Face Detection by Learned Affine Correspondences*

Miroslav Hamouz, Josef Kittler, Jiri Matas, and Petr Bílek**

Centre for Vision, Speech and Signal Processing
University of Surrey, United Kingdom
{m.hamouz,j.kittler,g.matas}@eim.surrey.ac.uk

Abstract. We propose a novel framework for detecting human faces based on correspondences between triplets of detected local features and their counterparts in an affine invariant face appearance model. The method is robust to partial occlusion, feature detector failure and copes well with cluttered background. Both the appearance and configuration probabilities are learned from examples. The method was tested on the XM2VTS database and a limited number of images with cluttered background with promising results – 2% false negative rate – was obtained.

1 Introduction

Human face detection and precise localisation play a major role in face recognition systems, significantly influencing their overall performance. In spite of the considerable past research effort it still remains an open problem. The challenge stems from the fact that face detection is an object-class recognition (categorisation) problem, where an object to be recognised is not just a previously seen entity under different viewing conditions, but rather an instance from a class of objects sharing common properties such as symmetry and shape structure. Existing face detection algorithms may be classified according to different criteria, but for our purpose the two following categories are appropriate.

Holistic Face Models In this approach, a representation of the image function defined over manually selected image region containing a face is learned from examples. During detection, probability of an image patch belonging to the face class is evaluated (or the patch is passed into a face - non-face classifier). This probability or classification must be computed for all possible positions, scales and rotations of the face. Typical examples of this method are the work of Moghaddam and Pentland [MP96, MP95] and Sung and Poggio [SP98]. Holistic methods suffer from at least three problems. During detection, a computationally exhaustive search through several scales and rotations has to be carried out in order not to miss any instances of a face in the image. Secondly, the human face is as a whole a highly variable object (class) from the appearance point of

* This work was supported by the EU Project BANCA

** Currently at CMP, Czech Technical University, Prague, bilek@cmp.felk.cvut.cz

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 566–575, 2002.

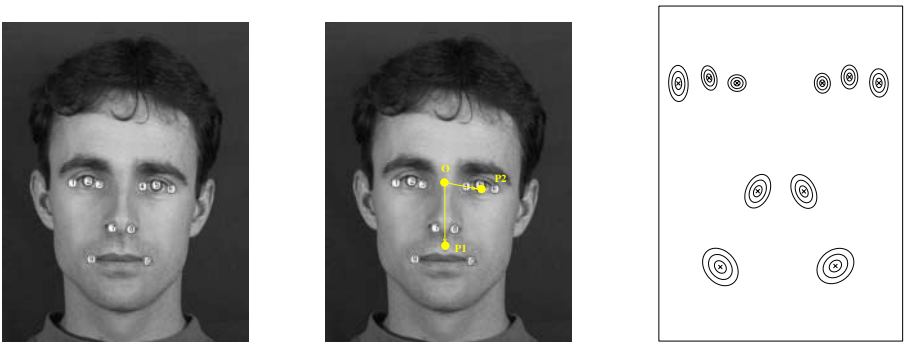
view thus making modelling of photometric effects difficult. Thirdly, effects of occlusion, e.g. by glasses, hair, beards or other objects are difficult to overcome.

Local Feature Face Models In this framework, local feature detectors are used. A face is represented by a shape (configuration) model together with models of local appearance. The most popular of such methods has been the Dynamic Link Architecture, where the preferred shape is defined by an energy function and local appearance is captured in the so-called jets, i.e. responses of Gabor filters [LVB+93, KP97]. Typically, the positions of local features are chosen manually and the appearance models are learned from examples. The work of Burl and Perona [BLP95, BP96, BWP98] is a rare example where an attempt is made to learn the local feature detectors automatically. Other approaches in this group include [VS00, SK98].

Some local methods formulate face detection as a search (minimisation) in the (in principle) continuous pose space. However, the problem can be formulated as a combinatorial search for correspondence between the (hopefully few) responses of the local detectors and the face model. Especially if the local feature detectors produce a small number of false positives, this could be significantly less computationally expensive.

