A Probabilistic Framework for Perceptual Grouping of Features for Human Face Detection

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

Abstract

Present approaches to human face detection have made several assumptions that restrict their ability to be extended to general imaging conditions. We identify that the key factor in a generic and robust system is that of exploiting a large amount of evidence, related and reinforced by model knowledge through a probabilistic framework. In this paper, we propose a face detection framework that groups image features into meaningful entities using perceptual organization, assigns probabilities to each of them, and reinforce these probabilities using Bayesian reasoning techniques. True hypotheses of faces will be reinforced to a high probability. The detection of faces under scale, orientation and viewpoint variations will be examined in a subsequent paper.

1. Introduction

Detecting human faces in a scene is an important problem for researchers in human face processing. Many face recognition algorithms have either assumed that the face has been cropped from the image (Craw *et.al.* [4], Turk and Pentland [15]), or they have assumed some constraints about the face and/or background such that the face detection process becomes trivial (Chow and Li [3]). In a general imaging environment these assumptions are certainly not true. As such, face detection still remains largely an unsolved problem.

2. Present approaches

The present approaches to face detection can be model-based, feature-based, neural network-based or colour-based. 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 [16], and Lanitis *et.al.* [6]. 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.* [7], Sumi and Ohta [12], and Yow and Cipolla [17] 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.

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 [13], and Rowley *et.al.* [9]. 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.* [2], Dai and Nakano [5]). Chen's approach 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 grey-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 grey-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 evidence for a feature to be present comes largely from 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, grey 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. The face model

We always need a model of the object in any object recognition task. A model of an object in terms of low level image features (such as edges, corners, etc.) is always very difficult because the image structure changes very drastically in different images due to noise and changes in imaging conditions. As such, models of explicit shape (e.g. deformable template models - Yuille *et.al.* [20]), only work well in high resolution and relatively noise-free images. However, a model of the object described in terms of higher level features (such as a face described in terms of eyes, nose and mouth), is usually quite stable and robust.

We model the face as a plane with 6 oriented facial features. To cope with occlusion and missing features (eyebrows, usually), we decompose the face model into components consisting of 4 features. These components are common occurrences of faces under different viewpoints and different identity, and are thus called Partial Face Groups (or PFGs). 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 (see fig. 1).

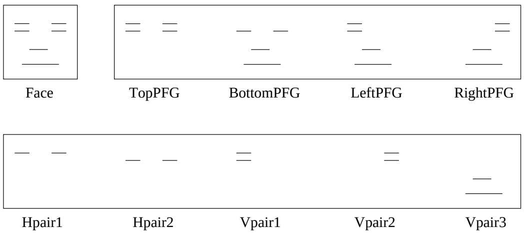

Figure 1. The face model and its components.

In order for the feature detection to be robust we have to use image features that are invariant to changes in scale and illumination intensity. We observe that at low resolutions, all the 6 facial features will appear only as dark elongated blobs against the light background of the face. And since edges are illumination invariant to a large extent, we model the 6 facial features as pairs of oriented edges as 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 are only used to provide evidence in labelling the facial feature and is not an important criteria in the detection of the feature.

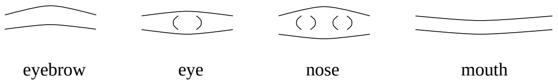

Figure 2. The facial feature models.

5. Perceptual grouping

With such a low resolution model, we can expect a lot of false positive feature candidates. We will therefore need a perceptual grouping framework that will group true feature candidates into faces based on geometrical, grey-level and spatial evidence. Feature candidates that cannot be grouped must be false and will be discarded.

We model our face detection process as a two stage model of perception (Triesman [14]). The first stage, which is described as pre-attentive perception, extracts image information into points and regions of interest, which directs the attention of processing efforts of the next stage. The second stage, the attentive stage, will perform grouping and reasoning activities based on the detection and identification of meaningful object groups in the image.

5.1. 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 is achieved by filtering the image using a preattentive filter (a second derivative of Gaussian used in Yow and Cipolla [17]) 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.

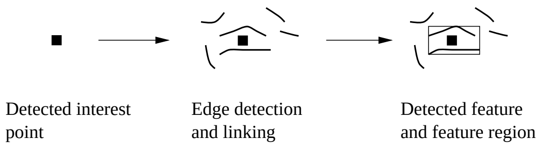
Detected interest point
Edge detection and linking
Detected feature and feature region

Figure 3. Preattentive feature selection.

