Object Localisation Using Generative Probability Model for Spatial Constellation and Local Image Features

J.-K. Kamarainen[1,2], M. Hamouz[1], J. Kittler[1], P. Paalanen[2], J. Ilonen[2], A. Drobchenko[2]
[1]Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK
[2]Machine Vision and Pattern Recognition Research Group
Lappeenranta University of Technology, Finland

Abstract

In this paper we apply state-of-the-art approach to object detection and localisation by incorporating local descriptors and their spatial configuration into a generative probability model. In contrast to the recent semi-supervised methods we do not utilise interest point detectors, but apply a supervised approach where local image features (landmarks) are annotated in a training set and therefore their appearance and spatial variation can be learnt. Our method enables working in purely probabilistic search spaces providing a MAP estimate of object location, and in contrast to the recent methods, no background class needs to be formed. Using the training set we can estimate pdfs for both spatial constellation and local feature appearance. By applying an inference bias that the largest pdf mode has probability one, we are able to combine prior information (spatial configuration of the features) and observations (image feature appearance) into posterior distribution which can be generatively sampled, e.g. using MCMC techniques. The MCMC methods are sensitive to initialisation, but as a solution, we also propose a very efficient and accurate RANSAC-based method for finding good initial hypotheses of object poses. The complete method can robustly and accurately detect and localise objects under any homography.

1. Introduction

Object and object category models based on local image features and their spatial constellation have been the main topic in many recent studies now considered as state-of-the-art in object detection, localisation and recognition [1, 17, 5, 6, 14, 22, 9]. Excluding [9] these methods first apply interest point detection to pinpoint local salient regions in image and then compute local image features by representing appearance of the interest point neighbourhood. Variation in spatial constellation of the local image features is captured by learning in a training phase or by simply computing frequency histograms of the features or their superclasses ("bag-of-features" models [22]). As a distinct advantage over other methods, the interest point driven approach enables semi-supervised learning where training images only need to be labelled in terms of their corresponding categories or in some methods roughly aligned to a same pose and coarsely segmented. However, no explicit selection and manual annotation of object landmarks are needed. The problem in this semi-supervised setting is that the training of the full model including the local features and spatial configuration becomes extremely complex and computationally intensive and typically only limited invariance can be realised (e.g. [1, 6]). Alternatively, the bag-of-feature based methods are more limited by their discriminative power [22].

In this paper we also utilise the feature-based approach which combines local image features and spatial constellation model. We adopt the generative model approach, which seems to have very advantageous properties for object detection and localisation [6], and which, when devised in a proper probabilistic framework, can take full advantage of the Bayesian data analysis. The main difference of our method and the above mentioned state-of-the-art methods is that we abandon the interest point detection stage, but use traditional supervised approach which requires an annotated training set. Our method for object detection, however, is not less general, but we consider the interest region selection as a separate problem to the one of recognition. By adopting the supervised training, we use an image feature localisation algorithm, based on multi-resolution Gabor features and their multiple hypothesis testing, which has been in our former studies found as distinctly robust and accurate in the presence of significant local appearance variation [12]. In addition to the feature localisation, the extraction method is able to rank the features based on their probability score (confidence). Probability scores also act as observations in the generative probability model. Furthermore, the training set and our probability score framework can be used to construct a probabilistic model for spatial configuration variation under any homography, i.e. the model copes with geometric distortions. In the generative model, the spatial constellation probability forms prior distribution and the local feature probabilities the observation information. Finally, in this fully probabilistic setting we are able to sample the posterior distribution with Markov chain Monte-Carlo (MCMC) methods and provide maximum a posterior (MAP) estimate of object location. Since the MCMC methods are very initialisation dependent, we further propose efficient and robust initial estimation method based on the random sample consensus (RANSAC) principle.

The main contributions of this study are: 1) a new supervised-manner-learnt generative probability model of object appearance, based on local image features (observations) and spatial constellation (prior), 2) a RANSAC-based method for initiating object hypotheses based on point patterns of best local image features, 3) MCMC sampling scheme converging to samples from the posterior distribution and providing a MAP hypothesis of object presence, and 4) exact localisation of an object in the image.

2. Feature-based object (category) model

Feature-based methods consider objects as instances from a certain object class where appearance may significantly vary, but all instances, in principle, contain a similar structure (same parts in a similar spatial configuration). An example from such category are human faces which vary in local part appearance and their relationships (Fig. 1).

