ACCV2002: The 5th Asian Conference on Computer Vision, 23–25 January 2002, Melbourne, Australia.

Discriminative Regions for Human Face Detection*

J. Matas[1,2], P. Bílek[1], M. Hamouz[2], and J. Kittler[2]
[1]Center for Machine Perception, Czech Technical University
{bilek,matas}@cmp.felk.cvut.cz
[2]Centre for Vision, Speech, and Signal Processing, University of Surrey
{m.hamouz,j.kittler}@ee.surrey.ac.uk

Abstract

We propose a robust method for face detection based on the assumption that face can be represented by arrangements of automatically detectable discriminative regions. The appearance of face is modelled statistically in terms of local photometric information and the spatial relationship of the discriminative regions. The spatial relationship between these regions serves mainly as a preliminary evidence for the hypothesis that a face is present in a particular position. The final decision is carried out using the complete information from the whole image patch. The results are very promising.

1 Introduction

Detection and recognition of objects is the most difficult task in computer vision. In many papers object detection and object recognition are considered as distinct problems, treated separately and under different names, e.g. object localisation (detection) and recognition. In our approach localisation of an object of a given class is a natural generalisation of object recognition. In the terminology that we introduce object detection is understood to mean the recognition of object's class, while object recognition implies distinguishing between specific objects from one class. Accordingly, an object class, or category, is a set of objects with similar local surface properties and global geometry. In this paper we focus on object detection, in particular, we address the problem of face localisation.

The main idea of this paper is based on the premise that objects in a class can be represented by arrangements of automatically detectable *discriminative regions*. Discriminative regions are *distinguished regions* exhibiting properties important for object detection and recognition. Distinguished regions are ”local parts” of the object surface, appearance of which is stable over a wide range of views and illumination conditions. Instances of the category are represented by a statistical model of appearance of local patches defined in terms of discriminative regions and by their relationship. Such a local model of objects has a number of attractive properties, e.g. robustness to partial occlusion and simpler illumination compensation in comparison with global models.

Superficially, the framework seems to be no more than a local appearance-based method. The main difference is the focus in our work on the *selection of regions where appearance is modelled*. Detectors of such regions are built during the learning phase. In the detection stage, multiple detectors of discriminative regions process the image. Detection is then posed as a combinatorial optimisation problem. Details of the scheme are presented in Section 3. Before that, previous work is revised in Section 2. Experiments in detecting human faces based on the proposed framework are described in Section 4. Possible refinements of the general framework are discussed in Section 5. The main contributions of this paper are summarised in Section 6.

2 Previous Work

Many early object recognition systems were based on two basic approaches:

- template matching — one or more filters (templates), representing each object, are applied to a part of image, and from their responses the degree of similarity between the templates and the image is deduced.

- measuring geometric features — geometric measurements (distance, angle ...) between features are obtained and different objects are characterised by different constraints imposed on the measurements.

*This research was supported by the Czech Ministry of Education under Research Programme MSM 210000012 Transdisciplinary Biomedical Engineering Research and by European Commission IST-1999-11159 project BANCA.

It is was showed by Brunelli et al. [3] that template matching outperforms measuring geometric features, since the approach exploits more information extracted from the image. Although template matching works well for some types of patterns, there must be complex solutions to cope with non-rigid objects, illumination variations or geometrical transformation due to different camera projections.

Both approaches, template matching and measuring geometric constraints, can be combined together to reduce their respective disadvantages. Brunelli et al. [3] showed that a face detector consisting of individual features linked together with crude geometry constraints have better performance than a detector based on ”whole-face” template matching.

Yuille [20] proposed the use of deformable templates to be fitted to contrast profiles by the gradient descent of a suitable energy function. A similar approach was proposed by Lades et al. [9] and Wiskott et al. [19]. They developed a recognition method based on deformable meshes. The mesh (representing object or object's class) is overlaid over image and adjusted to obtain the best match between the node descriptors and the image. The likelihood of match is computed from the extent of mesh deformation.

