Image feature localization by multiple hypothesis testing of Gabor features

Jarmo Ilonen, Joni-Kristian Kamarainen*, Pekka Paalanen, Miroslav Hamouz, Josef Kittler, *Member, IEEE*, Heikki Kälviäinen, *Member, IEEE*

***Abstract*— Several novel and particularly successful object and object category detection and recognition methods based on image features, local descriptions of object appearance, have recently been proposed. The methods are based on a localisation of image features and a spatial constellation search over the localised features. The accuracy and reliability of the methods depend on the success of both tasks: image feature localisation and spatial constellation model search. In this paper we present an improved algorithm for image feature localisation. The method is based on complex-valued multiresolution Gabor features and their ranking using multiple hypothesis testing. The algorithm provides very accurate local image features over arbitrary scale and rotation. We discuss in detail issues such as selection of filter parameters, confidence measure and the magnitude versus complex representation and show on a large test sample how these influence the performance. The versatility and accuracy of the method is demonstrated on two profoundly different challenging problems (faces and license plates).**

I. Introduction

Most of the research and development activities in biometric authentication, especially in face detection and face identification, address one of the two important image processing research topics: recognition of object classes (object categories) and recognition of object class instances. The first one is exemplified by face detection and the second one by face identification. In the recent contributions to these topics the focus of interest has shifted from the well-known image-based object recognition methods towards new and more generic feature-based approaches [4], [12], [25].

The idea of partitioning objects into their constituent parts (object primitives) and separately modelling the spatial relationships between them is not new. It was proposed by Fischler and Elschlager already in 1973 [9]. Since then the idea has been re-visited by many researchers and developed into more efficient forms e.g. by Burl, Weber and Perona et al. [4], [39], [38], [7], Lowe [24], [25], [14], Schmid et al. [31], and by Hamouz et al. [12]. The methods contain two distinct processing stages, image feature localisation and a spatial search. In the search step a spatial consistency checking process combines image features and matches their configuration to the stored object class constellation models allowing a certain degree of deformation in the spatial configuration.

The above methods significantly differ in their approach in the way they utilise image features. Generally, all of them further divide the localisation process into the detection of salient image features and their unique description [27], [28]. Accordingly, they first somehow pinpoint salient and discriminative regions in the scene (image feature detection) and then the detected regions are represented by a local descriptor (image feature description). The main difference between such image feature localisation methods and the one proposed in this study is the degree of supervision required in training the system. Whereas in the referenced methods image features in training images need not to be annotated in advance, the method in this study is supervised, requiring a sufficient set of image feature examples.

Image feature detection is based on the concept of distinctiveness in images. Many existing methods utilise Harris corner detectors (e.g. [28]), difference of Gaussian (DoG, scale-space, [25]) or the behaviour of local entropy [18]. The description of the detected image features is based on the properties of the associated local regions, the simplest solution being the grey-level histogram. The SIFT image features by Lowe have been shown to be distinctive, stable and discriminating [27], [28]. The perceived advantage of using such image feature detection and description methods is their simplicity due to their apparent unsupervised nature. However since a certain degree of supervision is required (segmentation and object labelling) this advantage is only relative rather than qualitative. In the following we shall refer to these methods as semi-supervised. Here we claim that a much more efficient approach can be devised if image feature detectors use supervised learning. It is clear that the semi-supervised methods are the most suitable for the cases where the object instance remains the same, such as in the interest point initialisation for object tracking [22], but they cannot tolerate appearance variation among different instances of the object class on the local level.

In this paper a novel supervised image feature localisation method based on Gabor feature representation is introduced. Terms localisation and detection are often used interchangeably, but by the term localisation we want to stress that we are interested in locating the position of image features exactly, not only detecting their general presence. As for most Gabor methods, the representation is inherently of multiresolution

Ilonen, Kamarainen, Paalanen and Kälviäinen are with the Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, P.O.Box 20, Lappeenranta, Finland; Tel: +358 5 62111; Fax: +358 5 6212899; E-mail: Jarmo.Ilonen@lut.fi, Joni.Kamarainen@lut.fi, Pekka.Paalanen@lut.fi, Heikki.Kalviainen@lut.fi. Hamouz and Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, Surrey, UK; Tel: +44 1483 686030; Fax: +44 1483 686031 M.Hamouz@surrey.ac.uk, J.Kittler@surrey.ac.uk

The authors would like to thank EPSRC Research Grant GR/S46543/01 and EU Project VidiVideo for support.

* Corresponding author

EDICS: OTH-RCGN

type. However, the proposed scheme differs from the conventional Gabor feature representation in several important aspects. First of all, it is supervised which means that both Gabor filter bank parameters as well as the models of class conditional probability distributions are optimised using training data. By virtue of being supervised it combines the image feature detection and description into a single learning step which captures the key characteristics of feature appearance and identifies the class this feature belongs to. Second, a low level description is constructed by complex valued Gabor filter responses in multiple resolutions and orientations where the low frequencies mainly contribute to the generality and the high frequencies to the specificity. The complex representation delivers very accurate localisation as it is demonstrated in the experimental part of the study. A preliminary version of this framework, called simple Gabor feature space, has been proposed by the authors [21]. Its efficiency derives from the invariance properties of Gabor responses [20]. Third, the local description of the image features is based on their class conditional probability density functions, which can also be used to indicate to the constellation model their best order for matching. Such ranking of the image features, which is based on the statistical hypothesis testing using likelihoods [29], enhances the efficiency of the subsequent processing. The probability functions are learnt by Gaussian mixture models. Previously, only the magnitude of Gabor responses has been used in hypothesis testing. Herein we extend this framework by adding phase information. This is inventively achieved by using complex-value representation. Complex-valued model implicitly incorporates both magnitude and phase information and facilitates high localisation accuracy. It can be argued that the statistical part can in principle handle arbitrary appearance variability, as the mixture model is able to approximate any probability function.

Fourth, the proposed image feature localisation is illumination, rotation (in-plane) and scale invariant and outputs a confidence value for each feature. The illumination invariance (to a constant factor) is achieved by means of an effective signal normalisation which is facilitated by the rich form of representation adopted. The method achieves high localisation accuracy by exploiting complex-valued arithmetic which also reduces the computation time. Last but not least, by its supervised nature it offers better class-identifiability than a semi-supervised detector. The proposed image feature extraction method and constellation model have been successfully applied to accurate face detection and localisation in [12].

In their previous work, the authors have proposed a Gaussian mixture model pdf based confidence measure for one-class classification [29], which is also applied in the proposed localisation method. In addition, early forms of the proposed image feature localisation method have been used in many of authors' studies, but previously briefly described only in [19]. In this paper, the localisation method is explained in detail together with the accompanying theory and algorithms. Its accuracy and versatility are demonstrated on two real-world challenging problems: on face detection and localisation and license plate localisation. The method is not tied to any specific problem or application and can be used in localisation of any object class. New facial landmark localisation results show that the accuracy is significantly improved as compared to our previous version of the algorithm [19]. This has been achieved by means of adaptive tuning of the system of Gabor filters, i.e. automatic parameter optimisation, used to generate the local image representation. In the case of license plates, the method reaches almost 100% accuracy level.

The paper is organised as follows: in Section 2 we briefly review state-of-the-art in feature based object detection and localisation. The complex-valued multiresolution Gabor features are introduced in Section 3 and their statistical ranking in Section 4. The methodologies discussed in the previous sections are developed in the practical algorithms for training and localising the image features in Section 5 and applied to problems of face landmark and license plate corner localisation in Section 6.

