Image and Vision Computing 47 (2016) 45–59

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Multi-view facial landmark detector learned by the Structured Output SVM☆

Michal Uřičář [a,]*, Vojtěch Franc [a], Diego Thomas [b], Akihiro Sugimoto [b], Václav Hlaváč [a]

[a] Center for Machine Perception, Czech Technical University in Prague, Czech Republic
[b] National Institute of Informatics, Tokyo, Japan

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form 29 December 2015
Accepted 18 February 2016
Available online 2 March 2016

Keywords:
Deformable Part Models
Structured output SVM
Facial landmarks detection

ABSTRACT

We propose a real-time multi-view landmark detector based on Deformable Part Models (DPM). The detector is composed of a mixture of tree based DPMs, each component describing landmark configurations in a specific range of viewing angles. The usage of view specific DPMs allows to capture a large range of poses and to deal with the problem of self-occlusions. Parameters of the detector are learned from annotated examples by the Structured Output Support Vector Machines algorithm. The learning objective is directly related to the performance measure used for detector evaluation. The tree based DPM allows to find a globally optimal landmark configuration by the dynamic programming. We propose a coarse-to-fine search strategy which allows real-time processing by the dynamic programming also on high resolution images. Empirical evaluation on "in the wild" images shows that the proposed detector is competitive with the state-of-the-art methods in terms of speed and accuracy yet it keeps the guarantee of finding a globally optimal estimate in contrast to other methods.

1. Introduction

The detection of facial landmarks in images is a crucial step in many computer vision applications involving faces. For example, the landmark positions are used for alignment and normalization of face having substantial impact on the overall accuracy of face recognition systems estimating e.g. age, gender or identity (for example [1,2]).

In this paper, we propose a real-time multi-view landmark detector based on Deformable Part Models (DPM) [3,4]. An exemplary output of the proposed detector is shown in Fig. 1. The detector is composed of a mixture of tree based DPM, each component describing landmark configurations in a specific range of viewing angles. The usage of view specific DPM allows to capture a large range of poses and to deal with the problem of self-occlusions. The estimation of the viewing angle and the landmark position is done simultaneously by a structured output classifier. The inference problem can be solved globally by the dynamic programming, hence the detector's output is independent on an initial estimate in contrast to majority of other methods. Parameters of the DPM based structured classifiers are learned from annotated examples by the Structured Output SVM algorithm [5]. The objective function of the learning algorithm is directly related to the performance measure used for detector evaluation. In order to obtain a real-time detector, we use several speedups. First, we use tree based shape model defined by separable pair-wise potential functions which allows to decrease the computational complexity of the dynamic programming procedure from quadratic to linear in terms of the number of landmark positions via employing the distance transform [6]. Second, we propose to use the MIPMAP technique [7] for fast computation of features used by the local landmark classifiers. Third, we propose a coarse-to-fine strategy in order to reduce the size of the search space on higher resolution images. Evaluation on the challenging "Annotated Faces in the Wild" (AFLW) database [8] and 300-W dataset shows, that the proposed detector achieves competitive localization error compared to the current state-of-the-art detectors [9–13], yet it keeps the guarantee of finding a globally optimal estimate.

The proposed detector significantly outperforms the detector of [9], on which we build our approach. The authors of [9] propose a multi-view tree based DPM detector, which simultaneously estimates face locations, landmark positions and viewing angles. The conceptual difference to our method is the objective function optimized by the learning algorithm. Their learning algorithm is a variant of a two-class Support Vector Machines [14] which, in this application, optimizes the detection rate of resulting face detector while the landmark positions serve only as latent variables not appearing in the loss function. In contrast, our method based on the Structured Output SVMs optimizes directly the average landmark localization error, being the evaluation metric of landmark detectors. Using the proper learning objective

☆ This paper has been recommended for acceptance by Georgios Tzimiropoulos.

* Corresponding author. Tel.: +420 224 355 778; fax: +420 224 357 385.
E-mail addresses: uricamic@cmp.felk.cvut.cz (M. Uřičář), xfrancv@cmp.felk.cvut.cz (V. Franc), diego_thomas@nii.ac.jp (D. Thomas), sugimoto@nii.ac.jp (A. Sugimoto), hlavac@cmp.felk.cvut.cz (V. Hlaváč).

http://dx.doi.org/10.1016/j.imavis.2016.02.004
0262-8856/© 2016 Elsevier B.V. All rights reserved.

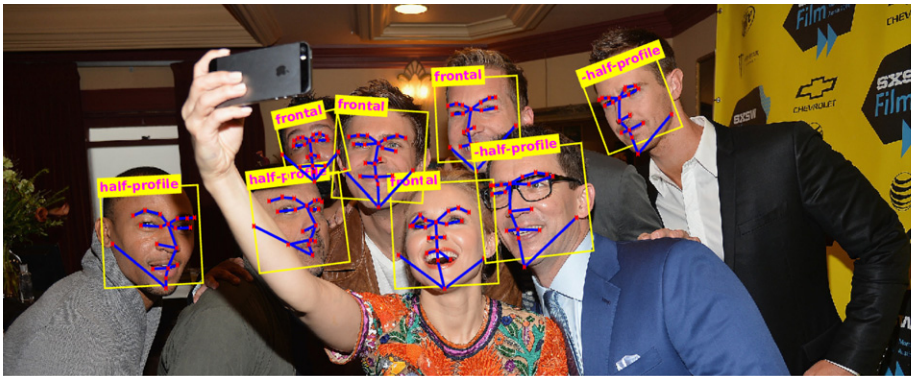


Fig. 1. The exemplary output of the proposed detector. The red dots represent the estimated landmark positions. The blue lines show the underlying graphical structure of the landmark configuration for a given view. A rough estimate of the viewing angle is shown in the yellow boxes at the top.

function leads to a significant improvement in the localization accuracy, as we demonstrate empirically.

The contributions of this paper are as follows:

- We treat the multi-view landmark detection as an instance of the structured output classification problem. The parameters of the detector are learned from examples by the Structured Output SVM algorithm [5]. Unlike the existing related method [9], the objective function of the learning algorithm is directly related to the performance measure commonly used for evaluation of landmark detectors.
- We implemented a coarse-to-fine strategy to decrease the computational complexity of the inference procedure based on the dynamic programing. The proposed strategy allows to keep a real-time performance of the proposed detector when it estimates a dense set of landmarks (for example 68 landmarks on the 300-W benchmark) on high resolution images.
- We propose to speed up the evaluation of a dense local feature descriptor via representing the base features (in our case Local Binary Patterns [15]) computed in multiple scales in the form of MIPMAP [7]. The MIPMAP representation avoids repetitive evaluation of the base features which significantly decreases the evaluation time without increasing the localization error.
- We experimentally show that a well tuned tree based DPM landmark detector with the guarantee to find globally optimal estimate is comparable in speed and accuracy to other methods using more complex shape models and local optimization strategies like [10–13].

This paper combines the results of our previously published conference paper [16] and a workshop paper [17]. In addition, this paper proposes a coarse-to-fine search strategy which is necessary for the estimation of a dense landmark sets in higher resolution images, such as images appearing in the 300-W competition.

The paper is organized as follows. Related work is discussed in Section 2. The proposed detector and its learning is described in Section 3. The coarse-to-fine strategy is outlined in Section 4. The experimental evaluation is given in Section 5. Finally, Section 6 concludes the paper.

2. Related work

2.1. Generative methods

The generative methods build a holistic parametric model of the face appearance. The shape and the texture are both represented by linear models learned from a set of aligned faces by the Principal Component Analysis. Fitting the generative model amounts to minimizing the error between the input image and the closest synthetic image generated by the model. Although the shape and the texture are described by linear models, the error function is highly non-linear with respect to the unknown poses and shape parameters. The resulting non-linear minimization problem is solved by iterative descent methods, finding a local optimum quality of which highly depends on an initial estimate. Among the most popular generative methods applied to facial landmark detection belong the Active Appearance Models (AAMs) [18,19] and the Morphable Models [2].

2.2. Discriminative methods

The discriminative methods learn predictors directly estimating pose, shape or the landmark positions from features computed on the input image. The advantages of the discriminative methods are their conceptual simplicity and a low test time. Nowadays, the most popular discriminative approach is a cascade of regressors, that were considered for example in [12,20,21]. Starting from an initial estimate, each regressor refines prediction of the previous one. The prediction in each stage is based on simple features extracted from patches located at positions determined by the output of the previous stage. Besides 2D landmark positions, [11] shows that the cascade of regressors can also accurately estimate pose and shape of a 3D face model. [22,23] proposed to use regression to estimate parameters of the AAMs. Regression methods combined with probabilistic graphical models were proposed in [24,25]. The graphical model is used to aggregate estimates of stochastically sampled local regressors into a single robust prediction.

2.3. Deformable Part Models

The Deformable Part Models perform alignment by searching the most likely configuration of local parts. The objective is to maximize the correlation of local parts with the image, simultaneously with the plausibility of their joint geometrical configuration. Instances of the DPM differ in shape model and optimization method used for fitting the model parameters. The Constrained Local Models (CLM) [13,26–29] employ the holistic PCA shape model like the AAMs. While [26] uses a generic optimization method, the works [13,27–29] propose optimization strategies tailored for specific models, which the methods use. For example, [29] proposed a non-parametric representation of the likelihood of landmark configurations and an optimization method resembling the mean-shift algorithm. Unlike the CLM, the Active Shape Models (ASMs) [30] separate the correlation of the local parts with the image and the regularization via a global PCA shape model into separate processes.

A specific category addressed in this paper are the DPM using a tree based graphical model to encode the shape prior [3,4]. The tree based DPM use relatively weak shape prior which can possibly result to anthropologically implausible landmark configurations. On the other hand, the weak shape prior requires less shape variation in the training examples and, most importantly, it allows to find the globally optimal landmark configuration by dynamic programming. The global optimization makes the method independent of an initial estimate which is the biggest advantage over other approaches. In addition, a mixture of DPM allows to model a large range of view angles in a principled way. A notable disadvantage is the high computational demand connected with the search for a globally optimal solution. The model parameters are typically trained from annotated examples by discriminative approaches [9,16], however generative methods can be used as well [31].

3. The proposed multi-view landmark detector

The tree based DPM approach [3,4,9,32] translates the estimation of landmark positions into an energy minimization problem. We follow this scheme by introducing a scoring function which is to be maximized w.r.t. the landmark positions and the viewing angle. The shape model is represented by an undirected graph $G=(V,E)$, where V is a finite set of vertices representing the landmarks and $E \subset \binom{V}{2}$ is a set of edges between pairs of landmarks, whose positions are related.[1] Examples of particular graphs used in the proposed detector are shown in Figs. 6 and 11.