Schmid et al. [14, 17] proposed detectors based on local-jets. The robustness is achieved by using spatial constraints between locally detected features. The spatial constraints are represented by angle and length ratios, that are supposed to be Gaussian variables each with their own mean and standard deviation.

Burl et al. [4, 5, 6] introduced a principled framework for representing possible deformations of objects using probabilistic shape models. The objects are again represented as constellations of rigid features (parts). The features are characterised photometrically. The variability of constellations is represented by a joint probability density function.

A similar approach is used by Mohan et al. [13] for the detection of human bodies. The local parts are again recognised by detectors based on photometric information. The geometric constraints on mutual positions of the local parts in the image are defined heuristically.

All the above mentioned methods make decisions about the presence or absence of the object in the image only from geometric constraints. Our proposed method shares the same framework, but in our work the local feature detector and geometric constraints define only a set of possible locations of object in the image. The final decision is made using photometric information, where the parts of object between the local features are taken into account as well.

There are other differences between our approach and the approach of Schmid [17] or Burl [4, 6]. A coordinate system is introduced for each object from the object class. This allows us to tackle the problem of selecting distinctive and well localisable features in a natural way whereas in the case of Schmid's approach, detectable regions were selected heuristically and a model was built from such selected features. Eventhough Weber [18] used an automatic feature selection, this was not carried out in an object-normalised space (as was in our approach), and consequently no requirements on the spatial stability of features were specified. The relative spatial stability of discriminative regions used in our method facilitates a natural affine-invariant way of verifying the presence of a face in the image using correspondences between points in the normalized object space and the image, as will be discussed into detail further.

3 Method Outline

Object detection is performed in three stages. First, the discriminative region detectors are applied to image, and thus a set of *candidate locations* is obtained. In the second stage, the possible constellations (hypotheses) of discriminative regions are formed. In the third stage the likelihood of each hypothesis is computed. The best hypotheses are verified using the photometric information content from the test image. For algorithmic details see Section 4.3.

In the following sections we define several terms used in object recognition in a more formal way. The main aim of the sections is to unify different approaches in the literature and different taxonomy.

3.1 Object Classes

For our purposes, we define an *object class* as a collection of objects which share characteristic features, i.e. objects are composed of several local parts and these parts are in a specific spatial relationship. We assume the local parts are detectable in the image directly and the possible arrangements of the local parts are given by geometrical constraints. The geometrical constraints should be invariant with respect to a predefined group of transformations. Under this assumption, the task of discrimination between two classes can be reduced to measuring the differences between local parts and their geometrical relationships.

3.2 Discriminative Regions

Imagine you are presented with two images depicting objects from one class. You are asked to mark corresponding points in the image pair. We would argue that, unless *distinguished regions* are present in the two images, the task is extremely hard. Two views of a white featureless wall, a patch of grass, sea surface or an ant hill might be good examples. However, on most objects, we find surface patches that can be separated from their surroundings and are detectable over a wide range of views. Before proceeding further, we give a more formal definition of distinguished region:

Definition 1 Distinguished Region *(DR) is any subset of an image that is a projection of a part of scene (an object) possessing a distinguishing property allowing its detection (segmentation, figure-ground separation) over a range of viewing and illumination conditions.*

In other words, the DR detection must be repeatable and stable w.r.t. viewpoint and illumination changes. DRs are referred to in the literature as 'interest points' [7], 'features' [1] or 'invariant regions' [16]. Note that we do not require DRs to have some transformation-invariant property that is unique in the image. If a DR possessed such a property, finding its corresponding DR in an other image would be greatly simplified. To increase the likelihood of this happening, DRs can be equipped with a characterisation computed on associated measurement regions:

Definition 2 A Measurement Region *(MR) is any subset of an image defined by a transformation-invariant construction (projective, affine, similarity invariant) from one or more (in case of grouping) regions.*

The separation of the concepts of DR and MRs is important and not made explicit in the literature. Since DRs are projections of the same part of an object in both views and MRs are defined in a transformation-invariant manner they are quasi view-point invariant. Besides the simplest and most common case where the MR is the DR itself, a MR may be constructed for example as a convex hull of a DR, a fitted ellipse (affinelly invariant, [16]), a line segment between a pair of interest points [15] or any region defined in a DR-derived coordinates. Of course, invariant measurements from a single or even multiple MRs associated with a DR will not guarantee a unique match on e.g. repetitive patterns. However, often DR characterisation by invariants computed on MR might be unique or almost unique.