II. Feature based object detection and recognition

Feature based object detection covers approaches which are based on first localising image features (local object descriptors) and then searching the most prominent configurations of the image features with a spatial constellation model representing spatial relationships between the descriptors. The process is demonstrated in Fig. 1. The most prominent configurations can be delivered as object hypotheses for the use of further processing stages, such as, object instance verification [12]. One of the main advantages of feature based methods is a natural tolerance to occlusion, as spatial models can cope with undetected image features. The deformable spatial constellation models also possess more favourable modelling properties of real objects as compared to the holistic models [4]. The feature based approach also provides reusability since the image feature localisation and spatial constellation model can be implemented and tailored independently for different applications. The problem of image feature localisation can itself be categorised as an object localisation problem since it exploits both low-level feature extraction and feature classification.

The local image feature approach to the general object detection has not been receiving the attention it deserves. The first proposal dates back to Fischler and Elschlager [9], and another notable example is the labelled graph matching method by von Malsburg et al. (e.g. [23], [36], [40]). Recently, more efficient methods have been introduced, e.g., the voting method by Lowe [25] and the statistical method by Weber [39]. Lowe has his own method for selecting and representing keypoints [25], but the applications of Weber's approach in [7] utilise unsupervised descriptors by Kadir [18]. This study does not concentrate on the detection of complete objects but only the localisation of image features.

In their earlier work, the authors have proposed multiresolution Gabor features which can be used to represent local image patches [21]. The expressive power of the proposed features is a key advantage of the representation since by increasing their dimension an arbitrary amount of image energy can be retained. Due to the expression power, the keypoints provided by the multiresolution Gabor features can be considered as

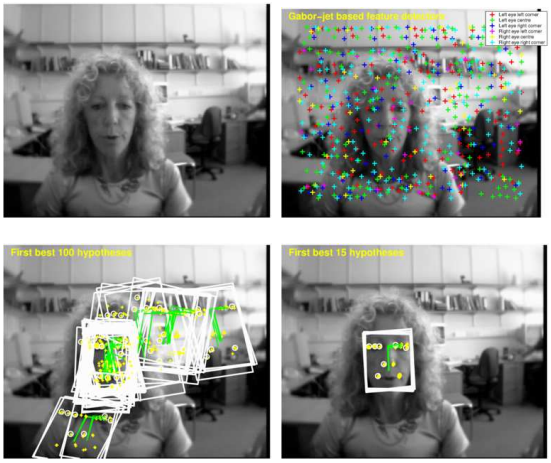


Fig. 1. Example of object class localisation based on image features and spatial constellation model search: (a) Original image; (b) Extracted image features in the image; (c) Object hypotheses generated by a spatial constellation model search; (d) 15 best object hypotheses. [12]

more general descriptors than SIFT descriptor [25] or scale descriptors [18], and thus, a simpler constellation model can be used to achieve the same detection and localisation performance.

A. From low-level features to localised image features

Detecting and localising image features is a pattern recognition problem and, in general, any feature extraction and classification method can be applied. However, an important consideration is what information should be provided in order to enable efficient processing also in upper layers. The options are either to use simple object features (e.g. corners [13]), but this gives rise to ambiguity in high level object interpretation; corners are shared by too many local image patches. As an alternative, the simple features can be used to select only some salient image patches ([18], [25]) which are then represented by unique descriptors, such as SIFT [25]. The unique local descriptors however are very object specific allowing no generalisation over other objects from the same object class. The alternative proposed in this work is to exploit the semantics of local image features and this requires unique image feature identification, in terms of class labels, shared by the same object class. The estimation of pose parameters (location, rotation and scale) may also speed up the subsequent processing, e.g., via pruning the constellation model search space. Another very useful piece of information is the level of confidence associated with the assignment of each image feature (the class label). The confidence information can be used to control the high level interpretation process by processing the most promising image features and discarding the most improbable ones.

If the detection is performed over different rotations and scales the confidence information may also be used to return the most probable pose parameters. Considering the given requirements, statistical methods seem to be the most applicable for classifying features to different keypoint classes. In statistical approaches points can be represented via their class conditional probability density functions (pdf's) and the identification and ranking can be based on the statistical probability values [29].

III. Multiresolution Gabor features

In this section, the construction and computation of multiresolution Gabor features will be described in detail. A multiresolution Gabor feature is based on image responses of 2-D Gabor filters. In addition to their well-known correspondence to the receptive field receptor profiles in the mammal visual system [6], the 2-D Gabor filter is a realisation of the general image processing operator proposed by Granlund [10]. The multiresolution structure in the frequency domain is similar to the wavelets, but without the orthogonality property. Gabor features are considered to span a frame, not a basis. The frame is a generalisation of basis, without the orthogonality and unique dual transform property. Frames however have many beneficial properties for object detection and recognition [20]. It should be noted that the exploitation of the redundancy allows a fast implementation of multiresolution Gabor features [17].

A. 2-D Gabor filter

A 2-D Gabor filter can be divided into an elliptical Gaussian and a complex plane wave. The filter in the 2-D spatial domain is [20]

$$\psi\left(x,y;f_0,\theta\right)=\frac{f_0^2}{\pi\gamma\eta}e^{-\left(\frac{f_0^2}{\gamma^2}x'^2+\frac{f_0^2}{\eta^2}y'^2\right)}e^{j2\pi f_0 x'}$$
$$x'=x\cos\theta+y\sin\theta\ \ ,\ \ y'=-x\sin\theta+y\cos\theta$$

where f_0 denotes the filter tuning frequency and the bandwidth is controlled by two parameters, γ and η, corresponding to the two perpendicular axes of the Gaussian. θ denotes the rotation angle of both the Gaussian and plane wave. It should be noted that this is not the most general form of the 2-D Gabor filter, but a form whose properties are the most useful in image processing [20].

The Fourier transformed version of the filter in the frequency domain is

$$\Psi(u,v;f_0,\theta)=e^{-\pi^2\left(\frac{u'-f_0}{\alpha^2}+\frac{v'}{\beta^2}\right)}$$
$$u'=u\cos\theta+v\sin\theta\ \ ,\ \ v'=-u\sin\theta+v\cos\theta$$

where it is evident that it is a single Gaussian band-pass filter.

B. Multiresolution structure

The multiresolution structure was originally introduced by Granlund as a general structure [10] and recently as specialised to Gabor filters by the authors [21]. The authors originally referred to it as a simple Gabor feature space, where the phrase "simple" refers to the fact that the feature space considers phenomena, here image features, at a single spatial location. A single spatial location does not straightforwardly correspond to a single pixel in digital images since the effective area, envelope, of a Gabor filter stretches over a substantially larger area; yet the local signal reconstruction accuracy is the highest

near the centroid. It is clear that complex objects cannot be represented by a simple Gabor feature which is representative only near its centroid but rather a spatial constellation model must be built based on several features.

The main idea in simple Gabor feature space is to utilise a response of Gabor filter $\psi(x, y; f, \theta)$ at a single location (x_0, y_0) of image $\xi(x, y)$

$$r_\xi(x_0, y_0; f, \theta) = \psi(x_0, y_0; f, \theta) * \xi = \iint_{-\infty}^{\infty} \psi(x_0 - x_\tau, y_0 - y_\tau; f, \theta)\xi(x_\tau, y_\tau)\mathrm{d}x_\tau \mathrm{d}y_\tau \quad . \tag{1}$$

The response is calculated for several frequencies f_l and orientations θ_l (l not necessarily the same).