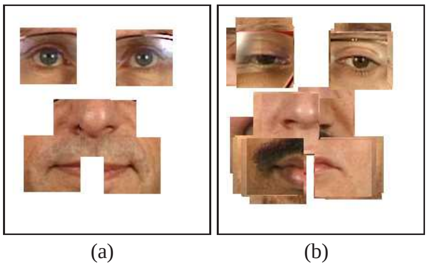


Figure 1. Objects from the category “human face”. Objects vary in local appearance (local image feature) and their spatial configuration (spatial constellation). Shown objects are transformed to isometry canonical space (rotation and translation removed): (a) 6 local image parts of a single object; (b) parts from several objects.

According to the Bayesian point of view, all observed and unobserved quantities are considered as random variables [8]. In the proposed generative probability model the local image features are direct observations of random variables and the spatial constellation model represents another indirectly observable (unobservable) random variable. These two parts will be described next.

2.1. Local image features – observations

An approach in accordance to state-of-the-art would first run an interest point detector to an input image, then remove geometric distortions and form local descriptors, and finally, assign automatically learnt class labels to the each location. The approach produces a large amount of equal local image features. Therefore, inference for object detection and localisation must strongly rely on the spatial constellation model. Our alternative and more traditional approach uses annotated local image landmarks. This approach does not require the interest point detection, but the local image patches are learnt and an invariant search over arbitrary poses is established. The method searches only the requested features and a statistical classifier can provide confidence measures to rank the features. To be as efficient as possible we should select accurate and efficiently computable local feature detectors.

In our earlier studies, we have introduced a local image feature extraction method based on multi-resolution Gabor responses (see [13]), $\boldsymbol{g}$, and efficient invariant (illumination, rotation, translation, scale) search by simple matrix shifts [12]. In our method, the classification and feature ranking are based on automatically estimated class specific probability density functions (pdfs), $p(\boldsymbol{g}|F_i)$, for each feature class F_i [18].

The likelihood values $p(\boldsymbol{g}|F_i)$ can be used to rank the features and return a fixed number for each class (see Fig. 2(a)). In addition, we have proposed a method for converting likelihood values ($p \in [0, \infty[$) to class specific probability scores ($Pr_{conf} \in [0, 1]$). The probability score can be formed by applying an inference bias that probability is certain, $Pr_{conf}(\boldsymbol{g}|F_i) = 1.0$, at the highest mode of the pdf. The probability score for some point is one minus density integral over the points with likelihoods greater or equal to the likelihood at the given point [18]. The score, also called as confidence, may be used to allow only features above preset confidence level to be further processed (see Fig. 2(b)), and the score values can also be used as probabilities. In our system purely probabilistic observation information is formed from an input image I by first computing the multi-resolution Gabor features, $\boldsymbol{g}(x, y)$, and then, computing likelihoods or probability scores for all feature classes. The probabilistic observation information for an object (category) C represented by $i = 0, \ldots, N-1$ local features is (utilising likelihoods)

$$\boldsymbol{T} \propto \prod_i p(\boldsymbol{g}|F_i) \ . \tag{1}$$

The model in (1) assumes that the local appearance of parts are independent which is actually over-generalisation and not a limiting factor for object localisation (see Fig. 1(b)).

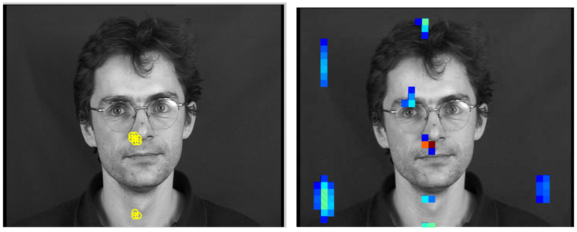
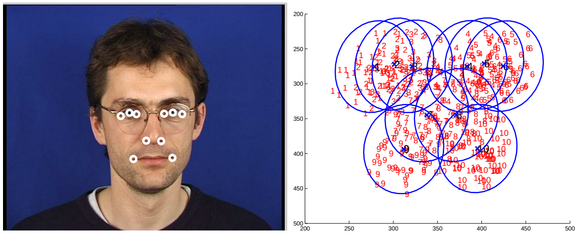

(a) (b)

Figure 2. Extracted image features and their observation likelihoods: (a) left nostril (10 best marked as yellow circles); (b) pdf values within 0.95 confidence, i.e. probability score ≥ 0.95 (red denotes high, yellow moderate and blue low likelihood).

