Object Evidence Extraction Using Simple Gabor Features and Statistical Ranking*

J.-K. Kamarainen[1], J. Ilonen[1], P. Paalanen[1], M. Hamouz[2], H. Kälviäinen[1], and J. Kittler[2]

[1] Dept. of Information Technology,
Lappeenranta University of Technology, Finland
[2] University of Surrey, United Kingdom

Abstract. Several novel methods based on locally extracted object features and spatial constellation models have recently been introduced for invariant object detection and recognition. The accuracy and reliability of the methods depend on the success of both tasks: evidence extraction and spatial constellation model search. In this study an accurate and efficient method for evidence extraction is introduced. The proposed method is based on simple Gabor features and their statistical ranking.

1 Introduction

By object evidence extraction we refer to the detection of local descriptors and salient sub-parts of objects. This approach can recover from object occlusion in a natural way; occlusion prevents the detection of all features, but the detection can still be based on a sub-set of features. Thus, it seems that the approach is a good candidate for general object detection and recognition. The idea of partitioning an object into smaller pieces which together represent the complete object is not new (e.g. [1]), but existing implementations have lacked sufficient accuracy and reliability until recently.

In 2D object detection and recognition local object feature detectors must perform reliably in a rotation, scale, and translation invariant manner. For real applications they should also exhibit sufficient robustness against noise and distortions. The problem of extracting local descriptors can be divided into two categories: 1) unsupervised and 2) supervised. The unsupervised approach is more challenging since it must first solve a more general problem of what is really "important" in images - the question which intrigues brain and cognitive science researchers as well. In the literature, several unsupervised descriptors have been proposed, e.g., combined corner and edge detectors by Harris and Stephens [2], but only very recently more representative and theoretically sound methods such as salient scale descriptors by Kadir [3] and SIFT (scale invariant feature transform) features by Lowe [4] have been introduced. The major advantage of unsupervised local descriptors is the unsupervised nature itself and the

* Academy of Finland (#204708) and EU (#70056/04) are acknowledged for support.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 119–129, 2005.

main disadvantage is the disability to exclusively label the findings; an object is described by a spatially connected distribution of non-unique labels. However, unsupervised descriptors may provide information about position, scale, and rotation, and thus, object detection can be based on an inspection of both the configuration and the properties of extracted evidence making the detection more efficient (see, e.g., [5]).

Unsupervised descriptors have recently been a more popular topic, but this study promotes the supervised approach. It seems improbable that either of the two approaches would have an overall superiority since they possess distinct advantages and disadvantages and enable different approaches in upper processing layers. Supervised detection of local descriptors is based on a detection scheme where important image sub-parts (evidence), are known in advance, and thus, detectors can be optimized. It is clear that since a supervised detector is more specific it can be made more reliable and accurate, but a new problem is how to select which image parts to use. The supervised descriptor detection (evidence extraction) is a similar problem to object detection itself, but an explicit assumption is made that local image patches are less complex than a complete object. Consequently simpler feature detection methods can be applied. Furthermore, since supervised descriptors are more reliable and accurate than unsupervised, simpler spatial models can be used to detect objects - a single detected evidence creates already a hypothesis that an object is situated in that location (see, e.g., [6]). Respectively, in the unsupervised descriptors based detection the number of descriptors required is often large. Several occurrences of descriptors in the vicinity of a correct spatial configuration compensates the low reliability of detecting a single descriptor. The selection of image sub-parts in the supervised detection is an application specific task, but it can also be automated if evidence items which are most typical for specific objects are selected; the theory of unsupervised detection can be utilized.

In this study a novel supervised evidence extraction method is introduced. The method is based on simple Gabor features introduced by the authors [7] and statistical ranking using Gaussian mixture model probability densities proposed by the authors in [8]. The method has been successfully applied in face localization [6]. This study describes the approach in more detail, introduces accompanying theory and algorithms and presents the latest experimental results.

2 Simple Gabor Features

The simple Gabor feature space and its properties have been introduced in [7]. Here the properties are explained more carefully in order to demonstrate the practical use.

2.1 Structure of Simple Gabor Feature Space

The phrase “simple” in the context of simple Gabor feature space refers to a fact that the feature space considers phenomena, here evidence, at a single spatial location. A single spatial location does not straightforwardly correspond to a single pixel in digital images since effective area, envelope, of Gabor filter stretches over a substantially larger area; yet the reconstruction accuracy is highest near the centroid. It is clear that complex objects cannot be represented by a simple Gabor feature which is concentrated near a single location but a spatial (constellation) model must be built upon the features and combine them (see, e.g., [6]).