In this paper we present a novel method which addresses these issues. We introduce a detection framework which exploits the advantages of both approaches and at the same time avoids their drawbacks. We use small local detectors to promote fast processing. Face location hypotheses are generated by the correspondence between the detected features and the model. In hypothesis verification all the available photometric information is exploited. The crucial concept of our method is a face space in which all the faces are geometrically normalised and consequently photometrically correlated. In such space, both the natural biological shape variability of faces and distortions introduced by scene capture are removed. As a result of face normalisation, the face features become tightly distributed and the whole face class very compact. This greatly simplifies the face detection process.

In contrast to holistic methods, our search for a face instance in the image is navigated by the evidence coming from the local detectors. By using correspondences between the features in the face space and features found by local detectors in the image a full affine invariance is achieved. This is substantially different from the holistic methods, where face patch is not affinely aligned and the detector has to be trained to cope with all the variability. Additionally, the search through subsets of features which invoke face hypotheses is reduced using geometric feature configuration constraints learned from the training set. The method is robust to occlusion since any triple of features is sufficient to instantiate a face hypothesis. In contrast to existing local methods, as the feature distributions in the face space are very compact, the geometric constraints become very selective. Moreover, all the photometric information is used for hypotheses verification. In other words any geometrically consistent configuration of features gets a favourable score, but serves only as a preliminary evidence of



(a) Manually selected features

(b) Co-ordinate system

(c) Variability of features in the face space

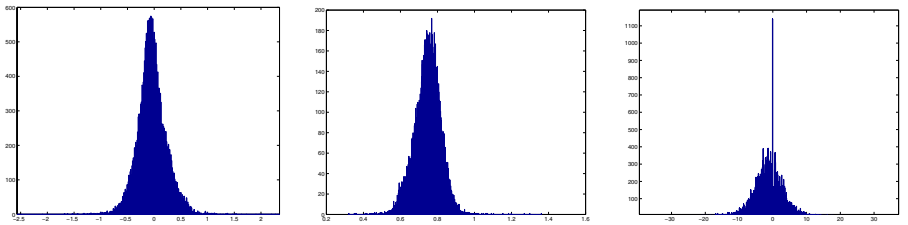
Fig. 1. Construction of face space co-ordinate system

a face being present. The final decision whether a face is present or not is made by comparing the underlying photometric information (or its function) against a fully affinely aligned appearance model employed by the detector.

The aim of this paper is to present the framework of the proposed approach. First, the concept of the face space is introduced in Section 2. Next we show how geometric constraints derived from the probabilistic model learned on training data can be used to prune the list of face hypotheses significantly. The probabilistic feature configuration model and "feature location confidence regions" used for pruning are described in Section 3. The speed up is demonstrated experimentally on the commonly used XM2VTS database [MMK+99] in Section 4. The paper is concluded in Section 5.

2 Proposed Methodology

Face Space In order to reduce the inherent face variability, each face is registered in a common coordinate system. We assume that the geometric frontal human face variability can be, to a large extent, modelled by affine transformations. To determine an affine transformation from one face to another, we need to define correspondences of three reference points. As a good choice we propose the midpoint between the eye coordinates, one eye coordinate and the point on the face vertical axis of symmetry half way between the tip of the nose and the mouth as illustrated in Figure 1(b). We refer to the coordinate system defined by these facial points as "face space". We established experimentally that the total variance of the set of XM2VTS training images registered in the face space was minimised by this particular selection of reference points. Note that two of these reference points are not directly detectable. However, their choice is quite effective at normalizing the width and the length of each face. Consequently



(a) Histogram of shear (b) Histogram of squeeze (c) Histogram of rotation (in degrees)

Fig. 2. Histograms of transformation parameters on XM2VTS

in the coordinate system defined by these points, the ten detectable face features, shown in Figure 1(a) (left and right eye corners, eye centres, nostrils and mouth corners) become tightly clustered. The compactness of the class of faces in the face space was verified experimentally, see Figure 1(c) where the position variance of the detectable face features is depicted.

