Combined model for detecting, localizing, interpreting and recognizing faces

Leonid Karlinsky, Michael Dinerstein, Dan Levi, and Shimon Ullman
{leonid.karlinsky,michael.dinerstein,dan.levi,shimon.ullman}@weizmann.ac.il

Weizmann Institute of Science, Rehovot 76100, Israel

Abstract. This work describes a method that combines face detection, localization, part interpretation and recognition, and which is capable of learning from very limited data, in a semi-supervised or even fully unsupervised manner. Current state-of-the-art techniques for face detection and recognition are subject to two major limitations: extensive training requirement, often demanding tens of thousands of images, and detecting faces without explicitly detecting relevant facial parts. Both of these limitations hinder the recognition task, since for a specific face the number of available examples is usually small, and because the strongest cues for identity lie in the specific appearance of facial parts. The proposed method alleviates both these limitations by effective learning from a small training set and by detecting the face through, and together with, its main parts. This is obtained by a novel unsupervised training method, which iterates phases of part geometry and part detector learning, to incrementally learn an object category from a set of unlabeled images, containing both class and non-class examples given in unknown order. We tested our method on face detection and localization tasks both in a set of 'real life' images collected from the web as well as in LFW and MIT-CMU databases. We also show promising results of our method when applied to a face recognition task.

1 Introduction

The focus of this work is on a method that combines face detection ('what is it?'), localization ('where is it?'), part interpretation ('where are the facial parts?') and recognition ('who is it?'), and which is capable of learning from very limited data, in a semi-supervised or even fully unsupervised manner. The method was developed for general object detection and localization and in this work we extend and apply it to the task of detecting and localizing faces. In addition, we extend it to perform the recognition task.

The face detection and localization problem has a long history in the literature, but a significant dividing line was the influential work by Viola and Jones [1]. Based on a survey by [2], before [1] the state-of-the-art methods for face detection were divided into several approaches: knowledge-based [3], feature invariant [4], template matching [5], and appearance based [6–8]. Following [1] the focus of face detection approaches turned towards efficient real-time face

detection. In [1], the face is represented by a cascade of linear classifiers each using simple low-level 'Haar features' and trained using boosting methods [9]. Currently, many state-of-the-art face detectors are variants of the approach in [1], with different improvements and extensions. For example, [10] extend the set of Haar features, [11] improve feature computation efficiency and introduce a coarse-to-fine technique, [12] improve the optimization of the cascade, [13] offer an alternative to the boosting technique used in [1] and provide efficient methods for multi-view detection, and [14] suggest an improved set of low-level features and improve the computational efficiency. Currently, the best results were reported by [14].

Despite all the developments listed above, there are still two main limitations shared by current state-of-the-art face detection techniques. These limitations are especially relevant if one wants to extend such techniques to face recognition. The methods based on [1] usually have an extensive training requirement. For instance [13] train on 75,000 face examples, and [14] have around 23,000 faces and 30,000 non-faces in their training set. While general face examples are abundant, this is usually not the case for face recognition tasks, where only a handful of images may be available for a specific individual. The second limitation is non-specificity to facial parts. The Haar-like features based methods detect faces without explicitly detecting facial parts and thus are capable of face detection and localization, but are not suited for recognizing and discriminating between different shapes of face parts such as eyes, nose, ears, mouth, chin and etc. The ability to detect specific facial parts plays a key role in face recognition task, as the distinction between two individuals is often based on fine differences in specific face parts appearance such as types of nose, eyes, hair, etc.

The proposed method addresses both of these limitations. It is capable of learning from a small number of examples, even in the challenging unsupervised setting, in which the object examples appear at unknown locations and only in an unknown subset of the training images. The subset of images containing class examples may even be as small as 10% of the training set. In addition, our method is part-based, meaning that the object is detected through first detecting its parts. The parts and their spatial configuration are learned automatically. Moreover, each part is represented by a (learned) set of its so-called semantic equivalents. For example, the nose part is represented by a set of image fragments representing the nose viewed with varying poses, illuminations, and other types of variations. These semantic equivalents are learned without supervision, and they are combined using an efficient representation (the CNOR model below) to form a robust detector for a corresponding part.