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

$$r_\xi(x, y; f, \theta) = \psi(x, y; f, \theta) * \xi(x, y) = \iint_{-\infty}^{\infty} \psi(x - x_\tau, y - y_\tau; f, \theta)\xi(x_\tau, y_\tau)dx_\tau dy_\tau \quad (1)$$

The response is calculated for several frequencies f_k and orientations θ_l.

The frequency corresponds to scale which is not an isotropic variable, the spacing of frequencies must be exponential [7]

$$f_k = c^{-k} f_{max}, \;\; k = \{0, \ldots, m-1\} \quad (2)$$

where f_k is the kth frequency, $f_0 = f_{max}$ is the highest frequency desired, and c is the frequency scaling factor ($c > 1$).

The rotation operation is isotropic, and thus, it is necessary to position filters in different orientations uniformly [7]

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

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 on angles $[\pi, 2\pi[$ are complex conjugates of responses on $[0, \pi[$ for real valued signals.

Feature Matrix. The Gabor filter responses can be now arranged into a matrix form as

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

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

2.2 Feature Matrix 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). It is thus possible to represent and consequently also recognize evidence using the Gabor feature matrix.

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

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

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

where * denotes complex conjugate.

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

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

It should be noted that responses on new low frequencies f_m must be computed and stored in advance while the highest frequency responses on f_0 vanish in the shift.

3 Statistical Classification and Ranking of Features

In general, any classifier or pattern recognition method can be used to train and to classify features into evidence classes. However, certain advantages advocate the use of statistical methods. Most importantly, not only class labels for observed features are desired but also it should be possible to rank evidence items in a scene and to sort them in the best matching order for returning only a fixed number of the best candidates. The ranking reduces search space of a spatial model (e.g., [9]), and furthermore, rank values can be integrated into a statistical spatial model as well. Ranking requires a measure for confidence, that is, a quantitative measure which represents the reliability of classification into a certain class. It is possible to introduce ad hoc confidence measures for the most classifiers, but statistical measures, such the value of the class-conditional probability density function (pdf) are more sound [8].

In order to apply statistical classification and ranking it is necessary to estimate class conditional pdf's for every evidence. Since Gabor filters are Gaussian shaped in both spatial and frequency domains they typically enforce observations into a form of Gaussian distribution in the feature space [10]. However, a single Gaussian cannot represent class categories, such as eyes, since they

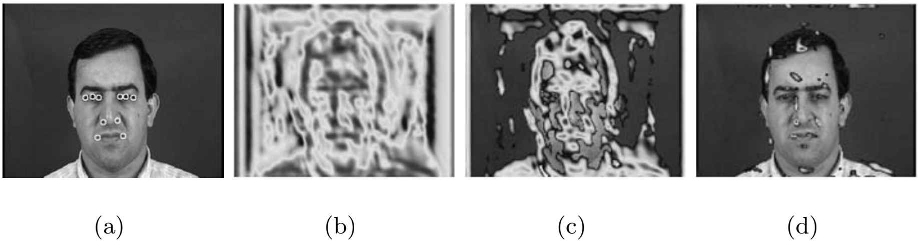


Fig. 1. Example of using density quantile and pdf values as confidence : (a) Face image and 10 evidence classes; (b) Pdf surface for the left nostril (left in image); (c) Pdf values belonging to 0.5 density quantile; (d) Pdf values belonging to 0.05 density quantile

may contain inherited sub-classes, such as closed eye, open eye, Caucasian eye, Asian eye, eye with eye glasses, and so on. Inside a category there are instances from several sub-classes which can be distinct in the feature space. In this sense Gaussian mixture model is a more effective principal distribution to represent the statistical behavior of simple Gabor features.

There are several methods to estimate parameters of Gaussian mixture models (GMM's) and for example the unsupervised method by Figueiredo and Jain [11] seems to be an accurate and efficient method [8]. The Figueiredo-Jain algorithm is unsupervised in the sense that it automatically estimates the number of components in a GMM. The original method can be extended to complex vectors constructed from the Gabor feature matrices in (4) as [8]

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

Using estimated pdfs it is possible to assign a class for features extracted at any location of an image by simply applying the Bayes decision making. However, as posteriors do not act as inter-class measures but as between-class measures for a single observation, class-conditional probability (likelihood) is a prefered choice to act as a ranking confidence score [8]. It is a measure of how reliable the class assignment of the evidence is. Now, evidence with the highest confidence can be delivered for consistency analysis first. The use of confidence values may reduce search space by an arbitrary degree by discarding evidence beyond a requested density quantile [8]. In Fig. 1 the use of density quantile for reducing the search space is demonstrated; it is clear that the correct evidence is already within 0.05 (0.95 confidence) density quantile.

