Object Class Detection Using Local Image Features and Point Pattern Matching Constellation Search

Alexander Drobchenko[1], Jarmo Ilonen[1], Joni-Kristian Kamarainen[2], Albert Sadovnikov[1], Heikki Kälviäinen[1], and Miroslav Hamouz[2]

[1] Machine Vision and Pattern Recognition Research Group
Lappeenranta University of Technology
[2] Centre for Vision, Speech and Signal Processing
University of Surrey

Abstract. Several novel methods based on locally extracted image features and spatial constellation models have recently been introduced for invariant object class detection and recognition. The accuracy and reliability of the methods depend on the success of both tasks: image feature extraction and spatial constellation model search. In this study a novel method for object class detection is introduced. It combines supervised Gabor-based confidence-ranked image features and affine invariant point pattern matching. The method is able to deal with occlusions and its potential is demonstrated on a standard face database.

1 Introduction

Object class (category) recognition has recently become a popular research topic in computer vision. The popularity probably originates from the problem of face detection where the faces establish an object class. The traditional detection methods can be considered as image or window based approaches where a scene is exhaustively scanned with a window and delivered as an input to a classifier system (e.g., template matching or support vector machine). Lately, the image based approaches have faced competition in the form of feature based methods (e.g., [1,2,3]). These methods utilize locally detected features which along with their spatial configuration are combined using "a constellation model" to establish a complete representation of an object.

Feature based methods yield certain advantages over image based methods, but hitherto most of them have been based on simple key points (e.g., [4]). The advantages of the key points are their generality (shared by many classes) and their semi-supervised nature, that is, objects must be only segmented, named and aligned. The main disadvantages are their incapability to generalize over varying presences of the same feature (e.g., human eye) and to specialize for a specific object class. Since the advantage of semi-supervised training is quite artificial an alternative approach can be utilized by labeling image features and training them in a supervised manner. That would enable a more representative

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 273–282, 2007.

set of features allowing a computationally lighter realization of the spatial model (e.g., [1]); The detection load is shared by the both, the image feature detection and the constellation model.

In this study, we propose an object class detection and localisation method, which utilizes the supervised feature detection in [5] and an affine transformation based point pattern matching constellation model.

2 Related Research

Partitioning an object to more easily detectable local patches and combining the patches using a spatial constellation model is not a new approach but originally introduced by Fishler and Elschlager in 1973 [6]. Since then a well-known graph matching method utilizing a similar structure was proposed by Lades et al. [7], but it cannot be used as a general object class detection method since it requires a sufficient initial guess of the object pose. Modern and currently state-of-the-art spatial constellation models appeared recently along with efficient methods for key point detection, e.g., by Lowe [8,3] and Burl and Perona et al. [9,1].

Lowe uses an approach resembling Hough transform for object detection based on SIFT features [3,8]. SIFT features with similar scale, orientation and translation (relative to the model) are grouped in bins. Then bins are sorted according to the number of hits and each bin is verified using an approximated affine mapping of the model onto features in the bin. Outliers are determined by a threshold on the difference in scale, rotation and translation from the parameters obtained in the affine model mapping.

Simultaneously, a probabilistic constellation model was developed by the Perona's group. The core of the system is the estimation of how normal noise is distorted by affine and projective transforms. A breadth first search (reviewing most probable models first) is then used for locating the most likely affine correspondence of the model and extracted image features.

The main difference of these state-of-the-art methods is in their use of the unsupervised key points by Lowe [8] or Kadir [10]. In this study we propose to use supervised image features and a similar spatial search method as proposed by Hamouz et al. [2], with the difference that the spatial constellation model is replaced by a direct affine point pattern matching. The method provides the global optimum over the given image features and is capable to estimate locations of missing features.

3 Supervised Image Feature Extraction

The extraction method was introduced by the authors in [5] and is based on simple Gabor features [11] and feature ranking based on confidence information derived from Gaussian mixture model pdf's [12]. In the following section the method will be shortly reviewed.

