gaussian derivative filters used in Leung *et.al.* [6] performs remarkably well as a low-intensity bar detector (also called ridge detectors or line detectors). This filter and its surface plot are shown in figure 1. This filter is a second derivative of Gaussian in one direction, and is a Gaussian in the orthogonal direction. The length of the filter is also elongated at three times the width, giving it better orientation selectivity.



(a) (b)

Figure 1: (a) Gaussian derivative filter. (b) Surface plot.

The second derivative of Gaussian detects bars of low intensity, and the Gaussian smoothes out any intensity variations in the orthogonal direction. This makes it an excellent detector for the brows, eyes and the nose. Also, the 3:1 elongation of the filter corresponds approximately to the length-width ratio of the eyes and nose, thus obtaining maximal response when the scale and orientation of the filter matches that of the features.

A simple thresholding and non-maximal suppression operation on the convolution output will then enable us to extract the position of features. The convolution output of the filter (of a matching scale and orientation) with two face images and the detected features are shown in figure 2.

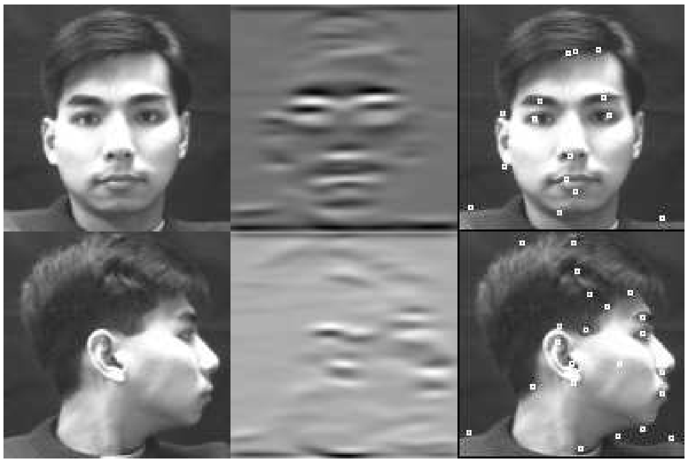


Figure 2: Original image, convolution output and detected features of two face images. We observe that the facial features are detected even for a profile view.

2.2 Steerable-Scalable Decomposition

An efficient way of searching for the features under different scale and orientation is to decompose the filter into a set of "steerable-scalable" basis filters (Perona [8]) and then interpolate the results to obtain the filter response at any scale and orientation. Although this decomposition is not exact, a larger number of basis filters can be used to improve the accuracy of the interpolated response. Perona had shown that for a 3-octave 1% approximation error (1% is the specified tolerable error, actual error measured is 2.5%), the number of filters required is 16 (rotation) x 8 (scale) = 128 filters. If a 10% approximation is allowed, the number of filters decreases approximately by a factor of 4 to 32.

Using the steerable-scalable decomposition technique, we can convolve the input image with a fixed set of basis filters and then interpolate the output to obtain the response at different scales and orientation. Since all the features that we want to detect have roughly the same size and orientation in a single image, we would obtain many high responses when the correct size and orientation is found.

Depending on the application, some prior information is usually known about the scale or orientation of the faces in the image (e.g. the face is always upright in ID type photographs). Hence, in such cases, it may be more efficient or more accurate to perform an exhaustive search using a set of Gaussian derivative filters at the expected scale and orientation. To improve the efficiency of search, we should fix the scale of the filter to the smallest allowable by the sampling theorem and instead vary the scale of the image by sub-sampling the image at the coarsest scale. Subsequently, the scale is refined until we have covered the range of scales desired.

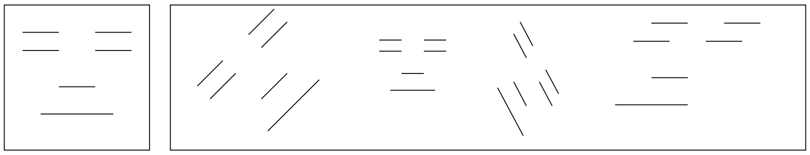
3 Geometric Grouping of Features

The feature detection process results only in a set of points that could be the actual features. We cannot use a full graph matching technique as in Leung *et.al.* [6] because it would be difficult or impossible with the number of features that we have detected. Moreover, it is usual for some of the facial features to be occluded when looking at the subject from a different viewpoint. We thus propose a way of grouping the feature points into partial face groups using affine invariants.

3.1 Affine Invariance

For most applications including ours, the subject is sufficiently far away from the viewer such that the depth variation in the face is small compared to the distance between the face and the camera. Under this condition the weak-perspective approximation [9] holds and we will have three affine invariants.