Note that, any set of pixels, not necessarily continuous, can posses a distinguishing property. Many perceptual grouping processes detect such arrangements, e.g. a set of (unconnected) edges lying along a straight line form a DR of maximum edge density. The property is viewpoint quasi-invariant and detectable by the Hough Transform. The 'distinguished pixel set' [10] would be a more precise term, but it is cumbersome.

The definition of ”local part” (sometimes also called ”feature”, ”object component” etc.) is very vague in the recent literature. For our purpose it is important to define it more precisely. In the following discussion we will use the term ”discriminative region” instead of ”local part”. In this way, we would like to emphasise the difference between our definition of discriminative region and the usual sense of local part (a discriminative region is a local part with special properties important for its detection and recognition).

Definition 3 A Discriminative Region *is any subset of an image defined by* discriminative descriptors *computed on measurement region. Discriminative descriptors have to have the following properties:*

- **Stability under change of imaging conditions**. *A discriminative region must be detectable over a wide range of imaging conditions (viewpoint, illumination). This property is guaranteed by definition of a DR.*
- **Good intra-category localization**. *The variation in the position of the discriminative region in the object coordinate system should be small for different objects in the same category.*
- **Uniqueness**. *A small number of similar discriminative regions should be present in the image of both object and background.*
- **High incidence**. *The discriminative region should be detectable in a high proportion of objects from the same category.*

Note, there exists a trade-off between the ability to localise objects and the ability to discriminate between. A very discriminative part can be a strong cue, even if it appears in an arbitrary location on the surface of the object. On the other hand, a less discriminative part can only contribute information if it occurs in a stable spatial relationship relative to other parts.

3.3 Combining Evidence

This is a rather important stage of the detection process, which significantly influences the overall performance of the system and makes it robust with respect to arbitrary geometrical transformations. The combination of evidence coming from the detected discriminative regions is carried out in a novel way, significantly different from approaches of the Schmid et al. [14, 17] or Burl et al. [4, 5, 6].

In most approaches, a shape model is built over the placement of particular discriminative regions. If an admissible configuration of these regions is found in an image, an instance of object in the image is hypothesised. It means that all the information conveyed by the area that lies between the detected discriminative regions is discarded. If you imagine a collage, consisting of one eye, a nostril and a mouth corner placed in a reasonable manner on a black background, this will still be detected as a face, since no other parts of the image are needed to accept the ”face-present” hypothesis.

In our approach the geometrical constraints are modelled probabilistically in terms of spatial coordinates of discriminative regions. But these geometrical constraints are used only to define possible positions (hypotheses) of object in the image. The final decision about object presence in the image is deduced from the photometric information content in the original image.

4 Experiment

We have carried out the experiment on face localisation [2] with the XM2VTS database [11]. In order to verify the correctness of our localization framework, several simplifications to the general scheme are made. In the experiment the discriminative regions were semi-automatically defined as the eye-corners, the eye-centers the nostrils and the mouth corners.

4.1 Detector of discriminative regions

As a distinguished region detector we use the improved Harris corner detector [8]. Our implementation [2] of the detector is relatively insensitive to illumination changes, since the threshold is computed automatically from the neighborhood of the interest point. Such a corner detector is not generally invariant to scale change, but we solve this problem by searching for interest points through several scales.

We have observed [2] that the distribution of interest points coincide with the manually labelled points. It means, these points should define discriminative regions (here we suppose, that humans often identify interest points as most discriminative parts of object).

Further, we have assumed that all potential in-plane face rotations and differences in face scale are covered by the training database.

