Real-time Multi-view Facial Landmark Detector Learned by the Structured Output SVM

Michal Uřičář[1], Vojtěch Franc[1], Diego Thomas[2], Akihiro Sugimoto[2], and Václav Hlaváč[1]
[1] Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, 166 27 Prague 6, Technická 2 Czech Republic
[2] National Institute of Informatics, Tokyo, Japan

***Abstract*— While the problem of facial landmark detection is getting big attention in the computer vision community recently, most of the methods deal only with near-frontal views and there is only a few really multi-view detectors available, that are capable of detection in a wide range of yaw angle (e.g. $\phi \in (-90^\circ, 90^\circ)$). We describe a multi-view facial landmark detector based on the Deformable Part Models, which treats the problem of the simultaneous landmark detection and the viewing angle estimation within a structured output classification framework. We present an easily extensible and flexible framework which provides a real-time performance on the "in the wild" images, evaluated on a challenging "Annotated Facial Landmarks in the Wild" database. We show that our detector achieves better results than the current state of the art in terms of the localization error.**

I. Introduction

The face recognition is one of the most successful applications of image analysis and pattern recognition. The facial landmark detection is a crucial step of the face recognition pipeline (by face recognition, we refer to an arbitrary algorithm like identity recognition, gender detection or age estimation), since the correct face alignment has substantial impact on the overall accuracy of the face recognition system (e.g. [19], [6]).

Apart from the facial recognition pipeline, the facial landmark detection can be used also as a pre-processing step to some stand-alone application, to mention a few, e.g. head-pose estimation [29], 3D face reconstruction, face tracking, facial expression analysis or Human-Computer Interaction (HCI).

Recently, there is increasing interest in the facial landmark detection field, However, most of available methods work only on the near-frontal face poses or on the very limited yaw range, requiring all detected landmarks to be visible and not allowing self-occlusions [18],[33], [32], [27] or work on the profile poses only [23]. We believe, that one of the reasons which contributes to this state is a lack of a properly annotated databases with a large range of face poses.

We describe a multi-view facial landmark detector based on deformable part models (DPM) which simultaneously treats landmark detection and viewing angle estimation within a structured output classification framework. We present an easily extensible and flexible framework and report settings leading to a real-time performance on a real-life images, evaluated on a challenging "Annotated Facial Landmarks in the Wild" (AFLW) database [17]. We show that the proposed detector has a smaller localization error than the state of the art methods [34], [3]. The Figure 1 depicts the exemplary output of the proposed detector.

Fig. 1. The exemplary output of our proposed detector. The yellow boxes are detections provided by the face detector. The red dots represents the landmark, the blue lines shows the organization of the underlying graph for a detected view (i.e. the discretized yaw angle), which is written on the top of the detected face bounding box in magenta.

The contributions of this paper are as follows: first, we model the multi-view facial landmark detection problem within the structured output classification framework which allows us to directly optimize the detector's evaluation metric, i.e. the average localization error, during the learning phase. This is the main difference between the proposed method and the work of [34] where the objective is to learn an accurate face detector while the accuracy of the estimated landmark position is not optimized in contrast to our work. Second, we deal with the problem of self-occlusions by using view-specific DPM for different range of yaw angles, where the actual yaw angle is simultaneously estimated with the landmark positions. In contrast, the detector of [34] uses different DPMs only for frontal and non-frontal faces. Third,

MU was supported by The Grant Agency of the CTU in Prague project SGS15/201/OHK3/3T/13. VF was supported by the the Grant Agency of the Czech Republic under Project P202/12/2071. VH was supported by The Technology Agency of the Czech Republic under Project TE01020197 Center Applied Cybernetics.

we empirically evaluate the localization accuracy of the proposed detector and two state-of-the-art methods [34], [3] on a very challenging AFLW [17] dataset and on the Multi-PIE [15] dataset. Four, we manually corrected imprecisely annotated examples from AFLW dataset [17] and provide the corrected annotation to the community.

We provide an open-source implementation of the proposed detector.

The paper is organized as follows. Section II summarizes the related work. The proposed method is described in Section III. The experimental evaluation is given in Section IV. Finally, Section V concludes the paper.

II. Related Work

In this section, we briefly summarize existing methods and describe works most related to the proposed approach.

The existing methods can be categorized depending on whether they use a strong or weak shape model and whether they use generatively or discriminatively learned appearance models.