If we represent the human face as a plane and its six features (eyebrows, eyes, nose and mouth) by line segments (figure 3(a)), we can verify that the affine invariants exist for different viewpoints of the face. Figure 3(b) shows the views of our face plane under different affine transformation :



(a) (b)

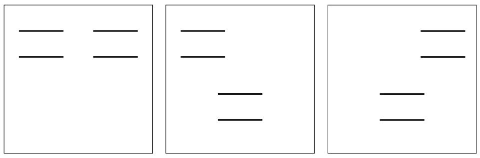
Figure 3: (a) The face plane. (b) Different affine transformations of the face.

3.2 Grouping Features into Partial Face Groups

Under large changes in viewpoint, some of the features will be occluded (e.g. the left brow and left eye are occluded in a right profile view). In such circumstances, we may only see four of the six features in a face. Therefore, we represent the face as a set of partial face groups (or PFGs) shown in figure 4. At large changes in viewpoint, some of these PFGs will still be detected depending on which viewpoint is being taken.

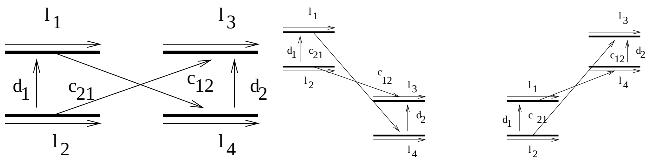
We replace each detected feature with a line that corresponds to the longitudinal axis of the feature, based on a region-connectivity analysis of the thresholded convolution output (Ballard and Brown [1]). For every group of four line segments in the image, we can derive length vectors $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{l}_3$, $\mathbf{l}_4$, separation vectors $\mathbf{d}_1$, $\mathbf{d}_2$, and cross vectors $\mathbf{c}_{12}$, $\mathbf{c}_{21}$ to be used in our grouping. The geometric relationship between these vectors in a PFG are shown in figure 5.

From the affine invariants, we obtain the following constraints to group a PFG:



(a) (b) (c)

Figure 4: (a) Top partial face group (PFG). (b) Left PFG. (c) Right PFG.



(a) Top PFG (b) Left and Right PFGs

Figure 5: Geometry of partial face groups (PFGs).

1. The length vectors $\mathbf{l}_1$ should be parallel to $\mathbf{l}_2$, $\mathbf{l}_3$ parallel to $\mathbf{l}_4$.
2. The magnitude of the separation vectors $|\mathbf{d}_1|$ and $|\mathbf{d}_2|$ should be proportional to the scale of the filter used.
3. The ratio of magnitude $|\mathbf{d}_1|$ to magnitude $|\mathbf{d}_2|$ should be constant.
4. The component of the cross vectors $\mathbf{c}_{12}$, $\mathbf{c}_{21}$ in the direction of the separation vectors $\mathbf{d}_1$, $\mathbf{d}_2$ should be a constant multiple of $|\mathbf{d}_1|$, $|\mathbf{d}_2|$ respectively.
5. The component of the cross vectors $\mathbf{c}_{12}$, $\mathbf{c}_{21}$ in the direction of the length vectors $\mathbf{l}_1$, $\mathbf{l}_2$ should be a constant multiple of $|\mathbf{l}_1|$, $|\mathbf{l}_2|$ respectively.

Each pair of line segments is examined to see if the first three constraints are valid. If they are, then this pair is labelled and stored as a valid pair. Each valid pair is then compared with other valid pairs to see if the last two constraints are violated. If they are not violated then these four line segments are grouped into a PFG. The constants used in the last three constraints are different for each PFG and are obtained from averaging measurements in a prior set of face images.

After we formed a PFG, we assign a *certainty* value to it based on the filter response of each feature and the errors in the geometric constraints. This *certainty* value is normalized to a value between 0 to 1 and is assigned as follows:

certainty = (Normalized sum of filter response of each feature in the PFG) $\times$ (1 - Normalized sum of error in the geometric constraints)

4 Probabilistic Reasoning

We will now use the PFGs we obtained as evidences to determine the probability of the presence of a face. Rather than just summing up the evidences in an ad-hoc fashion, we propose to model the face as a belief network (also called causal networks or influence diagrams) and propagate the evidences through the network to obtain the belief of the presence of a face.

4.1 Modelling the Face as a Belief Network

A belief network is a way of representing the conditional independence relationship between a set of variables and gives a concise specification of the joint probability distribution. A belief network is a directed, acyclic graph, or DAG, where the nodes represent a set of random variables and the links represent the influence of a parent node over a child node (Russell and Norvig [10]).