The MRs was defined very simply, as rectangular regions with the centre at the interest points. We select ten positions (the left eye centre, the right eye centre, the right left-eye corner, the left left-eye corner, the right right-eye corner, the left right-eye corner, the left nostril, the right nostril, the left mouth corner, the right mouth corner), which we further denote as regions 1–10. All properties of a discriminative region are then determined by the size of the region. As a descriptor of a region we use the normalised colour information of all points contained in the region.

Each region was modelled by a uni-modal Gaussian in a low-dimensional sub-space and the hypothesis whether the sample belongs to the class of faces is decided from the distance of this sample from the mean for a given region. The distance from the mean is measured as a sum of the in sub-space (DISS) and the from sub-space (DFSS) distances (Moghaddam et al. [12]).

4.2 Combining Evidence

The proposed method is based on finding the correspondences between generic face features (referred to as discriminative regions) that lie in the face-space and the face features detected in an image. This correspondence is then used to estimate the transformation that a generic face possibly underwent. So far the correspondence of three points was used to estimate a four or six parametric affine transformation.

When the the transformation from the face space to image space determined, the verification of a ”face-present” hypothesis becomes an easy task. An inverse transformation (i.e. transformation from the image space into the face-space) is found and the image patch (containing the three points of correspondence) is transformed into the face-space. The decision whether the ”face-present” hypothesis holds or not is carried out in the face-space, where all the variations introduced by the geometrical transformation (so far only affine transformation is assumed to be the admissible transformation that a generic face can undergo) are compensated (or at least reduced to a negligible extent). The distance from a generic face class [12] is computed for the transformed patch and a threshold is used to determine whether the patch is from a face class or not.

Moreover, many possible face patches do not have to be necessarily verified, since certain constraints can be put on the estimated transformation. Imagine for instance that all the feasible transformations that a face can undergo are the scaling from 50% to 150% of the original size in the face space and rotations up to 30 degrees. This is quite a reasonable limitation which will cause most of the correspondences to be discarded without doing a costly verification in the face space (in our experiments the pruning reached about 70%). In case of the six parametric affine transform both shear and anisotropic scale is incorporated as the admissible transformation.

4.3 Algorithm summary

Algorithm 1: Detection of human faces

1. **Detection of the distinguished regions**. For each image from the test set, detect the distinguished regions using the illumination invariant version of the Harris detector

2. **Detection of the discriminative regions**. For each detected distinguished region determine to which class the region belongs using the PCA-based classifier in the colour space from among ten discriminative region

classes (practically the eye corners, the eye centres, the nostrils and the mouth corners). The distinguished regions that do not belong to any of the predefined classes are discarded.

3. **Combination of evidence**.

    - Compute the estimate of the transformation from the image space to the face space using the correspondences between the three points in the face space and in the image space.
    - Decompose this transformation into rotation, scale, translation and possibly shear and test whether these parameters lie within a predefined constraints, i.e. make the decision, whether the transformation is admissible or not.
    - If the transformation derived from the correspondences is admissible, transform the image patch that is defined by the transformation of the face outline into the face space.

4. **Verification**. Verify the ”face present” hypothesis using a PCA-based classifier.

4.4 Results

Results of discriminative regions detector are summarised in Tab. 1. Note that since the classifier is very simple, the performance is not very high. However, even with such a simple detector of discriminative regions the system is capable of detecting faces with very low error, since we need only a small number of successfully detected discriminative regions (in our case only 3).

Several extensive experiments were conducted. Image patches were declared as ”face” when their Mahanalobis distance based score lied below a certain threshold. 200 images from the XM2VTS database were used for training a grayscale classifier based on the Moghaddam method [12], as mentioned earlier.

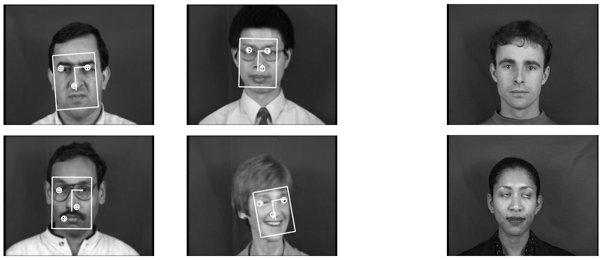
The detection rate reached 98% in case of XM2VTS database - see Fig. 1 for examples. Faces in several images containing cluttered background were successfully detected as shown in Fig. 2.