3.1 Simple Gabor Features

The simple Gabor feature space and its properties have been originally introduced by the authors in [11]. The features are based on responses of complex Gabor filters on multiple scales and orientations, thus forming a multi-resolution Gabor frame structure.

Responses of Gabor filters, $\psi(x, y; f, \theta)$, over the whole image $\xi(x, y)$,

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

are calculated for several frequencies f_k and orientations θ_l and arranged into a matrix form as $\mathbf{G} =$

$$\begin{pmatrix} r(x_0,y_0;f_0,\theta_0) & r(x_0,y_0;f_0,\theta_1) & \cdots & r(x_0,y_0;f_0,\theta_{n-1}) \\ r(x_0,y_0;f_1,\theta_0) & r(x_0,y_0;f_1,\theta_1) & \cdots & r(x_0,y_0;f_1,\theta_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_0,y_0;f_{m-1},\theta_0) & r(x_0,y_0;f_{m-1},\theta_1) & \cdots & r(x_0,y_0;f_{m-1},\theta_{n-1}) \end{pmatrix} \tag{2}$$

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°. This kind of feature structure is capable to accurately represent local image patches and facilitates invariance operations for image feature search over arbitrary rotations, scale and translation and by normalization also achieves illumination invariance [13,11].

3.2 Classification and Ranking of Features

In general, any classifier can be used to learn and to classify features into image feature 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 features in a scene and to sort them in the best matching order returning only a fixed number of the best candidates.

In order to apply statistical classification and ranking it is necessary to estimate class conditional pdf's for every feature class. However, a single Gaussian cannot represent class categories, such as eyes, since they may contain inherited sub-classes, such as closed eye, open eye, Caucasian eye, Asian eye, eye with 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 an effective principal distribution to represent the statistical behaviour of simple Gabor features.

There are several methods to estimate parameters of Gaussian mixture models (GMMs) and some of them can automatically estimate the number of components in a GMM [12]. Pdf's are estimated separately for different image feature types from the complex vectors of Gabor feature matrix in (2) as

$$\boldsymbol{g} = [r(x_0, y_0; f_0, \theta_0) \;\; r(x_0, y_0; f_0, \theta_1) \ldots r(x_0, y_0; f_{m-1}, \theta_{n-1})] \quad . \tag{3}$$

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 preferred choice to act as a ranking confidence score [12]; it is a measure of how reliable the class assignment is. Ranking facilitates an efficient search, image features with the highest confidences can be processed first.

The algorithms utilizing simple Gabor features and Gaussian mixture model feature ranking have been given in [5].

4 Affine Transform Based Spatial Constellation Model

In this as well as in the aforementioned studies ([9,1,6,2,3]) the detection is applied to real 3-D objects spanning real 3-D surfaces, but for simplicity, the objects have been treated as planar, that is, they can be uniquely represented in two dimensions. It is well known that for example frontal human faces with a low degree of in-depth rotation can be accurately detected using 2-D image processing techniques not utilizing the 3-D shape of faces. Extracted image features are 2-D projections of planar point sets in 3-D vector space, that is, we consider 2-D projective geometry invariant to affine transforms. The properties of affine space will be considered next and then a spatial search method utilizing them will be introduced.

4.1 Affine Transform

Affine transformation in N-dimensional vector space $\mathbb{R}^N$ can be represented as a matrix multiplication in homogenous coordinates as $T : \mathbb{R}^N \rightarrow \mathbb{R}^N,\ T(\boldsymbol{x}) = A\boldsymbol{x}$ where, in case of 2-D coordinates, the transform matrix becomes

