Detection of Facial Features in Scale-Space

P. Hosten, M. Asbach

This paper presents a new approach to the detection of facial features. A scale adapted Harris Corner detector is used to find interest points in scale-space. These points are described by the SIFT descriptor. Thus invariance with respect to image scale, rotation and illumination is obtained. Applying a Karhunen-Loeve transform reduces the dimensionality of the feature space. In the training process these features are clustered by the k-means algorithm, followed by a cluster analysis to find the most distinctive clusters, which represent facial features in feature space. Finally, a classifier based on the nearest neighbor approach is used to decide whether the features obtained from the interest points are facial features or not.

Keywords: clustering methods, face recognition, feature extraction, interest points, Karhunen-Loeve transforms, object detection, pattern classification.

1 Introduction

Face detection is one of the most challenging tasks in object recognition, because of the high variance among human faces, including facial expression. Furthermore, the typical challenges of object detection, such as variability in scale, orientation and pose as well as occlusion and lighting conditions have to be treated by a face detector.

In recent years, several methods have been developed for face detection that deal with some but not all sources of variance. Holistic approaches such as Eigenfaces [12] (sometimes called image-based [4] or appearance-based [14] methods in context of detection and recognition)or the boosted cascade of simple features [13] perform object detection by classifying image regions through a sliding window [15]. While they can be made invariant to scale and lighting, the major drawback of these techniques is that they cannot deal with different rotation or pose. In addition, facial expression and occlusion must be treated as intra-class variance, decreasing the performance of the classifier/detector. On the other hand, feature-based approaches handle pose, expression and even partial occlusion very well. Rotation invariance is only limited by the properties of the features used to build the detector.

This paper presents a new approach to the detection of facial features. The local features used for detection are invariant to scale, rotation and change in illumination, and are robust to changing viewpoints. We show that the features chosen are well suited to detect several meaningful parts of the human face like pupils, nostrils and the corner of the mouth. The paper is organized as follows: In section two, the feature extraction process is described. Feature space reduction and feature selection are explained in section three. Section four presents the classifier, and the results are shown in section five. In section six, the results are discussed.

2 Feature extraction

This section explains feature extraction from an image and its description in feature space, The features have to be invariant to affine transformations and illumination. While there have been numerous proposals for appropriate feature points, they have usually been chosen on the basis of human observation of face characteristics. In contrast to such knowledge-based methods, several methods have been developed to detect structures that are generally easy to locate, can be computed with high reliability and satisfy the demand of scale, rotation and illumination invariance [6, 10]. The local area around these so-called interest points is then extracted and modeled.

A 3D scale-space representation of an image is usually taken to detect local features and their corresponding scales (see Fig. 1). Given any image $I(\boldsymbol{x})$, its scale-space representation $L(\boldsymbol{x}, \sigma)$, is defined by

$$L(\boldsymbol{x}, \sigma) = g(\boldsymbol{x}, \sigma) * I(\boldsymbol{x}) \tag{1}$$

where $g(\boldsymbol{x}, \sigma)$ denotes the Gaussian kernel function given by

$$g(\boldsymbol{x}, \sigma) = \frac{1}{2\pi\,\sigma^2} \exp\left[-\frac{\boldsymbol{x}^2}{2\pi\sigma^2}\right]. \tag{2}$$

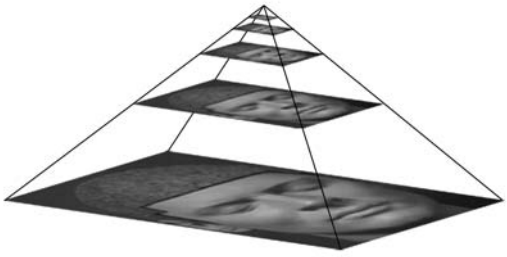

Fig. 1: The image $I(\boldsymbol{x})$ is embedded into a continuous family $L(\boldsymbol{x}, \sigma)$ of gradually smoother versions of it. Hereby the original image corresponds to the scale $\sigma = 0$ [11]. Interest points are detected in this scale-space.