Let $I \in \mathcal{I}^{H\times W}$ be a fixed-size image (denoted as the *normalized frame* in the sequel), let $\phi \in \Phi$ be a discretized yaw angle corresponding to a particular view, let $\boldsymbol{s}=(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{|V|-1}) \in (H\times W)^{|V|}$ be a configuration of landmark locations and, finally, let $\boldsymbol{w}$ denote the vector of parameters composed of parameters $\boldsymbol{w}_i^{\phi q} \in \mathbb{R}^{n_i^{\phi q}}$ and $\boldsymbol{w}_{ij}^{\phi g} \in \mathbb{R}^{n_{ij}^{\phi g}}$ ($n_i^{\phi q}$ and $n_{ij}^{\phi g}$ denote the number of parameters) associated with the unary and pair-wise potentials, respectively. Then, the scoring function and the proposed detector $h: I^{H\times W} \rightarrow \Phi \times \boldsymbol{S}$, are defined as follows:

$$f(I,\phi,\boldsymbol{s};\boldsymbol{w}) = \sum_{i \in V} q_i^{\phi}\left(\boldsymbol{s}_i, I; \boldsymbol{w}_i^{\phi q}\right) + \sum_{(i,j)\in E} g_{ij}^{\phi}\left(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{w}_{ij}^{\phi g}\right)$$

$$h(I;\boldsymbol{w}) = \arg \max_{\phi \in \Phi, \boldsymbol{s} \in \mathcal{S}} f(I,\phi,\boldsymbol{s};\boldsymbol{w}). \quad (1)$$

The first part of the scoring function, denoted as the *appearance model*, is composed of unary potentials $q_i^{\phi}(\boldsymbol{s}_i, I; \boldsymbol{w}_i^{\phi q})$ measuring the quality of the fit of individual landmark positions $\boldsymbol{s}_i$, $i \in V$, to the image I. The second part, denoted as the *deformation cost*, is composed of pair-wise potentials $g_{ij}^{\phi}(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{w}_{ij}^{\phi g})$ measuring the likelihood of the mutual position of the connected pairs of landmarks.

The normalized frame, serving as an input of the detector, is constructed from a response of a face detector. The face detector provides an estimate of the position, the scale and the in-plane rotation of the face. In order to compensate the imprecision of the face detector, we extend the face box by a multiple of its size. Finally, we apply a similarity transformation to obtain the normalized frame of a fixed size. The process of preparing the normalized frame is illustrated in Fig. 2.

The landmark configuration $\boldsymbol{s}$ is restricted to be from a predefined area, $\boldsymbol{s} \in \mathcal{S} = \mathcal{S}_0 \times \ldots \times \mathcal{S}_{|V-1|}$, where $\mathcal{S}_i \subset \{1,\ldots,H\} \times \{1,\ldots,W\}$ denotes the *search space* of the *i*-th landmark serving as a hard constraint on the landmark location.

3.0.1. Appearance models

The appearance model is a linearly parameterized function

$$q_i^{\phi}\left(\boldsymbol{s}_i, I; \boldsymbol{w}_i^{\phi q}\right) = \left\langle \boldsymbol{w}_i^{\phi q}, \boldsymbol{\Psi}_i^{\phi q}(I, \boldsymbol{s}_i)\right\rangle, \quad (2)$$

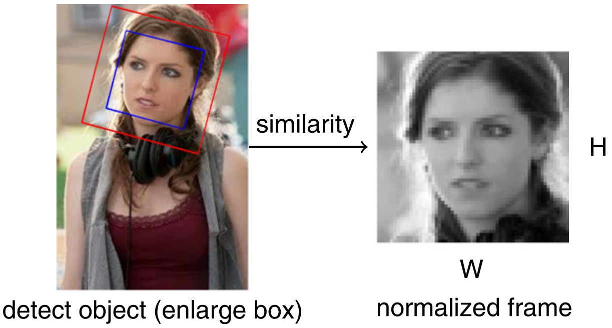


Fig. 2. The acquisition of the normalized frame. Blue box is a detection as provided by the face detector, red box is the detection box enlarged by a defined margin. The similarity transformation (removing the possible in-plane rotation and scaling the image to a fixed size) is applied on the red box and the normalized frame is obtained.

where $\boldsymbol{\Psi}_i^{\phi q}(I, \boldsymbol{s}_i) : \mathcal{I} \times \mathcal{S}_i \rightarrow \mathbb{R}^{n_i^{\phi q}}$ denotes a feature descriptor of a patch cropped from the image I around the position $\boldsymbol{s}_i$. Our approach allows to use an arbitrary feature descriptor. We have experimented with several descriptors including normalized intensity values, their derivatives and HOGs [33]. In the experiments, we use the multi-scale pyramid of Sparse Local Binary Patterns (S-LBP) [34,16], mainly because of a favorable trade-off between the speed and the resulting localization accuracy. The weight vectors $\boldsymbol{w}_i^{\phi q} \in \mathbb{R}^{n_i^{\phi q}}$, $i \in V$, are learned from examples.

The S-LBP descriptor evaluates standard 3 × 3 Local Binary Pattern (LBP) [15] in each position of the original patch. Each 8bit LBP code is represented by a binary vector composed of all zeros and a single one, whose position is determined by the LBP code. Then the patch is downscaled by a factor of two and the LBPs are computed again in all positions. This process is repeated until the resolution of the downscaled patch is below 3 × 3 pixels. The resulting sequence of binary vectors is concatenated to a column vector forming the final descriptor. The resulting sparse high-dimensional S-LBP descriptor can be best represented by indices of its components equal to one. To give an example of its dimensionality, let us consider a patch of size 15 × 15 pixels. The number of all 3 × 3 pixels sub-windows in all levels of the scale pyramid is $13\times13+5\times5+1\times1=195$. Since each LBP is represented by a 256-dimensional binary vector, the resulting descriptor has $n_i^{\phi q}=195\cdot256=49{,}920$ components.

Finding the optimal landmark location requires computation of the S-LBP features in patches centered in all searched positions. A naïve implementation results in a large number of repetitive evaluations of the base LPB feature descriptor since the search patches are highly overlapped. We propose to pre-compute the base LBP in all scales of the entire normalized frames. The resulting LBP codes are represented in the form of MIPMAP [7], which allows efficient indexing of corresponding features in different scales. The final S-LBP descriptor is then compiled from the MIPMAP on the fly (See Fig. 3.). This approach makes the feature computation independent of the number of sought landmarks (assuming that the computational demand of feature compilation can be neglected), leading to about 40% speedup compared to the naïve implementation. More importantly, this approach allows us to share the pre-computed features among different views making the final structured classifier only sub-linearly slower compared to the naïve strategy evaluating individual DPM detector from a scratch. Note that the feature descriptor evaluated via the MIPMAP representation is not exactly the same as the original S-LBP descriptor. Using the MIPMAP representation leads to skipping some base LBP features computed in lower scales, however, we found that it has no impact on the detectors accuracy.

[1] The notation $\binom{V}{2}$ means a set of edges of a fully connected graph with nodes V.

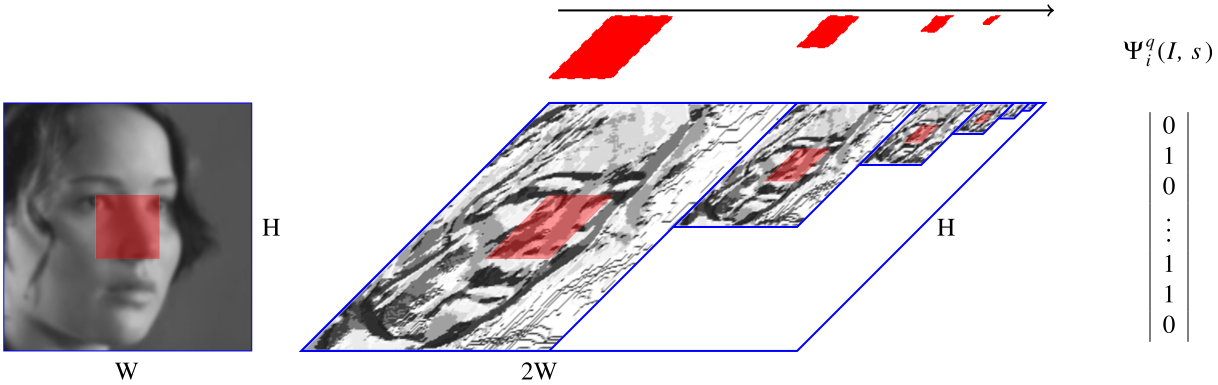


Fig. 3. Features are pre-computed in all positions and scales of the normalized frame and stored in form of a MIPMAP. The final S-LBP descriptor $\Psi_i^{\phi q}(I,\boldsymbol{s})$ is compiled from the MIPMAP on the fly by stacking the corresponding features.

3.0.2. Deformation costs

The deformation cost is also a linearly parametrized function

$$g_{ij}^{\phi}\left(\boldsymbol{s}_i,\boldsymbol{s}_j;\boldsymbol{w}_{ij}^{\phi g}\right) = \left\langle \boldsymbol{w}_{ij}^{\phi g}, \boldsymbol{\Psi}_{ij}^{\phi g}(\boldsymbol{s}_i,\boldsymbol{s}_j)\right\rangle, \tag{3}$$

where $\boldsymbol{\Psi}_{ij}^{\phi g}(\boldsymbol{s}_i,\boldsymbol{s}_j) : \mathcal{S}_i \times \mathcal{S}_j \rightarrow \mathbb{R}^{n_{ij}^{\phi g}}$, which similar to [4], is defined as a quadratic function of the displacement vector, namely,

$$\boldsymbol{\Psi}_{ij}^{\phi g}(\boldsymbol{s}_i,\boldsymbol{s}_j) = \begin{bmatrix} \delta x \\ \delta y \\ \delta x^2 \\ \delta y^2 \end{bmatrix}, \quad \text{where} \quad \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = \boldsymbol{s}_i - \boldsymbol{s}_j = \begin{bmatrix} x_i - y_i \\ x_j - y_j \end{bmatrix}. \tag{4}$$

The $n_{ij}^{\phi g} = 4$-dimensional parameter vectors $\boldsymbol{w}_{ij}^{\phi g} \in \mathbb{R}^{n_{ij}^{\phi g}}$, $(i,j) \in E$, are learned from examples.

The main advantage of having the deformation cost in the form of a separable quadratic function is the possibility to use the distance transform [6] to solve the max-sum problem (1) in time depending linearly on the number of searched positions. The only requirement for application of the distance transform is the concavity of the functions g_{ij}^{ϕ}. By examining the principal minors of the matrix form of g_{ij}^{ϕ}, we see that this can be enforced by linear constraints involving $\boldsymbol{w}_{ij}^{\phi g}$. In particular, we need to keep the 3rd and the 4th components of all 4-dimensional vectors $\boldsymbol{w}_{ij}^{\phi g}$, $(i,j) \in E$, negative. We denote the corresponding set of indices of the 3rd and the 4th components of $\boldsymbol{w}_{ij}^{\phi g}$, $(i,j) \in E$, within the joint parameter vector $\boldsymbol{w}$ by symbol J^-.

3.1. Learning of parameters of the detector by the SO-SVM algorithm

Thanks to the used parameterization of the unary and pair-wise potentials, the proposed DPM detector (1) is an instance of a linear classifier. Therefore we can learn the parameters by the SO-SVM framework [5]. The *joint parameter vector* $\boldsymbol{w}$ to be learned is given by a concatenation of parameter vectors of individual appearance models $\boldsymbol{w}_i^{\phi q}$, $i \in V$, as well as parameter vectors of all deformation costs $\boldsymbol{w}_{ij}^{\phi g}$, $(i,j) \in E$. We define a joint feature map $\boldsymbol{\Psi}(I,\phi,\boldsymbol{s})$ as a concatenation of the feature maps $\boldsymbol{\Psi}_i^{\phi q}(I,\boldsymbol{s}_i)$, $i \in V$, and $\boldsymbol{\Psi}_{ij}^{\phi g}(\boldsymbol{s}_i,\boldsymbol{s}_j)$, $(i,j) \in E$. It is seen that with these definitions, the scoring function of the detector (1) can be written as a dot product of the joint parameter vector and the joint feature map, such that $f(I,\phi,\boldsymbol{s};\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{\Psi}(I,\phi,\boldsymbol{s})\rangle$.

