Finding Initial Estimates of Human Face Location

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

Abstract

This paper describes a method to find initial estimates of face location in an image where the orientation and viewpoint of the faces are not known. Features such as eyes, nose and mouth are detected from the image using quadrature phase filters and grouped into potential face candidates. Affine invariants are used in grouping to overcome the problem of variation in viewpoint. An efficient searching algorithm is proposed to group these features based on the constraints in the geometry. Each face candidate is then evaluated using a belief network which assigns probabilities to each face candidate and rejects improbable ones. A result of 93% accuracy in detecting viewpoint variations is obtained.

1. Background and Motivation

Human face recognition has been an interesting problem attempted by numerous researchers over the years. There are many important applications such as criminal identification, visual surveillance and human computer interfacing which continuously provide the drive and motivation for research in this area.

Most human face recognition algorithms have assumed that the location of human face in the image is known, or that the face can be easily extracted from the background. However, this is not true of most applications. Hence, face detection and localization still remains as an important problem to be solved.

Recent work on face detection are attempted using various techniques: neural networks (Rowley *et al.* [9]), shape statistics (Leung *et al.* [5]), bandpass filtering (Graf *et al.* [3]), ellipse fitting (Jacquin and Eleftheriadis [4]), and colour (Wu *et al.* [11]). The neural networks and shape statistics approach works only for fronto-parallel faces with little variation in viewpoint. The methods based on bandpass filtering and ellipse fitting works only for head and shoulder images with very little background clutter. Wu *et al.* 's method of using colour and fuzzy logic works for more general scenes but it cannot cope with different hair colour, or when the skin coloured regions in the image doesn't form an elliptical shape of a face.

Yow and Cipolla [12]'s work on face detection under different viewpoints makes use of Gaussian derivative filters to detect features. This technique has been shown to work well but it has the problem of having too many false candidates and being unable to reject false candidates. In this paper, we describe the use of quadrature phase filters in feature detection which can significantly improve the robustness of the system and reduce the number of false detection. A new and more efficient method of grouping feature points is also proposed by a new vector representation of the partial face groups.

2. Feature Detection

In Yow and Cipolla [12], the usefulness of Gaussian derivative filters in detecting facial features under different viewpoints was shown. These Gaussian derivative filters, corresponding to low-intensity bar detectors, are able to detect the facial features even though the subject is being viewed from a very large angle from the front. Features from a profile view can be detected too.

However, these filters are not too robust as the facial features often produce a weaker response to these detectors in comparison with some features in the background. A high threshold applied to the filter response to eliminate background features might cause us to lose the facial features as well. A low threshold will let too many background points pass through to the grouping stage, increasing the computational load of the system and the false alarm rate.

To improve robustness, we use a pair of filters in quadrature phase instead of just a single filter for feature detection. The Gaussian derivative filter and its Hilbert transform is used (see figure 1). The quadrature phase filter extracts additional phase information about the features in the image and this additional information can be used to give a stronger response at the desired features.

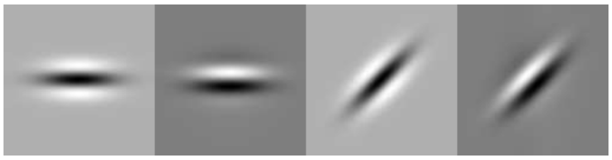

Figure 1. The Gaussian derivative filter and its Hilbert transform, shown at different orientations.

We use the Gaussian derivative filter and its Hilbert transform to produce an oriented energy output $E_\theta = G_\theta^2 + H_\theta^2$ where θ is the orientation of the filter. The convolution output of the filter at orientation θ is G_θ, and the convolution output of the Hilbert transform is H_θ. Freeman and Adelson [2] have made use of this oriented energy approach in conjunction with steerable filters to detect fine edges in an image.

However, the peaks in the energy response E_θ corresponds to the edges in the image instead of the center of the facial features. Hence, we cannot seek for the maximum response using E_θ alone. We observe that the Gaussian filter gives us a peak in the center of the features, and at those locations the energy response is sufficiently high too. Hence, we perform non-maximal suppression in the response G_θ, and at the same time we ensure that the feature points exceeds a certain threshold in the energy response as well.

Faces of different scale and orientation are detected using a family of Gaussian derivative filters and their Hilbert transforms at different scales and orientations. Since the facial features all lie in the same orientation and have approximately the same size, the filter of the right scale and orientation will generate many strong responses in G_θ and thus E_θ.