Our approach uses the scale adapted Harris Corner detector to detect interest points. This detector is based on the second moment matrix [7]:

$$\mu(\boldsymbol{x}, \sigma_i, \sigma_D) = \sigma_D^2\, g(\sigma_i) * \begin{bmatrix} L_D^2(\boldsymbol{x}, \sigma_D) & L_x L_y(\boldsymbol{x}, \sigma_D) \\ L_x L_y(\boldsymbol{x}, \sigma_D) & L_D^2(\boldsymbol{x}, \sigma_D) \end{bmatrix} \tag{3}$$

This matrix describes the gradient distribution in a local neighborhood of a point. Its eigenvalues λ_1, λ_2 represent the two principal signal changes. Thus it is possible to extract points with a significant signal change in both orthogonal directions indicating edges or junctions, for example. Since it

is easier to compute the trace and the determinant of the second moment matrix than its eigenvalues, the Harris detector uses the following measure to determine the location of interest points [3]:

$$\begin{aligned} cornerness &= \lambda_1\lambda_2 - \alpha(\lambda_1 + \lambda_2)^2 \\ &= \det(\boldsymbol{\mu}(\boldsymbol{x},\sigma_i,\sigma_D)) - \alpha\, trace^2(\boldsymbol{\mu}(\boldsymbol{x},\sigma_i,\sigma_D)). \end{aligned} \quad (4)$$

This measure is not suitable for detecting the maximum over scales in a scale-space representation. Thus the normalized Laplacian-of-Gaussian is used for automatic scale selection [5].

$$LoG(\boldsymbol{x},\sigma_n) = s_n^2 \left| L_{xx}^2(\boldsymbol{x},\sigma_n) + L_{yy}^2(\boldsymbol{x},\sigma_n) \right| \quad (5)$$

The region around interest points is described using local descriptors. Recently, several descriptors have been developed [8]. This paper uses the SIFT descriptor based on the gradient distribution in the detected region around the interest point [6]. The size of the region being described depends on the detection scale σ. Thus a scale invariant description is obtained. The resulting descriptor maps each interest point into a 128-dimensional vector $\boldsymbol{m}$. In addition, a dominant orientation is calculated for each interest point and descriptors are calculated on the basis of this angle to gain rotation invariance.

3 Feature reduction and selection

In this section, dimensionality reduction and feature selection are explained. The Karhunen-Loeve transform is applied to reduce the dimensionality. Then the clustering process and the cluster analysis is described.

3.1 Karhunen-Loeve transform

It is a well-known problem in the machine learning domain that the number of required training samples increases exponentially with the dimensionality of the feature space. This is called the "curse of dimensionality" [1]. Since the SIFT descriptor forms a 128 dimensional vector, the Karhunen-Loeve transform is applied to reduce dimensionality. Positive side effects of this are a reduction in computing time for the classification and the fact that the elements of the feature vector m become uncorrelated.

Hence the covariance matrix Σ of N training feature vectors $\boldsymbol{m}$ is used to find a linear subspace for better representation.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}_i$$

$$\boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{m}_i - \boldsymbol{\mu})(\boldsymbol{m}_i - \boldsymbol{\mu})^T$$

Therefore the covariance matrix $\boldsymbol{\Sigma}$ has to be orthogonalized by an eigenvector transform:

$$\boldsymbol{\Phi}^T \boldsymbol{\Sigma} \boldsymbol{\Phi} = \Delta = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ \vdots & \lambda_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_N \end{bmatrix}. \quad (6)$$

The columns of transform matrix Φ are the eigenvectors of Σ. These eigenvectors are orthonormal and represent the basis vectors. The use of only a subset of these eigenvectors as transformation matrix $\boldsymbol{\Phi}^T$ leads to reduced dimensionality of the feature space [2]:

$$\boldsymbol{f} = \hat{\boldsymbol{\Phi}}^T \cdot (\boldsymbol{m} - \boldsymbol{\mu}) \quad (7)$$

3.2 Clustering

In contrast to knowledge-based methods, the exact facial features are not defined beforehand. For clustering the training feature vectors $\boldsymbol{f}$ in the reduced feature space, we need, however, to estimate their number. A single cluster represents a distribution of a number of feature vectors $\boldsymbol{f}$ that belongs to the same facial feature. The number of clusters must therefore be big enough to offer at least one cluster centroid per facial feature. Choosing too many clusters, however, leads to overtraining, i.e. a single facial feature will be represented by a multitude of cluster centroids that are adapted to the training set instead of generalizing a given feature. Hence an evaluation with different numbers of clusters was conducted (see section 4) to find the optimum value ex post based on the classifier performance. A detailed description of the clustering process is given in the following paragraphs.