4 Evidence Extraction

By combining simple Gabor features in Section 2 and statistical classification and ranking in Section 3 a novel evidence extraction method can be devised. Next, Algorithms 1 and 2, one for estimating evidence specific pdfs using the training set images and the other for extracting evidence, are introduced on a general level and discussed in detail.

Algorithm 1 *Train evidence classifier*

```
for all Training images do
    Align and normalize image to represent an object in a standard pose
    Extract simple Gabor features at evidence locations
    Normalize simple Gabor features
    Store evidence features $P$ and their labels $T$
end for
Estimate GMM pdf for each evidence with data in $P$
```

In Algorithm 1 the fundamental steps to generate a pdf-based classifier for evidence extraction are shown. First, training images must be aligned to a standard pose. The standard pose corresponds to a pose where objects have roughly the same scale and orientation. In the supervised evidence extraction the normalization and aligning is possible since keypoint locations are known. In the standard pose, simple Gabor features in (4) are then computed at evidence locations. Feature matrices can be energy-normalized if a complete illumination invariance is required. Each feature matrix is reformatted into a vector form in (7) and stored in a sample matrix P along with corresponding labels, T. Finally, complex pdfs are estimated for each evidence separately, e.g., utilizing GMM and the FJ algorithm.

Algorithm 2 *Extract K best evidences of each type from an image I*

```
Normalize image
Extract simple Gabor features $G(x, y; f_m, \theta_n)$ from image $I(x, y)$
for all Scale shifts do
    for all Rotation shifts do
        Shift Gabor features
        Normalize Gabor features
        Calculate confidence values for all classes and for all $(x, y)$
        Update evidence confidence at each location
    end for
end for
Sort evidences for each class
Return $K$ best evidences for every evidence class
```

In Algorithm 2 the main steps to extract evidence from an image are shown. First, the image is normalized, that is, scale and grey levels are adjusted to correspond to average object presence used in the training. From a normalized image simple Gabor features are extracted at every spatial location and confidence values are computed for all requested invariance shifts. If features were energy normalized in the training phase the same normalization must be applied before calculating confidence values from GMM pdfs. In a less memory requiring implementation, confidence values can be iteratively updated after each shift in order to store only the best candidates of each evidence at each location. After the shifts have been inspected it is straightforward to sort them and return the best candidates. In this approach one location may represent more than one evidence, but each evidence can be in one pose only.

5 Experiments

In this section we present the results of an application of the algorithm to a practical problem of detecting facial evidence in images from XM2VTS database.

5.1 XM2VTS Database

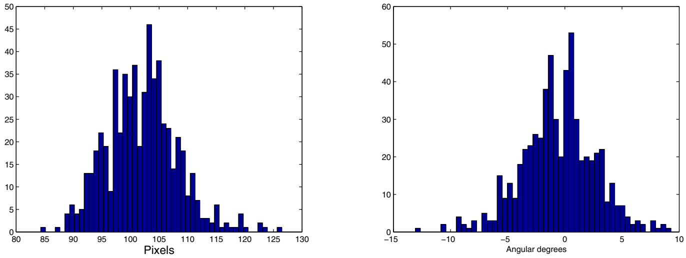
XM2VTS facial image database is a publicly available database for benchmarking face detection and recognition methods [12]. The frontal part of the database contains 600 training images and 560 test images of size 720 × 576 (width × height) pixels. Images are of excellent quality and any face detection method should perform well with the database.

To train the evidence detectors a set of salient face regions must be selected first. The regions should be stable over all objects from the same category, but also discriminative comparing to other object regions and backgrounds. For facial images ten specific regions (see Fig. 3(a)) have been shown to contain favourable properties to act as evidence [9].

Selecting Simple Gabor Feature Parameters. The first problem in the parameter selection is the number of frequencies, m, and orientations, n, to be used in feature matrix in (4). Many factors contribute to the final performance, but generally the more frequencies and orientations are used, the better is the representation power of the simple Gabor feature. By increasing the numbers, shift sensitivity increases as well, allowing a more accurate determination of evidence pose. Generally, sharpness values of the filter, which also affect to the representation power, 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 later evaluated experimentally.