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix} \quad . \tag{4}$$

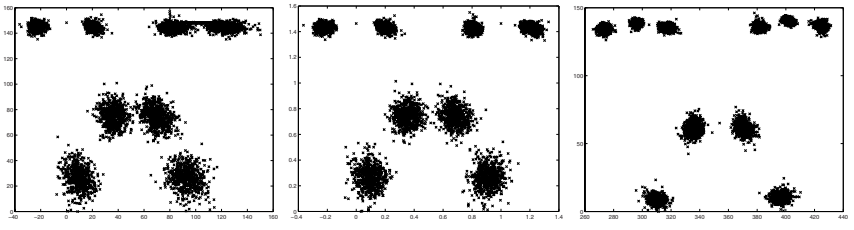
It is difficult to interpret the parameters in this form. There are various possible decompositions into a set of geometrically meaningful parameters. Parameters such as rotation, scale, shear, squeeze, scaling along first and second axis (dilation) may be involved. In this study we apply the following decomposition which is one of the easiest to derive and interpret

$$A = \underbrace{\begin{pmatrix} 1 & 0 & c \\ 0 & 1 & f \\ 0 & 0 & 1 \end{pmatrix}}_{\text{translation}} \underbrace{\begin{pmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\text{rotation}} \underbrace{\begin{pmatrix} p & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\text{scaling/squeeze}} \underbrace{\begin{pmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\text{shear}} \tag{5}$$

Since it is often computationally most convenient to operate on the original transform matrix A, the motivation is to utilize the transform parameters a, b, c, d, e, and f in (4) as functions of the decomposed parameters ϕ, p, q and n in (5).

4.2 Model Representation and Training

Local image features can be detected by the method proposed in Sec. 3, but in addition their configuration topology must be restricted in order to apply spatial constellation model search. In the following we assume planar objects and their 3-D projections on 2-D image plane. This assumption provides a sufficient framework for analysis where the spatial relationship of image features, the constellation, remains affinely almost fixed, e.g., for every two facial images there exists an affine transform which maps the features from one image to the corresponding features in another. The suitability of affine mapping for roughly frontal facial images is demonstrated in Fig. 1 where 10 image features are represented for two different fixed coordinates and for an affinely mapped space where the smallest variance is achieved.



(a) (b) (c)

Fig. 1. Image features from 600 training images in the XM2VTS database: (a) left eye center and eyes' angle fixed; (b) left and right eye fixed; (c) affine mapped (LSQ) to features of one face

A spatial model can be generated by storing the feature configuration from a single image. Once having such a model any other object can be mapped to the model, for example by the least mean square fit (LSQ), and accepted as a detection if the deviation from the model is within acceptable limits. However, in practice spatial variability of face features is not completely affine and selection of the proper representative for "ideal face" affects the model's performance. An iterative scheme presented in Algorithm 1. can be used to generate the model from a training set.

4.3 Spatial Constellation Search

In this section we solve a specific task of Point Pattern Matching (PPM). The task of PPM is to find a transform of a given type that best fits one set of points onto another in terms of a given metric. The role of the metric is important and different fits are provided depending on the selected metric (including for example subset fitting).

In our case we are fitting image feature coordinates with the affine model representative. In the simplest case the metric (or distance between the two point sets) can be defined as a sum of squares of distances between points with

Algorithm 1. Spatial constellation model training

1: Assign the model representation to the first element in the training set: $M_1 = X_1$
2: **for** $n = 1 \dots N$ {Over all training set} **do**
3: **for** $m = 1 \dots n$ **do**
4: Find an affine transform mapping $F_m : \boldsymbol{X}_m \rightarrow Y_m$, for which $\parallel Y_m - M_n \parallel^2$ achieves its minimum (RSS)
5: **end for**
6: Map first n elements to the current model M_n using the corresponding mappings F_n
7: Update the current model as the mean of the mappings: $M_{n+1} = \frac{\sum_{i=1}^{n} Y_i}{n}$, where summations and divisions by scalar are done elementwise
8: **end for**
9: Return the current model, M_{N+1}, as the result

the same label in different sets. In a such method the inputs would be: 1) a point set containing labeled candidate image feature locations, $\boldsymbol{S}_{i,j}$, such that $\boldsymbol{S}_{i,j}$ is the location (x,y) of the j-th most probable candidate for i-th image feature, and 2) the model M, obtained from the training step. The output of the method is an object hypothesis index vector, $\boldsymbol{I}_i$, denoting numbers of candidate locations used for object hypotheses generation, $\boldsymbol{S}_{i,I_i}$, $\forall\ i :\ I_i \neq 0$. Zero values in the index denote omitted features (omission handling described later).