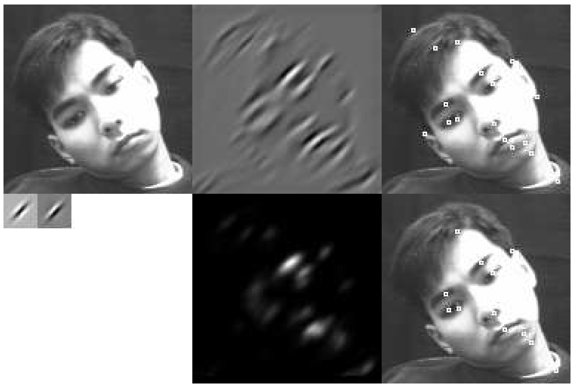

Figure 2. The image, filter output G_θ, energy output E_θ, filters used, features detected using G_θ only, and features detected using G_θ and E_θ.

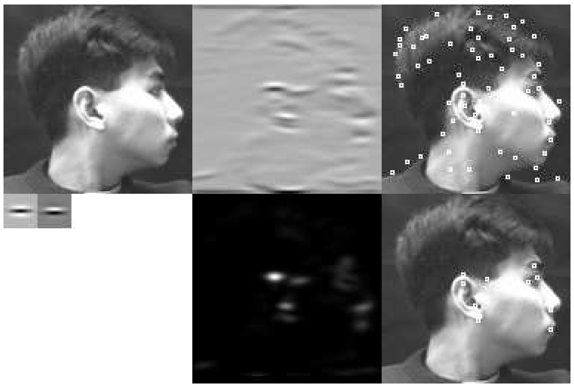

Figure 3. The output as in figure 2 for a different viewpoint. We can see a significant reduction in the number of features to be processed by the next stage of the algorithm.

Figure 2 shows the image, filter output G_θ, E_θ, filter used, extracted features using G_θ only, and extracted features using both G_θ and E_θ. Figure 3 shows the output of the same operations but of a face at a different viewpoint. A very low threshold is used in the non-maximal suppression stage to allow weak features to be retained. We can see from the figures how much background points can be eliminated by introducing the quadrature phase filter.

The advantage of using Gaussian derivative filters is that it can be decomposed into a set of steerable-scalable basis (Freeman and Adelson [2], Perona [7]), thus it can be efficiently implemented to search the image at various scale and orientation.

3. Affine Invariant Grouping

After the feature points are detected, it is necessary to group them into potential face candidates. Under the weak perspective assumption (Roberts [8]), the face can be considered as a plane with the features lying on the plane. The geometric distances between the features will also vary affinely under different scales, orientations and viewpoints.

Under different viewpoints, some of the facial features will be occluded (e.g. an eyebrow and an eye will be occluded when in profile view). Hence, in order to cope with different viewpoints, features are grouped into partial face groups (PFGs) which consists of groups of four features shown in the top of figure 4. Under different viewpoints, at least one of these partial face groups will be seen in the image.

To do an exhaustive combinatorial search of all groups of four points would be computationally expensive, if not impossible. Hence we exploit the knowledge of the geometry of the facial features (the mouth is be-

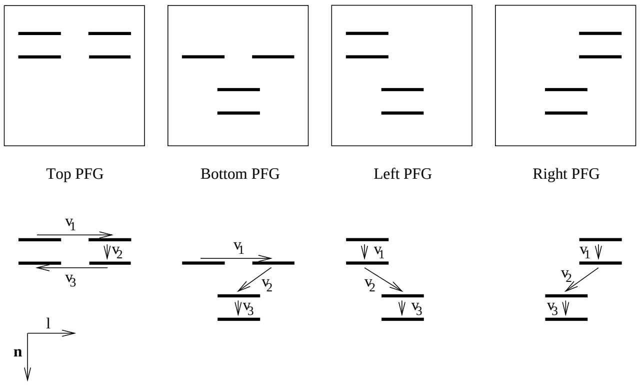

Figure 4. (Top) The different partial face groups (PFGs) used to group the features. (Bottom) The vectors associated with each partial face group.

low the nose, the nose is below the eyes, etc.) and the scale and orientation information (from the quadrature phase filters) to reduce the search space.

The basic idea is that the features should be ordered in the direction perpendicular to the longitudinal direction of the features. In this way, the ordering of the features will indicate the distance from the top of the head. Hence, there is no need to examine groups that have their feature positions in the reverse order.

We define two unit vectors, longitudinal vector $\mathbf{l}$ and the normal vector $\mathbf{n}$ which are respectively parallel and perpendicular to the longitudinal axis of the Gaussian derivative filter that was used to produce the convolution output. These define the coordinate vectors for the image. For every group of four feature points, we derive three vectors, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ as shown at the bottom of figure 4.

The feature points which are detected from the previous section are sorted in order of increasing values of the coordinate in the normal $\mathbf{n}$ direction. If we define the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ such that they will always point from the lower $\mathbf{n}$ coordinate to the higher $\mathbf{n}$ coordinate, the possible combinations of four points are greatly reduced. Moreover, by examining the total length along the $\mathbf{n}$ direction of a group of four points, we can immediately discard the group of points if the total length of the group exceeds the length of a human face at the appropriate scale. In this way, the combinatorial search is further reduced.