The strong shape models like the Point Distribution Models [7] provide generally a better prior on the landmark configuration. However, the strong shape model is usually represented by a highly non-convex function binding all landmark coordinates. In turn, the estimation leads to a non-convex optimization problem typically solved by a local method. Such detectors are thus inherently sensitive to a good initialization. The main representatives of this category are methods like the Active Shape Models (ASM) [8], Active Appearance Models (AAM) [7] and their derivatives like e.g. the STASM [20] or the Constrained Local Models (CLM) [9]. In contrast, the weak shape models are less strict in penalizing unlikely landmark configurations. On the other hand, the weak shape model is typically represented by a set of simple functions which in turn allows to solve the estimation problem globally. Hence using the weak shape model avoids the problem of getting stuck in a local optima. A prominent representative are variants of the Deformable Part Models (also called Pictorial Structures) [14], [12], [11] which have been applied to the landmark detection problem e.g. in [10], [34], [27], [33].

A local appearance is either represented by a generative models or by a disriminatively learned local detectors. An example of the method with a generative appearance model is the AAM [7], where the appearance is captured by the PCA model of the pixel intensities (or colors) in a triangulated net of image patches. The discriminatively learned local detectors have been used e.g. in the ASM [8] and the DPM [11]. The local detector is usually fed with a feature description of a rectangular patch cropped around the landmark position. A large variety of features have been used for landmark detection, e.g. the Gabor features in [30], [16], [2], [4], SIFT [20], SURF [22], HOG [34] or the LBP [27]. The local detectors are typically learned independently by methods like the AdaBoost algorithm (e.g. [10], [25]) or the Support Vector Regression algorithm [28].

The proposed detector belongs to the category of discriminatively learned DPM. We use the Structured Output SVM algorithm [26] to simultaneously learn parameters of the shape model and the local detectors. In contrast, learning of the shape model and the local detectors is in most existing works done independently, which is computationally simpler but has an impact on the accuracy.

The most related to our work are the methods of [34] and [27]. The authors of [34] propose a multi-view DPM based detector which simultaneously estimates the face location, the landmark positions and the viewing angle. The main difference if compared to our paper is in the learning objective. Their learning algorithm optimizes the detection rate of the resulting face detector while the landmark localization error is not taken into account. The proposed method in contrast optimizes directly the average landmark localization error, being the evaluation metric of the landmark detector. The work of [27] uses a similar learning algorithm, however, it is a single view detector working only on near-frontal faces.

III. Multi-view detector Learned by the Structured Output SVM

In this section, we describe the multi-view DPM based facial landmark detector learned by the Structured Output SVM (SO-SVM) algorithm.

A. Multi-view detector Based on Deformable Part Models

The DPM approach [14] translates estimation of the 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. Examples of particular graphs used in the proposed detector are shown in Figure 4.

Let $I \in \mathcal{I}^{H\times W}$ be an image of a fixed-size (we call it *normalized frame* in the sequel), let $\phi \in \Phi$ be a discretized yaw angle (i.e. it corresponds to a particular view defined by a range of yaw angle), let $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{|V|-1})$ be a configuration of landmark locations (i.e. the x, y coordinates of landmarks, $\boldsymbol{s}_i = (x_i, y_i)$) and, finally, let $\boldsymbol{w}$ be the vector of parameters composed of parameters $\boldsymbol{w}_i^{\phi q}$ and $\boldsymbol{w}_{ij}^{\phi g}$ associated with the unary and pair-wise potentials, respectively. Then, the scoring function and the proposed detector are defined as follows:

$$f(I, \phi, \boldsymbol{s};\ \boldsymbol{w}) = \sum_{i \in V} q_i^{\phi}(\boldsymbol{s}_i, I;\ \boldsymbol{w}_i^{\phi q}) + \sum_{(i,j)\in E} g_{ij}^{\phi}(\boldsymbol{s}_i, \boldsymbol{s}_j;\ \boldsymbol{w}_{ij}^{\phi g})$$
$$(\hat{\phi}, \hat{s}) = \arg \max_{\phi \in \Phi, \boldsymbol{s} \in \mathcal{S}} f(I, \phi, \boldsymbol{s};\ \boldsymbol{w}). \qquad (1)$$

The first part of the scoring function, denoted as the *appearance model*, is composed of unary potentials $q_i^{\phi}(\boldsymbol{s}_i, I;\ \boldsymbol{w}_i^{\phi q})$ which measure quality of the fit of individual landmarks to the image. The second part, denoted as the *deformation cost*,