The frequency corresponds to the scale which is not an isotropic variable. The spacing of frequencies must be exponential [21]

$$f_l = k^{-l} f_{max}, \;\; l = \{0, \ldots, m-1\} \tag{2}$$

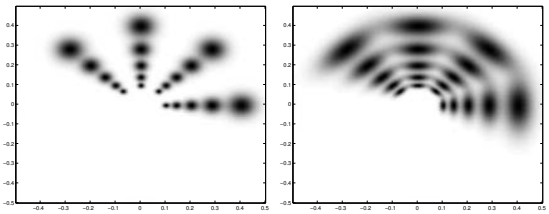
where f_l is the lth frequency, $f_0 = f_{max}$ is the highest frequency desired, and k is the frequency scaling factor ($k > 1$ $k \in \mathbb{R}$).

The rotation operation is isotropic, and thus, it is necessary to position the filters in different orientations uniformly as [21]

$$\theta_l = \frac{l2\pi}{n}, \;\; l = \{0, \ldots, n-1\} \tag{3}$$

where θ_l is the lth orientation and n is the number of orientations to be used. The computation can be reduced to half since responses at angles $[\pi, 2\pi[$ are complex conjugates of responses at $[0, \pi[$ for real valued signals.

Examples of multi-resolution filter banks are shown in Fig. 2. An optimal construction can be based on a selection of several combinations of filter parameter values while the other ones can be analytically derived [17].



(a) $\gamma = \eta \approx 2.35$ (b) $\gamma \approx 2.35$, $\eta \approx 1.03$

Fig. 2. Examples of Gabor filter banks in the frequency domain, both use $m = 5$ frequencies, $n = 4$ orientations. The envelope overlap is $p = 0.2$ and frequency spacing factor $k = \sqrt{2}$ (for more information on variables see [17]); (a) $\gamma = \eta \approx 2.35$; (b) $\gamma \approx 2.35, \eta \approx 1.03$.[17]

1) Gabor feature matrix: Gabor filter responses in a single location (simple Gabor features) can be conveniently arranged into a matrix form as

$$\mathbf{G} = \begin{pmatrix} r(x_0,y_0;f_0,\theta_0) & r(x_0,y_0;f_0,\theta_1) & \cdots & r(x_0,y_0;f_0,\theta_{n-1}) \\ r(x_0,y_0;f_1,\theta_0) & r(x_0,y_0;f_1,\theta_1) & \cdots & r(x_0,y_0;f_1,\theta_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_0,y_0;f_{m-1},\theta_0) & r(x_0,y_0;f_{m-1},\theta_1) & \cdots & r(x_0,y_0;f_{m-1},\theta_{n-1}) \end{pmatrix} \tag{4}$$

where rows correspond to responses at the same frequency and columns correspond to responses at the same orientation. The first row is the highest frequency and the first column is typically the angle $0°$.

C. Feature manipulation for invariant search

From the responses in the feature matrix in Eq. (4) the original signal $\xi(x, y)$ can be approximately reconstructed near the spatial location (x_0, y_0) [16], [20].

The additional property which makes multiresolution Gabor features useful is the fact that linear row-wise and column-wise shifts of the response matrix correspond to the scaling and rotation in the input space. Thus, an invariant search can be performed by simple shift operations over all spatial locations (spatial shift).

Rotating an input signal $\xi(x, y)$ anti-clockwise by $\frac{\pi}{n}$ equals to the following shift of the feature matrix

$$\begin{pmatrix} r(x_0,y_0;f_0,\theta_{n-1})^* & r(x_0,y_0;f_0,\theta_0) & \Rightarrow & r(x_0,y_0;f_0,\theta_{n-2}) \\ r(x_0,y_0;f_1,\theta_{n-1})^* & r(x_0,y_0;f_1,\theta_0) & \Rightarrow & r(x_0,y_0;f_1,\theta_{n-2}) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_0,y_0;f_{m-1},\theta_{n-1})^* & r(x_0,y_0;f_{m-1},\theta_0) & \Rightarrow & r(x_0,y_0;f_{m-1},\theta_{n-2}) \end{pmatrix} \tag{5}$$

where * denotes the complex conjugate.

Downscaling the same signal by a factor $\frac{1}{k}$ equals to the following shift of the feature matrix

$$\begin{pmatrix} r(x_0,y_0;f_1,\theta_0) & r(x_0,y_0;f_1,\theta_1) & \cdots & r(x_0,y_0;f_1,\theta_{n-1}) \\ r(x_0,y_0;f_2,\theta_0) & r(x_0,y_0;f_2,\theta_1) & \cdots & r(x_0,y_0;f_2,\theta_{n-1}) \\ \Uparrow & \Uparrow & \ddots & \Uparrow \\ r(x_0,y_0;f_m,\theta_0) & r(x_0,y_0;f_m,\theta_1) & \cdots & r(x_0,y_0;f_m,\theta_{n-1}) \end{pmatrix} \tag{6}$$

It should be noted that responses on new low frequencies f_m in (6) must be computed and stored in advance while the highest frequency responses on f_0 vanish. It is important to notice that the proposed simple linear shift operations are much more efficient than any exhaustive search where original signal is being rotated and scaled.

D. Selecting filter parameters

Since the image feature localisation proposed in this work is a supervised approach it requires a training set of image features. The same set can also be used for selecting the feature parameters, most importantly the number and values of Gabor filter frequencies. There is no general rule how the frequencies $f_0, \ldots, f_{m-1}$ should be selected. In many studies the frequencies are optimised using a certain score function, such as a maximal signal energy restoration or maximal separation between two input classes [2], [3], [5], but they often lead to non-homogeneous parameter sampling, violating Eqs. (2) and (3), which in turn makes invariant processing difficult because signal rotation and scaling cannot be handled by simple matrix manipulations as in Eqs. (5) and (6). The parameter selection restrictions can be embedded in the optimisation process, but the optimisation is still a computationally expensive task, and furthermore, it is very difficult to optimise without negative examples. The maximal discrimination between different image features does not guarantee optimality in actual scenes because only positive training examples are used and they are generally not presentative of all possible image features. If the optimisation is however

desired the cross-validation seems to be the only realisable option. In the experimental part of this study the efficiency of such a simple technique as cross-validation is demonstrated.

1) General guidelines: The first problem in the parameter selection is the number of frequencies, m, and orientations, n, to be used in the feature matrix in Eq. (4). Many factors contribute to the total performance; the more frequencies and orientations are used the better is the representational power of the multiresolution Gabor feature. By increasing these numbers the shift sensitivity increases too, allowing a more accurate determination of an image feature pose. However, the representational power is also influenced by the effective areas of Gabor filters controlled by the bandwidth parameters γ and η. As a rule of thumb, as large number of filters as possible within given computational resources should be used. Generally, the bandwidth values can be set to $\gamma = \eta = 1.0$ and a good initial number of filters are four orientations $n = 4$ on three frequencies $m = 3$ making the feature matrix of size 3×4. The effect of changing parameter values can be evaluated experimentally. It should be noted that it is not necessary to compute features at all locations due to an enormous redundancy (overlap) of filters. A sufficient spacing of the filters can be computed via an effective shiftability measure, e.g., in [32], and the redundancy can also be utilised in an efficient computation of the multiresolution Gabor features [17].

IV. Statistical feature classification and ranking

In general, any classifier or pattern recognition method can be used with Gabor features to classify extracted low-level features into image feature classes. However, certain advantages suggest that the use of statistical methods is preferable. One of these, namely the ability to provide not only the class assignments for the observed features but also to rank scene points in order to return only a fixed number of best features is of particular importance. The purpose of ranking is to reduce the computational load in the spatial model search.

The between-class ranking of features is the traditional problem of classification where, for example, Bayesian inference has been effectively applied. Within-class ranking however requires an information measure describing the degree of certainty of a feature belonging to a specific class. A certainty measure can be realised with most classification methods, but statistical methods usually provide certainty information in an interpretable form along with solid mathematical background [29]. In that sense, statistical methods possess superior properties as compared to other methods; the decision making has an interpretable basis from which the most probable or lowest risk (expected cost) option can be chosen and a within-class comparison can be performed using statistical hypothesis testing [29]. The information measure reflecting statistical uncertainties will be referred to as confidence.