Object search is based on yet another assumption, which allows to reduce the search complexity greatly. Once images for three points (triplet) in the model have been selected, the affine transform is uniquely defined and other points from hypothesis can be selected as the closest corresponding candidate locations. We are assuming that if we try all possible triplets of model points, we would not miss the best possible, globally optimal, hypothesis.

For reducing computation time it might be useful to check only a subset of all the possible triplets, since time complexity is linearly dependent on the number of mapping triplets. The number of triplets to be checked can thus be from 1 to $\frac{n!}{(n-3)!n!}$ where n is the number of image features. The amount of triplets to check depends on the time constraints and desired omission resistance. The search is described in Algorithm 2..

Handling missing image features. We call an image feature omitted if its correct coordinates on the image are not well enough described by one of the candidate locations for the corresponding feature. The reason for the features being omitted can be for example partial occlusion or a feature detection failure.

Although massive omissions decrease method performance, it is possible to recover the correct hypothesis if enough features are remaining. A crucial point for the correct hypothesis recovering is that at least one matching triplet exists.

A simple omission detection approach exists: if there are no extracted image features which contribute to the overall error (RSS) for less than a given threshold, the point is considered to be omitted. Another parameter is the omission penalty the amount which is added to the overall error for each point

Algorithm 2. Spatial constellation search

1: **for all** selected triplets o, p, q **do**
2: **for all** possible label values i, j, k **do**
3: Find affine mapping F of triangle $[M(o)M(p)M(q)]$ on triangle $[S(o,i)S(o,j)S(q,k)]$ (equivalent to solving system of 6 linear equations).
4: Create image from the model using F: $M^F = FM$.
5: Select points closest to model points with equal labels for each label and store their indices in $I_{o,p,q,i,j,k}$
6: Calculate the sum of squared distances (or Residual Square Sum): $RSS_{o,p,q,i,j,k} = \sum_t \parallel M^F(t) - S(t, I_{o,p,q,i,j,k}) \parallel$
7: **end for**
8: **end for**
8: Sort values in $RSS_{o,p,q,i,j,k}$ and return corresponding $I_{o,p,q,i,j,k}$.

considered as omitted. The introduction of these two parameters is justified by two reasons: 1) the distance from the predicted feature location to the actual local feature response should be discarded if it is much greater than the feature size – the found feature is most likely not related to the current model mapping; 2) It is useful to be able to control the balance between hypotheses with a few omitted features, high error and a substantial number of omissions or lower error for the rest of the image features.

After the points are checked for omission, algorithm proceeds in the same way, with the only difference that omission penalty is added to the final hypotheses error instead of a squared distance of each omitted image feature.

Applying affine transformation restrictions. A model of the frontal human face will not represent only frontal faces if it is subject to an unrestricted affine transform. A transformation between two frontal images of the same face is a similarity transform (a combination of shift, rotation and scaling) which is only a custom case of affine transform. Affine transform is a similarity transform if both shear and squeeze parameters in its decomposition are fixed to zero. Since in the real world applications faces cannot be absolutely frontal, some degree of shearing and squeezing should be still allowed in the transformation model for better performance. Another example is restricting possible rotations - often getting an upside-down image is something completely not acceptable and thus means false detection with a very high probability.