The affine invariants under the weak-perspective assumption determine certain geometric relationships between the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ for the different partial face groups. The contraints used to group the features are shown in table 1.

In table 1, s is a scale factor determined by the scale

Top PFG	*Bottom PFG*	*Left & Right PFGs*
$v_1 \parallel v_3$.	$v_1 \perp v_3$.	$v_1 \parallel v_3$.
$\frac{\lvert v_1 \rvert}{\lvert v_3 \rvert} = 1$.	$\frac{\lvert v_1 \rvert}{v_2.l} = k_3$.	$\frac{\lvert v_1 \rvert}{\lvert v_3 \rvert} = k_6$.
$\frac{v_1.l}{s\lvert l \rvert} = k_1$.	$\frac{v_2.n}{\lvert v_3 \rvert} = k_4$.	$\frac{v_2.n}{\lvert v_3 \rvert} = k_4$.
$\frac{v_2.n}{s\lvert n \rvert} = k_2$.	$\frac{v_1.l}{s\lvert l \rvert} = k_1$.	$\frac{v_2.l}{s\lvert l \rvert} = k_7$.
	$\frac{v_2.n}{s\lvert n \rvert} = k_5$.	$\frac{v_2.n}{s\lvert n \rvert} = k_5$.

Table 1. The geometric constraints used for grouping features.

of the Gaussian derivative filter used in the convolution. k_1, k_2, k_3, k_4, k_5, k_6 and k_7 are constants which are obtained from averaging measurements in a prior set of face images.

If any of the constraints are violated, the PFG is discarded immediately. If two or more valid PFGs overlap, they are combined to form a face candidate consisting of six points if the feature positions are coherent.

4. Belief Network Reasoning

Due to the fact that many false features may be detected, many of the face candidates that is obtained may be false. Therefore, it is necessary to perform some reasoning process to reject the false face candidates.

A belief network (Russell and Norvig [10]) is utilized to propagate evidence and evaluate belief in the face candidates. The human face is modelled as a direct, acyclic graph (DAG) consisting of one parent node (the face candidate) and four child nodes (each of the four partial face groups). As partial face groups are detected from the image, the total belief in the presence of a face is updated. A diagrammatic representation is shown in figure 5.

An elegant evidence propagation algorithm is given by Pearl [6] in which no ad hoc adjustments or unfounded assumptions of conditional independence is made about the system. This method is superior to simply adding up the initial probabilities of the partial face groups.

The likelihood that a partial face group obtained from the previous section is part of an actual human face, can be measured from the strength of the filter response of each feature and the geometric errors in the grouping. The initial probabilities of each partial face group are thus obtained by summing the normalized filter response and the inverse of the normalized geometric errors in each PFG in the grouping stage:

$$P = \frac{\Sigma R_i}{k_1}\left(1 - \frac{\Sigma \epsilon_j}{k_2}\right)$$

where

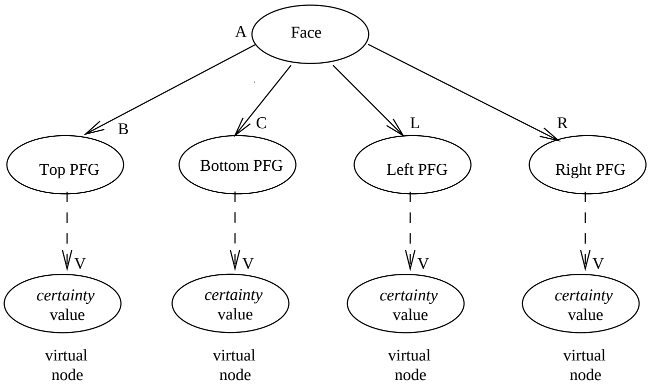

Figure 5. The face modelled as a belief network. Each of the child node will propagate evidence to the parent node depending on whether the associated partial face group (PFG) is present.

R_i is the filter response of the ith feature.
ϵ_j is the jth geometric error in the PFG, and
k_1 and k_2 are normalizing constants.

Letting a_j be the possible values that a node A can take, λ be the message passing from a child node to its parent, and π be the message passing from a parent node to all 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 advantage of using a belief network is that high-level knowledge is built into the system through the conditional independence relationship. The evidence in each variable can be evaluated independently (and in parallel) and combined in a non ad-hoc manner using Pearl's algorithm.