First of all, the training feature vectors $\boldsymbol{f}$ are split up into two subsets. The first subset contains all feature vectors describing the facial features (class $a = 1$). In the second subset all feature vectors describing the rest of the image are aggregated (class $a = 0$):

$$F_{\text{face}} = \left\{ \forall \boldsymbol{f} \middle| a = 1 \right\}, F_{\text{non face}} = \left\{ \forall \boldsymbol{f} \middle| a = 0 \right\} \quad (8)$$

Each subset is clustered with the k-means algorithm. Thereby the number of clusters in each subset is kept proportional to the respective number of features $\boldsymbol{f}$. In other words the average number of features represented by each cluster is chosen to be identical for both subsets.

3.3 Cluster analysis

A cluster analysis is applied to F_{face} to find the most characteristic clusters of facial features. The cluster precision introduced in [9] is a measure of the representativeness of a cluster:

$$p_{ja} = \frac{\#F_{ja}}{\#F_{ja=0} + \#F_{ja=1}} \quad (9)$$

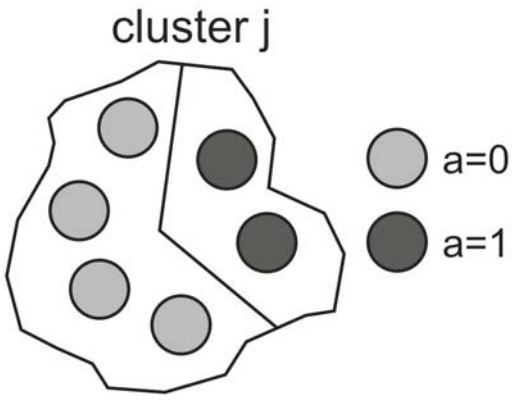

Fig. 2: Cluster j represents the feature of class $a = 0, 1$. The cluster precision is $p_{ja=0} = 0.67$ and $p_{ja=1} = 0.33$.

This results in the probability p_{ja} that a feature of class a is represented by cluster j.

In order to determine which feature is represented by which cluster, a vector quantization is applied to all training feature vectors $\boldsymbol{f}$. Thus the cluster precision for each cluster is obtained. Finally only clusters that represent facial features ($a = 1$) and whose cluster precision is higher than 0.9 are kept to form the classifier. In this way the most distinctive clusters representing the facial features in feature space are determined. In the following, these clusters will be denoted as target clusters.

4 Classifier

The distances of the feature vectors $\boldsymbol{f}$ to the target clusters determined in section 3.3, are a measure of the performance of the classification process. As shown in Fig. 3, the training features assigned to a face ($a = 1$) are closer to the target cluster than the other features ($a = 0$). A threshold can therefore be calculated from the training features for a given false-positive rate. In this context, false-positive denotes all features wrongly classified as facial features, whereas true-positive denotes all features correctly classified as facial features.

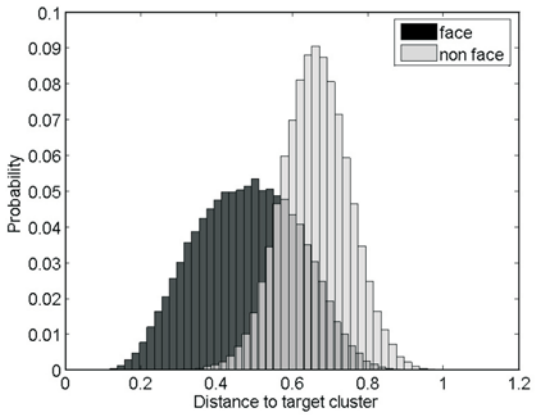

Fig. 3: Distances of the training feature vectors $\boldsymbol{f}$ to the target cluster.