2.2. Constellation of point patterns - prior model

Probabilistic spatial constellation model can be seen as “a prior” model of a part configuration for an object category. For any given configuration the model provides a probability value that the configuration represents an object, i.e. the model describes part locations in terms of probability. The configuration information cannot be directly observed. In our case the parts are modelled by points of image features (landmarks) and the configuration is thus a point pattern of the labelled points. An example of point patterns generated by 10 facial landmarks is shown in Fig. 3.

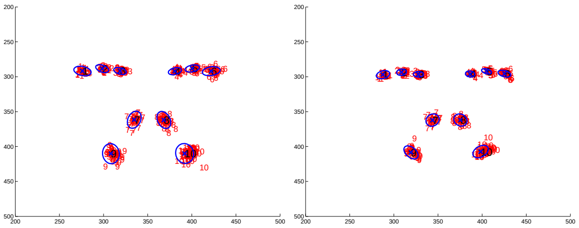


(a) (b)

Figure 3. Point patterns of 10 facial landmarks: (a) 10 annotated image features; (b) 50 randomly selected faces without alignment (red numbers), mean model (black numbers) and 2 std. deviation curves of Gaussian pdfs. Avg. error to the mean model is 349 over 600 training images from XM2VTS/frontal database ([16]).

To construct the probability model we adopt a similar point distribution model as introduced by Cootes et al. [4]. In their model points are first aligned and then variation in the aligned space is represented with respect to a “mean shape”. We adopt the mean shape algorithm from [4] with exceptions that patterns are iteratively (not batch training) transformed to the mean shape space which is iteratively updated, and in addition, the alignment is not restricted to approximate similarity, but can be any 2-D homography. In this study we do not approximate homographies using any singular value decomposition based methods [10] or iterative methods [4], but provide the results under exact homographies by the exact formula. The non-exact estimations could improve the results, but we wanted to study the performance under different exact homographies. The main difference between our approach and Cootes et al. [4] is that we do not represent the part appearance variation using the PCA components, but directly estimate the class conditional probability density functions using a Gaussian mixture model which can cope with arbitrary variability, in contrast to the PCA.

For local image features $F_0, F_1, \ldots, F_{N-1}$ of an object (category) C and corresponding feature coordinates $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$ the probabilistic model is the probability density function $p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}|C)$, that is, the probability of having a spatial configuration defined by the feature coordinates for an object from the category C. This probability density function does not use any observable information and is therefore prior model. However, for any configurations it can be used to select the best one. To compute likelihood, a point set is first transformed to the mean model space. The canonical mean space and the transformation can be generated under any selected homography, e.g. isometry (translation and rotation), similarity (+isotropic scale) or affinity (+unisotropic scale). Spatial models for real objects captured using real cameras would need a 3-D to 2-D model via projective transformation, but for simplicity we restrict ourselves to the 2-D transformations. 2-D transformations have been successful in many computer vision applications since affine transformation approximates the projective transformation in simplified settings. In addition, our experimental results show that in applications over-restricted homography can be a regulating factor. Estimated pdfs under similarity and affinity for XM2VTS/frontal are shown in Fig. 4 (compare to Fig. 3(b)).



(a) (b)

Figure 4. Estimated point pattern spatial constellation pdfs: (a) isometry (mean model error 50); (b) similarity (err. 36).

2.2.1 Formulation of prior model

The probability densities for each image feature are estimated in the aligned space, e.g. the 2-D Gaussian pdfs illustrated in Fig. 4. Assuming independence between the locations of N image features a joint pdf can be formed as

$$\boldsymbol{Y} \propto p(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}|C) = p(\boldsymbol{x}_0|C)p(\boldsymbol{x}_1|C)\cdots p(\boldsymbol{x}_{N-1}|C) \ . \tag{2}$$