As in any supervised learning problem, a set of instances of known observations (a training set) are provided and the necessary statistics must be inferred to classify new unknown observations and to estimate the classification confidence. A class is typically represented in terms of a class conditional probability density function (pdf) over feature space. It should be noted, that finding a proper pdf estimate has a crucial impact on the success of the classification and ranking. Typically, the form of the pdf's is somehow restricted and the estimation is reduced to a problem of fitting the restricted model to the observed features. Often simple models such as a single Gaussian distribution (normal distributed random variable) can efficiently represent features but a more general model, such as a finite mixture model, must be used to approximate more complex pdf's; arbitrarily complex probability density functions can be approximated using finite mixture models. The finite mixture representation is a natural choice for certain kinds of observations: observations which are produced by a randomly selected source from a set of alternative sources belonging to the same main class. This kind of task arises when we deal with object categories (classes) rather than object instances. For example, features from eye centres are partitioned into closed eye and open eye, or Caucasian and Asian eye sub-classes. The problem, which will be considered next, is how the probability densities should be approximated with finite mixture models and how the model parameters should be estimated.

A. Class conditional pdf estimation using Gaussian mixtures

Finite mixture models can approximate a wide variety of pdf's and are thus attractive solutions for cases where single function forms, such as the normal distribution, fail. However, from a practical point of view it is often sound to form the mixture using one predefined distribution type, a basic distribution. Generally the basic distribution function can be of any type but the multivariate Gaussian distribution is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many areas [35]. For example, in multivariate analysis most of the existing inference procedures have been developed under the assumption of normality and in linear model problems the error vector is often assumed to be normally distributed. The multivariate normal distribution also appears in multiple comparisons, in studies of the dependence of random variables, and in many other related fields. If no prior knowledge of the pdf of a phenomenon exists, only a general model can be constructed and the Gaussian distribution is a good candidate. For a more detailed discussion on the theory, properties and analytical results of multivariate normal distributions we refer the reader to [35].

The multiresolution Gabor feature computed in a single location can be converted from the matrix form in (4) to a feature vector as

$$\vec{g} = [r(x_0, y_0; f_0, \theta_0) \;\; r(x_0, y_0; f_0, \theta_1) \ldots r(x_0, y_0; f_{m-1}, \theta_{n-1})] \tag{7}$$

Since the feature vector is complex valued the complex Gaussian distribution function (e.g. [29]),

$$\mathcal{N}^{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\pi^D |\Sigma|} \exp\left[-(\boldsymbol{x} - \boldsymbol{\mu})^* \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \quad , \tag{8}$$

where Σ denotes the covariance matrix, must be used in the mixture model. It should be noted that the pure complex form of the Gaussian in (8) provides computational stability in the parameter estimation as compared to a concatenation of real

and imaginary parts to two real numbers as the dimensionality of the problem doubles in the latter case [29]. Now, a Gaussian mixture model (GMM) probability density function can be defined as a weighted sum of Gaussians

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c \mathcal{N}^{\mathbb{C}}(\boldsymbol{x};\boldsymbol{\mu}_c,\Sigma_c) \tag{9}$$

where α_c is the weight of the cth component. The weight can be interpreted as *a priori* probability that a value of the random variable is generated by the cth source, and thus, $0 \leq \alpha_c \leq 1$ and $\sum_{c=1}^{C} \alpha_c = 1$. The Gaussian mixture model probability density function can be completely defined by a parameter list

$$\boldsymbol{\theta} = \{\alpha_1, \boldsymbol{\mu}_1, \Sigma_1, \ldots, \alpha_C, \boldsymbol{\mu}_C, \Sigma_C\} \ . \tag{10}$$

The main question remains how the parameters in (10) can be estimated from the given training data. The most popular estimation method is the expectation maximisation (EM) algorithm, but the EM algorithm requires the knowledge of the number of Gaussians, C, as an input parameter. The number is often unknown and this is a strong motivation to apply unsupervised methods, such as that of Figueiredo-Jain (FJ) [8] or the greedy EM algorithm [37]. The unsupervised methods may provide accurate and reliable results despite the fact that the standard EM algorithm outperforms them if the correct number of Gaussians is known [29]. Of the two unsupervised methods the Figueiredo-Jain method provides more accurate results and its complex extension can be directly applied to pdf's of complex multiresolution Gabor feature vectors in (7) [29].

B. Likelihood as confidence measure

The term confidence may have different meanings and definitions, but in our case confidence is used to measure the reliability of a classification result where a certain class is assigned to an observation. If the confidence is low, it is more probable that a wrong decision has been made. Intuitively, the value of a class conditional pdf at an observation reflects the confidence that the observation is consistent with that class: the higher the pdf value, the more instances of the class will be similar to the observation. While the posteriori probability is a between-class measure for a single observation, pdf value is an intra-class measure that can be used to select the best representative from a single class [29].

A confidence value measure that exhibits the properties of true probabilities should satisfy $\in [0,1]$. For any finite or infinite support region $\mathcal{R} \subseteq \Omega$, where Ω is the definition space of the pdf, it holds that $0 \leq p(\boldsymbol{x}) < \infty$, $\forall \boldsymbol{x} \in \Omega$. Now the value κ can be defined via non-unique confidence region $\mathcal{R}$ such that [29]

$$\int_{\Omega \setminus \mathcal{R}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \kappa \ . \tag{11}$$

The confidence value we propose is easily interpretable via the confidence region $\mathcal{R}$. The confidence region is a region which covers a proportion $1-\kappa$ of the probability mass of $p(\boldsymbol{x})$ since for all probability distributions $\int_{\Omega} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$. It is clear that $\kappa = 1$ for $\mathcal{R}$ containing only a finite number of individual points and $\kappa = 0$ for $\mathcal{R} = \Omega$. It should be noted that it makes no sense to talk about confidence value until the region $\mathcal{R}$ is defined as the minimal volume region which satisfies the confidence definition. The minimum volume region is also called the highest density region (HDR) in the literature [15]. The confidence value we propose, $1-\kappa$, corresponds to the smallest set of points (region) which includes the observation $\boldsymbol{x}$ and has the probability mass κ, κ being smallest possible.

A class specific confidence value κ_j can be defined for each image feature class $j = 1, \ldots, J$, but intuitively the same value should apply for all classes. A confidence value corresponds to the proportion of the probability mass that remains in region $\mathcal{R}_j$. In a classification task where a certain confidence for decision making is required, the confidence value is not used but the confidence region $\mathcal{R}_j$ itself is important since a sample vector $\boldsymbol{x}$ is allowed to enter the class ω_j only if $\boldsymbol{x} \in \mathcal{R}_j$. If a sample is not within the confidence region of any of the classes, it must be classified to a garbage class. The garbage class is a special class and samples assigned to the class need special attention; for example, more information is needed for the observations falling into the garbage class or in a two-class problem where data is available only from one class the garbage class may represent the other class with an unknown distribution.

The probability distribution values can be directly used to rank image features in an image or even over several images, but if the probabilistic property is required, that is, normalised values between $[0,1]$, then the confidence must be solved by computing the minimum highest density region which includes the observation. The calculation of the highest density region is not trivial, but efficient and accurate approximation methods with convergence do exist, e.g., in [29].