Restrictions are implemented as multipliers for the final hypothesis residual square sum, based on the parameters of transform used for model mapping in current hypothesis generation. The final coefficient was combined of four functions: $P_{transform} = P_{shear} \times P_{squeeze} \times P_{scale} \times P_{rotation}$. Each of these functions is an inverse of the estimated probability density for corresponding transform parameters, with a small added value limiting maximum penalty: $P_{param}(t) = \frac{1}{pdf_{param}(t)+\Delta}$.

5 Experiments

The XM2VTS facial image database used in the experiments is a publicly available database for benchmarking face detection and recognition methods [14]. The frontal part of the database contains 600 training images and 560 test images of size 720×576 (width $\times$ height) pixels. 10 marked image feature detectors were trained and searched as described in [5]. The localisation accuracy was measured by d_{eye}, which is defined as maximum over distances between detected features and groundtruth normalized by groundtruth eye distance [15]. The eye-distance normalization makes d_{eye} scale independent. d_{eye} is standard and recommended face localisation error measure [16].

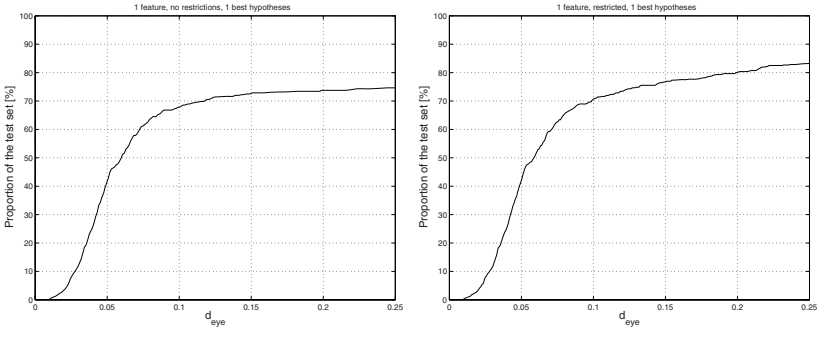


Fig. 2. Results for 1 best hypothesis using 1 best image feature (for each 10 feature classes)

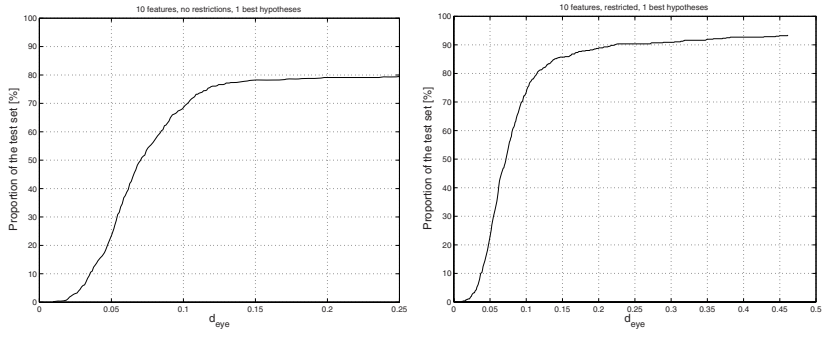


Fig. 3. Results for 1 best hypothesis using 10 best image features

Fig. 2 shows the results from an experiment where only one feature per class was extracted and only the best face hypothesis was accepted. The experiment was performed with and without affine restrictions estimated from the training set images. The results merely represent the accuracy and reliability of detected image features and in 70% of images already the feature detector provides the correct face ($d_{eye} = 0.1$).

In the second experiment, results shown in Fig. 3, the number of image features was increased to 10 (total of 100 for 10 different classes) which improved the results, but also revealed a problem; the best hypothesis is often misaligned with respect to the ground truth.

Hypotheses close to the misaligned best hypothesis were included by allowing detection of 10 best hypotheses, which significantly increased the detection accuracy to 90% (Fig. 4).