The independence assumption in the mean space is fair since the mutual dependence is already embedded in the alignment procedure. The form in (2) does not however hold if all image features are not visible. Due to image feature detection failures or object occlusion all image features cannot be necessarily detected, and thus, a pdf covering also occlusions is needed. Instead of a complex formulation including all possible configurations and to avoid modelling the background class as done in [2, 6, 20] we again adopt the probability score framework. In this framework, where the major mode provides absolute certainty, samples near the mean model have a high confidence (≈ 1) and samples further away have a low confidence (≈ 0). Now, we may approximate the constellation pdf covering also occlusions using the simple sum rule

$$p(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}|C) \propto \frac{1}{N}\sum_{i=0}^{N-1} Pr_{conf}(\boldsymbol{x}_i|C) \ , \tag{3}$$

which assumes the same occlusion probability for all features. This assumption holds despite varying sizes of local features if they can be partially detected. The form in (3) cannot be derived from (2) without applying the inference bias and using the probability scores instead of the true distribution values. The pdfs for each feature can be estimated using a Gaussian mixture model (GMM) to support multimodal distributions, but in our simplified case we use a single Gaussian.

3. Hypothesis initialisation using variant of random sample consensus

In this section we describe how the probabilistic point pattern model in Sec. 2.2 and extracted image features in Sec. 2.1 are used to efficiently find a set of object hypotheses. Input for the initialisation method are N_{bestF} best image features of each type (Fig. 2) and the constellation pdf in (3) estimated under selected homography (Fig. 4).

Our method operates in the spirit of random sample consensus (RANSAC) [7], but modifies it to be more appropriate for our task. The three parameters of the standard RANSAC are the number of random iterations, the required minimum number of inliers and a distance threshold for inlier detection. In our case, the input image may contain several objects, and therefore, we cannot define any minimum required set of inliers but instead want to return all good hypotheses occurred during a certain number of iterations - an image may for example contain several objects giving rise to a multimodal posterior distribution. The number of iterations affects the efficiency, but since this part of the method is not a bottleneck we can safely fix the number of iterations. Depending on the problem characteristics, the number of iterations could be optimised (e.g. [15]), but this applies only to problems where a single solution exists. However, we want to point out that the confidence scores could be used to improve the sampling procedure similarly to [3]. The most important difference between our method and the standard RANSAC is the fact that we do not need to distinguish between inliers and outliers and thus define a threshold. Our method operates completely in the probability space defined by the constellation pdf in (3) and what is considered as an outlier only contributes insignificantly to the probability score. The method is sketched in Algorithm 1.

Algorithm 1 *Find initial object hypotheses*

```
Initialise the set of $N_{bestH}$ best hypotheses to null and set score values to zero
for Maximum number of iterations do
    Randomly select the minimum number (dep. on homography) of image features from the $N_{bestF}$ best features
    Estimate the homography H from the image space to the mean shape space
    Transform all image features to the mean space using $H$
    For each transformed feature compute the probability score
    For each feature class select the corresponding feature with the highest score
    Compute the total constellation score as a sum of the highest scores (Eq. 3)
    if the total score is higher than the worst in the set of best hypotheses then
        Estimate the inverse homography $H_I$ from the mean shape space to the image space using the selected features whose score is $\geq P_{featconf}$
        Transform the mean shape to the image space using $H_I$
        Add hypothesis to the best hypothesis set and remove the current worst
    end if
end for
Return $N_{bestH}$ best hypotheses
```

The provided algorithm is very straightforward and its only parameters are the maximum number of iterations and $P_{featconf}$. The maximum number of iterations depends on the quality of features, but can be fixed for a specific application. The confidence level $P_{featconf}$ can also be fixed for example to 0.05 which allows any feature within the 0.95 probability mass area (two standard deviation for the normal distribution) be included into the estimation (inlier) set. The total score does not require any threshold since an outlier would only contribute negligibly. The usage of probability scores thus changes the traditional RANSAC to work in purely probabilistic manner. The progress of the method is demonstrated in Fig. 5 and the best results after 100 random samples are demonstrated in Fig. 6.