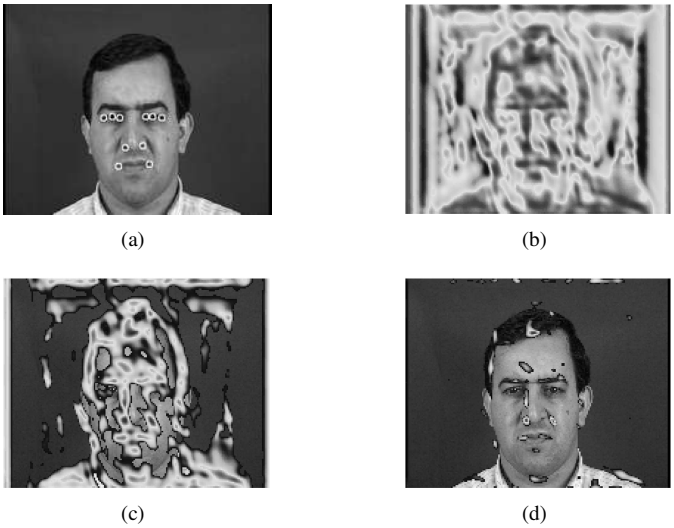
1) Confidence score example: In the previous section it was argued that the class-conditional probability density (likelihood) value is the preferred choice as a ranking confidence score. It is a measure of how reliable a class assignment of a single image feature is. Image features with the highest scores can be delivered to the spatial model first. The use of confidence values may reduce the search space considerably by discarding image features beyond a requested density quantile [29]. In Fig. 3 the use of density quantile for reducing the search space is demonstrated; it is clear that the spatial domain corresponding to the 0.05 (0.95 confidence) density quantile contains the correct image feature.

V. Image feature localisation

In this section we apply the findings from the previous sections and propose algorithms for supervised training and localisation of image features.

A. Training algorithm

Based on the given results it is straightforward to establish a supervised training algorithm for detecting instances of image feature classes. For a set of training images, every image should first be aligned, that is, object instances are represented approximately in the same scale and orientation. With the training set images where groundtruth (image feature



(a) (b) (c) (d)

Fig. 3. Example of using density quantile and pdf values as confidence: (a) Face image and 10 image feature classes; (b) Pdf surface for the *left nostril* class; (c) Pdf values belonging to 0.5 density quantile; (d) Pdf values belonging to 0.05 density quantile.

locations) are available the alignment is a simple task. Then the multiresolution responses are computed at the groundtruth locations and pdf's are estimated. An algorithm to perform the supervised training is presented in Algorithm 1.

Algorithm 1: Train image feature classifier

1: **for all** Training images **do**
2: Align and normalise image to represent an object in a standard pose
3: Extract multiresolution Gabor features at given locations
4: Normalise the features
5: Store the features to the sample matrix P and their corresponding class labels (class numbers) to the target vector T
6: **end for**
7: Using samples in P estimate class conditional pdf's separately for each class using Gaussian mixture models and the FJ algorithm

In Algorithm 1 we give the fundamental steps to generate a pdf-based classifier for image feature extraction. First, the training images must be aligned to a standard pose, i.e. the pose representing objects in the same scale and orientation. In the standard pose, multiresolution Gabor features in Eq. (4) are then computed at groundtruth image feature locations. Feature matrices can be energy-normalised, e.g. to unity matrix norm, if robustness to illumination change is required [20]. The normalisation makes the feature invariant to (local) multiplication by a constant factor. Each feature matrix is reformatted into a vector of complex numbers form in Eq. (7) and stored in the sample matrix P along with the corresponding image feature labels, T. Finally, pdfs over the complex feature vectors are estimated for each image feature class separately with GMM's and the FJ algorithm.

B. Localisation algorithm

After the training phase the image feature specific pdf's can be used to detect and localise image features in input images. In Algorithm 2 the main steps to extract the features from an image are shown.

Algorithm 2: Extract K best image features of each class from an input image I

1: Normalise image
2: Compute multiresolution Gabor features $G(x, y; f_m, \theta_n)$ for the whole image $I(x, y)$
3: **for all** Scale shifts **do**
4: **for all** Rotation shifts **do**
5: Shift Gabor features
6: Normalise Gabor features
7: Calculate confidence values for all classes and for all (x, y)
8: Update feature class confidence at each location
9: **end for**
10: **end for**
11: Sort the image features for each class
12: Return the K best features of each image feature class

First, the image is normalised, that is, scale and grey levels are adjusted to correspond to average object presence used in the training. From a normalised image multiresolution Gabor features are extracted at every spatial location and confidence values are computed for all requested invariance shifts. If Gabor features were energy normalised in the training phase the same normalisation must be applied before calculating the confidence values of GMM pdf's. In a less memory intensive implementation, confidence values can be iteratively updated after each shift in order to store only the best image features of each class at each location. After the shifts have been inspected it is straightforward to sort them and return the best image feature candidates. In this approach one location may represent more than one image feature, but each feature can be assigned to one pose only.

VI. Experiments

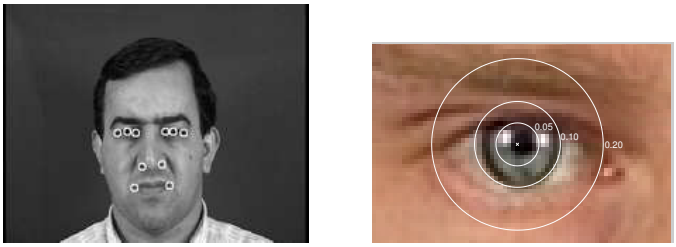
With the following experimental results we validate the theoretical basis and demonstrate the accuracy of our devised image feature localisation method.

A. XM2VTS face database

XM2VTS facial image database is a publicly available and popular database for benchmarking face detection and recognition methods [26]. The frontal part of the database contains 600 training images and 560 test images of size 720×576 (width $\times$ height) pixels. Images represent front pose on constant background and are of excellent quality, and thus, any face detection method should perform very well with the database.

The appearance of a set of salient facial features is selected as the image features. The regions corresponding to these image features should be stable over the whole object category, but at the same time must be discriminative enough to set the object apart from other object categories and backgrounds. For

facial images ten specific regions (see Fig. 4(a)) have been shown to have favourable properties to act as keypoints [12]. A normalised distance between the eyes, 1.0, will be used as measure of image feature detection accuracy. This kind of measure is considered as the most appropriate for evaluating localisation methods in [30]. The distance measure is illustrated in Fig. 4(b).



(a) (b)

Fig. 4. (a) XM2VTS face and 10 salient keypoints (left and right outer eye corners, left and right inner eye corners, left and right eye centres, left and right nostrils, and left and right mouth corners). Left and right correspond to the image left and right here. (b) Demonstration of accuracy distance measure.

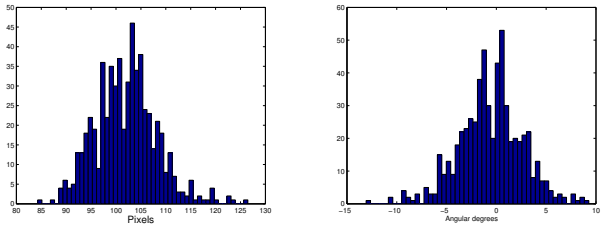
1) Gabor feature parameter selection: The multiresolution Gabor filter parameters can be easily selected by optimisation procedures, such as the simple cross-validation used in this study, over an evaluation set. Since it is, however, also easy to tune the parameters manually, the manual selection principles are next explained as they help to understand the properties of the multiresolution Gabor features. Results with this manual selection used in our previous studies are demonstrated in the experiments.