Our method was tested on unconstrained frontal face detection and localization tasks in a challenging set of 'real-life' images, collected randomly from the web. The faces are usually small relative to the images, appear at random locations in the images and exhibit strong scale, lighting, pose and expression variations. The set also contains partially occluded faces (e.g. by sunglasses or other objects) and multiple face instances. We also tested detection performance on the LFW database [15] and detection and localization performance on the

MIT-CMU dataset [16]. Examples of the output of our method on images taken from various datasets appear in figure 7. Finally, we applied our method to the simultaneous object detection, localization and recognition task on a dataset consisting of the set of 'real-life' images mentioned above mixed with an additional set of about 130 images containing a face of a specific person. The person images also exhibit large scale, pose, lighting, location, expression and occlusion variations. Our system showed good performance on this combined detection, localization and recognition task despite being trained on a limited set of only 10 images of the specific individual. Examples of combined person detection and recognition appear at the bottom row of figure 7. In addition, we tested the contribution of various aspects of the proposed method by removing some of them and evaluating the performance on the same detection and localization task.

The rest of the paper is organized as follows. Section 2 describes the method and the learning algorithms used, section 3 provides details of the experimental validation, and section 4 contains the summary and proposes possible future research directions.

2 Method

This section presents our method for combined detection, localization, part interpretation and recognition of human faces. The method builds upon and extends a general method for unsupervised category learning which is presented in a companion paper [17]. The unsupervised learning algorithm, called Unsupervised Consistency Amplification (UCA) is briefly overviewed in section 2.1 and its extensions introduced for face detection and recognition are described in section 2.2. For a more detailed description of the UCA algorithm please refer to [17].

2.1 Unsupervised Consistency Amplification (UCA)

The basic UCA unsupervised training algorithm is described in detail in a companion paper [17], here we give only a brief overview. UCA alternates between model learning and data partitioning. Given an image set S, an initial model (learned using initial generic features) is used to induce an initial partitioning by identifying highly likely class members. The initial partitioning is then used to improve both the appearance and the geometrical aspects of the model, and the process is iterated. In this manner the process exploits intermediate classification results at a given stage to guide the next stage. Each stage leads to an improved consistency between the detected features and the model, which is why the process is termed Unsupervised Consistency Amplification (UCA). Each UCA iteration consists of two phases of learning: the feature learning Appearance-phase (A-phase) followed by the part model learning Geometry-phase (G-phase). The approach and the order of the phases are summarized in Figure 1.
Here we briefly describe the phases of the algorithm:

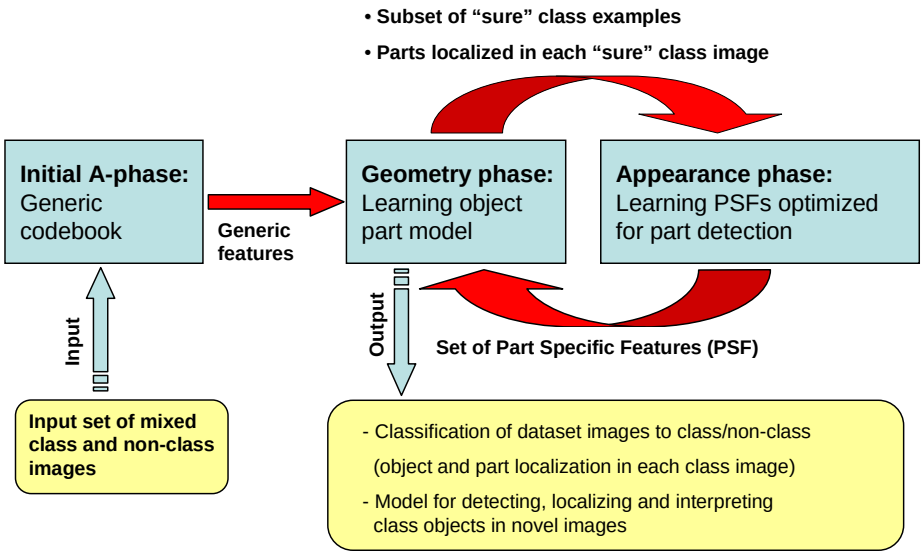

Fig. 1. Schematic diagram of the UCA algorithm

Initial Appearance-phase: We use a generic codebook of quantized SIFT descriptors of 40 x 40 patches for the initial (appearance) features. This codebook is computed by a standard technique [18] from all the images in given set S . The codebook descriptors are compared to the descriptors at all points of all the images in S and storing the points of maximal similarity (either one or several, see below) in each image.

Geometry-phase: The detection of parts using the generic features is usually noisy, due to detections in non-class images, and at some incorrect locations in the class images. The goal of the geometric part model learning is to both distinguish between class and non-class images and between the correct and incorrect part detections, based on consistent geometric relations between the features. This is accomplished by the G-phase of the algorithm, which is also used for the selection of the most useful features and the automatic assignment of each of their detections in every image in S to either object or background model. During training, we model the background by a distribution of the same family as the class object distribution, which allows preventing from the spurious geometric consistency on the background to be accounted for by the learned class model. In our experiments we found that modeling the background distribution is better than assuming it to be uniform. The learned background model is then discarded after the training and is not used for classifying new images. Thus, the generative probabilistic model used in the G-phase is a mixture of two star-models, one for object and the other for background. It is learned without supervision from all the images in S using a novel probabilistic graphical model formulation explained in [17]. After the geometric structure has been learned,

a subset $H \subset S$ of images which contain class objects with high confidence is selected. In these images the object centers and parts are localized.

Appearance-phase: The A-phase constructs the part detectors used as features by the following G-phase. Each part-detector constructed in the A-phase represents an object part by extracting several typical appearance patches of the part, from different images. Part-patches can be extracted, because the locations of the parts in the images of the subset H are already estimated from the previous G-phase. An optimal subset of these part patches is learned by a novel probabilistic discriminative model, the Continuous Noisy OR (CNOR), described in [17]. The model parameters are optimized for part detection in correct location on faces.

Applying the model to new images: As in [17] we apply the model to new images by voting for the object reference point from the locations detected by the learned part detectors. Each part P_i has two learned parameters defining its spatial location relative to the object reference point - offset μ_i and covariance matrix Λ_i defining the uncertainty region of the offset. Both these parameters are used during voting. Following the voting, the parts are detected and localized by back-projection: each part that was detected within one STD from its expected location is declared as 'detected'. In the current paper we extend the voting method used in [17] towards multi-scale and multi-object detection. To handle multiple scales we iterate over candidate horizontal sizes of the face, in our case from 20 to 150 pixels with 10 pixel increments. The search over different scales produces only a few false alarms due to high precision of the UCA classifier (see [17]). We also apply non-maximal suppression to suppress overlapping detections (bounding boxes) with lower scores. To handle multiple objects we replace the voting method used in [17]. In [17] only five highest scoring detections of each part were used in the voting. Since an image may contain more then five faces (as is the case in many of our examples), we replaced this method by voting using the entire set of part detections for each part. Given an image, detector for part P_i is applied to a grid of locations on the image (we used regular grid with a step of 5 pixels) producing a response map R_i , then R_i is shifted by μ_i , and dilated with Λ_i and added to a cumulative voting mask. The local maxima of the voting mask after all part detectors have voted constitute the candidate locations of the faces. Applying non-maximal suppression on voting masks gathered from all the candidate scales and thresholding the result produces the final face detections.

The training of the G-phase and A-phase probabilistic models was achieved by applying Expectation Maximization (EM) algorithm [19] and its structure learning variant [20]. The following section explains how UCA was used in faces detection, localization, part interpretation and recognition tasks.