The SO-SVM algorithm translates the learning of the parameter vector of a linear structured classifier into the following convex program

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} F(\boldsymbol{w}) := \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} r_i(\boldsymbol{w})\right] \tag{5}$$

$$\text{s.t.} \quad w_i \le c^-, i \in J^-.$$

where $r_i(\boldsymbol{w})$ is a loss incurred by the classifier on the i-th training example $(I^i, \phi^i, \boldsymbol{s}^i)$ and $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ is a quadratic regularizer introduced to prevent over-fitting. The optimal setting of the regularization constant $\lambda > 0$ is tuned on a validation set. Recall, that the inequality constrains are used to ensure the concavity of functions g_{ij}^{ϕ}. To this end, we set c^- to a small negative constant. The loss $r_i(\boldsymbol{w})$ is the margin-rescaling convex proxy (c.f. [5]) of the true loss $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$ and it reads

$$r_i(\boldsymbol{w}) = \max_{\phi \in \Phi, \boldsymbol{s} \in \mathcal{S}} \left[\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}') + \left\langle \boldsymbol{w}, \boldsymbol{\Psi}\left(I^i,\phi,\boldsymbol{s}\right)\right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{\Psi}\left(I^i,\phi^i,\boldsymbol{s}^i\right)\right\rangle\right]. \tag{6}$$

The form of the true loss $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$ is discussed later in Section 3.1.1. Evaluation of the proxy loss $r_i(\boldsymbol{w})$ amounts to running the classifier with the scoring function augmented by the true loss $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$.

We solve the problem (5) approximately by the Bundle Methods for Regularized Risk Minimization (BMRM) algorithm [35], which we have slightly modified to accept the inequality constraints on $\boldsymbol{w}$. The BMRM algorithm is outlined in Algorithm 1. The core idea is to approximate the original hard problem (5) by a reduced problem

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} F_t(\boldsymbol{w}) := \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + r_t(\boldsymbol{w})\right] \tag{7}$$

$$\text{s.t.} \quad w_i \le c^-, i \in J^-,$$

whose objective function $F_t(\boldsymbol{w})$ is obtained by replacing the risk term $r(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} r_i(\boldsymbol{w})$ by its cutting plane model

$$r_t(\boldsymbol{w}) = \max_{i=0,1,\ldots,t-1} \left[r(\boldsymbol{w}_i) + \langle \boldsymbol{r}'(\boldsymbol{w}_i), \boldsymbol{w} - \boldsymbol{w}_i\rangle\right] \tag{8}$$

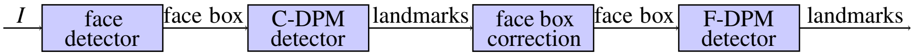


Fig. 4. The figure visualizes the proposed coarse-to-fine strategy used to improve localization accuracy and to keep the processing time of the DPM based detector low. The C-DPM detector operates on a low dimensional image which is localized by a face detector. The resulting rough estimate of the landmarks helps to obtain a corrected face localization. The corrected face localization allows to compute narrow search spaces of the F-DPM operating on higher resolution images.

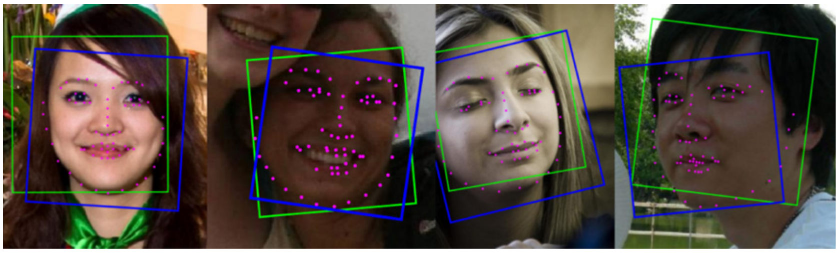


Fig. 5. A few examples of the corrected face box computed from the response of the C-DPM. The original face box detected by face detector is green, the corrected face box is blue. The landmark positions predicted by C-DPM and used for the face box correction are depicted in magenta color.

where $\mathbf{r}'(\mathbf{w}_i) \in \mathbb{R}^n$ denotes a sub-gradient of $r(\mathbf{w})$ evaluated at $\mathbf{w}_i \in \mathbb{R}^n$. Starting from an initial guess $\mathbf{w}_0 = 0$, the BMRM algorithm computes a new iterate $\mathbf{w}_t$ by solving the reduced problem (7). In each iteration t, the cutting plane model (8) is updated by a new cutting plane computed at the intermediate solution $\mathbf{w}_t$ leading to a progressively tighter approximation of $F(\mathbf{w})$. It was proved, that the BMRM converges to an ϵ-precise solution satisfying $F(\mathbf{w}_t) \le F(\mathbf{w}^*) + \varepsilon$ in $\mathcal{O}(\frac{1}{\epsilon})$ iterations for arbitrary $\epsilon > 0$. The BMRM accesses the objective via the first order oracle, which for a given query $\mathbf{w}_t$ evaluates $r(\mathbf{w}_t)$ and the sub-gradient $\mathbf{r}'(\mathbf{w}_t)$. Components of the sub-gradient $\mathbf{r}'(\mathbf{w}_t) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{r}'_i(\mathbf{w}_t)$ can be computed by the Danskin's theorem (see e.g. [36]) as follows:

$$\mathbf{r}'_i(\mathbf{w}_t) = \mathbf{\Psi}\left(I^i, \hat{\phi}, \hat{\mathbf{s}}\right) - \mathbf{\Psi}\left(I^i, \phi^i, \mathbf{s}^i\right),$$

where

$$\left(\hat{\phi}, \hat{\mathbf{s}}\right) = \arg \max_{\phi \in \Phi, \mathbf{s} \in \mathcal{S}} \left[\Delta^{\phi,\mathbf{s}}(\phi, \mathbf{s}, \phi', \mathbf{s}') + \left\langle \mathbf{w}, \mathbf{\Psi}\left(I^i, \phi, \mathbf{s}\right)\right\rangle\right].$$

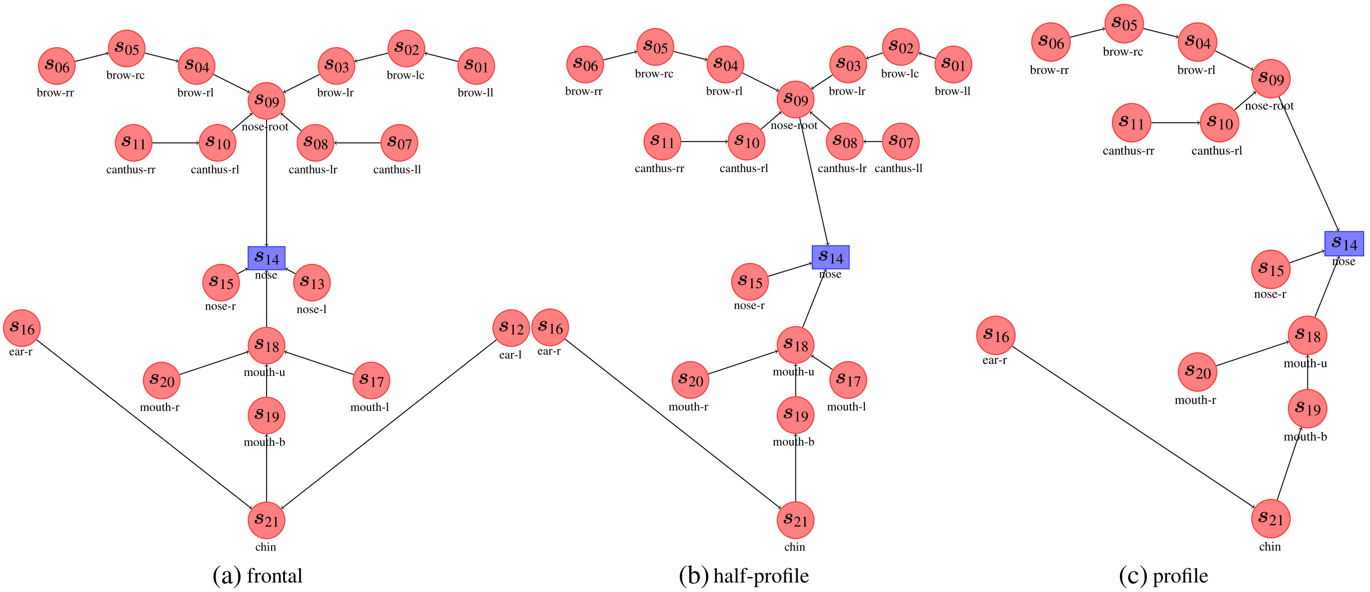
The BMRM translates the original problem (5) to a sequence of reduced problems (7). The reduced problem can be expressed as an equivalent convex quadratic program, the dual form of which has only t variables. Hence the reduced problem is amenable by off-the-shelf QP solvers. The computational bottleneck is the evaluation of risk $r(\mathbf{w})$ and its sub-gradient $\mathbf{r}'(\mathbf{w})$. Fortunately, both quantities are sums of simpler terms, hence their evaluation can be efficiently parallelized.

Algorithm 1. BMRM algorithm.

```
Require: $\epsilon$, first order oracle evaluating $r(\boldsymbol{w})$ and $\boldsymbol{r}'(\boldsymbol{w})$
Initialization: $\boldsymbol{w} \leftarrow \mathbf{0}, t \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    Call oracle to compute $r(\boldsymbol{w}_t)$ and $\boldsymbol{r}'(\boldsymbol{w}_t)$
    Update the cutting plane model $r_t(\boldsymbol{w}_t)$
    Solve the reduced problem (7)
until $F(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_t) \le \epsilon$
```