5 Discussion and Future Work

We proposed a method for face detection using discriminative regions. The detector performance is very good for the case when the general face detection problem is constrained by assuming a particular camera and pose position. We also assumed that the parts that appear distinctive to the human observer will be also discriminative, and therefore the discriminative regions were selected manually. In general, the correlation between distinctiveness and discriminativeness cannot necessarily be assumed and therefore the discriminative regions should be ”learned” from the training images. The training problem was addressed in this paper only partially. As an alternative the method proposed by Weber et al. [18] can be exploited.

Table 1. Performance of discriminative region detectors

	false negative		false positive	
	%	#	%	#
Region 1	31.89	191	72.26	3831
Region 2	10.68	64	37.88	1342
Region 3	57.76	346	33.03	433
Region 4	54.92	329	19.85	218
Region 5	15.03	90	22.34	538
Region 6	13.69	82	62.33	3260
Region 7	15.53	93	4.00	78
Region 8	12.52	75	5.07	104
Region 9	48.75	292	6.27	70
Region 10	33.56	201	14.90	233



Correctly detected

False rejections

Figure 1. Experiment results

The admissible transformation, which a face can undergo has so far been restricted to affine transformation. Nevertheless, the results showed even in such a simple case, that high detection performance can be achieved. Future modifications will involve the employment of more complex transformations (such as general non-rigid transformations). The PCA based classification can be replaced by more powerful classifiers, such as Neural Networks, or Support Vector Machines.

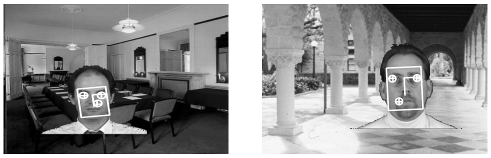


Figure 2. Experiments with cluttered background

6 Conclusion

In the paper, a novel framework for face detection was proposed. The framework is based on the idea that most real objects can be decomposed into a collection of local parts tied by geometrical constraints imposed on their spatial arrangement. By exploiting this fact, face detection can be treated as recognition of local image patches (photometric information) in a given configuration (geometric constraints). In our approach, discriminative regions serve as a preliminary evidence reducing the search time dramatically. This evidence is utilised for generating a normalised version of the image patch, which is then used for the verification of the ”face present” hypothesis.

The proposed method was applied to the problem of face detection. The results of extensive experiments are very promising. The experiments demonstrated that the proposed method is able to solve a rather difficult problem in computer vision. Moreover we showed that even simple recognition methods (with a limited capability when used alone) can be configured to create powerful framework able to tackle such a difficult task as face detection.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. of Computer Vision and Pattern Recognition*, pages I:774–781, 2000.

[2] P. Bílek, J. Matas, M. Hamouz, and J. Kittler. Detection of human faces from discriminative regions. Technical Report VSSP–TR–2/2001, Department of Electronic & Electrical Engineering, University of Surrey, 2001.

[3] R. Brunelli and T. Poggio. Face recognition: Features vs. templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(10):1042–1053, 1993.

[4] 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.

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

[6] 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.

[7] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proc. of Computer Vision and Pattern Recognition*, pages I:612–618, 2000.

[8] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Alvey Vision Conference*, pages 147–151, 1988.

[9] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. 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.

[10] J. Matas, M. Urban, and T. Pajdla. Unifying view for wide-baseline stereo. In B. Likar, editor, *Proc. Computer Vision Winter Workshop*, pages 214–222, Ljubljana, Sloveni, February 2001. Slovenian Pattern Recorgnition Society.

[11] 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.

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

[13] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[14] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[15] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proc. of European Conference on Computer Vision*, pages 754–760, 2000.

[16] T. Tuytelaars and L. van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. of British Machine Vision Conference*, pages 412–422, 2000.

[17] 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.

[18] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. of European Conference on Computer Vision*, pages 18–32, 2000.

[19] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[20] A. L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.