2.2 UCA-based face detection and recognition

The model for face detection (for brevity detection = detection + localization + part interpretation) and recognition is trained in three stages, each stage training a separate UCA model. In the first stage, M_1 - a model for detection is trained on

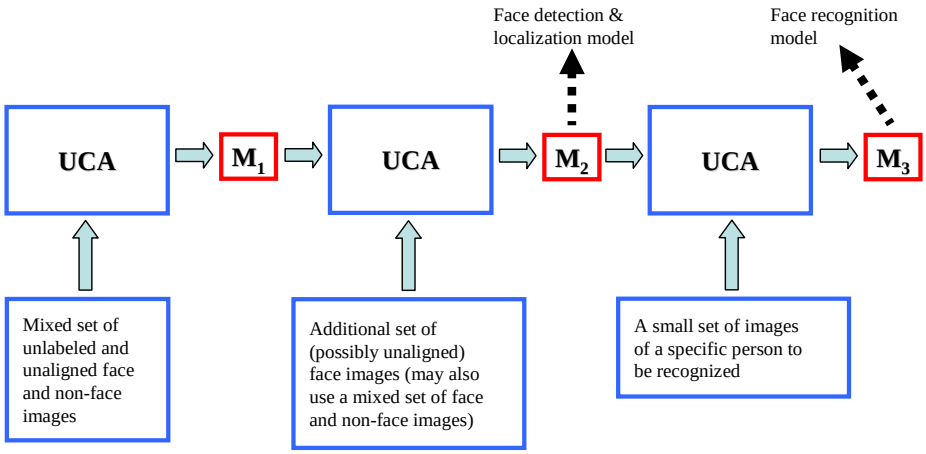

Fig. 2. A schematic illustration of the proposed approach showing the inputs and the outputs of all the training stages

an unlabeled set of mixed face and non-face images in an entirely unsupervised manner. In the second stage an improved model for detection M_2 is trained on an additional training set, in which the faces are detected using M_1. In the third stage, a model for recognition M_3 is trained using several examples of a specific face to be recognized detected and localized, based on the application of M_2.

Stage 1: The first stage is fully unsupervised, its training set consisting of mixed non-face and face images (faces appear at unknown locations). This stage was performed in the experiments of [17] and here we use the model it learned (denoted M_1) for the second stage. Fifty face examples from Caltech faces dataset were present in the unsupervised training set of the first stage (together with 450 non-face images). In general this stage may be used in cases we train on a semi-supervised set (having only images that contain the learned object), but when object examples vary in viewing direction (or other viewing conditions), e.g. if we get a set of mixed frontal and profile faces. Although it is possible to try to capture all the viewing directions by a single model, it is clearly harder and will typically cost in loss of ability to detect facial parts. So in this case it is beneficial to first apply unsupervised learning to separate different views and only later to learn a separate model for each view. Experiments along these lines were performed in [17] for automatically separating different car views.

Stage 2: The second stage is learning an improved model using weak supervision. The goal of this stage is to improve the performance of the model learned during the unsupervised stage by providing it with more object (face) examples. This stage could also be performed by the system in autonomous (unsupervised) online manner by crawling on web images and detecting instances of the object using the M_1 model. In our case we gave the system 300 additional image examples randomly chosen from the LFW database and ran the M_1 model on them to detect and localize faces on these images. The output of M_1 were face bounding

boxes detected in the training images. We then train the model M_2 using the UCA method by assuming the output of M_1 to be the output of the G-phase in one of UCA iterations (see section 2.1 for the definition of the G-phase). The model M_2 is then used to perform the face detection and localization in section 3. The second stage can be seen as additional iterations of the UCA method applied in the first stage. The main difference is that the standard UCA loops over the same image set over and over again, while in M_2 training stage new images are provided either in a weakly supervised or in online unsupervised manner.
Stage 3: The third stage, training a recognition model for a specific face, is performed when we receive a (limited) set of examples of a face of a certain individual (that we wish to recognize in the future) and train a UCA model on these images. The training examples need not be cropped or aligned and can exhibit any scale variation, all we require them to be is 'roughly frontal' (as the example images in Figure 7). The training is organized along the lines of the second stage. The resulting model M_3 can be used by itself for simultaneous detection, localization and recognition of a specific person, but it is better used in conjunction with M_2, since M_3 had only limited training on the few examples provided for the specific person. We investigate both of these options in the section 3.