The belief network formalization also includes an inference mechanism that allows us to recompute the "beliefs" in the nodes based on the combination of evidences propagating through the network. An elegant propagation solution for trees is discussed in Pearl [7] and a solution for general networks is given in Lauritzen and Spiegelhalter [5].

We model the human face as a DAG consisting of one parent node and three child nodes (figure 6). The conditional probability table for each node is shown beside the node. There are only two possible values at each node, i.e. Present or NotPresent. Hence the columns in the conditional probability table must sum to one. The uncertainty in the presence of the node is modelled by a virtual child node at the node. We specify the values in the conditional probability table based on our knowledge about the relationships in the system.

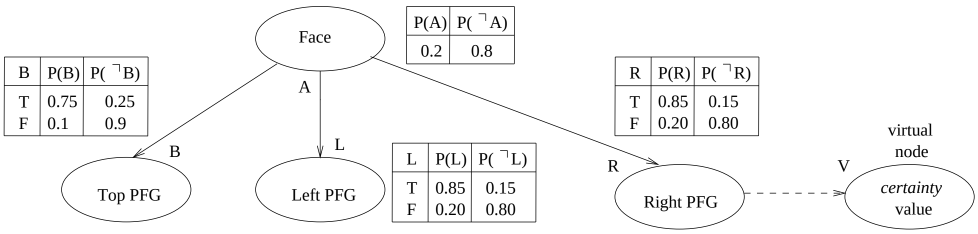


Figure 6: The face modelled as a belief network.

The advantage of using a belief network is that each piece of evidence can be computed independently (and thus in parallel) and the evidences are combined in a non-ad hoc manner. It is easy to expand the network to include other nodes without having to re-specify the conditional probabilities of the existing links.

4.2 Propagating Evidence in the Belief Network

We will now describe and apply Pearl's [7] method for propagating probabilities in trees to compute the belief of the presence of a face in out belief network. For the DAG shown in figure 6, let A be the variable at a particular node, and A can have values a_j where $j = 1, 2, \ldots, m$ for all m possible values of the variable A. In our case A can have only two values, hence $a_1 = 1$ and $a_2 = 0$. Let each node also contains the belief $P(a_j)$ for each possible value a_j.

Evidence is propagated by means of passing a numerical value from a node to its adjacent parent node or child node. The value passing from a child node to a parent node is always called a λ message, and the value from a parent node to a child node is always called a π message. We define $\lambda_{BA}(a_j)$ as the λ message propagating from the child B to parent A for the value a_j and $\pi_{AB}(a_j)$ the π message propagating from the parent A to child B for the value a_j.

When a node is instantiated, its belief $P(a_j)$ for value a_j is set to 1. It will then send λ messages to all its parents and π messages to all its children. However, if a node is not instantiated but receives a λ message, it would update its belief $P(a_j)$ and sends new λ messages to its parents and π messages to its children. If however a π message is received, it would update its belief $P(a_j)$ and sends only new π messages to its children.

The new belief at a node B with parent A and child C is given by $P'(b_i) = \alpha\lambda(b_i)\pi(b_i)$ where

$\lambda(b_i) = \prod_i \lambda_{CB}(b_i)$ (Product of all the λ messages from B's children)

$\pi_(b_i) = \sum_j P(b_i|a_j)\pi_{AB}(a_j)$ (Sum of the all the π messages from B's parent $\times$ its conditional probability)

α is a normalizing variable so that all the $P'(b_i)$ at node B sum to 1.

The uncertainty in the evidence of a node B can be modelled as a virtual node which is attached as a child to node B. When evidence is found from the image but with uncertainty, the virtual node is instantiated instead of node B itself. This causes the virtual node to send λ messages containing the *certainty* value to node B, causing a propagation of values through the network.

5 Results

A face candidate is formed from the combination of all the PFGs supporting it. A face in the fronto-parallel view will have up to three PFGs supporting it, and thus will be represented by six feature points. A profile view has only one supporting PFG and thus only four feature points. If two face candidates formed in this way share a common feature location, the candidate with the lower probability is discarded immediately.

Figure 7 shows all the possible faces detected for a particular case where the viewing angle is large. The candidates are ranked in decreasing order from left to right, with associated probability values of 0.7245, 0.5526, 0.2386, and 0.2060 respectively. The results indicate a good discriminating margin between the correct and incorrect faces. The discriminating margin becomes very high for fronto-parallel views, but gets quite weak in the presence of cluttered background.