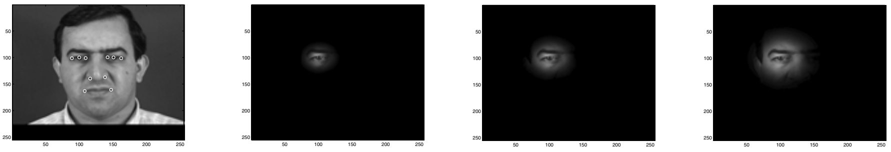
Using only 4 orientations affects the angular discrimination to be 45°, which is much broader than the rotations in the XM2VTS training set (Fig. 2(b)). The selection of frequencies is a more vital question. First of all in Fig. 2(a) it can be seen that in the XM2VTS database the mean distance between eyes is 102 pixels and the distribution is approximately normal. Thus, for optimal accuracy, training images should be normalized to the eye center distance of 102 pixels. Alternatively for recognizing also the smallest faces the training distance can be normalized to 84 pixel distance and the frequency factor c set to $\frac{102}{84} \approx 1.2$ in order to have exactly the mean value for the first scale shift. Second, shift would correspond to the eye distance 122 which is near the maximal value of eye center distances (126) and now the whole interval is covered. The interval can be sub-divided further, but this increases the computational complexity and does not infinitely increase the accuracy due to the scale sensitivity.

Setting the frequency factor to 1.2 would be optimal, but it would be a very small value causing a significant overlap of Gabor filters. The amount of overlap can be controlled by adjusting the filter sharpness, γ and η, but still, the smaller the frequency factor is, the more frequencies are needed to cover a broad



(a) (b)

Fig. 2. Scale and orientation contents of XM2VTS training data computed using coordinates of left and right eye centers: a) Distribution of eye center distances (min. 84 pix, max. 126 pix, mean 102 pix); b) Distribution of eye center rotation angles (abs. min. 0°, abs. max. 13.0°, abs. mean 2.5°, mean −0.5°)

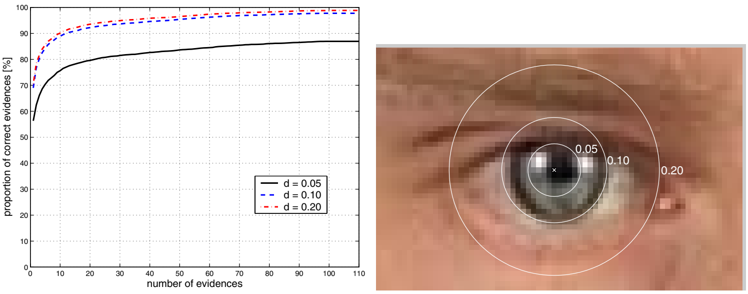


(a) (b) (c) (d)

Fig. 3. Normalized facial image and effective areas of Gabor filters on different frequencies: (a) 10 salient evidences (left and right outer eye corners, left and right inner eye corners, left and right eye centers, left and right nostrils, and left and right mouth corners); (b) $f_0 = \frac{1}{1 \cdot 15}$, (c) $f_1 = \frac{1}{\sqrt{2} \cdot 15}$, (d) $f_2 = \frac{1}{2 \cdot 15}$

frequency range and to represent objects accurately. In the case of XM2VTS database the whole scale variation can be covered without any scale shifts and by just selecting filters that can efficiently represent the various evidence. Thus, the frequency factor c was set to $\sqrt{2}$. In Fig. 3 an example of aligned image for extracting Gabor features is shown. The distance of the eye centers is normalized to 51 which is half of the mean value, and thus, test images can be processed in a half scale for faster computation. Furthermore, the angle between the eye centers is rotated to 0°, which roughly corresponds to the expectation. Images are cropped to the size of 256×256. In Fig. 3 effective areas of selected filters are also shown and it can be seen that they extract information on several different scales providing distinct information. With the given heuristics it can be assumed that the represented parameter values could perform well for the XM2VTS.

Furthermore, it seems that the simple Gabor features form smooth probability distributions for facial evidences, and thus, the methods for estimating



(a) (b)

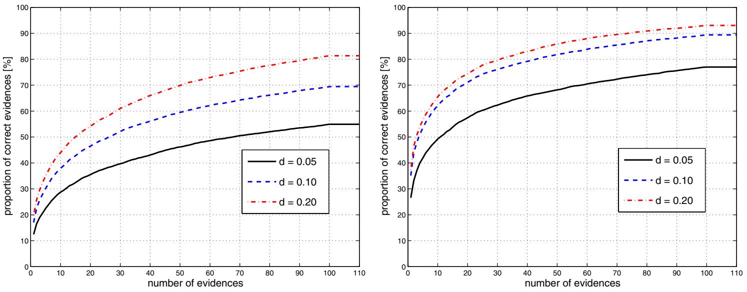
Fig. 4. Results for evidence extraction from XM2VTS test images: (a) Accuracy; (b) Demonstration of accuracy distance measure

parameters of pdf's perform accurately and robustly converging to the same estimates repeatedly with random initializations.