Figure 2 summarizes the proposed approach. Section 3 describes the experimental validation of our method.

3 Results

To test the proposed method, we applied it to frontal face detection, localization and recognition tasks in several databases, namely: people-containing images gathered from Google image search (denoted WEB database); Labeled Faces in the Wild (LFW) database [15]; MIT-CMU dataset [16]; and a set of unconstrained images of a person taken under different viewing conditions, and at different locations and times (denoted PERSON database). The MIT-CMU dataset combines all the images from tests A, B and C in the dataset description [16]. The statistics of all the datasets are summarized in Table 1.

Name	# images	# frontal faces	image size (rows x columns + STD)	Faces size (min – max)	Has multiple people	Has scale variations	Faces are aligned
WEB	354	477	626×593±197×220	31×20–180×130	+	+	–
LFW	13233	13233	250×250±0×0	135×95	–	–	+
MIT-CMU	130	511	429×440±227×208	21×14–253×168	+	+	–
PERSON	162	162	480×640±0×0	36×28–209×160	–	+	–

Table 1. Provides statistics of the datasets used in our experiments

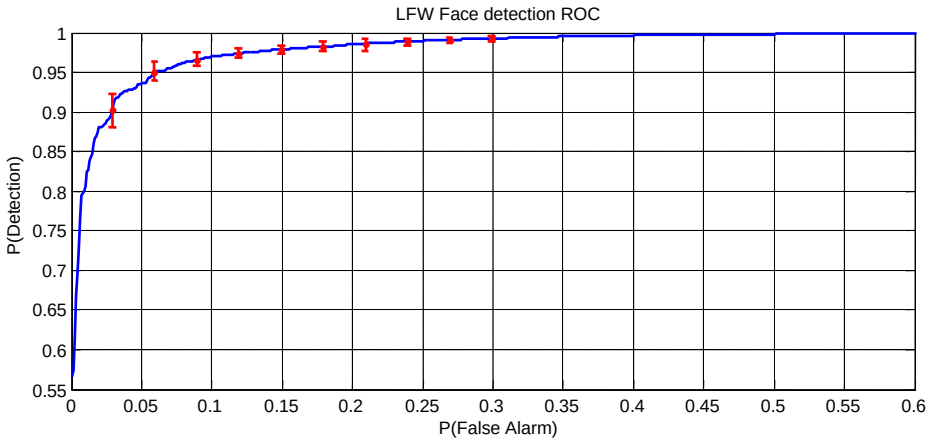

Fig. 3. Summary of the LFW face detection experiments. Although faces in LFW are of the same scale, a full method including scale search was applied both to face-containing and background images

In the first experiment, the model M_2 was applied to the face detection task in the LFW database. Since faces in the LFW are cropped, roughly aligned and equally scaled, there was no reason to test localization on that database, although in several dozens of images that we manually checked the localization was perfect. The LFW database is comprised only of images containing faces detected by Viola & Jones face detector, so in order to test face detection, maximal response of the UCA was computed for all images in the LFW and in additional 916 background images which were a union of the Caltech and Google background sets. These measurements were used to build the ROC curve for face detection depicted on Figure 3. The error bars of the ROC were computed by 10-fold cross validation.

In the second experiment we tested the combined detection and localization performance of our method. To this end, model M_2 was applied to the WEB and MIT-CMU datasets. The results are summarized in ROC curves in Figures 4a and 4b respectively. We also compared our performance on the WEB database with the performance of the OpenCV [21] implementation of Viola & Jones face detector [1] and shown a significant performance gain (see figure 4a). The face was considered correctly localized if the detected bounding box exceeded the standard Jaccard index overlap score of 0.5 with the manually marked ground truth bounding box:

$$Jaccard\ Index\,(BB_1, BB_2) = \frac{|BB_1 \cap BB_2|}{|BB_1 \cup BB_2|} \tag{1}$$

