Finding Initial Estimates of Human Face Location

Kin Choong Yow and Roberto Cipolla

CUED/F-INFENG/TR 239

October 1995

Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
England

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 human 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. 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 93% 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.* [14]), 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. Most human face recognition algorithms have also 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.

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], Chow and Li [1], Yang and Huang [16]) have assumed a roughly fronto-parallel viewpoint and were unable to cope with significant changes in viewpoint.

More recent work on face detection are attempted using various techniques: neural networks (Rowley *et al.* [12]), shape statistics (Leung *et al.* [8]), bandpass filtering (Graf *et al.* [5]), ellipse fitting (Jacquin and Eleftheriadis [6]), and colour (Wu *et al.* [15]). 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 [17]'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

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.* [18]) 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.

In Yow and Cipolla [17], 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.

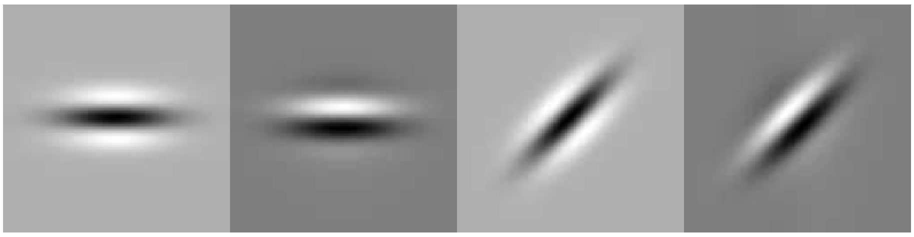

Figure 1: The Gaussian derivative filter and its Hilbert transform, shown at different orientations.

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 [3] 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 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

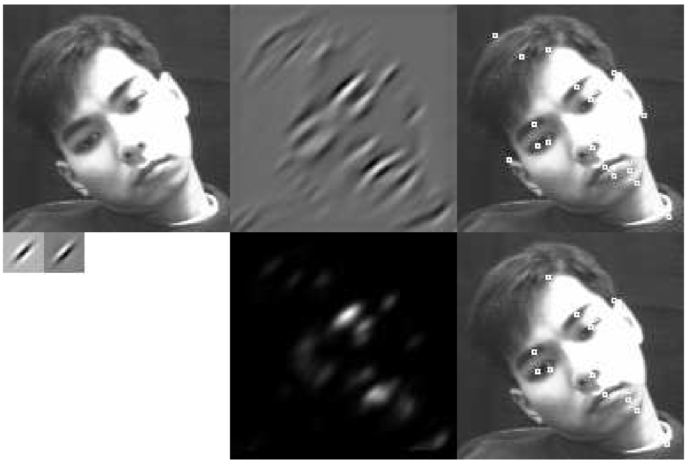

Figure 2: The image, filter output G_θ, energy output E_θ, filters used, features detected using G_θ only, and features detected using G_θ and E_θ.

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.

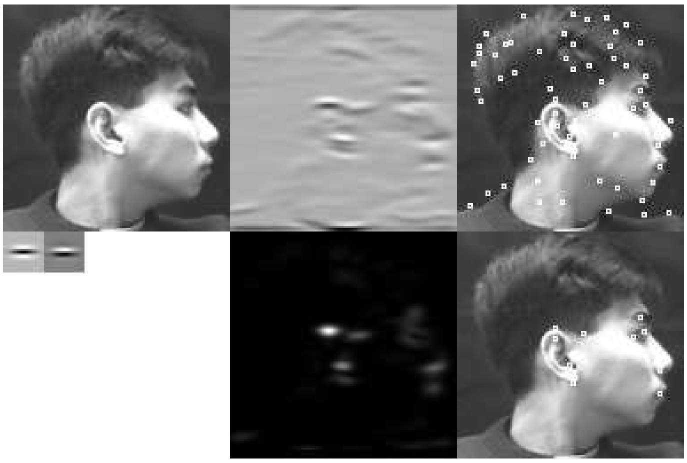

Figure 3: The output as in figure 2 for a different viewpoint. We can see a significant reduction in the number of features to be processed by the next stage of the algorithm.

The advantage of using Gaussian derivative filters is that it can be decomposed into a set of steerable-scalable basis (Perona [10]), thus it can be efficiently implemented to search the image at various 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.