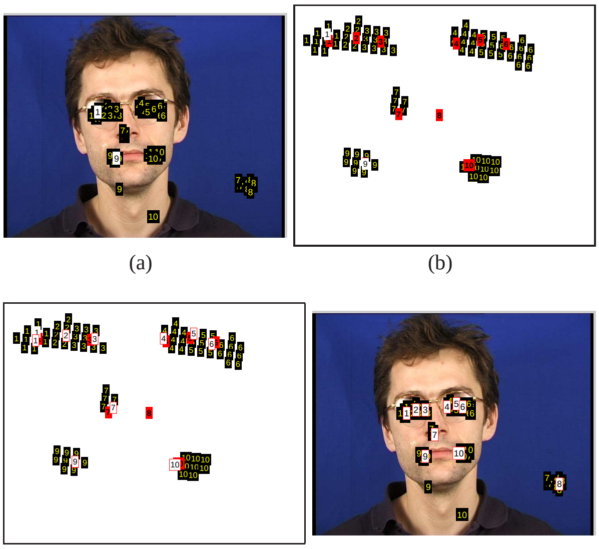

(a) (b) (c) (d)

Figure 5. Point pattern search of initial hypotheses: (a) 10 best features (yellow numbers on black) - 2 selected for homography (H) estimation; (b) all features transformed to the mean space (black numbers on red); (c) features with the highest probability score (white bg); (d) selected transformed back (H_I) to the image space.

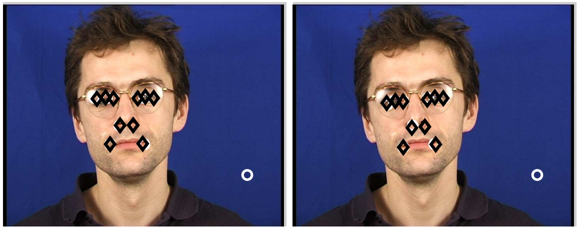


Figure 6. 2 best hypotheses after 100 random samples (best features denoted by white circles and hypotheses by black diamonds).

4. Refining hypotheses using generative probability model and MCMC sampling

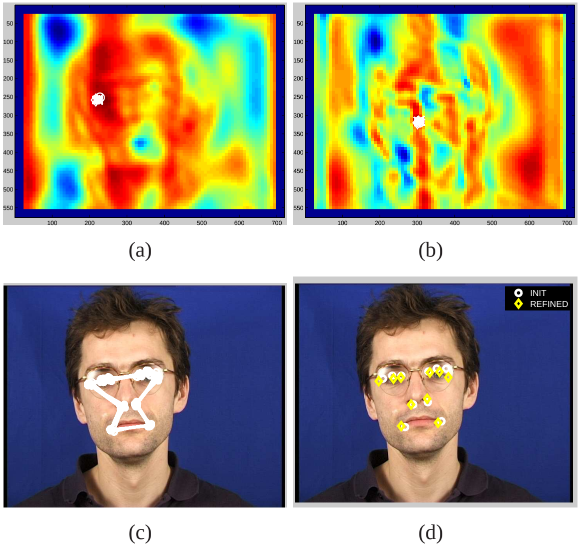
The initialisation method proposed in the previous section provides a requested number of hypotheses H_0, H_1, …, $H_{N_{bestH}}$, where $H_k = \bar{\boldsymbol{x}}^k = (\boldsymbol{x}_0^k, \boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_{N-1}^k)$, i.e. hypotheses are represented by coordinate tuples for N image features of the object class C. A true Bayesian estimate should be influenced by observations and the prior model (configuration), but it should be noted that observations were already implicitly used in the initialisation method since it utilised N_{bestF} best features selected by their observation likelihoods. However, to complete the work, we form a posterior distribution for the parameters of object presence in an observed image. Our tools are the constellation (prior) model in (2) and the probabilistic observation model in (1). The posterior distribution can be sampled, for example, using the standard Metropolis algorithm [8], while the seeds for the method are the initial hypotheses H_i.

A similar Bayesian data analysis based methods to object detection and localisation have been proposed by Sullivan et al. [20] and Tamminen et al. [21]. The first is based on a very simple observation model and needs to estimate a general background class, and therefore, cannot compete with our method in general. Tamminen et al. also utilise multi-resolution Gabor response vector, but form only an ad hoc descriptor [21]. Their descriptor cannot generalise well over complex landmarks occurring in real images. As a more powerful solution, our method described in Sec. 2.1 estimates the complete pdf. In addition, they utilise a complex ad hoc parametrised and tuned configuration model ($\boldsymbol{Y}$ in [21]) while ours is directly learnt from the training examples. Furthermore, the sequential Monte Carlo algorithm in [21] (particle filtering) is not anymore necessary due to our initialisation method which tolerates multimodal posterior distributions, and in addition, works under any homography while they are limited to translation. For compatibility we utilise the same terms, $\boldsymbol{T}$ and $\boldsymbol{Y}$, for the corresponding random variables.