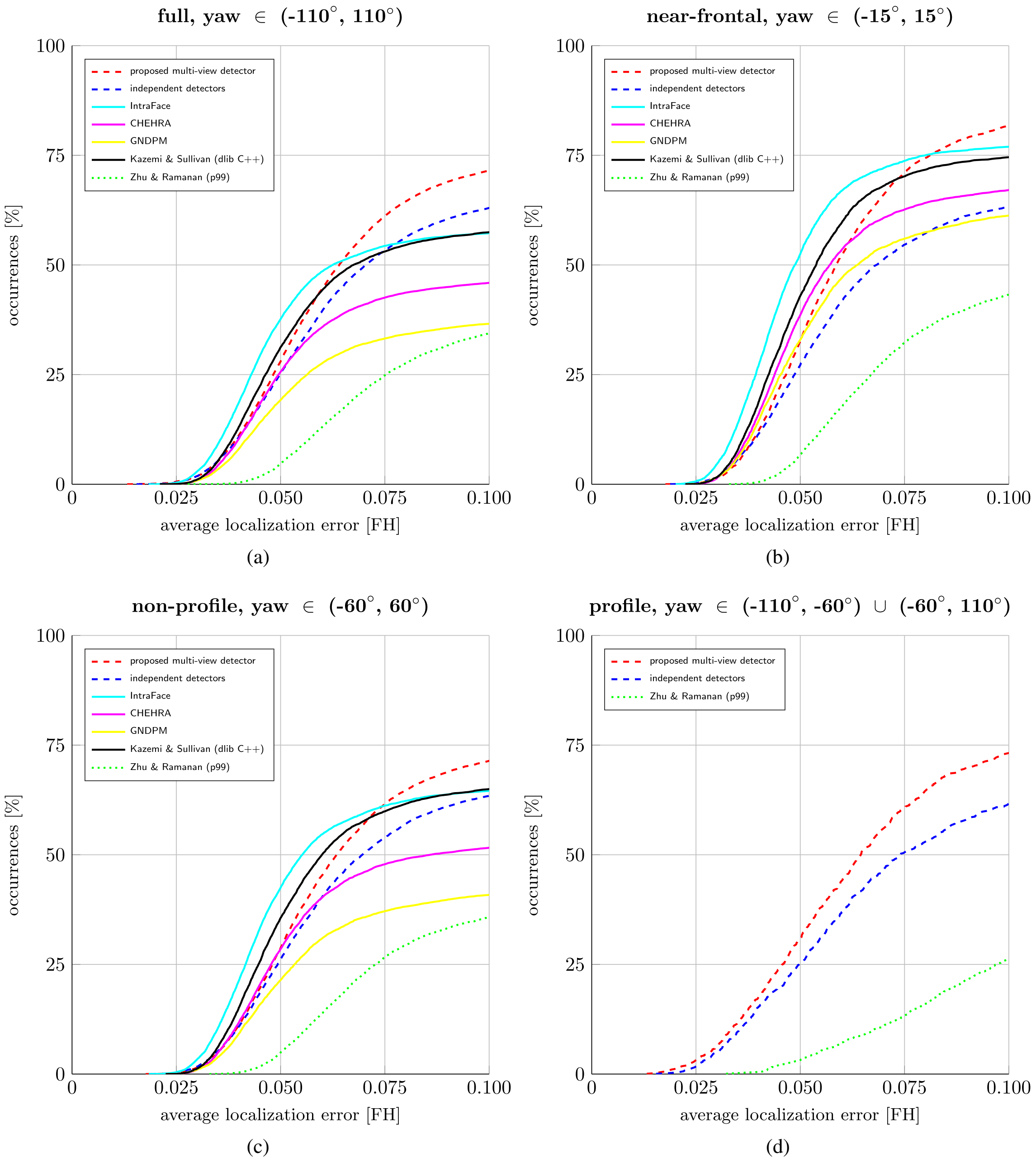
3.1.1. Loss function

The learning algorithm (5) optimizes a convex surrogate of the true loss $\Delta^{\phi,\mathbf{s}}(\phi, \mathbf{s}, \phi', \mathbf{s}')$. The loss $\Delta^{\phi,\mathbf{s}}(\phi, \mathbf{s}, \phi', \mathbf{s}')$ is designed to measure a



(a) frontal (b) half-profile (c) profile

Fig. 6. The underlying graph structure of individual view-specific detectors, depicted for positive viewing angles (negative views are mirrored). The vertices are represented by the red circles, edges by black arrows connecting them. Each vertex is denoted by its identification number and the name of the corresponding landmark. All graphs are trees which allows to solve the inference problem globally by the dynamic programming. The root vertex is represented by the blue square. Note that the core of all graphs is the same and just the self-occluded landmarks with their incident edges are removed in non-frontal cases.



(a)

(b)

(c)

(d)

Fig. 7. Cumulative histograms of the average localization error measured on the testing subset of the AFLW dataset. The localization error is normalized by the face height computed as a distance between the root of the nose and the chin. Individual sub-figures contain error measured on a subset of test images with the ground truth yaw in a corresponding range.

discrepancy between the true and the estimated landmark positions on a given training example. We define the loss function as follows

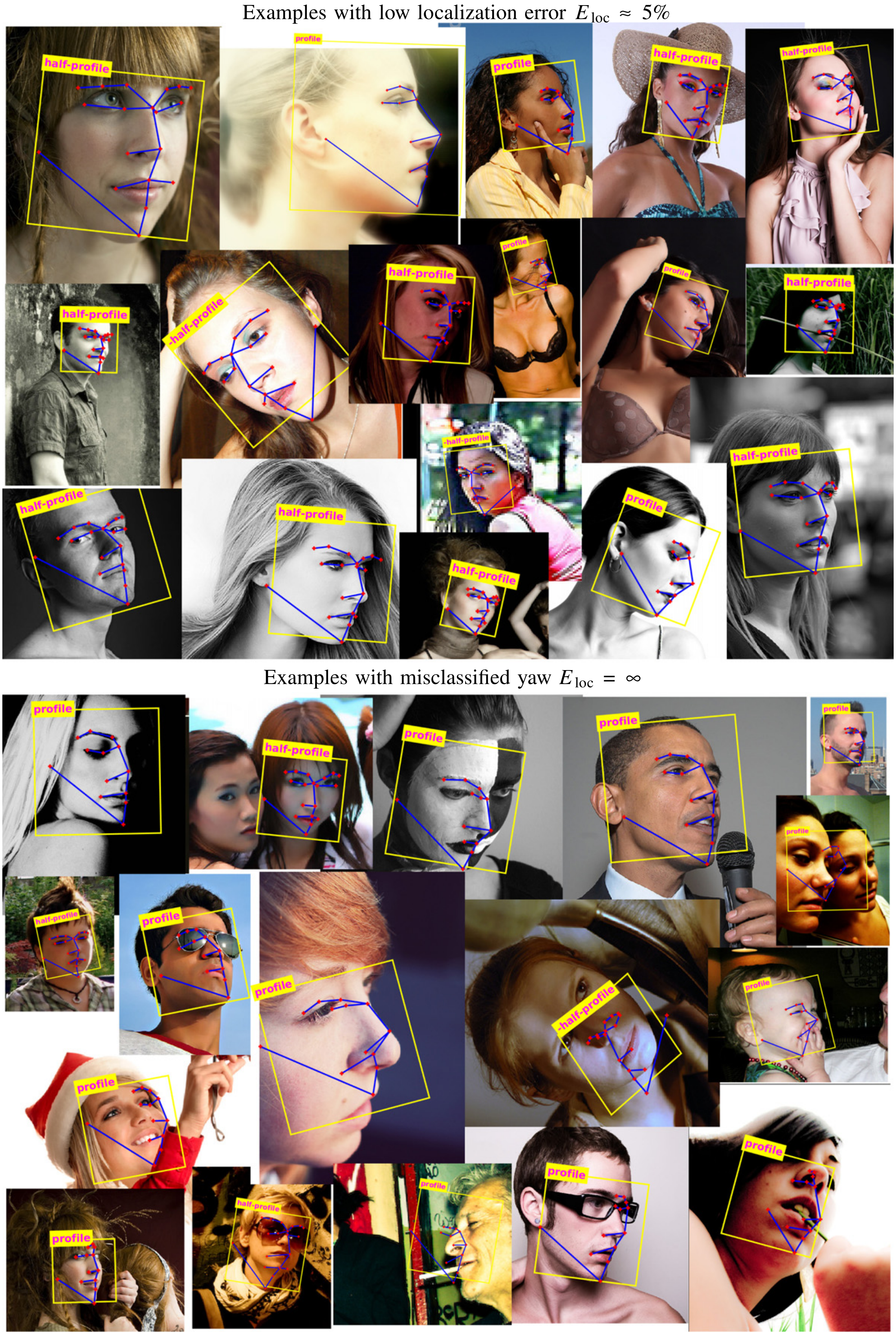
$$\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}') = \begin{cases} \kappa(\boldsymbol{s})\dfrac{1}{|V|}\sum_{j=1}^{|V|}\|\boldsymbol{s}_j - \boldsymbol{s}'_j\|, & \text{if } \phi = \phi' \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

where the normalization constant $\kappa(\boldsymbol{s})$ is a reciprocal to the inter-ocular distance or the distance between the root of the nose and the chin. The inter-ocular distance is a common normalization suitable for near-frontal faces. For large yaw angles the inter-ocular distance goes to zero in which case the root of the nose to chin distance, independent to yaw angle, seems to be a better choice. We use both mentioned normalization constants in experiments. For details see Sections 5.1.1 and 5.3.1.

The penalty for confusing the viewing angle is set to 1, which is much larger value than an acceptable localization error. In turn, the loss function penalizes mistakes in the viewing angle more than the landmark misplacement. Note also that the loss $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$ is non-negative and 0 iff $(\phi,\boldsymbol{s}) = (\phi',\boldsymbol{s}')$ as commonly required by the SO-SVM framework.

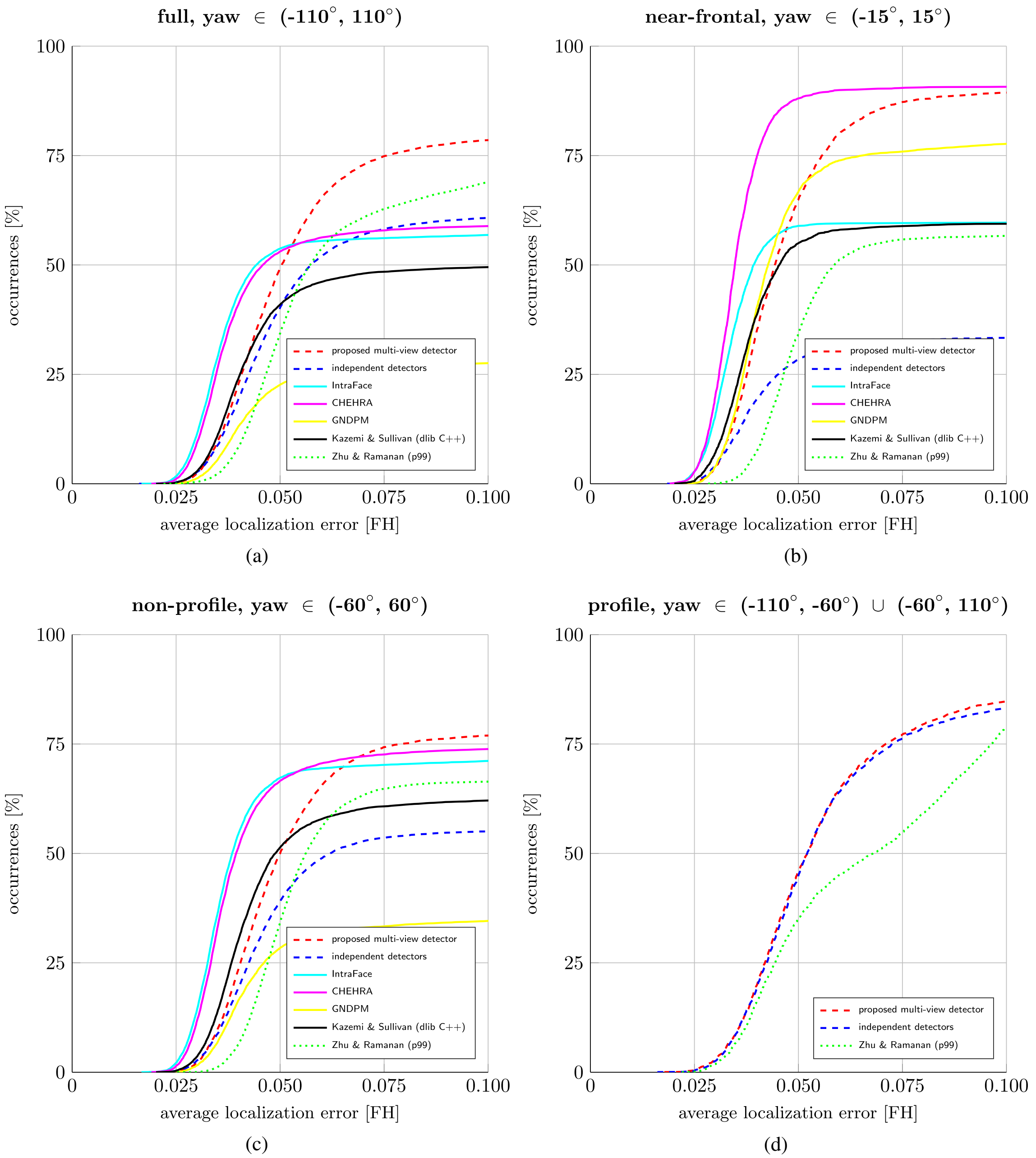
3.2. Inference problem

Evaluation of the detector (1) amounts to solving an instance of the max-sum problem for each view ϕ separately and then taking the landmark configuration with the overall highest score. A tractability of the max-sum problem depends on the graph G. While our framework does not limit the graph structure, for the sake of speed we set the graph G to be a tree. In this case the global solution can be found in polynomial time by the dynamic programing. For a general graph the