Face Detection Method Our method uses local feature detectors, which are optimised to detect the 10 key face features shown in Figure 1(a). Any triplet of face features detected in the image allows us to determine the affine transformation that would map the features and therefore the whole face into the face space. As shown in the previous section, the distribution of the key features is very compact and therefore it can be adequately represented by their mean vectors. Mapping the image features onto these mean positions will enable us to register the hypothesised face in the face space and perform appearance-based hypothesis verification. In contrast to Schmidt [VS00], Burl [BLP95, BP96], and others, in our approach the face is not detected simply as an admissible (i.e. conforming to a model) configuration of several features found in an image, rather a face position is hypothesized by a triplet of local features and all the photometric information is used to verify whether face is present or not. It can be argued that this framework is quite close to the way humans are believed to localise faces [BY98].

In the simplest case, the matching of a hypothesised face with the prototype face could be accomplished by normalised correlation. We adopted a PCA representation of the face class in the face space. However, any appearance model, e.g. exploiting a neural network or SVMs, could be used. The novelty of our approach lies in the hypothesis generation part, not in the verification. Our hypothesis verification method is in essence similar to that of Moghaddam and Pentland, originally proposed for face recognition [MP95, MP96].

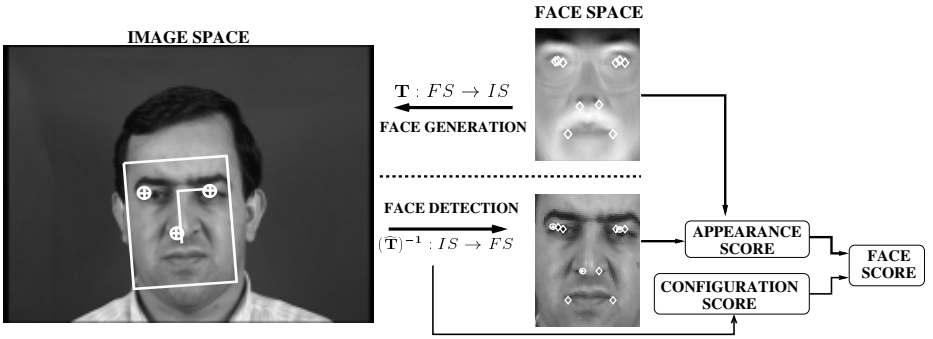


Fig. 3. Schematic diagram of the proposed method

3 Probabilistic Feature Configuration Model

An essential prerequisite of the proposed approach is that the face features in the image are detected and identified with manageable false positive and false negative rates. Clearly this is not an easy task, in spite of the redundancy. To hypothesise a face we need to detect at least three face features of different type. As there are 10 features on each face that our detectors attempt to extract, there are potentially many triplets that may generate a successful hypothesis. However, some of the configurations would be insufficient to define the mapping to the face space, such as a triplet composed of eye features. The two types of error of the feature detectors have different effects. If less than three features are detected, the face will not be detected. On the other hand, false positives will increase the number of triplets, and thus face hypotheses, that have to be verified.

Rather than setting a restrictive threshold, the number of false hypotheses is controlled by geometric pruning. This is achieved as follows. Firstly, we eliminate all the combinations of features which would lead to a degenerate solution for the affine transformation or the transform would be sensitive to errors. The condition number (ratio of the biggest and the smallest singular value) of the matrix made up by putting the face space coordinates of features as columns (homogeneous) was used to determine the well-posedness of triplets. Second, we set out to identify and filter out the triplets that would yield an unfeasible (unrealistic) affine transformation - such that are outside an approximation of the convex hull of transformation encountered in the training set. To construct the probabilistic transformation model, the 6-parameter affine transformations $\widehat{T}$ was decomposed in the following way:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi & t_x \\ -\sin\phi & \cos\phi & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} t & 0 & 0 \\ 0 & \frac{1}{t} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^R \quad (1)$$