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 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 [11]), 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 below 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

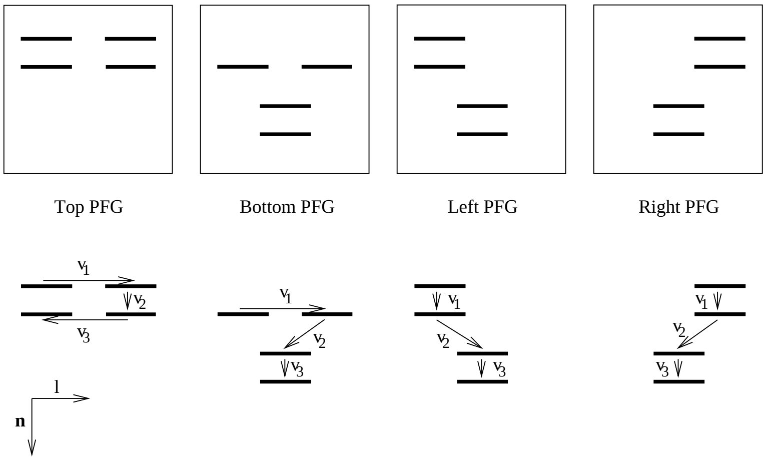

Figure 4: (Top) The different partial face groups (PFGs) used to group the features. (Bottom) The vectors associated with each partial face group.

in order of increasing values of the coordinate in the normal **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 **n** coordinate to the higher **n** coordinate, the possible combinations of four points are greatly reduced. Moreover, by examining the total length along the **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.

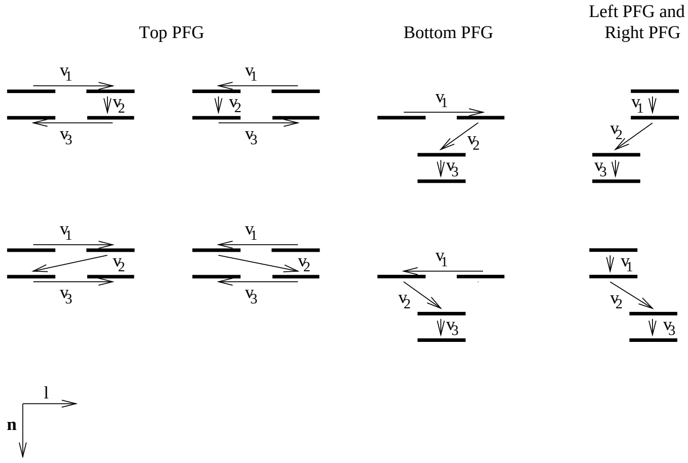

Figure 5: The different configurations of the vectors that forms the different PFGs. From the figure we notice that for the affine invariants, we need to use the length of the vector projection instead of the length of the vector itself.

The possible configurations of the vectors that forms the different partial face groups are shown in figure 5. We observe that we cannot simply use the magnitude of the vectors for comparisons, but should use the projection of these vectors onto the coordinate vectors **l** and **n**.

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.

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.

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

To reject false face candidates, we make use of a belief network to propagate evidence and evaluate belief. 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 [13]).

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 [9] and a solution for general networks is given in Lauritzen and Spiegelhalter [7].

We model the human face as a DAG consisting of one parent node (the face candidate) and four child nodes (each of the four partial face groups). A diagrammatic representation is shown in 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.

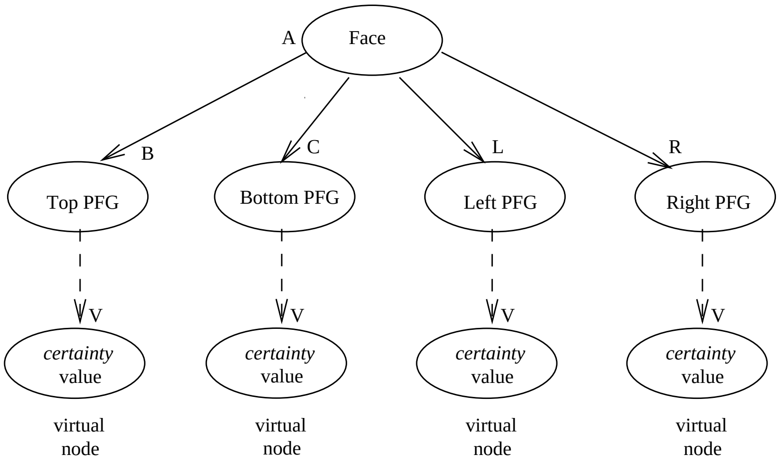

Figure 6: The face modelled as a belief network. Each of the child node will propagate evidence to the parent node depending on whether the associated partial face group (PFG) is present.

The evidence propagation algorithm given by Pearl [9] makes no ad hoc adjustments or unfounded assumptions of conditional independence 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

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.

For the DAG shown in figure 7, let A be the variable at a particular node, and A can have values a_j where $j = 1, 2, ..., m$ 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 × 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.

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 7.