Examples with low localization error $E_{\mathrm{loc}} \approx 5\%$

Examples with misclassified yaw $E_{\mathrm{loc}} = \infty$

Fig. 8. Exemplary images from the AFLW testing set with the average localization error not higher than $E_{\mathrm{loc}} \approx 5\%$ (top), and with the misclassified yaw angle $E_{\mathrm{loc}} = \infty$ (bottom). The yellow box represents face detection as provided by the face detector (i.e. the input of proposed detector), the discretized yaw category is written on the top edge of the face box. The landmarks are denoted by red crosses. The underlying graph corresponding to the yaw category is shown by blue lines connecting landmarks.



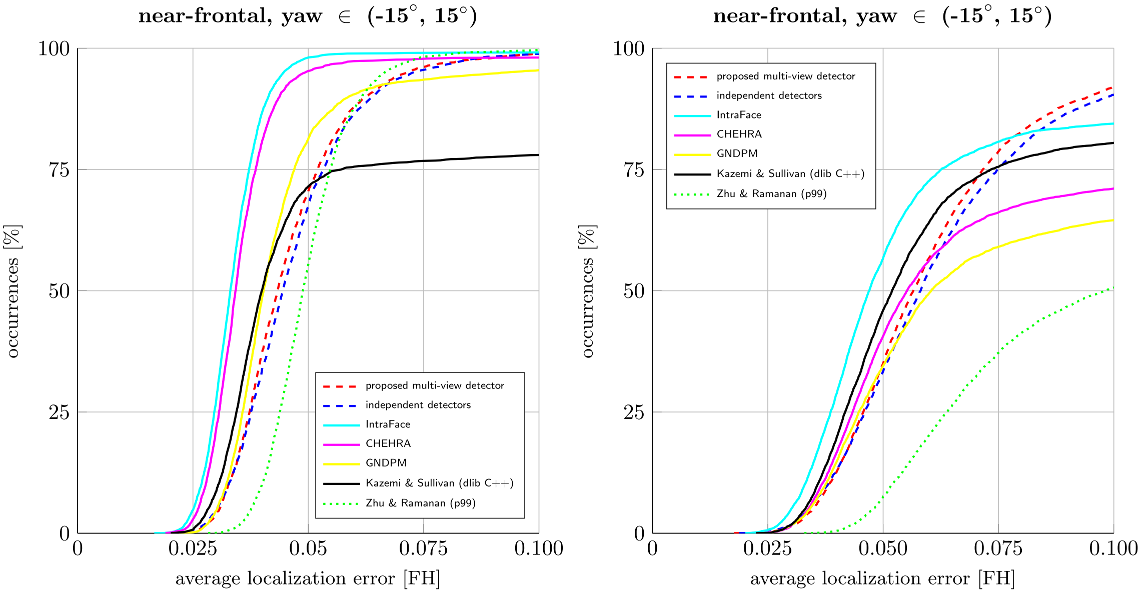
(a) (b) (c) (d)

Fig. 9. Cumulative histograms of the average localization error measured on the testing subset of the Multi-PIE dataset. The localization error is normalized by the face height computed as a distance between the root of the nose and the chin. Individual sub-figures contain error measured on a subset of test images with the ground truth yaw in a corresponding range.

max-sum problem is known to be NP-hard. The dynamic programing solver proceeds as follows. The graph is topologically sorted so that the root landmark is evaluated last. Then the leafs are evaluated (that is, the maximizing label is found for each possible label in the adjacent node) and cut away from the graph. The same evaluation procedure propagates till the root landmark is solved. The evaluation in the root landmark provides the optimal value and the optimal landmark position at the root. The remaining landmark positions are obtained by backtracking. For more details see e.g. [37].

The computational time required by a plain dynamic programming scales quadratically with the number of searched landmark positions (given by $\mathcal{S}_0 \times \cdots \times \mathcal{S}_{|V-1|}$) and linearly with the number of landmarks. This starts to be impractical for high resolution normalized frames and a dense set of landmarks, which is the case of the 300-W challenge. To alleviate the problem, we use the generalized distance transform (DT) [6], whose computational time scales only linearly with the number of landmark positions. The DT exploits the fact that the pair-wise potentials are concave separable functions of the x and y coordinates. This allows to perform the maximization over a grid search space by effectively maximizing over x and y coordinates separately. For more details on the distance transform we refer to [6].

The DT can be also used in evaluation of the loss $r_i(\boldsymbol{w})$ and its subgradient $\boldsymbol{r}'_i(\boldsymbol{w})$, which is the computational bottle-neck of the BMRM. Note that the BMRM maintains the non-negativity constraints necessary for the application of the DT during the whole course of the algorithm. The usage of DT leads to a substantial speed-up of the learning procedure. Detailed experimental evaluation of speed-ups due to the usage of DT is discussed in Section 5.2.



(a) Multi-PIE (b) AFLW

Fig. 10. Cumulative histograms of the view insensitive average localization error measured on the near-frontal examples of testing subsets of AFLW and Multi-PIE. The localization error is normalized by the face height computed as a distance between the root of the nose and the chin.

4. Coarse-to-fine strategy to speed up DPM detector

A practical limitation of the DPM detectors is their computational cost scaling with the size of the search spaces of the individual landmark positions, $\mathcal{S}_i$, $i \in V$. The size of the search space is a function of the resolution of the normalized frame and the a priori knowledge of the landmark's position. The a priori landmark position depends on the accuracy and robustness of the used face detector. That is, an imprecise localization provided by the face detector has to be compensated by a large search space, in order not to miss the correct landmark position. The search is done in the normalized frame and the found landmark location is projected into the original image. Therefore the resolution of the normalized frame lower bounds the accuracy of the landmark localization. In turn, improving localization accuracy increases the search time.

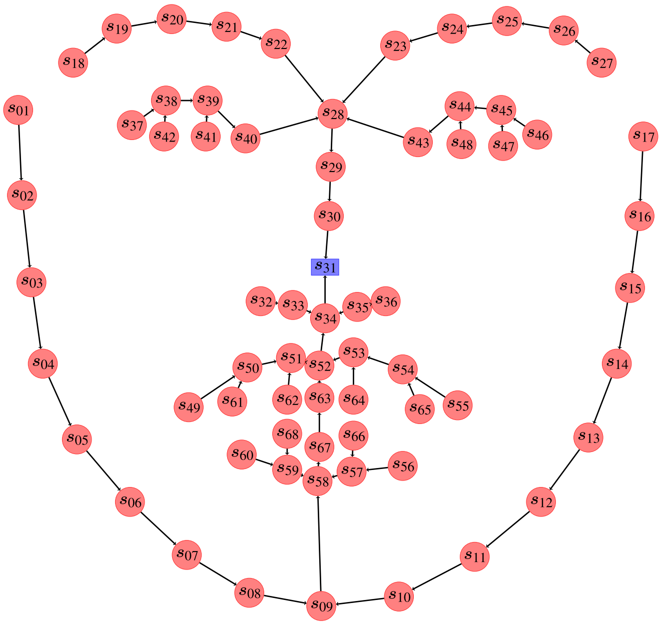


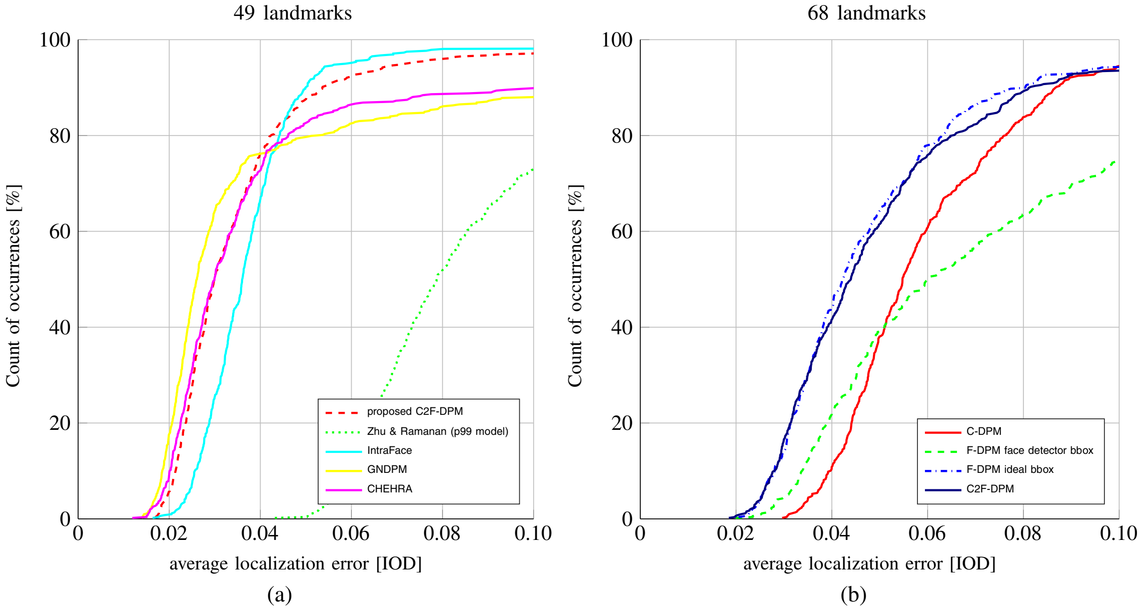
Fig. 11. The underlying graph (V,E) of the proposed DPM detector used on the 300-W benchmark. The C-DPM and the F-DPM detectors use the same graphs. The nodes V are denoted by the red circles (root node is the blue square) while the edges E are shown as black arrows.