In the case of the XM2VTS database all faces are in almost the same pose, there are no large variations in the distance between eyes (Fig. 5(a)) and the angle between the eyes (Fig. 5(b)). This means that invariant searches are not needed, and the filter bank parameters should be selected to cover any residual variations. Angular variations are limited to $\pm 10°$, therefore up to eight filter orientations, $n \leq 8$, can be used (angular discrimination is then $22.5°$).

The lower bound for filter frequency spacing, k, can be determined by examining the scale differences in the training images (Fig. 5(a)). The highest eye distances are approx. 120 pixels, and the lowest approx. 90 pixels. Filters should include at least that much scale variations, therefore $k \geq \frac{120}{90} \approx 1.33$. To be on the safe side a slightly larger value can be selected, for example, $k = \sqrt{2}$. Higher values still can be useful, depending on the characteristics of image features. As for selecting the number of filter frequencies no clear guidelines can be given, however, the value should generally be $m \geq 3$ to provide enough discriminative frequency information.

A more thorough explanation on how the set of parameters called "old parameters" in this document were selected is presented in [19] ($n = 3$, $m = 4$, $k = \sqrt{2}$ and $f_{high} = 1/30$).

The parameters called "tuned parameters" were experimentally selected by using a cross-validation procedure over the training and evaluation sets in the database. These parameters were $n = 4$, $m = 6$, $k = \sqrt{3}$ and $f_{high} = 1/40$. Compared to the old parameters the total number of filters was doubled and



(a) (b)

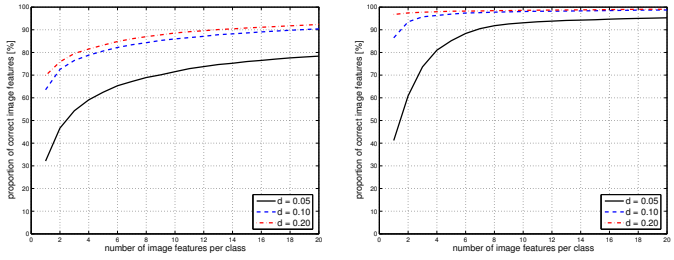
Fig. 5. Scale and orientation contents of XM2VTS training data computed using coordinates of left and right eye centres: a) distribution of eye centre distances (min. 84 pix, max. 126 pix, mean 102 pix); b) distribution of eye centre rotation angles (abs. min. $0°$, abs. max. $13.0°$, abs. mean $2.5°$, mean $-0.5°$).

because of larger filter spacing, k, filter bank includes much lower frequencies.

2) Training mixture model pdf: The Gabor filter responses were computed at 10 spatial locations in every image in the training set and arranged as feature vectors which can be used with any one-class classifier. We opt for one-class classification in order to avoid modelling the background class; the background class should include a very comprehensive set of features collected from various images, which is impractical. A classifier based on Gaussian mixture models (GMM) was used. The pdf estimation was performed with the unsupervised FJ algorithm because the number of mixture components is in general unknown [29]. During the classification phase Gabor filter responses were computed in all locations of the image, and classified in each of the 10 classes. For each class the classification results, i.e. pdf values, were sorted and a number of the highest ranked were considered potential image features. It must be noted that with certain types of signals Gabor filter responses can be highly non-Gaussian (for example, responses change very rapidly if the location of the filters is offset from the centre of a perfect circle), and another type of classifier, such as support vector machine (SVM) [33], may perform better than a classifier based on GMMs.

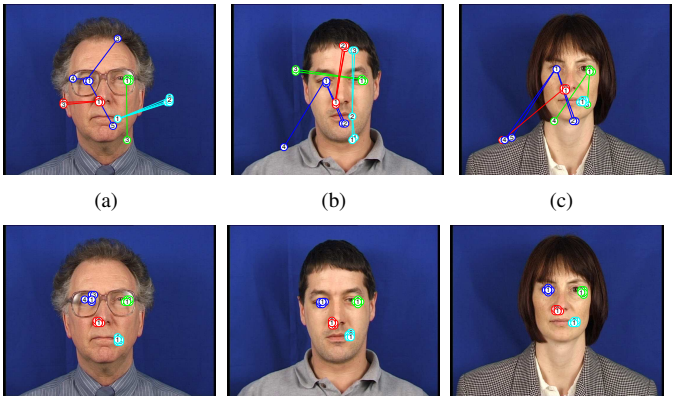
3) Results for original images: Image features were extracted in a ranked order and a keypoint was considered to be correctly extracted if it was within a pre-set pixel distance from the correct location. The results with XM2VTS are presented in Fig. 6. The distances are scale normalised, so that the distance between centres of the eyes is 1.0 (see Fig. 4(b)). With the old parameters, Fig. 6(a), not all image features can be extracted within the distance of 0.05 on average, but at least 3 correct image features were included in the first 10 image features (1 per each class) and by increasing the number to 100 a significant improvement was achieved (7 for 0.05 , 9 for 0.10 and 0.20). With the tuned parameters, Fig. 6(b), on average 4 correct image features were among the first 10 image features within the distance limit of 0.05, but a significant improvement was noted as the number of features was increased to 100: over 9 for 0.05 and almost all features found for 0.10 and 0.20. It should be noted that accuracies of 0.10 and 0.20 are still very good (Fig. 4(b)). Increasing the number of image features over 100 (10 per class) did not improve the results significantly,

but for the tuned parameters, relaxing the distance threshold to 0.10, almost perfect results were obtained with only 10 first image features from each class. The accuracy is more evident in the example images and the extracted features shown in Fig. 7, where the non-optimal parameters typically provide very accurate best feature, but the next best candidates spread over the image to false locations. This problem does not occur with the tuned parameters, that recognise the landmark from its "neighbourhood". It should be noted that the constant background appearing in Fig. 7 generates many false alarms and in general is more difficult than a textured background since the illumination normalisation of the feature matrix tends to produce undesirable artifacts in this case.



(a) (b)

Fig. 6. Accuracy of image feature extraction from XM2VTS test images: (a) Old parameters; (b) Tuned parameters.

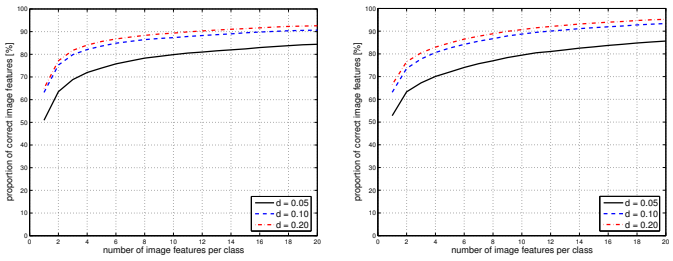


(a) (b) (c)

(d) (e) (f)

Fig. 7. Examples of extracted features (left eye centre: blue, right eye outer corner: green, left nostril: red, right mouth corner: cyan, 5 best feature for each landmark numbered from 1 to 5): (a),(b),(c) Old parameters from [19]; (d),(e),(f) Tuned parameters.

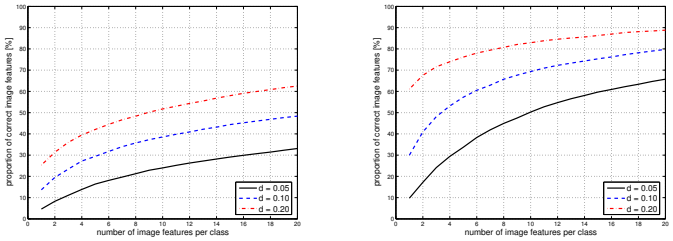
4) Other features: To compare our Gabor features to other popular and widely used methods another multiresolution method, steerable filters [34], and recently introduced very accurate (state-of-the-art results in face recognition) feature descriptor, local binary patterns (LBPs) [11], were used to replace the Gabors in our image feature localisation method. It should be noted that the fast shift procedures can be established for the steerable filters, but LBP requires slow exhaustive search. Parameters of the two methods were correspondingly optimised and executed to produce comparable results. The results are shown in Fig. 8. Both methods performed very well, but the proposed method using Gabor features still remained superior to them (Fig. 6(b)). The SIFT descriptor was also tested, but due to its poor performance, explained by its incompatibility for the given task, the results are not reported.