Measurements of the region's image characteristics (such as edge length, edge strength, grey-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 4 classes of facial features, namely, eyebrow, eye, nose, and mouth. If the facial feature does not belong to any of them, it is discarded from the list.

5.2. Attentive feature grouping

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

The rules for grouping the facial components are divided into 2 groups. One group encodes geometric information such as length, orientation, inter-feature distance, etc., and the other group encodes spatial information about whether there should be edges of a particular strength and orientation at some spatial location in the feature region.

These rules are represented by values in a geometric feature vector $\mathbf{x}_g$ and a spatial feature vector $\mathbf{x}_s$, in the same fashion as the facial feature vector $\mathbf{x}$ in the earlier section. The Mahalanobis distance $\mathcal{M}_{ij}$ of these feature vectors are used to determine its membership in the class.

For efficiency, the geometric feature vector $\mathbf{x}_g$ is examined first. If the feature vector fails to be a valid instance of the geometric class, it is discarded, saving the more expensive computation of the spatial feature vector $\mathbf{x}_s$.

The measurements we choose for the vector $\mathbf{x}_g$ are :

1. the ratio of feature lengths to image size.
2. the ratio of feature lengths between features.
3. the aspect ratio of the feature region.
4. the ratio of inter-feature distances.
5. the difference in orientation between features.

and the measurements we choose for vector $\mathbf{x}_s$ are :

1. the number of directional edgels in a region.
2. the ratio of edge strengths in a region to edge strengths of facial features.
3. the mean grey level of a region.
4. the variance in the grey level distribution of a region.

This grouping process is effective in removing false positives because a lot of geometric and grey-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 itself (see fig. 4).

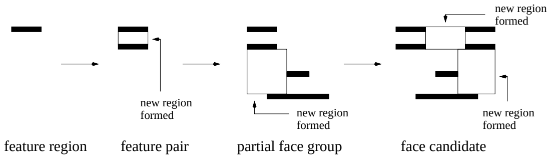

Figure 4. Attentive feature grouping.

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, so the processing time is kept small throughout the whole algorithm.

6. Probabilistic framework

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. We thus propose a probabilistic framework to assign and propagate probabilities among the facial features and face groups so that we will achieve a high confidence rate for true positive faces.

We make use of Bayesian networks (or belief networks), which are directed acyclic graphs, to propagate evidence. Belief networks have nodes representing random variables and arcs signifying direct dependencies specified in terms of conditional probabilities (Sarkar and Boyer [11]). Each node can take either of 2 values, True or False, and has a conditional probability table (or CPT) describing the conditional probability of each value given each possible combination of the values of the parent nodes.

The entries in the CPT can be estimated directly by using the statistics of the set of examples (Russell *et.al.* [10]). The crucial value here is the prior probability of the "face" node, and this is often hard to estimate. The choice of an appropriate prior clearly depends on the complete space of hypothesis, and we may assume an uniform prior for our case.

The belief network used in our previous approach [17] has a root node (the "face" node) and 4 child nodes (one for each PFG). This was shown to be highly effective for fronto-parallel view of faces because all 4 PFGs can be detected in this view, giving a large amount of evidence for true face candidates. However, for profile views, the probability of the face remained low because only one PFG can be found in the image.

To overcome this, we propose a new belief network, using the facial features as child nodes instead of the PFGs (fig. 5). The belief network now has 6 child nodes instead of 4. Profile view of faces will thus have 4 pieces of evidence (facial features) out of 6, instead of 1 (face group) out of 4 previously. This leads to a better capability of detecting profile views of faces.

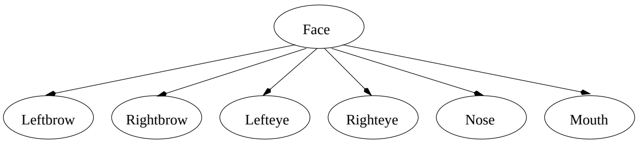

Figure 5. Belief network.

So how do we update and improve the probabilities of these child nodes using model knowledge ? As mentioned earlier, one source of evidence that is often overlooked is the presence of a neighbouring feature (e.g. presence of another eye next to an eye candidate). To harness this extra piece of evidence, we build a second belief network (fig. 6) to reinforce the belief of each feature based on the presence of neighbouring features.

When evidence for a facial feature becomes available, a virtual node is created (the "evidence" node) and instantiated, allowing the evidence, specified in the form of a probability, to propagate through the entire network and update all the other nodes. The resulting effect is a large increase in the probabilities of the feature candidates which are true facial features.