To alleviate the problem, we propose a coarse-to-fine strategy (denoted also as C2F-DPM) with two stages. In the first stage, we use a DPM detector, denoted as C-DPM, which operates in a low-resolution normalized frame. The output of the C-DPM detector is used to compute a better estimate of the face location than the one provided by the face detector. Hence, the C-DPM detector serves as a precise face detector. In the second stage we apply a DPM detector, denoted as F-DPM, which searches for the landmarks in a high resolution normalized frame. The initial estimate by the C-DPM allows to set much smaller search spaces in the high resolution normalized frame of the F-DPM detector without a danger of overlooking the landmarks. The scheme of the proposed coarse-to-fine strategy is outlined in Fig. 4.

The precise face box used to initialize F-DPM is constructed from the response of the C-DPM detector as follows. The center of the precise face box is computed as the mean of the estimated landmarks. Then the centers of both eyes $\boldsymbol{C}_l, \boldsymbol{C}_r$ are computed (again as the mean position of the corresponding estimated landmarks). The size of the precise face box is defined as $2.7 \cdot \|\boldsymbol{C}_l - \boldsymbol{C}_r\|_2$. Finally, the in-plane rotation of the precise face box is computed as the deviation of the (least squares optimal) line l fitted to the eyes landmarks and the x-axis. A few examples of the corrected face box are depicted in Fig. 5.

5. Experiments

We split the experiments into two parts. In the first part, Section 5.1, we evaluate the proposed multi-view landmark detector on the AFLW [8] and Multi-PIE [38] datasets. These experiments demonstrate the ability to estimate landmarks robustly in a large range of viewing angles. The measured timing statistics are summarized in Section 5.2. In the second part, Section 5.3, we evaluate the detector on the public part of the 300-W [39,40] datasets and we also present results on the non-public test set obtained from the organizers of the 300-W competition. The experiments in the second part evaluate the ability of the detector



(a)

(b)

Fig. 12. Cumulative histograms of the average localization error evaluated on the public test images of the 300-W dataset. (a) Shows results for the proposed C2F-DPM detector and the compared methods. (b) Depicts the localization error of the F-DPM detector initialized by different approaches (see description in the text).

to estimate a dense set of landmarks from near-frontal faces in high resolution images.

5.1. Multi-view experiments

5.1.1. Implementation details of the proposed multi-view detector

We discretize the viewing (yaw) angle as follows $\Phi = \{$-profile, -half-profile, frontal, half-profile, profile$\}$. For each view we detect a different number of landmarks visible from the given range of yaw angles. The particular ranges of the yaw angles and the corresponding number of estimated landmarks are listed in Table 1. The underlying graphs $G_\phi = (V, E)$ of the DPM for individual views are depicted in Fig. 6.

The size of the normalized frame is set to 60 × 60 pixels, which provides sufficient localization accuracy for the AFLW dataset. Since the normalized frame has a relatively small resolution, we do not use the coarse-to-fine strategy in this experiment. A face box found by the face detector is enlarged by a factor of 1.5 and the enlarged box is affinely transformed into the normalized frame. In particular, we use a commercial implementation[2] of the Wald-Boost face detector proposed in [41]. In the case when the face detector fails to find a face, the detector returns empty set symbols ($\boldsymbol{s} = \emptyset, \phi = \emptyset$) to denote the failure (c.f. definitions of the evaluation metrics in Section 5.1.2). We use the S-LBP features computed by the MIPMAP for the appearance models (2). Each landmark's descriptor was computed from a patch of size 9 × 9 pixels except for the root landmark, the tip of the nose, which used a bigger patch of 15 × 15 pixels. As the deformation cost (3) we use a separable quadratic function of the displacement vector as defined in Eq. (4). This leads to the overall dimensionality of the joint parameter vector $\boldsymbol{w} \in \mathbb{R}^n$ equal to $n = 1{,}335{,}360$. As the normalization factor $\kappa(\boldsymbol{s})$, present in the true loss (9), we use the face size computed as the distance between the root of the nose and the chin (namely, the distance $\|\boldsymbol{s}_{09} - \boldsymbol{s}_{21}\|$ using the notation from Fig. 6).

The entire learning procedure composed of tuning the regularization constant λ took around 5 days on a machine with a 12 cores CPU.

5.1.2. Datasets and the evaluation protocol

We use the AFLW [8] database for both training and evaluation and the Multi-PIE [38] database just for the evaluation. Both datasets come with annotation of 21 facial landmarks (see Fig. 6(a)). We used a subset of 12,525 images from the Multi-PIE for which we have precise ground truth annotation. The original AFLW database consists of 24,686 images, however, the annotation of a large number of images is either inconsistent (confused landmarks) or imprecise. In order to correct the annotation, we fitted a 3D face model proposed in [42] to the manually annotated landmarks. The projected landmarks of the 3D model were then manually inspected and corrected when necessary. The process reduced the number of images to 21,688 (mainly due to failures of the face detector involved in the semi-automatic annotation procedure), but it significantly improved the quality of ground truth annotation.

We randomly selected ≈25% of images for training, ≈10% for validation and ≈65% for testing. The number of training examples is relatively small taking into account the number of model parameters, which is $\dim(\boldsymbol{w}) = 1{,}335{,}360$. Surprisingly, the test accuracy of the learned detector is quite high, which we attribute to the generalization ability of the SO-SVM algorithm.

Once the joint parameter vector $\boldsymbol{w}$ is learned, we evaluate the detector on the hold out test examples. For the evaluation we use three metrics, namely, the average localization error E_{loc} (sometimes also called point-to-point error), the yaw misclassification rate E_{yaw} and the face detector error E_{fd} defined as follows

$$E_{\text{loc}}(\phi, \boldsymbol{s}, \phi^*, \boldsymbol{s}^*) = \begin{cases} \dfrac{\kappa(\boldsymbol{s})}{|V|} \displaystyle\sum_{j=1}^{|V|} \|\boldsymbol{s}_j - \boldsymbol{s}_j^*\|, & \text{if } \phi = \phi^* \\ \infty, \text{if } \boldsymbol{s} = \emptyset \quad \text{or} & \phi \neq \phi^* \end{cases} \tag{10}$$

$$E_{\text{yaw}} = \frac{1}{m} \sum_{i=i}^{m} [\![\phi_i \neq \phi_i^*]\!] \ , \tag{11}$$

$$E_{\text{fd}} = \frac{1}{m} \sum_{i=i}^{m} [\![\phi_i = \emptyset]\!] \ , \tag{12}$$

where the brackets $[\![\cdot]\!]$ denote the Kronecker delta, $(\phi_i, \boldsymbol{s}_i)$ is the detector response on the i-th test image and $(\phi_i^*, \boldsymbol{s}_i^*)$ denotes the ground truth annotation. We report the cumulative histogram of the average localization error E_{loc} and the single number statistics E_{yaw}, E_{fd}. Since most of the competing detectors come with their own integrated face

[2] Courtesy of Eyedea Recognition Ltd., http://www.eyedea.cz/.

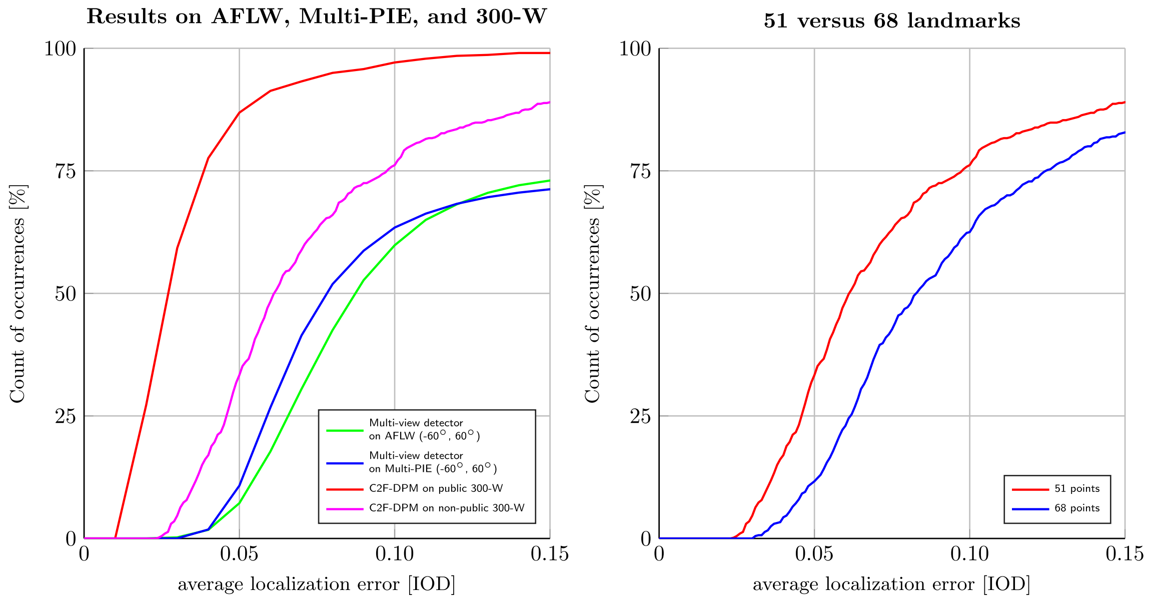


Fig. 13. (a) Shows comparison of the proposed detectors on all four test datasets. Results on non-profile facial images of the AFLW and the Multi-PIE are processed by the multi-view detector estimating 21 landmarks. Results on the public and the non-public part of the 300-W are obtained by the C2F-DPM detector estimating 51 landmarks. (b) Shows difference in the localization accuracy evaluated on all 68 landmarks compared to the subset of 51 landmarks not containing the contours.

detectors, we cannot separate contributions of the face detector failure from the definition of the evaluation metrics. That is, a face detector failure incurs maximal localization error E_{loc} as well as the yaw misclassification error E_{yaw}.

To avoid a possible confusion of the reader and to make the evaluation comparable with previous works, we also introduce the view insensitive average localization error E_{loc2} defined as

$$E_{\text{loc2}}(\boldsymbol{s},\boldsymbol{s}^*) = \frac{\kappa(\boldsymbol{s})}{|D|}\sum_{j}^{|D|}\|\boldsymbol{s}_j - \boldsymbol{s}_j^*\|. \qquad (13)$$

In contrast to (10), $E_{\text{loc2}}(\boldsymbol{s},\boldsymbol{s}^*)$ does not penalize the misclassification of the viewing angle and it computes the point-to-point error only for the set of detected landmarks (denoted as D). The error $E_{\text{loc2}}(\boldsymbol{s},\boldsymbol{s}^*)$ has been the standard evaluation metric used so far, however, it is not suitable for evaluation of truly multi-view detectors which may return different sets of landmarks depending on the viewing angle like the proposed one (e.g. more landmarks are visible for frontal view than for the profiles).

Since we consider the speed of the detector as its important aspect, we measure timings for all stages of the proposed detector and provide a comparison to other methods.

5.1.3. Competing methods

In this section we describe a list of existing methods against which we compare the proposed detector. Namely, we compare to the tree based DPM detector of Zhu & Ramanan [9], which is the most related to our work. Despite a relatively low localization accuracy, it is up to our best knowledge the only publicly available detector working in the full range of the yaw angle. We also compare against the IntraFace [10] detector, considered to be the current state-of-the-art in both precision and speed, the detector of Kazemi & Sullivan [12], Gauss-Newton Deformable Part Models (GN-DPM) [13], and Chehra [11]. In addition, we evaluate a baseline tree based DPM detector composed of a set of independent view-specific detectors in order to demonstrate the advantage of the proposed structured output model.