where R is reflection (either 0 or 1), n shear, t squeeze, ϕ rotation, r scale

Data : Local detector positions and labels, face-to-image transformation model $p(\widehat{T})$, face appearance model, input image, translation-free feature position confidence regions
Result : Frames of detected face instances in the image
for *All the three-tuples of feature types X,Y,Z defining a well-posed triplet* **do**
  for *All responses of feature detector X* **do**
    Compute feature position confidence regions for remaining features.;
    for *All responses of feature detector Y,Z in confidence regions* **do**
      estimate face to image space affine transformation $\widehat{T}$;
      compute probability $p(\widehat{T})$ of transformation ;
      if $p(\widehat{T}) > threshold1$ **then**
        map image patch into the face space using $\widehat{T}^{-1}$;
        compute probability of appearance $P(A)$;
        if $p(face) = p(\widehat{T}) \cdot p(A) > threshold2$ **then**
          Face detected, record face position defined by $\widehat{T}$;
        end
      end
    end
  end
end

Algorithm 1: Summary of the Face Detection Algorithm

and t_x, t_y translation. This decomposition provides an intuitive feel for the given mapping. We assumed that these parameters are independent and this has been confirmed experimentally. The probability of a given transformation $\widehat{T}$

$$p(\hat{T}) \approx p(n) \cdot p(t) \cdot p(\phi) \cdot p(r) \cdot p(t_x, t_y) \tag{2}$$

is used, together with the appearance model, to reject or accept a face hypothesis. The components $p(n)$, $p(t)$, and $p(\phi)$ of $p(\hat{T})$ are assumed Gaussian, which is in good agreement with experimental data, see Figure 2; $p(t_x, t_y)$ and $p(r)$ are assumed uniform. This assumption reflects the understanding that the useful information content is carried only by shear, squeeze and rotation. Scale which corresponds to the size of a face and translation that represents the position in an image are chosen to lie within a predefined interval (i.e. certain range of face sizes and positions are allowed). Such explicit modelling of probability of face to image transformations may be useful in applications. It is easy to make the system ignore e.g. small faces in the background. The final likelihood of a face hypothesis whose location is defined by a feature triplet is

$$p(face|triplet) = p(A|triplet) \cdot p(\widehat{T}|triplet) \tag{3}$$

where p(A) is the likelihood of appearance. In fact, instead of computing probabilities, we work in the log space, where the product becomes a sum and instead of probability we get a score based on Mahalanobis distance.

Fig. 4. Five types of detected features, roughly corresponding to a left eye corner, right eye centre, left nostril, left mouth corner and right mouth corner

Efficiency In case that all well-posed triplets would be checked (i.e. at least $\widehat{T}$ computed) the complexity of the search algorithm would be $O(n^3)$, where n is the total number of features detected in the image. In our approach, the set of face-to-image transformations from the training set is used to restrict the number of triplets tested. After picking a first feature of a triplet, a region where the other two features of the triplet may lie is established as an envelope of all positions encountered in the training set. Such regions are called *feature position confidence regions*. Since the probabilistic transformation model and these regions were derived from the same training data, triplets formed by features taken from these regions will have nonzero transformation probability. All other triplets are false alarms, since such configurations did not appear in the training data. The detection algorithm is summarised in Alg. 1. The structure of the detection process is graphically depicted in Figure 3. We assume that an instance of the face in the image is obtained by a randomly chosen appearance and transformation T with probability as in the training set (face generation). In the detection stage, first a linear estimate of T, denoted $(\widehat{T})$ is obtained from a triplet of different types of face features. Next, an image patch is mapped back from the image to face space by an inverse transformation $\widehat{T}^{-1}$ and the probability of the image function being an instance of face appearance is evaluated. The 'appearance score' and the 'configuration score' (the probability of $\widehat{T}$) is combined to make a decision on the presence of face.

4 Experiments

In the experiments reported here, local feature detectors were implemented via the PCA-based classification of neighbourhoods of local maxima of the Harris corner detector. At the detected interest points a scale-invariant face feature detectors based on Moghaddam's and Pentland's probabilistic matching [MP95, MP96] were trained. The face-space coordinate of each feature was defined as the position with the highest frequency of occurrence produced by the respective feature detectors in the training set. Details of the detection process are described

Fig. 5. Confidence regions for the second and third feature after left eye corner detection

in [MBHK02]. An example of the result of detection performed on a test image is shown in Figure. 4

To make the testing for admissible transformation fast and simple, the feature position confidence regions were approximated by bounding boxes. An illustration of confidence regions on XM2VTS data for a detected left eye corner is depicted in Figure 5.