In order to test the contribution of various components of our method to the final performance, Figure 4a also contains ROC curves of performance after removing different components. Specifically, we tested our method without

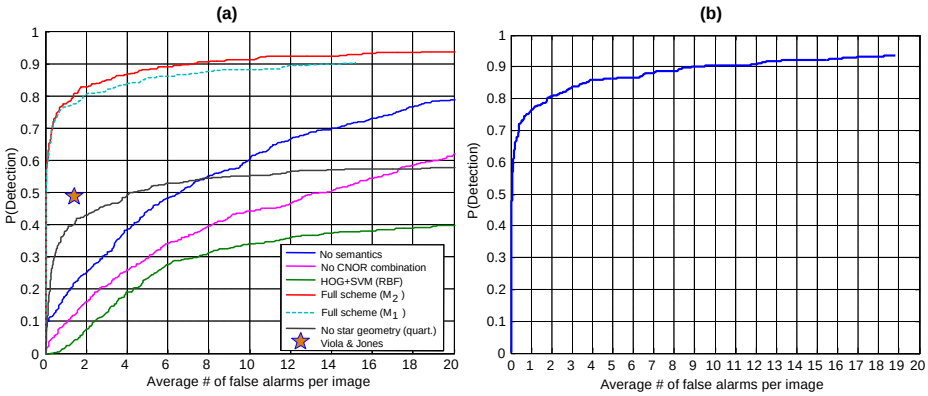

Fig. 4. (a) Summary of the WEB face detection and localization experiments. Additionally provides comparison with OpenCV implementation of [1], HOG+SVM and performance after removing different components of our method. **(b)** Summary of the MIT-CMU face detection and localization experiments, ROC curve

using semantic equivalents for part detection ('No semantics' in the figure), without using the CNOR model for combining the semantic equivalents ('No CNOR combination') and without using the learned geometry of parts ('No star geometry'). In all cases, removal of these components resulted in significant drop in performance. When semantic equivalents were not used for part detection, a single best image fragment was chosen to represent the part. When CNOR model was not used for combining semantic equivalents, they were combined using a simple summation. When part geometry was not used, the parts were split into four groups corresponding to the four quarters of the face. The face was then detected as a combined vote of the four quarter detections, while each of the quarters was detected by the bag-of-features method. Localizing faces directly by a bag-of-features of the whole face (without using this four quarters scheme) failed to produce reasonable results due to the limitation imposed by the overlap score (eq. 1) and the currently used non-maximal suppression scheme.

Additionally, figure 4a contains an ROC curve of HOG + SVM classifier implemented along the lines of [22]. HOG descriptors were computed for all the face bounding boxes detected by M_1 in all the images used to train M_2. For these HOG descriptors RBF kernel SVM was trained against HOG descriptors of maximal score bounding boxes detected by M_1 in background images (Caltech + Google backgrounds).

In the third experiment combined detection, localization and recognition were tested on the combination of the PERSON and the WEB datasets. The results are summarized in figures 5a and 5b. Ten images of a person were used to train the model M_3. The models M_2 and M_3 were used in a cascade-like sequence. First, M_2 was applied to detect candidate faces in the image and then M_3 was applied to search for the face of the specific person at locations and scales close to the ones detected by M_2. We also tested the combined detection, localization