Table 1
The discretization of the viewing angle (yaw).

Viewing angle names ($\phi\in\Phi$)				
-profile	-half-profile	frontal	half-profile	profile
Viewing angle ranges				
(-110°, -60°]	(-60°, -15°]	(-15°,15°)	[15°,60°)	[60°,110°)
Number of landmarks detected in ϕ				
13	19	21	19	13

To have a fair comparison with the competing methods, we crop the test images around the face box enlarged by 30%. This should minimize the dependency on a specific face detector used by the competing approaches. We also consider only a subset of landmarks common to all methods. In particular, the common subset contains 18 landmarks for the frontal view, leaving out the landmarks representing ears and chin (see Fig. 6a). Not all of the competing methods are capable of estimating the viewing angle. In such case, we use the head pose estimator based on fitting a 3D model to the estimated landmarks [43]. The estimated yaw is then rounded to the intervals defined in Table 1.

5.1.3.1. Independent detectors. To show the benefits of the proposed multi-view detector simultaneously estimating the view angle and the landmark locations, we use the following baseline approach. We learn a set of independent single-view DPM detectors, each for a different view $\phi\in\Phi$. The particular single-view detector is then selected based on the response of the face detector providing a rough estimate of the yaw ϕ. The individual single-view detectors have exactly the same structure and use the same features as the proposed multi-view detector.

5.1.3.2. Detector of Zhu & Ramanan [9]. We use the code provided by the authors with the fully shared model "p99". This detector simultaneously works as the face detector and detector of facial landmarks. The detector returns 68 or 39 landmarks for the near-frontal or profile views, respectively. The Zhu & Ramanan detector uses a part of the Multi-PIE database for training, which is not consistent with our split. Hence the corresponding results on the Multi-PIE dataset can be positively biased.

5.1.3.3. Chehra [11]. We use the implementation of the recently published facial landmark detector provided by the authors. The detector uses a cascade of discriminatively trained regressors estimating the pose and shape parameters of a 3D face model. The detector was trained on the 300-W dataset [40]. The detector returns 49 landmarks.

5.1.3.4. IntraFace [10]. We use the code kindly provided by the authors. This detector uses supervised descent method (SDM) learning a descent direction from the training data in order to minimize an objective function formulated as a non-linear least squares matching of the face model to the image. It detects 49 landmarks. The detector comes with an estimator of the yaw angle. The detector was learned on a subset of Multi-PIE and LFW [44] datasets. Hence the results on the Multi-PIE dataset can be positively biased, since we use a different split.

5.1.3.5. GN-DPM [13]. We use a code provided by the authors. This detector is an instance of a generative DPM, where the optimization of the appearance and global shape model is done simultaneously by the Gauss–Newton algorithm. It detects 49 landmarks. The detector is initialized from a response of the Zhu & Ramanan detector (using the *p*204-*Wild* model). The detector was trained on the LFPW [45] dataset being a part of the 300-W benchmark.

5.1.3.6. Kazemi & Sullivan [12]. We used the implementation from the "dlib C++" library. This detector is based on a gradient boosting of ensemble of regression trees. It detects 68 landmarks. The detector is trained on the iBUG dataset which overlaps with testing part of the 300-W benchmark. Hence we compare this method only on the AFLW and Multi-PIE datasets.

5.1.4. Results

This section summarizes the comparison of the proposed landmark detector with competing methods on the AFLW and Multi-PIE datasets. In Figs. 7 and 9, we show the cumulative histograms of the mean localization error E_{loc} evaluated on the test images of the AFLW and the Multi-PIE dataset, respectively. Besides statistics computed on all test images, we also evaluate the detectors on subsets of test images with a specified range of the ground truth yaw angle:

- **near-frontal** images with $\phi \in (-15^{\circ}, 15^{\circ})$
- **non-profile** images with $\phi \in (-60^{\circ}, 60^{\circ})$
- **profile** images with $\phi \in (-110^{\circ}, -60^{\circ}) \cup (60^{\circ}, 110^{\circ})$

Most of the detectors are not trained or designed for the full range of yaw angle and thus they fail on the profile images. For this reason, we show results on the profile subset only for the proposed detector, the baseline independent DPM detector and the detector of [9], all operating in the full range of yaw angle. In Table 2, we also show the yaw misclassification error E_{yaw} and the face detector error E_{fd} on the non-profile images.

The results demonstrate that the proposed detector has consistently good localization accuracy in all views. For near-frontal and non-profile faces, the proposed detector does not provide the smallest localization error ($E_{\text{loc}} < 0.075$) with the highest frequency but it dominates the other methods in the regime with a tolerable error ($E_{\text{loc}} > 0.075$). This behavior is consistent over the AFLW and the Multi-PIE dataset. On profile images, the proposed detector significantly outperforms the only full multi-view competitor [9]. The proposed detector is also consistently better than the baseline composed of independent DPM detectors which demonstrates the benefits of using the structured classifier over the independent estimate of the view and the landmark locations.

The presented comparisons has the following limitations. The main drawback is that the evaluation criteria E_{loc} combines the landmark localization error with the face detector error E_{fd} which, as can be seen from Table 2, is non-negligible. Moreover, on the AFLW dataset our face detector has 0% error because the same detector was involved in the semi-automatic annotation procedure and hence the results of our method are positively biased. This is not the case on the Multi-PIE where, however, our face detector has also significantly lower error. On the other hand, our method and the detector of [9], unlike other competitors returning a single landmark set for all views, are additionally penalized for the incorrect yaw estimates which also contributes to the overall localization error E_{loc} (c.f. Eq. (10)). Another drawback is an inconsistent training set used by different method. The mentioned deficiencies are mitigated by using the 300-W benchmark as described in Section 5.3.

In order to make our evaluation compatible with previous works, we also evaluate the methods using the view insensitive localization error E_{loc2}. The cumulative histograms of E_{loc2} obtained on the near-frontal examples of the testing subset of both AFLW and Multi-PIE are shown in Fig. 10. It is seen that if the set of landmarks to be estimated is known a priori, the performance of all methods improves significantly. The best performing method is the IntraFace on both datasets. The results of the proposed DPM based detector remain competitive. For completeness, Table 3 summarizes the count of intrinsic face detector failures for each method, obtained on the near-frontal examples of the testing subset of both AFLW and Multi-PIE. Note that our commercial face detector is significantly outperforming all competitors, especially on the AFLW database.

Fig. 8 shows exemplary outputs of the proposed detector on a sample of test images from the AFLW dataset. We show both examples with small localization error, $E_{\text{loc}} \approx 5\%$, and the highest error, $E_{\text{loc}} = \infty$, that is, images on which the yaw estimate failed.

5.2. Timing

Table 4 presents average times required by competing methods to process a single image. The time is measured on the cropped images containing only the face in order to decrease time spent in the face detector which is an integral part of the methods [10,9,12]. We do not count initialization time and, if possible, we subtract the face detector time [12].

The fastest among the compared methods is the independent DPM detector using an external method for the yaw estimate. Otherwise, the proposed DPM detector is consistently significantly faster (by an

Table 2
The yaw mis-classifications error E_{yaw} and the face detector error E_{fd} for non-profile testing examples (that is, images with ground truth yaw in the range (-60°,60°)) from AFLW and Multi-PIE datasets, respectively.

	Database			
	AFLW		Multi-PIE	
Method	E_{fd}	E_{yaw}	E_{fd}	E_{yaw}
Proposed	0.00%	23.61%	0.12%	22.27%
Independent detectors	0.00%	30.34%	0.12%	44.08%
Zhu & Ramanan [9]	35.60%	56.47%	0.12%	33.43%
CHEHRA [11]	25.30%	40.52%	19.99%	25.39%
IntraFace [10]	19.10%	32.76%	12.00%	28.01%
Kazemi & Sullivan [12]	20.57%	31.80%	10.68%	32.85%
GN-DPM [13]	19.05%	48.09%	0.13%	63.01%

Table 3
Count of face detector failures on AFLW & Multi-PIE databases for near-frontal testing subset.

	Database	
Method	AFLW	Multi-PIE
Proposed	0/5663	0/3747
Independent detectors	0/5663	0/3747
IntraFace	705/5663	26/3747
Chehra	930/5663	48/3747
GNDPM	1050/5663	0/3747
Kazemi & Sullivan	932/5663	218/3747
Zhu & Ramanan	2002/5663	0/3747

Table 4
The average time (in seconds) required by competing methods to process a single face. We show the mean and the standard deviation in seconds computed over the test images. The results are computed separately for each dataset. The "proposed" stands for the multi-view detector on the AFLW (first column) and the Multi-PIE (second column) dataset and C2F-DPM detector on the 300-W dataset (third column), respectively.

	Database		
Method	AFLW	Multi-PIE	300-W
Proposed	0.011 ± 0.005	0.012 ± 0.002	0.1 ± 0.02
Independent detectors	0.003 ± 0.001	0.004 ± 0.001	–
Zhu & Ramanan [9]	60.4 ± 24.0	18.9 ± 11.4	73.9 ± 144.4
Chehra [11]	0.1 ± 0.08	0.2 ± 3.4	0.2 ± 2.6
IntraFace [10]	0.05 ± 0.1	0.03 ± 0.01	0.1 ± 0.2
Kazemi & Sullivan [12]	0.4 ± 0.4	0.4 ± 0.3	–
GN-DPM [13]	0.8 ± 0.4	0.5 ± 0.1	0.6 ± 1.8

order of magnitude at least) than the other methods on both AFLW and Multi-PIE datasets.

The processing time required by individual stages of the proposed detector is detailed in Table 5. It is seen that the computations are dominated by the feature evaluation, which depends on the resolution of the normalized frame and the size of the search space. On the other hand, the MAX-SUM inference takes, thanks to the distance transform, less than 20% of the overall time. To demonstrate the benefit of the distance transform, we also show the time required by the MAX-SUM inference when solved by a plain dynamic programming.

5.3. 300-W Experiments

5.3.1. Implementation details

The 300-W dataset contains near-frontal images with all 68 landmarks considered to be always visible. Therefore in this experiment we consider only the single-view variant of the DPM detector. On the other hand, in contrast to the experiments on AFLW and Multi-PIE, we estimate a dense set of 68 landmarks in images with considerably higher resolution. The graph with landmark configuration is depicted in Fig. 11. In order to keep the processing time of the detector reasonably low, we use the proposed coarse-to-fine search strategy, denoted as C2F-DPM detector, with the following settings.

5.3.1.1. The coarse C-DPM detector. The size of normalized frame of the C-DPM detector is set to 80 × 80 pixels. The normalized frame is obtained by affinely transforming an image cropped around the face box enlarged by a factor of 1.5. The patches used to compute features for the appearance model are set to 13 × 13 pixels for all landmarks except of the root landmark ($\boldsymbol{s}_{31}$), whose patch size is 21 × 21 pixels. The C-DPM detector has $\dim(\boldsymbol{w}) = 2{,}478{,}348$ parameters in total, which are learned from examples.

5.3.1.2. The fine F-DPM detector. The size of normalized frame of the F-DPM detector is set to 160 × 160 pixels. The face box is extended by factor of 1.25. The patches of the appearance model are 15 × 15 pixels for non-root landmarks and 21 × 21 pixels for the root landmark. The overall dimensionality of the parameter vector $\boldsymbol{w}$ is $\dim(\boldsymbol{w}) = 3{,}456{,}012$.