Our model observes an image $\boldsymbol{I}$, which is processed by the feature extraction in Sec. 2.1 providing feature vectors, $\boldsymbol{g}$. The vectors are further turned to likelihoods of each feature class at every image pixel, $\boldsymbol{I} \mapsto Gabor \mapsto \boldsymbol{T}$. The likelihood is actually $\boldsymbol{T}_i^C$, with the index i denoting the feature type i and C the object category. Any complete set of the features forms a spatial constellation, for which likelihood $p(\boldsymbol{x}|Y)$ can be computed from (2). Finally, the complete posterior for the presence of an object from class C at location $\bar{\boldsymbol{x}}$ is

$$p(\bar{\boldsymbol{x}}|\boldsymbol{T},\boldsymbol{Y}) = p(\bar{\boldsymbol{x}}|\boldsymbol{T})p(\bar{\boldsymbol{x}}|\boldsymbol{Y}) = \underbrace{\prod_i T(\boldsymbol{x}_i)}_{\text{observation}}\underbrace{\prod_i p(\boldsymbol{x}_i|C)}_{\text{prior}} \quad . \tag{4}$$

It is noteworthy that our model in (4) does not need to sample over any hyperparameters, such as the observation model parameters $\boldsymbol{G}$ and prior model parameters ξ in [21]. These are automatically learnt and embedded in the estimated pdfs for the local image features, $p(\boldsymbol{g}|F_i)$, and the canonic spatial constellation model, $p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}|\boldsymbol{C})$. The only sampled dimensions are the coordinates of the local image features, $\bar{\boldsymbol{x}}$. The provided samples have the same form and dimensions as the initial hypotheses H_i, and therefore, are called as the *refined hypotheses* $\hat{H}_i$ derived using the *initial hypotheses* H_i. Metropolis sampling in the posterior space is illustrated in Fig. 7.



(a) (b)

(c) (d)

Figure 7. Hypothesis refinement using MCMC: 200 samples in the observation space - (a) left outer eye corner and (b) left nostril; (c) all samples in the image space; (d) initial and refined hypotheses.

5. Experiments

Experiments were conducted using 2 different publicly available face data sets, XM2VTS/frontal (600 training and 560 test images) [16] and XM2VTS/non-frontal (592 training and 588 test images). It should be noted, however, that our method is not face specific, but for any object category for which local landmarks can be defined.

The local image feature extraction and likelihood computation utilised in this study have been described in detail in our previous study [12], where localisation results superior to any comparable method were reported.

5.1. Finding initial hypotheses

5.1.1 XM2VTS/frontal

From the frontal section of the XM2VTS database, following Protocol I, 600 independent training and 560 test images can be selected (size 720 × 576 pixels). The images are of good quality and contain frontal faces in roughly the same imaging distance and captured on a constant background. Any face detection and recognition method should perform well for XM2VTS/frontal. The publicly available ground truth contains eye centres, but we added the remaining 8 image features by manually annotating them.

To measure and report object localisation accuracies, we adopted the cumulative localisation accuracy measure d_{eye} from [11], which has been confirmed to be a good measure of face localisation accuracy [19]. To compute d_{eye}, the maximum of estimated left and right eye centre errors is selected and normalised by the true eye distance. $d_{eye} \leq 0.25$ is considered as a successful face localisation [11], but for a successful face recognition a more accurate alignment is needed. Thus we report the accuracies on $d_{eye} = \{0.05, 0.10, 0.20\}$. d_{eye} measures accuracy much more objectively and accurately than any "bounding box" measure used in the literature. To similarly report accuracies over all landmarks we used d_{all}, which selects the worst over all 10 landmarks.

In Fig. 8 are shown the results for the proposed method under isometry, similarity and affinity. For example, under isometry the initial hypotheses failed for only 1 test image (failure rate 0.18%) shown in Fig. 9 ($d_{eye} > 0.25$ for all 10 best hypotheses). There were only negligible differences between the accuracies under different homographies which results from the fact that the faces were captured within the fixed imaging setting and the local feature detection succeeded almost perfectly. The failure rate was less than 1.0% for all homographies.