(a) (b)

Fig. 8. XM2VTS results with: (a) Local binary patterns; (b) Steerable filters.

5) Complex vs. magnitude responses: The multiresolution Gabor features as described in Section III use complex valued feature responses. The majority of studies using Gabor features utilise only the magnitude information due to its simplicity in numerical computation. However, it can be demonstrated that the loss of phase information impoverishes the feature representation and results in the degradation of localisation accuracy. This fact was experimentally confirmed by performing the previous experiment with response magnitudes only. The results are shown in Fig. 9, where the decrease in the performance is evident as compared to Fig. 6. Although the magnitude only response works quite well provided that the system parameters are tuned for the application in hand, and this is the main reason for the persistent popularity of the magnitude representation, the complex representation is far superior.



(a) (b)

Fig. 9. Accuracy of image feature extraction from XM2VTS test images (response magnitudes used instead of complex responses): (a) Old parameters; (b) Tuned parameters.

The advantage of using complex (magnitude and phase) instead of magnitude-only representation can be clearly seen in Fig. 10, where responses of a single filter are plotted for the left and right eye corners. In the complex plot the two classes are clearly separable, but completely overlap in the magnitude-only plot.

6) Likelihood vs. a posteriori: In Section IV we argued that the likelihood value computed from the feature specific pdf's is the correct statistical measure to rank the features within image. Here we demonstrate how the other natural option of

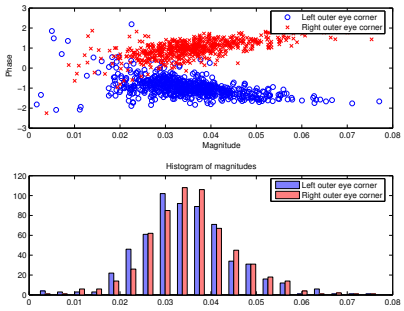
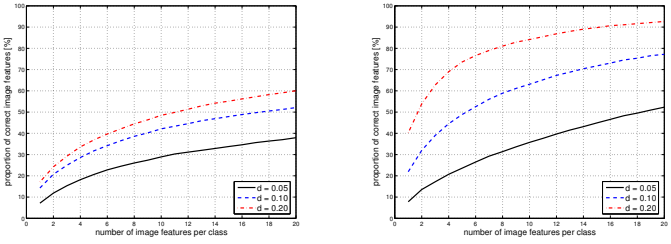


Fig. 10. Scatter plots of Gabor filter responses for left and right eye corners.

using a posteriori probability values derived from the Bayes rule fail to deliver good performance. This was achieved by repeating the first experiment but using a posteriori values to extract best image features instead. The results, shown in Fig. 11, were fairly good (tuned), but still far from the accuracy achieved using likelihood values in multiple hypothesis testing (Fig. 6).

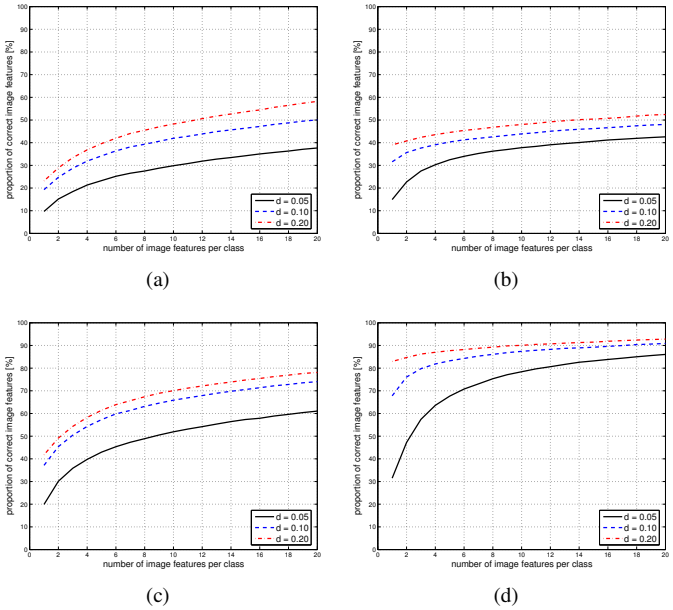


(a) (b)

Fig. 11. Accuracy of image feature extraction from XM2VTS test images (posteriori values used instead of likelihoods): (a) Old parameters; (b) Tuned parameters.

7) Results on artificially rotated and scaled images: The main problem with XM2VTS data set was that faces did not comprehensively cover different scales and rotations (see Fig. 5), and therefore, the invariance properties of the image feature extraction could not be verified. The results with artificially rotated and scaled images in XM2VTS database are presented in Fig. 12. The images in the test set were randomly rotated between $-45°$ to $45°$, and up-scaled by a random factor between 1 to $\sqrt{2}$. First, the image features were searched without using scale or in-plane rotation invariance manipulations. The results for these tests are presented in Figs. 12(a) and 12(b) for old and tuned filter bank parameters. Second, in the image feature detection phase one scale-shift and two orientation shifts (-1 step and $+1$ step) were applied. For the old filter bank parameters this means that the scale-shift is $\sqrt{2}$ and the orientation shifts $-45°$ and $45°$, and for the tuned parameters scale-shift is $\sqrt{3}$ and orientation shifts $-30°$ to $30°$. The results for these tests are presented in Figs. 12(c) and 12(d). Using scale and orientation shifts gives significantly better results. The difference is especially noticeable with the tuned parameters, Figs. 12(b) and 12(d).

8) Summary: It is noteworthy, that the accuracy of the results with the tuned parameters correspond to the natural



(a) (b)

(c) (d)

Fig. 12. Accuracy of image feature extraction from artificially rotated and scaled images from XM2VTS test set; (a) old parameters - no invariance shifts; (b) tuned parameters - no invariance shifts; (c) old parameters - invariance shifts; (d) tuned parameters - invariance shifts.

variation in accuracy of the manual marking of different humans, that is, the results cannot be improved as compared with the current XM2VTS protocol. Furthermore, the merits of using complex responses and likelihood in feature ranking were also experimentally confirmed.

B. Banca face database

This experiment was performed using the English section of a very challenging BANCA face database [1]. The English part includes 6240 test images of significantly varying quality, background and pose (including in-depth rotations). For the training, XM2VTS and worldmodel images from English, Spanish, Italian and French BANCA sections were used. The total number of training images was 1600.

The results are presented in Fig. 13. The settings for the filter bank were $n = 3$, $m = 6$, $k = \sqrt{3}$ and $f_{high} = 1/25$ ("tuned"). The difference to the optimised parameters with XM2VTS database was that higher frequencies were found and only three different frequencies provided the best result. The frequency changes were due to the larger scale variations in the BANCA database images and the filter bank had to be tuned for the smallest scales. The number of filter frequencies was decreased to prevent the lowest frequency filters including information from too large area, such as cluttered background. One scale shift was used in the experiment. It is evident that the BANCA database is much more difficult since the accuracy decreased from the average of more than 9 to less than 6 correct image features within the distance of 0.05 (10 features per class). The spatial search may still succeed since the minimum number of requested features is 3 (affine transformation). Some detections results including poorly detected features are shown in Fig. 14.