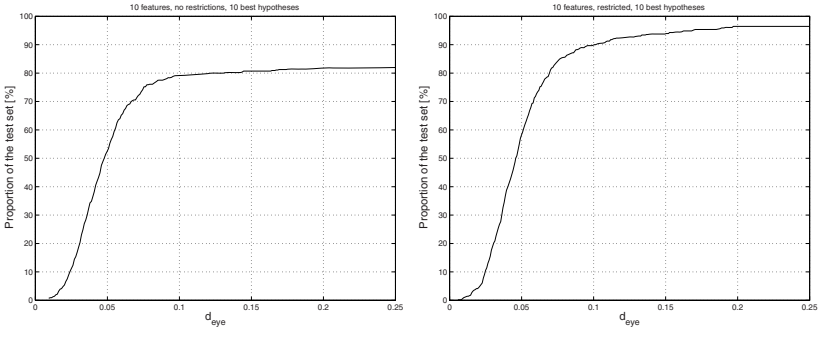


Fig. 4. Results for 10 best hypotheses using 10 best image features

It is noteworthy that the results of this simple method are comparable to the much more complicated but state-of-the-art method reported in [2].

6 Conclusions

In this study we proposed a new feature based method for the detection of object classes in gray-level still images. The proposed method follows state-of-the-art approaches by separating the process to image feature extraction and spatial constellation search. The image feature extraction is based on simple Gabor features and their statistical ranking providing very accurate and reliable results. The spatial search was formulated as a point pattern matching problem over affine invariant point sets. The method finds the globally optimal constellation of extracted image features, is robust to outlier features, and provides estimated location of missing image features.

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).

References

1. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2003)
2. Hamouz, M., Kittler, J., Kamarainen, J., Paalanen, P., Kalviainen, H., Matas, J.: Feature-based affine-invariant localization of faces. IEEE Trans. on Pattern Analysis and Machine Intelligence 27, 1490–1495 (2005)
3. Helmer, S., Lowe, D.: Object recognition with many local features. In: Workshop on Generative Model Based Vision. (2004)
4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on PAMI vol. 27 (2005)
5. Kamarainen, J.K., Ilonen, J., Paalanen, P., Hamouz, H., Kälviäinen, H., Kittler, J.: Object evidence extraction using simple gabor features and statistical ranking. In: Proc. of the 14th Scandinavian Conf. of Image Processing, Joensuu, Finland pp. 119–129 (2005)
6. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Trans. on Computers 22, 67–92 (1973)
7. Lades, M., Vorbrüggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers 42, 300–311 (1993)
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. of the International Conference on Computer Vision, Corfu, Greece pp. 1150–1157 (1999)
9. Burl, M.C.: Recognition of Visual Object Classes. PhD thesis, California Institute of Technology (1997)
10. Kadir, T.: Scale, Saliency and Scene Description. PhD thesis, Oxford University (2002)
11. Kyrki, V., Kamarainen, J.K., Kälviäinen, H.: Simple Gabor feature space for invariant object recognition. Pattern Recognition Letters 25, 311–318 (2003)
12. Paalanen, P., Kamarainen, J.K., Ilonen, J., Kälviäinen, H.: Feature representation and discrimination based on Gaussian mixture model probability densities - practices and algorithms. Pattern Recognition 39, 1346–1358 (2006)
13. Kamarainen, J.K., Kyrki, V., Kälviäinen, H.: Invariance properties of Gabor filter based features - overview and applications. IEEE Trans. on Image Processing 15, 1088–1099 (2006)
14. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended M2VTS Database. In: Chellapa, R. (ed.) Proc. of Second Int. Conf. on Audio and Video-based Biometric Person Authentication. pp. 72–77 (1999)
15. Jesorsky, O., Kirchberg, K., Frischholz, R.: Robust face detection using the hausdorff distance. In: Proc. of 3rd Int. Conf. on Audio- and Video-based Biometric Person Authentication. pp. 90–95 (2001)
16. Rodriguez, Y., Cardinaux, F., Bengio, S., Mariéthoz, J.: Measuring the performance of face localization systems. Image and Vision Computing 24, 882–893 (2006)