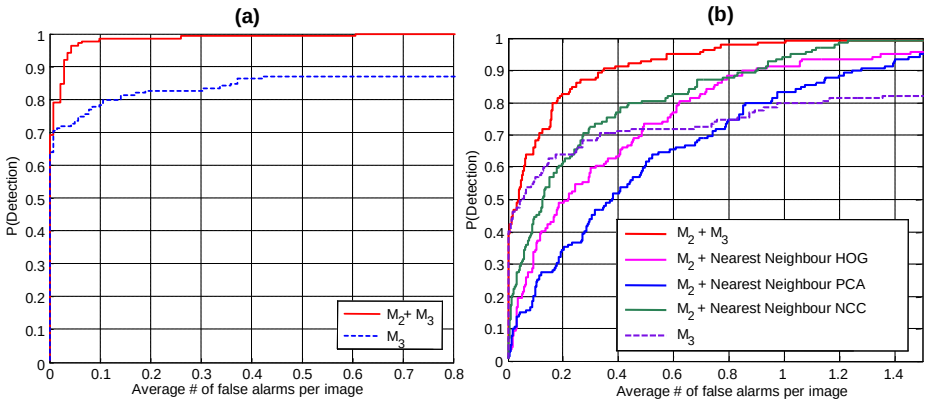

Fig. 5. **(a)** Summary of PERSON face detection and localization experiments. **(b)** Summary of PERSON + WEB combined detection, localization and recognition experiments

and recognition using a single model (M_3). When used by itself, it produces less good detection and localization results (see Fig. 5a) due to the limited training on only 10 images. In figure 5b we also compare our approach to several baseline approaches. Specifically, we tried to replace the M_3 in the $M_2 \rightarrow M_3$ sequence by Nearest Neighbor PCA, NCC or HOG, all trained on the same 10 images we used for training M_3. As can be seen in Figure 5b, the performance of these methods is significantly worse. This can be partially attributed to the fact that these methods try to detect the whole face using a single template, while our method applies part based detection which, as explained in the introduction, is more appropriate for the recognition task. Examples of face interpretation, that is detection of various facial parts, are given in Figure 6.

4 Discussion

This paper presented a method for combined faces detection, localization, part interpretation and recognition in unconstrained images, where faces may appear at any image location, scale and under a variety of difficult viewing conditions. The method shows promising performance in various experiments on different datasets, superior to several standard baseline methods and the standard implementation of the Viola & Jones face detector. The method requires only a few images for training and can be trained in a fully unsupervised manner. The underlying models employed by the method are general and not limited to face detection.

Even in cases of semi-supervised training, the ability to automatically separate different object views in a given training set (containing mixed views) can improve classification performance of the learned models. Moreover, this ability also facilitates learning part based models that are capable of detecting

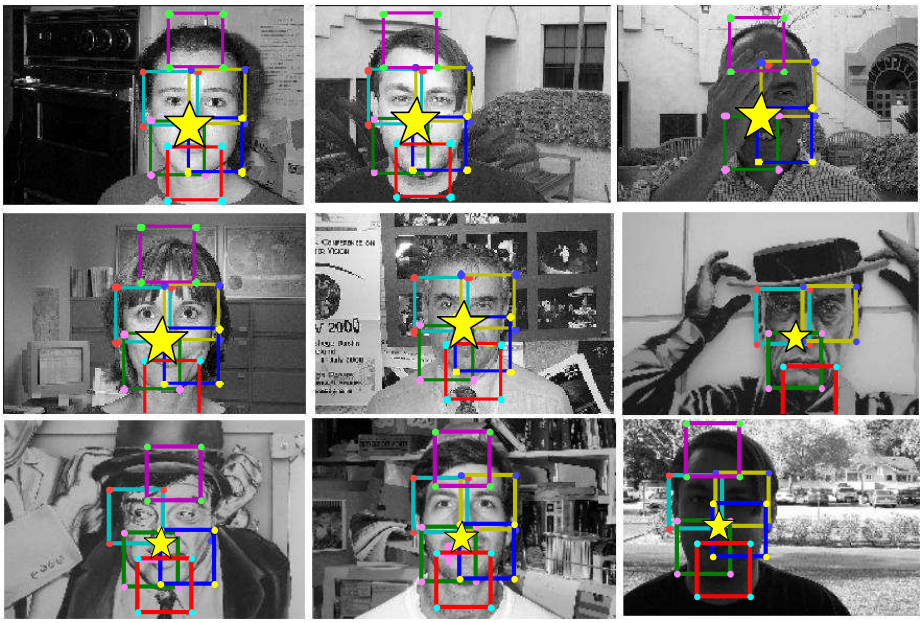

Fig. 6. Examples of detections of several of the (about 50) modeled parts. Yellow star shows the detected object model center location and the colored boxes show the parts that were detected by the model

'meaningful' object parts. Initial experiments performed in [17] for automatic separation of car views from the PASCAL 2007 dataset, show that our method has this ability. An interesting future research direction would be learning viewpoint invariant part based models for face detection from a mixed set of examples of different face views. Additional interesting extension may be linking the models of different views in terms of their detectable parts (such that semantically equivalent parts from different view models are linked) in order to facilitate view-invariant part interpretation.