In this way a classifier is obtained. Up this point, all evaluations have been based on the training set only. This is necessary for a cross evaluation study that requires that tests are only performed on data that has not been part of the training process. This dataset is split up into 46 subsets containing approx. 49 images each. During cross evaluation, a training corpus is formed of a random selection of 45 subsets, leaving a single subset as a test set. The results of such evaluation runs are averaged for the final results presented in section 5. A single run consists of the following steps:

1. All SIFT descriptors of the test set are projected into the feature space found in 3.1.
2. For all such features the respective nearest neighbors of the target clusters from 3.3 is calculated.
3. All features whose distance to their target cluster is smaller than a given threshold, are classified as facial features.
4. From the annotation of the dataset, true positives and false positives are discriminated.
5. By varying the threshold, classifier sensitivity is adjusted. A ROC curve over true positives vs. false positives results.

This procedure is repeated for a number of cluster densities (see Fig. 5) to find the ideal number of clusters.

5 Results

In this section the results of the previously introduced interest point detector and classifier are presented. Our dataset consists of 2254 images showing one to three people in cluttered backgrounds. As described in section 4 the dataset is randomly split into a training set containing 2205 images and a test set containing 49 images.

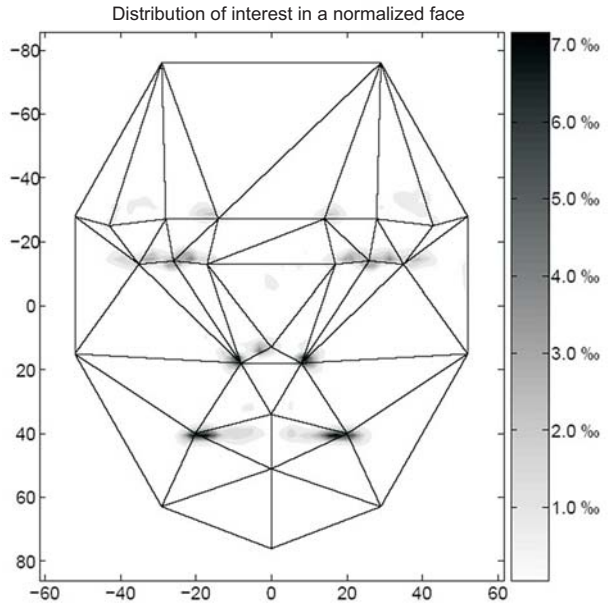

Fig. 4: This figure shows the distribution of interest points detected in a face by the scale adapted Harris corner detector. The interest points are mapped into the normalized coordinate system.

Fig. 4 depicts the distribution of the interest points over the face area. About 150,000 facial interest points have been extracted from all images of the whole dataset. They are mapped into a normalized coordinate system to achieve comparability. As can be seen, eyes, nostrils and corners of the mouth are extracted very reliably. Thus these interest points are suitable for facial analysis.

In order to determine the best setup of the classifier, the influence of the number of clusters on its performance has been analyzed. The result is shown in Fig. 5. Obviously, the effect of overfitting arises for a large number of clusters and a low cluster density, respectively.

Fig. 6 shows the distances of the target clusters to the features $\boldsymbol{f}$ extracted from the test set. Just as has been observed from the training set, the test features assigned to a face ($a = 1$) are closer to the target cluster than the other features ($a = 0$).

Since only a small part of the image area is covered by human faces, the number of interest points describing facial features is 20 times lower than the number of interest points found in the remaining image. That is why only really low false-positive rates (0.1 %–0.2 %) are acceptable. Setting a threshold for 0.2 % false positives results in a detection rate of about 20 %, which seems acceptable given the fact that this corresponds to about 10 correctly detected facial fea-

ture points per image, whereas approximately 2 features are wrongly classified. This aspect is exemplified in the following.

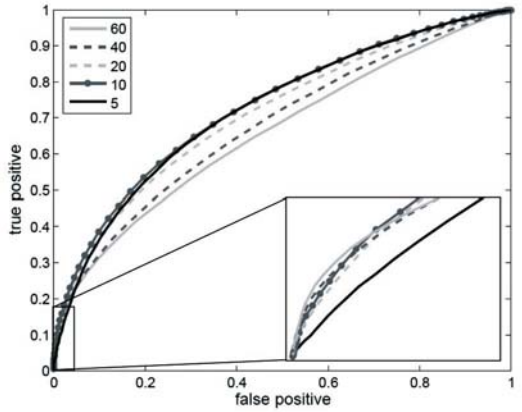

Fig. 5: This figure shows the influence of different cluster densities on the performance of the classifier. Thus the cluster density is a measure for the number of training features represented by a cluster. The smaller the value, the more clusters are computed.