An experiment on 400 test images from the XM2VTS database and several images with cluttered background was carried out in order to find out the detection rates and the efficiency of pruning. **The overall detection rate on XM2VTS database was 98%**. Since there can be more than 3 features detected on a face, more triplets can lead to a successful detection (at most 120 if all features are detected). In our experiment only the best face hypothesis with distance below a global threshold is taken as valid. Figure 6 shows a typical result.

The speed up achieved by search-pruning was measured. For the XM2VTS data, the pruning reduced the search by 55 per cent. **For images with cluttered background, the reduction was 92 %**, making the detection process more than 10 times faster. Clearly, the search reduction achieved on the XM2VTS database gives a very conservative estimate of potential gains as the background is homogeneous. In the presence of cluttered background, the reduction in the number of hypotheses is much more impressive, as many triplets involving false positives are present.

5 Conclusions

We proposed a novel framework for detecting human faces based on correspondences between triplets of detected local features and their counterparts in an affine invariant face appearance model. The method is robust to partial occlusion or feature detector failure since a face may be detected if only three out of ten currently used detectors succeed. Robustness with respect to cluttered

Fig. 6. Typical detection result. The best face hypothesis (left) and the worst-score face hypothesis below a distance threshold (right). Features that formed the triplets are marked. The worst face hypothesis has a quite high shear, and consequently high score because high shears were unlikely transformation in the training set

background is achieved via pruning, since background hypotheses lead to extremely unlikely face-to-image space transformations. Since both the appearance and transformation probability is learned from examples, the method can easily be tuned to incorporate application-dependent constraints. The application of the probabilistic model of feature configuration reduced the search complexity by more than 55% in uncluttered images and by 92% in images with complex background. The method was tested on the XM2VTS database and a limited number of images with cluttered background with very promising results – 2% false negative rate and 0 false positives - were obtained.

References

[BLP95] M. C. Burl, T. K. Leung, and P. Perona. Face localization via shape statistics. In *Proc. of International Workshop on Automatic Face and Gesture Recognition*, pages 154–159, 1995. 567, 569

[BP96] M. C. Burl and P. Perona. Recognition of planar object classes. In *Proc. of Computer Vision and Pattern Recognition*, pages 223–230, 1996. 567, 569

[BWP98] M. C. Burl, M. Weber, and P. Perona. A Probabilistic approach to object recognition using local photometry abd global Geometry. In *Proc. of European Conference on Computer Vision*, pages 628–641, 1998. 567

[BY98] V. Bruce and A. Young. *In the Eye of Beholder, The Science of face perception.* Oxford University Press, 1998. 569

[KP97] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pages 21–24, 1997. 567

[LVB+93] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, Ch. von der Malsburg, R. P. Würtz, and W. Konen. Distrotion invariant object recognition in

the dynamic link architecture. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 42(3):300–310, 1993. 567

[MBHK02] J. Matas, P. Bílek, M. Hamouz, and J. Kittler. Discriminative regions for human face detection. In *Proceedings of Asian Conference on Computer Vision*, January 2002. 573

[MMK+99] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In R. Chellapa, editor, *Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, Washington, USA, March 1999. University of Maryland. 568

[MP95] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proc. of International Conference on Computer Vision*, pages 786–793, 1995. 566, 569, 572

[MP96] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. In *Early Visual Learning*, pages 99–130. Oxford University Press, 1996. 566, 569, 572

[SK98] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. Computer Vision and Pattern Recognition, pages 45–51. IEEE*, 1998. 567

[SP98] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–50, January 1998. 566

[VS00] V. Vogelhuber and C. Schmid. Face detection based on generic local descriptors and spatial constraints. In *Proc. of International Conference on Computer Vision*, pages I:1084–1087, 2000. 567, 569