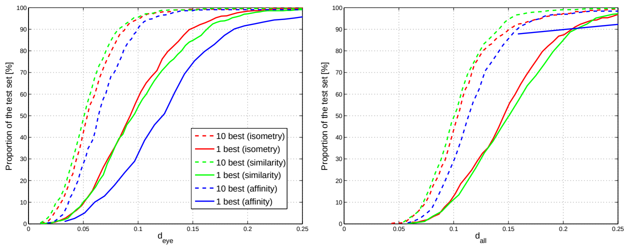


Figure 8. XM2VTS/frontal initial hypotheses accuracy (failure rates: iso. 0.18%, sim. 0.36% and aff. 0.89%) ($N_{bestH} = 10$, $N_{bestF} = 10$, $P_{featconf} = 0.05$, 200 iters): (a) d_{eye} ; (b) d_{all}.

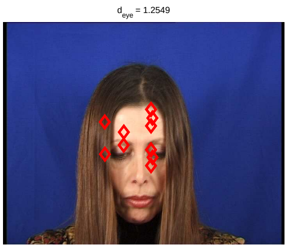
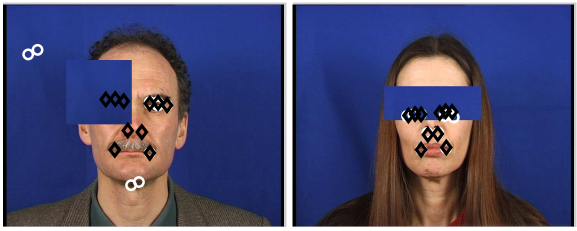


Figure 9. Failed localisation for XM2VTS/frontal (isometry).

To demonstrate the natural occlusion tolerance of the method results for two artificially occluded faces are shown in Fig. 10.

5.1.2 XM2VTS/non-frontal

The overall quality of XM2VTS/non-frontal (XM2VTS MPEG7) images is comparable to the XM2VTS/frontal, but clients were asked to watch to 4 different directions (see Fig. 11). The extreme poses in many images prevented all image features to be detectable.



(a) (b)

Figure 10. Successful localisation of occluded faces (similarity).

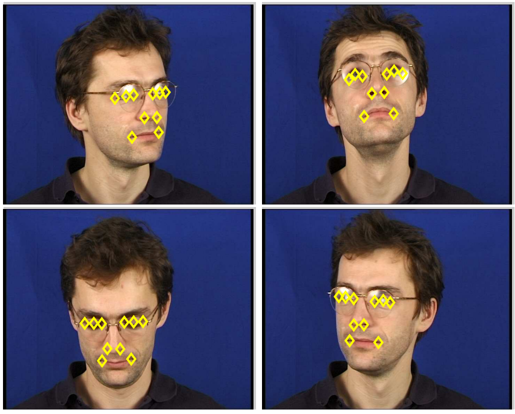


Figure 11. Poses for each XM2VTS/non-frontal face entry.

The results for the different homographies are shown in Fig. 12. For the non-frontal faces isometry and similarity performed almost equally well, but the affinity started to fail more severely. This was due to the fact that less image features were available for the estimation. Isometry and similarity need 2 features in the random sampling, and thus, require at least 3 features to reliably rank hypotheses. Affinity respectively requires 4 features, which was not satisfied for roughly 10% of images.

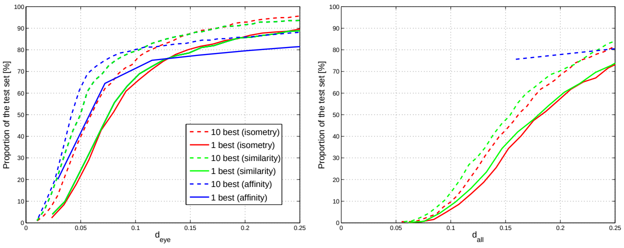


Figure 12. XM2VTS/non-frontal initial hypotheses accuracy (iso. 4.93%, sim. 6.46% and aff. 12.70%): (a) d_{eye}; (b) d_{all}.

5.2. Refining hypotheses

For demonstrating the accuracy of refined hypotheses using the MCMC sampling (the Metropolis method) we selected the initial hypotheses provided by the RANSAC initialisation under similarity since it performed well for the databases. First of all, it turned out that the accuracy of initial hypotheses was already sufficient for the most applications. The standard MCMC algorithms can improve the results only insignificantly within a reasonable amount of samples. However, in order to demonstrate the generative model sampling, we only aimed to improve the d_{eye} accuracy which is important for the face recognition algorithms.