In its current version, the method does not make full use of the detailed part interpretation obtained during the detection process for the purpose of subsequent individual face identification. Since part detection obtained by our method is usually highly accurate, for each facial part it is possible (in future work) to build a universal dictionary of part appearances, such as semantic equivalent image fragments proposed in [23]. Then, one may learn an association between each of the part appearances and the conditions (lighting, pose, expression, etc.) under which the part is viewed. Given even a single image of a specific individual, her facial parts may be detected (by our model) and categorized according to the learned part appearance dictionaries. Subsequently, when confronted with a new image of the same individual taken under different conditions, the learned association between the conditions and the part appearances in the new image

(known following the detection, localization and part interpretation performed by our model) could be used to recognize the individual. Some ideas along these lines were explored in [24] (but without having a method for reliable part interpretation) and extending it might be an interesting future research direction towards individual recognition under highly varying viewing conditions.

Aknowledgment: This work was supported by EU IST Grant FP6-2005-015803 and ISF Grant 7-0369.

References

1. Jones, M.J., Viola, P.A.: Robust real-time face detection. ICCV (2001)
2. M.-H. Yang, D.K., Ahuja, N.: Detecting faces in images: A survey. PAMI, vol. 24 (2002)
3. Yang, G., Huang, T.S.: Human face detection in complex background. Pattern Recognition, vol. 27 (1994)
4. Yow, K., Cipolla, R.: Feature-based human face detection. Image and Vision Computing, vol. 15 (1997)
5. Sinha, P.: Object recognition via image invariants: A case study. Investigative Ophthalmology and Visual Science, vol. 35 (1994)
6. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. PAMI, vol. 20 (1998)
7. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. PAMI, vol. 20 (1998)
8. Schneiderman, H., Kanade, T.: A statistical method for 3d object detection applied to faces and cars. CVPR (2000)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory (1995)
10. Lienhart, R., Maydt, J.: An extended set of haar features for rapid object detection. ICIP (2002)
11. Schneiderman, H.: Feature-centric evaluation for efficient cascaded object detection. CVPR (2004)
12. Luo, H.: Optimization design of cascaded classifiers. CVPR (2005)
13. Huang, C., Ai, H., Li, Y., Lao, S.: High-performance rotation invariant multi-view face detection. PAMI (2007)
14. Yan, S., Shan, S., Chen, X., Gao, W.: Locally assembled binary (lab) feature with feature-centric cascade for fast and accurate face detection. CVPR (2008)
15. Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49 (2007)
16. Online: http://www.vasc.ri.cmu.edu/idb/html/face/frontal_images. CMU VASC Frontal Face Database (1997)
17. Karlinsky, L., Dinerstein, M., Levi, D., Ullman, S.: Unsupervised classification and part localization by consistency amplification. ECCV (2008)
18. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. ICCV (2005)
19. Neal, R., Hinton, G.: A view of the em algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models (1998)

20. Friedman, N.: The bayesian structural em algorithm. UAI (1998) 129–138
21. Online: http://www.opencvlibrary.sourceforge.net. OpenCV (2006)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR (2005)
23. Epshtein, B., Ullman, S.: Identifying semantically equivalent object fragments. CVPR (2005)
24. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. ICCV (2005)

Fig. 7. Examples of face detection, localization and recognition obtained by our method. The databases used for experiments are explained in the Results section. Examples inside the red box are from the WEB database, examples in the green box are from the MIT-CMU database and blue box contains examples of combined detection, localization and recognition of a specific person in the PERSON database. The current algorithm does not use color information, all the color images were processed in grayscale by the algorithm