5. Results

The algorithm is tested on a subject against a simple background at different viewpoints. The up-and-down 'nodding' viewpoints are tested at $-45°$, $0°$ and $+45°$ vertically. The left-to-right 'turning' viewpoints ranges between $\pm 90°$ at $45°$ interval. Three different scales and three different orientations are also used. Of the 135 test images, 9 failed completely, giving a success rate of 93%. Some of the results are shown in figure 6.

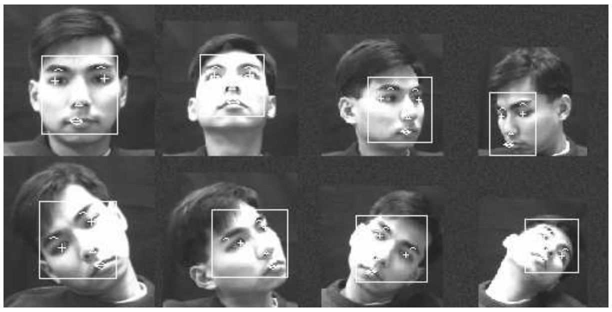

Figure 6. Results of algorithm on a single subject against a simple background. Different scale and orientation are also used. A detection rate of 93% is achieved for the 135 images at different scale, orientation and viewpoints.

The algorithm is further tested on different subjects against complex background. The size of the faces are kept more or less constant but the viewpoint of the subject is allowed to vary. 2 images of 10 different subjects were used of which 3 failed. Some of the results are shown in (figure 7). The results show that the algorithm is able to detect faces at large angles of viewpoint against a cluttered background. The algorithm is also not affected by partial occlusions such as the subject wearing spectacles. However, as the algorithm groups features purely due to geometric constraints, features detected from the background clutter may turn out to be in the geometrically correct position. This will cause false faces to be detected.

The failed cases are shown in figure 8. The main reason that cause the algorithm to fail is when features detected in the background falls into geometrically correct position, and that these features have stronger responses compared to the actual features. The use of quadrature phase filters have significantly reduced the number of falsely detected features but when the size of the face is small compared to the image, the number of background features detected will still be large compared to the facial features.

6. Discussion and Future Work

The main advantage of this algorithm is that it is capable of detecting face locations within a large range of viewpoints, especially when all six facial features can be seen. For larger angles, some partial face groups can still be detected but it will be harder to eliminate false partial faces from actual ones.

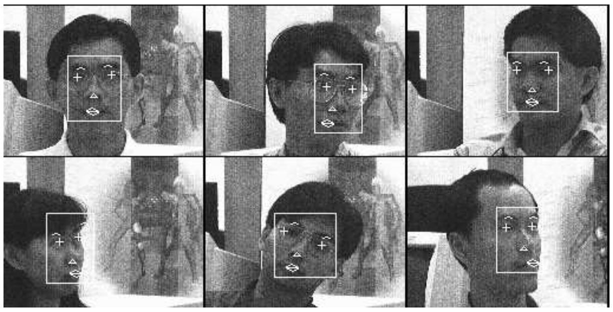

Figure 7. Results of algorithm on various subjects against complex background. Faces with large angle of viewpoint are also detected (bottom left).

Figure 8. Some failed cases. Features are falsely detected due to background clutter or failed to be picked up by the feature detector. Other types of constraints (such as motion) will be needed to reject false faces.

Although the Gaussian derivative filter provides some invariance to illumination, the algorithm will in general fail for extreme illumination conditions. Also, occlusions and different facial expression are not addressed in this paper.

To eliminate false faces, other types of contraints (such as motion) will be needed. We will study the use of deformable templates (Yuille *et al.* [13]) and active shape models (Cootes and Taylor [1]) to verify each feature detected. We will also make use of higher level features will be used to improve the perceptual grouping stage.

7. Acknowledgements

The author would like to thank Dr. John Daugman for helpful discussions on quadrature phase filters. The author would also like to thank all the staff of the Centre for Graphics and Imaging Technology, Nanyang Technological University for providing the test images in this paper. Kin Choong Yow is funded by the 1993 Nanyang Technological University Overseas Scholarship.

References

[1] 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.

[2] 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.

[3] H. P. Graf, T. Chen, E. Patajan, and E. Cosatto. Locatin faces and facial parts. In *International Workshop on Automatic Face and Gesture Recognition*, pages 41–46, Zurich, 1995.

[4] A. Jacquin and A. Eleftheriadis. Automatic location tracking of faces and facial features in video signal. In *International Workshop on Automatic Face and Gesture Recognition*, pages 142–147, Zurich, 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 Computer Vision*, pages 637–644, MIT, Boston, 1995.

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

[7] 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.

[8] 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.

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

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

[11] H. Wu, Q. Chen, and M. Yachida. An application of fuzzy theory: Face detection. In *International Workshop on Automatic Face and Gesture Recognition*, pages 314–319, Zurich, 1995.

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

[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.