The 10 best hypotheses were refined by drawing 100 samples from the posterior distributions and by selecting the ones with the highest posterior value. The results (d_{eye} only) are shown in Fig. 13. For the both data sets the accuracies of eye localisation were significantly improved.

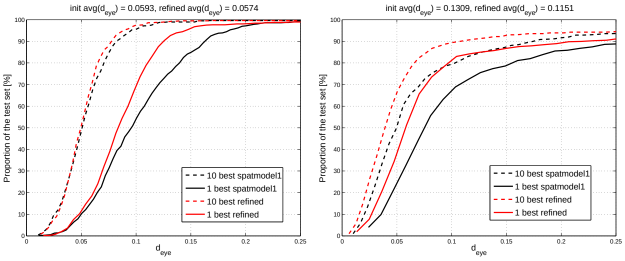


Figure 13. Initial hypotheses (similarity) refined using the MCMC sampling (eye centre observations only, 100 samples, Gaussian proposal distribution with variance 1.0, refined marked as red): (a) XM2VTS/frontal d_{eye}; (b) XM2VTS/non-frontal d_{eye}.

This part of the work revealed that the quality of the initial hypotheses were already very good and the refinement was hardly needed. Sampling posterior distributions is very computing intensive and cannot be used in real-time or near real-time applications until better and faster methods are applied. However, our initial method clearly provided very accurate high posterior value samples which were further refined with the MCMC sampling, which ultimately provided the true MAP estimates.

6. Conclusions and further work

We proposed a novel purely statistical method for detection and localisation of objects under any homography. The experimental results verified very accurate localisation performance as our general method was able to compete with state-of-the-art tailored face localisation methods. The method was devised under the Bayesian data analysis framework, and therefore, it does not only provide an object pose estimate, but samples from posterior distribution of object presence. It is justified to call the estimate as the maximum a posterior estimate for an object presence based on local image features and their spatial constellation.

The automatic method for learning pdfs of image features was introduced in our previous study [12] and the present one complements the research by introducing the statistical model for the spatial constellation. In the future, the proposed strong statistical framework will be further complemented with an automatic method to select the local image features, and then, the method can be used to automatically learn and recognise object categories using efficient fully probabilistic generative probability model.

Acknowledgements

This work was supported by EPSRC project “2D+3D=ID” (GR/S98528/01), with contributions from EU project BIOSECURE, and Academy of Finland (204708, 112631).

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in image via a sparse, part-based representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004. 1

[2] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. of the ECCV*, pages 628–641, 1998. 4

[3] O. Chum and J. Matas. Matching with PROSAC – progressive sample consensus. In *Proc. of the CVPR*, volume 2, pages 220–226, 2005. 4

[4] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1), 1995. 3

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the CVPR*, volume 2, pages 524–531, 2005. 1

[6] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *Int. J. of Computer Vision*, 71(3):273–303, 2007. 1, 4

[7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6), 1981. 4

[8] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. CRC Press, 2nd edition, 2004. 2, 5

[9] 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, 2005. 1

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge press, 2003. 3

[11] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the Hausdorff distance. In *Proc. of the AVBPA*, pages 90–95, 2001. 6

[12] 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 SCIA*, pages 119–129, 2005. 1, 2, 6, 7

[13] 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. 2

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of the CVPR*, volume 2, pages 2169–2178, 2006. 1

[15] J. Matas and O. Chum. Randomized RANSAC with sequential probability ratio test. In *Proc. of the ICCV*, volume 2, pages 1727–1732, 2005. 4

[16] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS Database. In *Proc. of the AVBPA*, pages 72–77, 1999. 3, 6

[17] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proc. of the CVPR*, pages 26–36, 2006. 1

[18] 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. 2

[19] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24:882–893, 2006. 6

[20] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Bayesian object localisation in images. *Int. J. Computer Vision*, 44(2):111–135, 2001. 4, 5

[21] T. Tamminen and J. Lampinen. Sequential Monte Carlo for Bayesian matching of objects with occlusions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(6):930–941, 2006. 5

[22] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. of Computer Vision*, 73(2):213–238, 2007. 1