5.3.2. Evaluation protocol

We use the public part of the 300-W dataset [39] for training and evaluation of the proposed methods. The public part contains 6,193 images in total. We use the original split of the images to the training and the test part. Since our learning algorithm requires a validation set for tuning the regularization constant, we further split the original training set into training and validation part. This results in 3 subsets: 518 images for testing, 551 for validation (tuning the reg. parameter λ) and 5,124 for training the weights $\boldsymbol{w}$. We use the average localization error normalized by the interocular distance. The face detector failures are penalized by ∞. That is, we use the evaluation metric entirely consistent with the 300-W challenge.

Since the available implementations of most of the competing methods do not detect all 68 landmarks we use a subset of 49 landmarks (without the landmarks on the cheek contour) common to all compared detectors. The two missing landmarks from the 51 landmark set are the inner corners of the mouth.

5.3.3. Results on public test images

Fig. 12(a) shows the cumulative histograms of the average localization error for all compared methods evaluated on the public test set part of the 300-W dataset. The best method appears to be the IntraFace detector [10] closely followed by the proposed C2F-DPM detector. Only marginally worse results are obtained for the Chehra [11] and the GN-DPM detector [13]. The only full multi-view competitor [9] provides reasonable localization error yet significantly worse than the other compared methods.

In Fig. 12(b), we show the localization error of the fine, second stage, F-DPM detector initialized by different methods. Namely, we consider initialization from an uncorrected face detector's bounding box and also from an ideal bounding box computed from the ground truth annotation. It is seen that the proposed initialization from the coarse C-DPM detector (whose error is also shown) yields the best results closely matching the initialization from the ideal bounding box.

The rightmost column of Table 4 reports an average time required by the compared methods to process as a single image of the 300-W dataset. The fastest methods are the IntraFace [10] (which detects 49 landmarks) and the proposed C2F-DPM detector (detecting 68 landmarks). The slowest method is the Zhu & Ramanan [9] which employs the DPM simultaneously for the face localization and the landmark localization unlike the other methods using much faster sliding window face detectors.

5.3.4. Results on non-public test images

The results of the proposed C2F-DPM detector on the non-public test images provided by the organizers of the 300-W competition are summarized in Fig. 13. In Fig. 13(a) we compare localization accuracy of the proposed detector on all four benchmark datasets. First, we

Table 5
The time requirements of individual stages of the proposed detector. The statistics are shown for the multi-view detector used on AFLW and Multi-PIE dataset and the C-DPM and F-DPM detectors evaluated on the 300-W datasets. We list the mean and the standard deviation computed over test images. Last row shows the time needed to compute MAX-SUM inference without using the distance transform for comparison. All times are shown in milliseconds.

	Type		
Stage	Multi-view detector (AFLW + Multi-PIE)	C-DPM (300-W)	F-DPM (300-W)
Normalized frame	0.009 ± 0.003	0.4 ± 0.1	1.4 ± 0.4
Feature computation	8.1 ± 4.1	35.5 ± 6.1	64.2 ± 9.0
MAX-SUM inference	2.4 ± 1.9	5.3 ± 0.8	6.4 ± 0.9
overall	10.5 ± 4.7	41.2 ± 6.8	72.0 ± 9.9
MAX-SUM inference without dist. transf.	93.0 ± 1.6 (38 × slower)	970.0 ± 30.1 (183 × slower)	2167 ± 38.8 (339 × slower)

include results on a subset of non-profile facial images (and thus comparable with 300-W dataset) from the AFLW and the Multi-PIE dataset processed by the multi-view DPM detector estimating 21 landmarks. Second, we include results obtained by the C2F-DPM detector on the public and the non-public test images of the 300-W estimating 51 landmarks. It is seen that estimation of landmark location jointly with the viewing angle in "multi-view" images from AFLW and Multi-PIE constitutes (not surprisingly) significantly harder problem. Furthermore, we observe that the results on the non-public test images of 300-W are much less optimistic compared to the public ones.

The face detector error E_{fd} on the non-public part was 1.33%.

Fig. 13(b) shows the difference in localization accuracy when evaluated on the full set of 68 landmarks and a subset of 51 landmarks not containing the cheek contour landmarks.

6. Conclusions

We have proposed a real-time, full multi-view landmark detector based on the Deformable Part Models. The detector uses a mixture of tree based graphical models to capture landmark configurations in a full range of yaw angle. The landmark positions and the viewing angle are estimated simultaneously by a global optimization method based on the dynamic programming. The objective function of the learning algorithm is tightly related to the evaluation metric. The benefits of using a proper objective function are demonstrated by empirical comparison with the Zhu & Ramanan detector [9], which has similar structure, but uses simpler two-class SVM algorithm for learning. To achieve a real-time performance we have implemented several speedups. First, we proposed a coarse-to-fine search strategy using an output of a fast low-resolution DPM detector to shrink a search space of the consequent precise DPM detector operating on a high resolution image. Second, we sped up the computation of LBP based dense feature descriptor by pre-computing base LBP features in multiple scales and representing them as a MIPMAP. Third, we use a DPM with decomposable pair-wise potentials, which allow to reduce the inference time by the distance transform. Experiments on public benchmarks with "in the wild" images show that the proposed detector is comparable in accuracy and speed with other approaches using more complicated shape models and local optimization methods for inference.

An open-source implementation of the proposed detector together with learned models can be downloaded from http://cmp.felk.cvut.cz/~uricamic/clandmark.

Acknowledgments

MU was supported by The Grant Agency of the CTU in Prague project SGS15/201/OHK3/3T/13. VF was supported by the Ministry of Education, Youth and Sports under project ERC-CZ LL1303. VH was supported by The Technology Agency of the Czech Republic under Project TE01020197 Center Applied Cybernetics.

References

[1] A.M. Martnez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Trans. Pattern Anal. Mach. Intell. 24 (6) (2002) 748–763.
[2] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1063–1074.
[3] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, IEEE Trans. Comput. C-22 (1) (1973) 67–92.
[4] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[5] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (2005) 1453–1484.
[6] P.F. Felzenszwalb, D.P. Huttenlocher, Distance transforms of sampled functions, Theory Comput. 8 (1) (2012) 415–428.
[7] L. Williams, Pyramidal parametrics, Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '83, ACM, New York, NY, USA 1983, pp. 1–11.
[8] M. Köstinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, IEEE International Conference on Computer Vision Workshops, ICCV'11 Workshops, Barcelona, Spain 2011, pp. 2144–2151.
[9] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, The 25th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'12, Providence, RI, USA 2012, pp. 2879–2886.
[10] X. Xiong, F.D. la Torre, Supervised descent method and its applications to face alignment, The 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'13, Portland, OR, USA 2013, pp. 532–539.
[11] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, The 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14, Columbus, OH, USA 2014, pp. 1859–1866.
[12] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, The 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14, Columbus, OH, USA 2014, pp. 1867–1874.
[13] G. Tzimiropoulos, M. Pantic, Gauss-Newton deformable part models for face alignment in-the-wild, The 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14, Columbus, OH, USA 2014, pp. 1851–1858.
[14] V.N. Vapnik, Statistical Learning Theory, Adaptive and Learning Systems, Wiley, New York, New York, USA, 1998.
[15] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[16] M. Uřičář, V. Franc, V. Hlaváč, Detector of facial landmarks learned by the Structured Output SVM, Proceedings of the International Conference on Computer Vision Theory and Applications, VISAPP'12, vol. 1, SciTePress - Science and Technology Publications, Rome, Italy 2012, pp. 547–556.
[17] M. Uřičář, V. Franc, D. Thomas, A. Sugimoto, V. Hlaváč, Real-time multi-view facial landmark detector learned by the Structured Output SVM, Proceedings of the 11th IEEE International Conference on Automatic Face and Gesture Recognition Conference and Workshops, BWILD'15, IEEE, Ljubljana, Slovenia, 2015.
[18] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[19] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[20] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, The 23rd IEEE Conference on Computer Vision and Pattern Recognition, CVPR'10, San Francisco, CA, USA 2010, pp. 1078–1085.
[21] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 FPS via regressing local binary features, The 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14, Columbus, OH, USA 2014, pp. 1685–1692.
[22] J. Saragih, R. Göcke, A nonlinear discriminative approach to AAM fitting, The 11th IEEE International Conference on Computer Vision, ICCV'07, Rio de Janeiro, Brazil 2007, pp. 1–8.
[23] J. Saragih, R. Göcke, Learning AAM fitting through simulation, Pattern Recogn. 42 (11) (2009) 2628–2636.
[24] M.F. Valstar, B. Martnez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, The 23rd IEEE Conference on Computer Vision and Pattern Recognition, CVPR'10, San Francisco, CA, USA 2010, pp. 2729–2736.
[25] B. Martnez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[26] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models, Proceedings of the British Machine Vision Conference 2006, BMVC'06, Edinburgh, UK 2006, pp. 929–938.
[27] D. Cristinacce, T.F. Cootes, Automatic feature localisation with constrained local models, Pattern Recogn. 41 (10) (2008) 3054–3067.
[28] Y. Wang, S. Lucey, J.F. Cohn, Enforcing convexity for improved alignment with constrained local models, The 21st IEEE Conference on Computer Vision and Pattern Recognition, CVPR'08, Anchorage, AK, USA, 2008.
[29] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[30] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[31] M. Everingham, J. Sivic, A. Zisserman, "Hello! My name is Buffy" — automatic naming of characters in TV video, Proceedings of the British Machine Vision Conference 2006, BMVC'06, Edinburgh, UK 2006, pp. 899–908.
[32] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (1) (2005) 55–79.
[33] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, The 18th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'05, San Diego, CA, USA 2005, pp. 886–893.
[34] S. Sonnenburg, V. Franc, COFFIN: a computational framework for linear SVMs, Proceedings of the 27th International Conference on Machine Learning, ICML'10, Haifa, Israel 2010, pp. 999–1006.
[35] C.H. Teo, S.V.N. Vishwanathan, A.J. Smola, Q.V. Le, Bundle methods for regularized risk minimization, J. Mach. Learn. Res. 11 (2010) 311–365.
[36] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1999.
[37] M.I. Schlesinger, V. Hlaváč, Ten Lectures on Statistical and Structural Pattern Recognition, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.
[38] R. Gross, I. Matthews, J.F. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.

[39] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, A semi-automatic methodology for facial landmark annotation, The 26th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR'13 Workshops, Portland, OR, USA 2013, pp. 896–903.

[40] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, Proceedings of IEEE International Conference on Computer Vision Workshops, ICCV'13 Workshops, Sydney, Australia, 2013.

[41] J. Šochman, J. Matas, WaldBoost — learning for time constrained sequential detection, The 18th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'05, San Diego, CA, USA 2005, pp. 150–156.

[42] J. Čech, V. Franc, J. Matas, A 3D approach to facial landmarks: detection, refinement, and tracking, 22nd International Conference on Pattern Recognition, ICPR'14, Stockholm, Sweden 2014, pp. 2173–2178.

[43] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, The 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'13, Portland, OR, USA 2013, pp. 3444–3451.

[44] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Tech. Rep. 07–49, University of Massachusetts, Amherst, October 2007.

[45] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'11, Colorado Springs, CO, USA 2011, pp. 545–552.