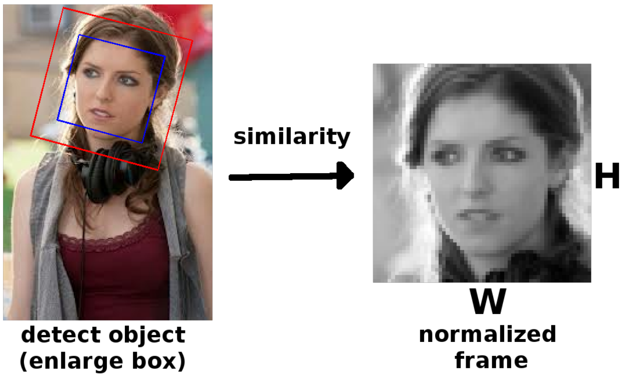


Fig. 2. The acquisition of the Normalized Frame (NF). Blue box is a detection as provided by the face detector, red box is the detection box enlarged by a defined margin. The similarity transformation (removing the possible in-plane rotation and scaling the image to a fixed size) is applied on the red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

is composed of pair-wise potentials $g^{\phi}_{ij}(\boldsymbol{s}_i, \boldsymbol{s}_j;\ \boldsymbol{w}^{\phi g}_{ij})$ which correspond to the likeliness of the mutual position of the connected pair of landmarks.

The normalized frame (NF) can be obtained from an arbitrary image by specifying a bounding box around the face. Such input is usually acquired by a face detector. To improve the scale invariance of the used face detector, there is a possibility to extend or shrink the bounding box by a pre-defined margin. Then the similarity transformation to compensate the scale change and the in-plane rotation is used to obtain the NF (see Figure 2). The restriction of the input image size is crucial due to the real-time performance requirement.

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

Appearance Models

The appearance model is a linearly parameterized function

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

where $\boldsymbol{\Psi}^{\phi q}_i(I, \boldsymbol{s}_i)\colon \mathcal{I} \times \mathcal{S}_i \to \mathbb{R}^{n^{\phi}_{iq}}$ is a feature descriptor of a patch cropped from the image I around the position s_i. Our approach allows to use arbitrary feature descriptor. In particular, in the experiments we use the Sparse Local Binary Pattern (S-LBP) pyramid proposed in [27]. The vector $\boldsymbol{w}^{\phi q}_i \in \mathbb{R}^{n^{\phi}_{iq}}$ is a weight vector which we learn from examples.

To speed up evaluation of the appearance model, we propose a different strategy to compute the features than was originally used in [27]. We propose to pre-compute the parts of the S-LBP features for the whole NF and to store them in a form of a mipmap [31] covering the whole scale space. The final features can be then computed on the fly when needed. The scheme is depicted in Figure 3. This approach makes the feature computation independent of the number of sought landmarks (the computational demand of compilation of features is negligible) leading to a speedup about 40% in our particular setting. More importantly this also allows us to share the pre-computed features among different yaw angles ϕ making the final classifier only sub-linearly slower if compared to the naïve strategy using an individual detector for each view.

Deformation Costs

The deformation cost is also a linearly parametrized function

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

where $\boldsymbol{\Psi}^{\phi g}_{ij}(\boldsymbol{s}_i, \boldsymbol{s}_j)\colon \mathcal{S}_i \times \mathcal{S}_j \to \mathbb{R}^{n^{\phi}_{ig}}$ which, following [11], is defined as a quadratic function of the displacement vector, i.e,

$$\begin{aligned} \boldsymbol{\Psi}^{\phi g}_{ij}(\boldsymbol{s}_i, \boldsymbol{s}_j) &= (dx, dy, dx^2, dy^2), \quad \text{where} \\ (dx, dy) &= (x_j, y_j) - (x_i, y_i), \end{aligned} \qquad (4)$$

The vector $\boldsymbol{w}^{\phi g}_{ij} \in \mathbb{R}^{n^{\phi}_{ig}}$ are parameters which we learn examples.