We test the algorithm using images of a face at different scale, orientation and

Figure 7: Possible candidates of faces detected, ranked from left to right.

viewpoints on a simple background. A total of 33 images of different orientations and viewpoints were used, covering an azimuthal angle of $\pm 90°$, vertical tilt of $\pm 45°$, and angular rotation of $\pm 45°$. Of the 33 images, 3 failed completely, giving a 91% success rate in detection for different viewpoints. In 5 of the remaining test images, two or more faces are detected. We consider these as successful detection because we want to be more conservative in detecting faces. Extra candidates can be eliminated by means of other types of constraints (such as motion). Figure 8 shows some of the results of the test.

Figure 8: Different viewpoints of a subject against a simple background.

We test the algorithm further on different subjects against complex background. Figure 9 shows the results. The algorithm works very well when all the six facial features can be seen and there is relatively little background clutter. For a smiling face (top row, middle image), the corner of the mouth is detected instead, though the face as a whole is still detected. Other facial expressions may cause the feature detector to miss a feature and the algorithm to fail.

Figure 10 shows some failed cases. The algorithm fails when the contrast in the image has gone too low due to shadows (left image) or when there is too much background clutter (right image). Also, the algorithm will fail when only three out

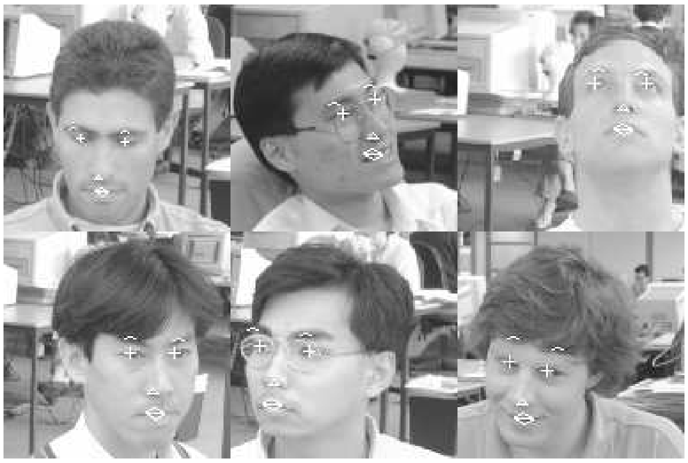


Figure 9: Different subjects against a complex background.

of four features needed to group a PFG are present (e.g. very thin or no eyebrows).

Figure 10: Some failed cases.

6 Conclusion

In this paper, we presented an algorithm which is capable of detecting and localizing human faces in an image given no prior information about the scale, orientation and viewpoint of the faces. The proposed algorithm detects features from the image, groups them using affine invariance, propagates them as evidences in a belief network, and compute the probability of each candidate being a face. The algorithm is shown to work well for fronto-parallel view of faces and is also able to cope with large viewpoint changes. The problems of occlusion, variations in illumination and facial expression were not dealt with in this paper.

7 Future Work

The algorithm that we have proposed is only able to give a rough indication of the presence of a face. It is unable to reject incorrect faces or localize features well.

Therefore, our main future direction of work will be to apply motion constraints to evaluate if a face deforms affinely over time.

We will also apply deformable template matching techniques (e.g. Yuille *et.al.* [13], Cootes and Taylor [3]) to accurately localize the features after our initial estimate is known, identifying and verifying each feature accurately.

References

[1] C. H. Ballard and C. M. Brown. *Computer Vision.* Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[2] J. Canny. A computational approach to edge detection. *IEEE Trans. Patt. Analy. and Machine Intell.*, 8(6):679–698, 1986.

[3] T. F. Cootes and C. J. Taylor. Active shape models — "smart snakes". In D. Hogg and R. Boyle, editors, *Proc. British Machine Vision Conference*, pages 266–275, Leeds, 1992. Springer–Verlag.

[4] V. Govindaraju, D. B. Sher, R. K. Srihari, and S. N. Srihari. Locating human faces in newspaper photographs. In *Proc. Conf. Computer Vision and Patt. Recog.*, pages 549–554, 1989.

[5] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–194, 1988.

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

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo, California, 1988.

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

[9] L.G. Roberts. Machine perception of three-dimensional solids. In J. T. Tippet, editor, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, Cambridge, Massachusetts, 1965.

[10] S. Russell and P. Norvig. *Introduction to Artificial Intelligence.* Prentice Hall, 1995.

[11] D. Tranel, A. R. Damasio, and H. Damasio. Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurobiology*, 38:690–696, 1988.

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

[13] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *Int. Journal of Computer Vision*, 8(2):99–111, 1992.