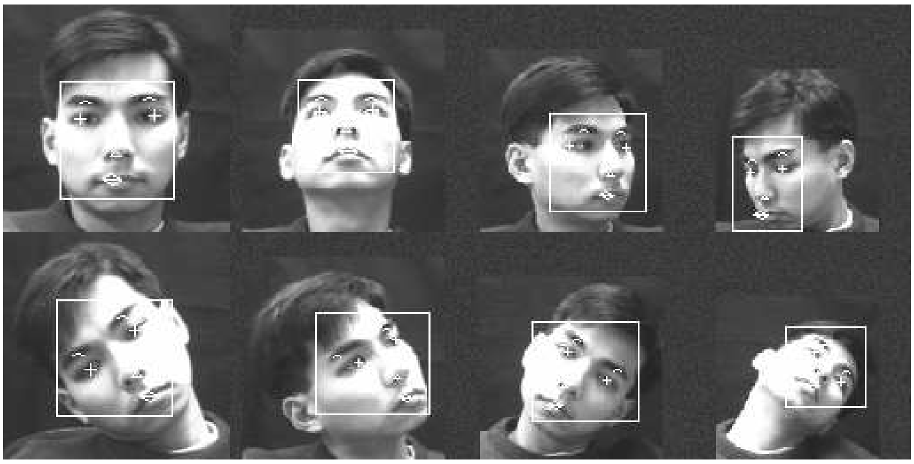

Figure 7: Results of algorithm on a single subject against a simple background. Different scale and orientation are also used. A detection rate of 93% is achieved for the 135 images at different scale, orientation and viewpoints.

Figure 8 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.

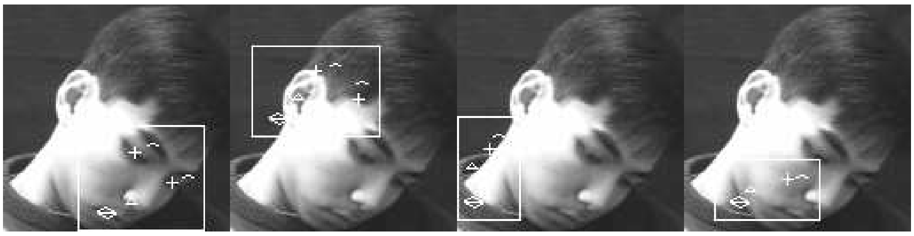

Figure 8: Possible candidates of faces detected, ranked from left to right.

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 9). 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 10. 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.

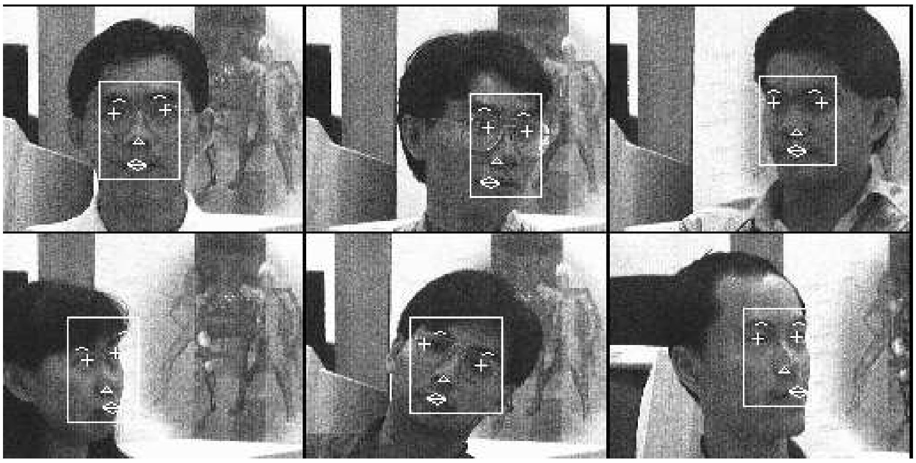

Figure 9: Results of algorithm on various subjects against complex background. Faces with large angle of viewpoint are also detected (bottom left).

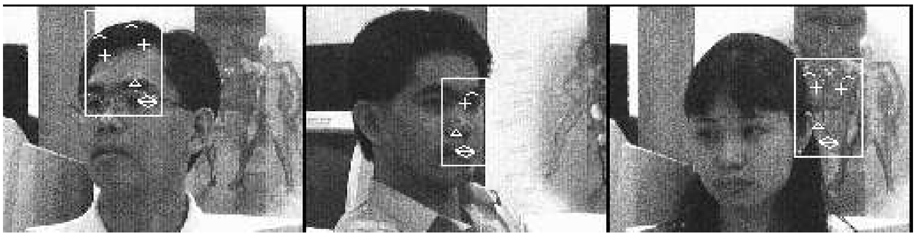

Figure 10: Some failed cases. Features are falsely detected due to background clutter or failed to be picked up by the feature detector. Other types of constraints (such as motion) will be needed to reject false faces.

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.

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.* [18]) and active shape models (Cootes and Taylor [2]) 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 authors would like to thank Dr. John Daugman for helpful discussions on quadrature phase filters. The authors 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] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[2] 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.

[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] 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] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.