We use a propagation algorithm for singly connected networks given by Pearl [8] which does not make any unfounded assumption of the conditional independence of the system. In Pearl's algorithm, each node when instantiated with a piece of evidence will modify its parent or child nodes based on the conditional probabilities between the nodes. These parent or child nodes will further modify their parent and child nodes, thus propagating the evidence throughout the network (see [19] for details). The main difference between this propagation algorithm and the one for trees (used in our previous work [17]) is that nodes in a singly connected network can have more than one parent. Our belief network structure in fig. 6 clearly requires this.

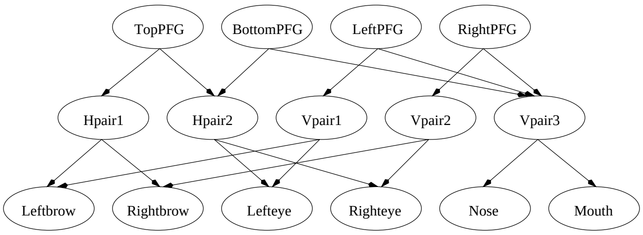

Figure 6. Reinforcement belief network.

After the perceptual grouping process described in the earlier section. Each face candidate will have between 4 to 6 features associated with it. A belief network is initialized for each face candidate and virtual nodes are created for each feature and component face groups that is found in the process. 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 :

$$P_i = (1 - \frac{\mathcal{M}_{ij}}{\tau_j})$$

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.

7. Implementation

In this paper, we are interested to evaluate the validity of the framework, rather than trying to solve the scale and orientation problem. Hence we implement the algorithm assuming that the orientation is vertical, the viewpoint is fronto-parallel, and we allow the user to specify the filter scale. The variations in scale, orientation and viewpoint are treated in a subsequent paper (see Yow and Cipolla [18] this volume).

7.1. Learning the feature class space and conditional probabilities.

A set of 40 images taken of different subjects under different scale and slightly different viewpoint is used as a training set. Facial features are marked by hand and the algorithm is run through these test images, making the necessary measurements to define each class space. The frequency of occurrences of each feature and the component face groups are also measured and entered into the conditional probability tables.

7.2. Perceptual grouping

Interest points are first detected by spatial filtering using a second derivative of Gaussian described in [17]. Edge detection is then performed using a Canny edge finder with both hysteresis threshold set to zero. A standard boundary following algorithm (Ballad and Brown [1]) is used to link the edges. The results after verification with each feature class are shown in fig. 7.

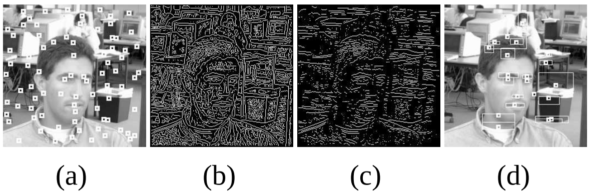

Figure 7. (a) Interest points obtained from spatial filtering (81 points). (b) Canny edge detection with zero threshold. (c) Linked edges of approximately horizontal orientation. (d) Feature regions detected (21 points).

The list of feature candidates is then examined to form pairs, and each horizontal pair and vertical pair is further examined to form partial face groups. If any two partial face groups have some component features that are the same, they are combined to form a 5- or 6-feature face candidate. If not, each PFG will become a 4-feature face candidate. The results for the perceptual grouping stage is given in fig. 8.

7.3. Evidence propagation

Each facial feature that is detected is assigned 4 probability values, P_{brow}, P_{eye}, P_{nose} and P_{mouth}. If the Mahalanobis distance of the facial feature in a particular feature class is greater than the admission threshold, the facial feature is given a probability value of zero for that feature class.

Each of these probabilities is propagated through the reinforcement network by creating and instantiating a virtual node in turn. Fig. 9 shows the face candidates and the final probabilities for 2 subjects found by the algorithm. As these faces cannot exist simultaneously because they overlap, only the face with the highest probability is selected.

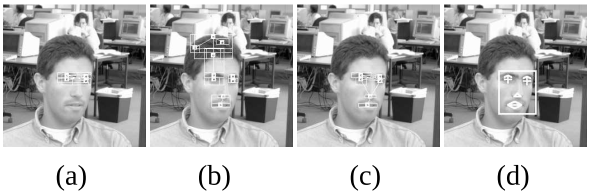

Figure 8. (a) Horizontal pairs (4 pairs). (b) Vertical pairs (7 pairs). (c) Partial face groups (1 top, 2 bottom, 1 left, 1 right). (d) Face candidates detected (1 face).