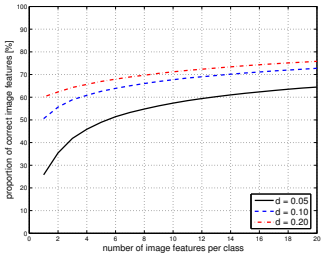


Fig. 13. Accuracy of image feature extraction from the English section of BANCA database (only tuned parameters).

(a) (b) (c)

Fig. 14. (a-c) Examples of extracted features (BANCA).

C. Commercial license plate database

To demonstrate the generality of the method, it was also applied to detect corners of car license plates captured using standard traffic camera for this purpose. Due to the lack of publicly available license plate databases a commercial database was used. 200 images captured during a randomly selected date were used as the training set where the groundtruth was manually marked (four corners, Fig. 15(a)). 400 images captured during another randomly selected date were used in testing. Similarly to the face detection example, the localisation distance was normalised according to the distance between two ultimate corners of the license plate. The measure is illustrated in Fig. 15 (3 small circles in the upper left corner).

(a) (b)

Fig. 15. License plate database: (a) Example image where corners marked with green circles (b) Demonstration of accuracy measure for license plate localisation (green circles in the upper left corner).

Since the corner provides a very salient image feature, the localisation after parameter tuning is extremely accurate (Fig. 16). By using the detected corners it is very easy to detect the whole license plate for further processing. Examples of detected license plate features are shown in Fig. 17.

VII. Conclusions

In this study, we proposed a method for accurate and efficient localisation of local image features which makes a significant contribution to the problem of feature based object detection and recognition. The proposed method is supervised and is based on multiresolution Gabor features and statistical ranking using multiple hypothesis testing. The input of the method are training images and manually marked locations of different image features in the training images. Using the training data all the parameters can be automatically adjusted. The trained system can be used to localise the features in observed images in a translation, rotation (in-plane), scale and illumination (constant factor) invariant manner, and the system is also robust to small pose (affine) changes. The method is not tied to any specific application and can be used for localisation of image features of any type. The authors believe that no competing method with a comparable localisation accuracy is currently available.

The main shortcoming of the method is its supervised nature requiring manually annotated image features (landmarks), and therefore, the problem of automatic selection of the best image features will be addressed in the future research. Accordingly, the next processing step, establishing an accurate and efficient spatial search over the extracted image features, will be under investigation.

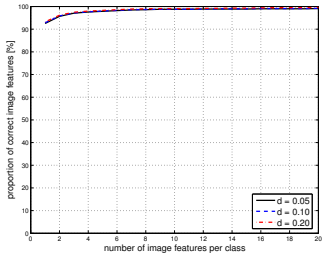
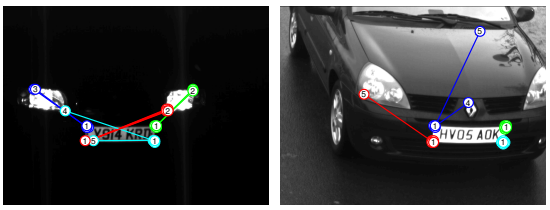


Fig. 16. Accuracy of image feature (license plate corner) extraction (only tuned parameters). Please note that for $d = 0.05$, the accuracy reaches 93% with only one (highest rank) image feature extracted.



(a) (b)

Fig. 17. Examples of extracted features (left upper corner: blue, right upper corner: green, left lower corner: red, right lower corner: cyan, 5 best features for each class numbered from 1 to 5): (a) night scene; (b) day scene.

References

[1] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *Proc. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 625–638, 2003.

[2] A. Bodnarova, M. Bennamoun, and S. Latham. Optimal Gabor filters for textile flaw detection. *Pattern Recognition*, 35:2973–2991, 2002.

[3] Manfred Bresch. Optimizing filter banks for supervised texture recognition. *Pattern Recognition*, 35:783–790, 2002.

[4] Michael C. Burl. *Recognition of Visual Object Classes*. PhD thesis, California Institute of Technology, 1997.

[5] Jian Chen, Yoshinobu Sato, and Shinichi Tamura. Orientation space filtering for multiple orientation line segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5):417–429, May 2000.

[6] John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 264–271, 2003.

[8] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, Mar 2002.

[9] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computers*, 22(1):67–92, 1973.

[10] Goesta H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8:155–173, 1978.

[11] A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages 797–804.

[12] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas. Feature-based affine-invariant localization of faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(9):1490–1495, September 2005.

[13] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the Fourth Alvey Vision Conf.*, pages 147–151, 1988.

[14] S. Helmer and D.G. Lowe. Object recognition with many local features. In *Workshop on Generative Model Based Vision (in conjunction with CVPR2004)*, Washington DC, MD, USA, 2004.

[15] R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, May 1996.

[16] J. Ilonen and J.-K. Kamarainen. Object categorization using self-organization over visual appearance. In *Proc. of the Int. Joint Conf. on Neural Networks*, Vancouver, Canada, 2006.

[17] J. Ilonen, J.-K. Kamarainen, and H. Kälviäinen. Efficient computation of Gabor features. Research report 100, Department of Information Technology, Lappeenranta University of Technology, 2005.

[18] Timor Kadir. *Scale, Saliency and Scene Description*. PhD thesis, Oxford University, 2002.

[19] J.-K. Kamarainen, J. Ilonen, P. Paalanen, H. Hamouz, H. Kälviäinen, and J. Kittler. Object evidence extraction using simple Gabor features and statistical ranking. In *Proc. of the 14th Scandinavian Conf. of Image Processing*, pages 119–129, Joensuu, Finland, 2005.

[20] J.-K. Kamarainen, V. Kyrki, and H. Kälviäinen. Invariance properties of Gabor filter based features - overview and applications. *IEEE Trans. on Image Processing*, 15(5):1088–1099, 2006.

[21] V. Kyrki, J.-K. Kamarainen, and H. Kälviäinen. Simple Gabor feature space for invariant object recognition. *Pattern Recognition Letters*, 25(3):311–318, 2004.

[22] V. Kyrki and D. Kragic. Integration of model-based and model-free cues for visual object tracking in 3d. In *IEEE International Conference on Robotics and Automation*, pages 1566–1572, Barcelona, Spain, 2005.

[23] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[24] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

[25] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[26] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS Database. In *Proc. of the 2nd Int. Conf. on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[27] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 257–263, 2003.

[28] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. of Computer Vision*, 60(1):63–86, 2004.

[29] P. Paalanen, J.-K. Kamarainen, J. Ilonen, and H. Kälviäinen. Feature representation and discrimination based on Gaussian mixture model probability densities - practices and algorithms. *Pattern Recognition*, 39(7):1346–1358, 2006.

[30] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24:882–893, 2006.

[31] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints. *International Journal of Computer Vision*, 66(3):231–259.

[32] J. Sampo, J.-K. Kamarainen, M. Heiliö, and H. Kälviäinen. Measuring translation shiftability of frames. *Computers and Mathematics with Applications*. In press.

[33] Bernhard Schölkopf. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[34] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

[35] Y.L. Tong. *The Multivariate Normal Distribution*. Springer Series in Statistics. Springer-Verlag, 1990.

[36] J. Triesch and C. von der Malsburg. Classification of hand postures against complex backgrounds using elastic graph matching. *Image and Vision Computing*, 20(13–14):937–943, 2002.

[37] J. J. Verbeek, N. Vlassis, and B. Kröse. Efficient greedy learning of Gaussian mixture models. *Neural Computation*, 5(2):469–485, Feb 2003.

[38] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. of the 6th European Conf. on Computer Vision*, pages 18–32, 2000.

[39] Markus Weber. *Unsupervised Learning of Models for Object Recognition*. PhD thesis, California Institute of Technology, 2000.

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