The main advantage of having the deformation cost in the form of separable quadratic function is the possibility to use the distance transform (DT) [13] to solve the max-sum problem (1). The only requirement needed for application of the DT is the concavity of the functions g^{ϕ}_{ij}. By examining the principal minors of the matrix form of g^{ϕ}_{ij}, we see that this can be enforced by adding additional constraints on $\boldsymbol{w}^{\phi g}_{ij}$. In particular, we need all $w^{\phi g}_{ij}(3)$ and $w^{\phi g}_{ij}(4)$ to be negative (i.e. the 3rd and 4th coordinates of $\boldsymbol{w}^{\phi g}_{ij} < 0,\ \forall (i,j) \in E$).

B. Learning of the Parameters by the SO-SVM algorithm

The proposed DPM detector (1) is an instance of a linear classifier. Therefore we can learn the parameters by the SO-SVM framework [26]. Note that the *joint parameter vector* $\boldsymbol{w}$ to be learned is given by a concatenation of the parameter vectors of the individual appearance models $\boldsymbol{w}^{\phi q}_i$ as well as parameters vectors of all deformation costs $\boldsymbol{w}^{g}_{ij}$. We define a joint feature map $\boldsymbol{\Psi}(I, \phi, \boldsymbol{s})$ as a concatenation of the feature maps $\boldsymbol{\Psi}^{\phi q}_i(I, \boldsymbol{s}_i)$ and $\boldsymbol{\Psi}^{\phi g}_{ij}(\boldsymbol{s}_i, \boldsymbol{s}_j)$. It can be seen that with these definitions, the scoring function can be written as a dot product of the joint parameter vector and the joint feature map, i.e. $f(I, \phi, \boldsymbol{s};\ \boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{\Psi}(I, \phi, \boldsymbol{s}) \rangle$.

We are interested in the parameter vector $\boldsymbol{w}^*$ defined as a solution of the following convex program

$$\begin{aligned} \boldsymbol{w}^* &= \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} \Big[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + R(\boldsymbol{w}) \Big], \\ &\quad \text{s.t.} \quad l_i \leq w_i \leq u_i, \quad i = 1, \ldots, n \end{aligned} \qquad (5)$$

where $R(\boldsymbol{w})$ is the training risk and $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ is a quadratic regularizer introduced to prevent over-fitting. The box constraints allow to set a prior on the parameter vector. In particular, we use the box constraints to enforce that the deformation cost to be concave.

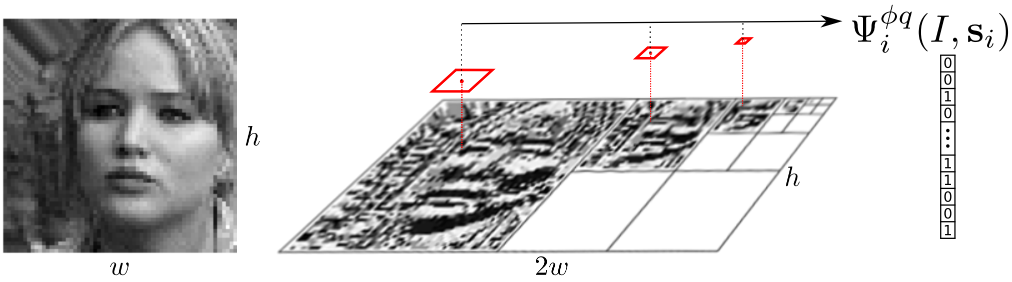


Fig. 3. Features are pre-computed on the whole normalized frame and stored in form of a mipmap. The final feature vector $\Psi_i^{\phi q}(I, \boldsymbol{s})$ (in this case S-LBP) is compiled on the fly when needed from the mipmap, by stacking features from the template window of the corresponding level of the scale space pyramid.

The risk $R(\boldsymbol{w})$ is defined as

$$R(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\phi\in\Phi, \boldsymbol{s}\in\mathcal{S}} \Big[\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}') + \big\langle \boldsymbol{w}, \boldsymbol{\Psi}(I^i,\phi,\boldsymbol{s})\big\rangle\Big] - \frac{1}{m}\sum_{i=1} \big\langle \boldsymbol{w}, \Psi(I^i,\phi^i,\boldsymbol{s}^i)\big\rangle\,, \tag{6}$$

which is a convex upper bound of the true training risk defined as an average of the loss function $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$ described in the following paragraph.

Note that the first maximization term in (6) corresponds to solving the estimation problem (1) with the scoring function augmented by the values of the loss function.