Results for Original Images. Evidence items were extracted in a ranked order and an evidence item was considered to be correctly extracted if it was within a pre-set distance limit from a correct location. In Fig. 4(a) are shown the accuracies for three different distance limits. The distances are scale normalized, so that the distance between the centers of the eyes is 1.0 (see Fig. 4(b)). From the figure it can be seen that all evidence cannot be extracted within the distance of 0.05, but on average 8 items of correct evidence are already included in the first ten items of evidence (one from each class) and by increasing the number to 100, only a small improvement can be achieved. However, within the distance 0.10 nine items of correct evidence were included already in the first ten items of evidence from each class and by extracting 100 items of evidence almost perfect detection rate was achieved. It should be noted that for constellation model methods it is possible to handle several thousands items of evidence (e.g. [9]).

Results for 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. 2), and thus, invariance of evidence extraction cannot be reliably verified. In the second experiment the same images were used, but they were artificially scaled by a uniform random factor between $[1, \sqrt{2}]$, which corresponds to the scale factor c, and rotated by $[-45°, 45°]$ where 45° corresponds to the angle between two adjacent filters. In Fig. 5 the results for an experiment where no invariance shifts were applied and for another experiment where shifts were applied are shown. It is clear that the shifts provided more invariance for the extraction since at the error $d = 0.05$ the accuracy increased from 45% to almost 70% when the total of 50 items of evidence were fetched.

A significant increase in the accuracy was achieved by adding only single shifts of features, but it is not necessary to tie shifts to the configuration of simple



(a) (b)

Fig. 5. Results for evidence extraction from artificially rotated and scaled XM2VTS test images: (a) No shifts; (b) $\{0, 1\}$ scale shifts and $\{-1, 0, 1\}$ rotation shifts applied

Gabor features in the training. In the extraction, the spacing can be tighter, e.g., orientations by $45^\circ/2 = 22.5^\circ\%$ and scales by $\sqrt{\sqrt{2}} = \sqrt[4]{2}$ to establish a double density. With the double density only every second feature in the feature matrix is used in the classification, but the invariance is further increased.

6 Conclusions

In this study, evidence based object detection was studied. We have argued that it is an accurate and reusable approach to general object detection. In the pursuance of this approach, a complete method and algorithms for invariant evidence extraction have been proposed. The proposed method is supervised by its nature and is based on simple Gabor features and statistical ranking. The analytical results were verified by experiments using real data of facial images. The method has been proved to be sufficiently accurate and reliable in practice and the future research will focus on developing a spatial model which can optimally utilize the provided evidence.

References

1. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Trans. on Computers **22** (1973) 67–92
2. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of the Fourth Alvey Vision Conf. (1988) 147–151
3. Kadir, T.: Scale, Saliency and Scene Description. PhD thesis, Oxford University (2002)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision **60** (2004) 91–110
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. (2003)

6. Hamouz, M., Kittler, J., Kamarainen, J.K., Paalanen, P., Kälviäinen, H.: Affine-invariant face detection and localization using GMM-based feature detector and enhanced appearance model. In: Proc. of the 6th Int. Conference on Automatic Face and Gesture Recognition, Seoul, Korea (2004) 67–72
7. Kyrki, V., Kamarainen, J.K., Kälviäinen, H.: Simple Gabor feature space for invariant object recognition. Pattern Recognition Letters **25** (2003) 311–318
8. Paalanen, P., Kamarainen, J.K., Ilonen, J., Kälviäinen, H.: Feature representation and discrimination based on Gaussian mixture model probability densities – practices and algorithms. Research report 95, Department of Information Technology, Lappeenranta University of Technology (2005)
9. Hamouz, M.: Feature-based affine-invariant detection and localization of faces. PhD thesis, University of Surrey (2004)
10. Kämäräinen, J.K.: Feature Extraction Using Gabor Filters. PhD thesis, Lappeenranta University of Technology (2003)
11. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 381–396
12. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended M2VTS Database. In: Proc. of the 2nd Int. Conf. on Audio and Video-based Biometric Person Authentication. (1999) 72–77