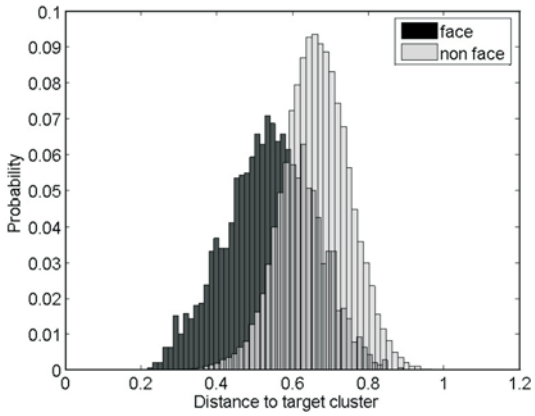

Fig. 6: Distances of the test feature vectors f to the target cluster

Fig. 7 shows all interest points detected by the scale-adapted Harris Corner detector. These points, indicated by circles, are especially located in corner like structures. The result of the classifier can be seen in Fig. 8. It shows only the interest points that have been classified as facial features. These features represent eyes, eye brows and corner of the mouth.

Fig. 7: This image of the test set shows all interest points detected by the scale-adapted Harris corner detector. The diameter of the circle indicates the appropriate scale on which the interest point has been detected.

Fig. 8: This image shows only the interest points that have been classified as facial features

6 Conclusion

This paper has introduced a new approach to the detection of facial features. Our approach builds on local image descriptions that are invariant towards affine image transformations and illumination. We have shown that the scale adapted Harris detector yields feature points that are suitable for the detection of specific facial features like eyes, nostrils and corners of the mouth, and that these features can be appropriately described by the SIFT descriptor. Our new method is able to detect on average 10 features per face, with a false-positive rate of 0.2 %, which corresponds to approximately 2 wrongly classified features per image.

Acknowledgements

The research described in this paper was supervised by Prof. J.-R. Ohm, Institute of Communications Engineering at RWTH Aachen University.

References

[1] Bellman, R.: *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[2] Fukunaga, K.: *Introduction in statistical pattern recognition*. 1972.

[3] Harris, C., Stephens: A Combined Corner and Edge Detector. *Alvey Vision Conference*, 1988.

[4] Hjelmas, E., Low, B. K.: Face Detection: A Survey. *Computer Vision and Image Understanding*,Vol. **83** (2001), No. 3, p. 236–274.

[5] Lindeberg, T.: Edge Detection and Ridge Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 1998.

[6] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.

[7] Mikolajczy, K., Schmid, C.: Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 2004.

[8] Mikolajczy, K., Schmid, C.: A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[9] Mikolajczyk, K., Leibe, B., Schiele, B.: Local Features for Object Class Recognition. *Technical report, Multimodal Interactive Systems TU Darmstadt*, Germany, 2005.

[10] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Van Gool, L., Kadir, T.: A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 2005.

[11] Ter Haar Romeny, B., Florack, L., Koenderink, J., Viergever, M. (editors): Scale-Space Theory in Computer Vision, *First International Conference, Scale-Space'97*, Utrecht, The Netherlands, July 2–4, 1997, Proceedings. Springer, 1997.

[12] Turk, M., Pentland, A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, Vol. **3** (1991)No. 1, p. 71–86.

[13] Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. *Conference on Computer Vision and Pattern Recognition 2001*, 2001.

[14] Yang, M.-H., Kriegman, D. J., Ahuja, N.: Detecting Faces in Images: A Survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. **24** (2002), No. 1, p. 34–58.

[15] Zhang, J., Yan, Y., Lades, M.: Face Recognition: Eigenface, Elastic Matching and Neural Nets. *Proceedings of the IEEE*, Vol. **85** (1997), No. 9.

Peter Hosten
e-mail: hosten@ient.rwth-aachen.de

Ing. Mark Asbach
e-mail: asbach@ient.rwth-aachen.de

Institute of Communications Engineering
RWTH Aachen University
52056 Aachen, Germany