For subject 1, in fig. 9a, the top PFG is not found in the process and so the computed probability of the face candidate is lower. Moreover since the hypothesized eye location (on the right) is actually a brow, the image evidence that is propagated in this case is actually P_{eye} which is very low compared to fig. 9b.

For subject 2, only the bottom partial face groups is found in the first case, resulting in a low probability. Clearly, without the use of the probabilistic framework and the reinforcing of evidence, the difference between the true and false positive candidates will be very close, making it very difficult to successfully reject the false candidates.

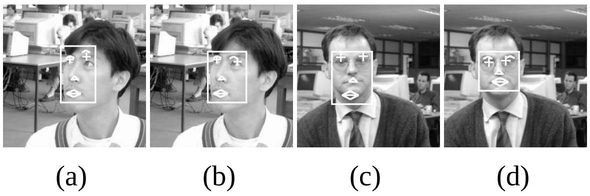

Figure 9. Face candidates found for 2 subjects. (a) Probability = 0.6578. (b) Probability = 0.9255. (c) Probability = 0.5045. (d) Probability = 0.9468.

8. Results

We test the algorithm on 100 256x256 images taken from subjects sitting in front of a workstation mounted with a Pulnix monochrome CCD camera. The user specifies the filter scale at run time, and the algorithm takes about 10 seconds to run on a SUNSparc20 workstation.

Of the 100 test images, 92 are successfully detected (92% detection rate). Some of the successful results are shown in fig. 10. We find that the algorithm is able to cope with small variations in scale, orientation and viewpoint, although the scale is specified and we have made the assumption that the orientation is vertical and the viewpoint fronto-parallel. Presence of glasses, occlusion and absence of facial features are also tolerated to some extent.

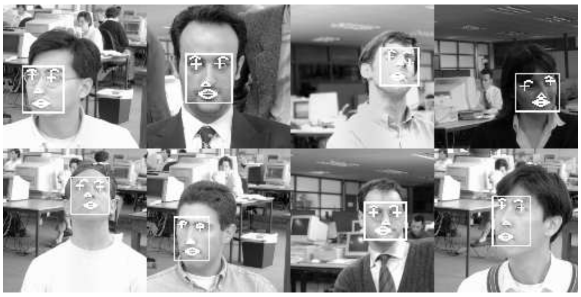

Figure 10. Result of face detection on various test images.

Some of the unsuccessful cases are shown in fig. 11. In the first image, the subject's eyebrows are too close to the eyes and are incorrectly located. The right eyebrow of the subject in the second image coincides nicely with a dark strip in the background, and is thus treated as a single long feature. In the third image, the face has rotated beyond the angle that the algorithm can cope.

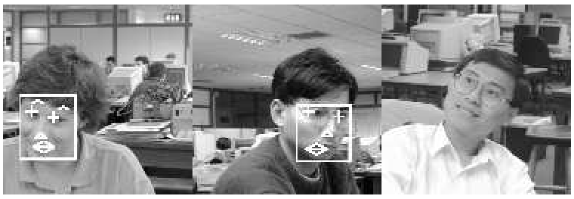

Figure 11. Some unsuccessful cases.

9. Conclusion

We have proposed a face detection framework which detects interest points using spatial filters, groups them into face candidates using geometric and grey-level information, and selects true faces using a belief network. The confidence of true positive faces is improved by using a large amount of evidence in a probabilistic framework. The algorithm is shown to be able to work for small variations in scale, orientation and viewpoints of the face.

References

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

[2] 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.

[3] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[4] I. Craw, D. Tock, and A. Bennett. Finding face features. In G. Sandini, editor, *Proc. 2nd European Conf. on Comp. Vision*, pages 92–96, Italy, 1992. Springer–Verlag.

[5] Y. Dai and Y. Nakano. Face-texture model-based on SGLD and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017, 1996.

[6] 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.

[7] 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.

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

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

[11] 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.

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

[13] 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.

[14] A. Triesman. Perceptual grouping and attention in visual search for features and objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):194–214, 1982.

[15] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[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. Finding initial estimates of human face location. In *Proc. 2nd Asian Conf. on Comp. Vision*, volume 3, pages 514–518, Singapore, 1995.

[18] K. C. Yow and R. Cipolla. Detection of human faces under scale, orientation and viewpoint variations. In *Proc. 2nd Int. Conf. on Auto. Face and Gesture Recog.*, Vermont, USA, 1996.

[19] K. C. Yow and R. Cipolla. Feature-based human face detection. Technical Report CUED/INFENG/TR249, University of Cambridge, August 1996.

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