We solve the problem (5) approximately by the Projected Stochastic Gradient Descent (P-SGD) algorithm [24] outlined in Algorithm 1. The function $P\colon \mathbb{R}^n \to \mathbb{R}^n$ denotes the orthogonal projector on the box-constraints defined as

$$\begin{aligned} P(\boldsymbol{w}) &= [P_1(w_1),\dots,P_n(w_n)]^\top, \text{where} \\ P_i(w_i) &= \max\{l_i, \min\{u_i, w_i\}\}\ . \end{aligned}$$

The scalar $\alpha > 0$ is a constant step-size of the inner SGD loop and its value is estimated on the 10% subset of the training examples. The concavity of functions g_{ij}^{ϕ} is enforced by setting upper bounds $u_a < C,\ \forall a \in \mathcal{A}$, where C is a small negative number and $\mathcal{A}$ represents the set of indices pointing to the 3rd and 4th coordinates of $\boldsymbol{w}_{ij}^{g}$ of the joint parameter vector $\boldsymbol{w}$. The rest of variables in $\boldsymbol{w}$ is unbounded, i.e. the box constraints $(l_a, l_b) = -\infty,\ u_b = \infty, a \in \mathcal{A}, b \notin \mathcal{A}$) are used. The sub-gradient $\boldsymbol{r}_i'(\boldsymbol{w}_t)$ can be computed by Danskin's theorem [5, Proposition B.25] as follows:

$$\boldsymbol{r}_i(\boldsymbol{w}_t) = \Psi(I^i,\hat{\phi},\hat{\boldsymbol{s}}) - \Psi(I^i,\phi^i,\boldsymbol{s}^i),\ \text{where} \tag{7}$$

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

Loss Function

The learning algorithm (5) optimizes a convex surrogate of the loss function $\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}')$ which measures discrepancy between the true and the estimated landmark positions on a given training example. We define the loss function as follows:

$$\Delta^{\phi,\boldsymbol{s}}(\phi,\boldsymbol{s},\phi',\boldsymbol{s}') = \begin{cases} \kappa(\boldsymbol{s})\frac{1}{|V|}\sum_{j=1}^{|V|}\|\boldsymbol{s}_j - \boldsymbol{s}_j'\|, & \text{if } \phi = \phi' \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

Algorithm 1 Projected SGD with averaging

Require: $\lambda > 0$, $\alpha > 0$, $(l_i \le u_i)$, $i = 1,\dots,n$

set $\boldsymbol{w}_0 := \boldsymbol{0}$, $\boldsymbol{v}_0 = \boldsymbol{0}$, $t := 0$
repeat
  for i in randperm(m) **do**
    compute sub-gradient $\boldsymbol{r}_i'(\boldsymbol{w}_t)$ of i-th example at $\boldsymbol{w}_t$
    $\boldsymbol{g}_t = \frac{\lambda}{m}\boldsymbol{w}_t + \frac{1}{m}\boldsymbol{r}_i'(\boldsymbol{w}_t)$
    $\boldsymbol{w}_{t+1} = P(\boldsymbol{w}_t - \alpha\boldsymbol{g}_t)$
    $\boldsymbol{v}_{t+1} = \frac{t-1}{t}\boldsymbol{v}_t + \frac{1}{t}\boldsymbol{w}_{t+1}$
  end for
until convergence

where the normalization constant $\kappa(\boldsymbol{s})$ is a reciprocal to the size of the face. The face size if defined to be the length between a segment connecting the center of eyes with the chin which we compute from the ground truth landmark positions. The penalty for confusing the viewing angle is set to 1 which is larger than common localization error due to the use of the normalization constant $\kappa(\boldsymbol{s})$. In turn, the loss function penalizes more mistakes in the viewing angle.

Note also that the loss $\Delta^{\phi,\boldsymbol{s}}\colon \Phi\times\mathcal{S}\times\Phi'\times\mathcal{S}' \to \mathbb{R}$ is non-negative and 0 iff $(\phi,\boldsymbol{s}) = (\phi',\boldsymbol{s}')$ as commonly required by the SO-SVM framework.

Max-sum problem

The max-sum optimization task (1) requires a solver depending on the structure of the graph G. We limit the structure of the graph to a tree, since this allows us to find the global solution in a reasonable time (the solution is organized in terms of a Dynamic Programming with DT [13]). However, we should point out that this is not the limitation of our framework, since it can be easily extended by a solver capable of solving more complicated structures (e.g. by applying some approximations).

IV. Experiments

In this section we describe the implementation details of the proposed landmark detector (sec IV-A), the evaluation protocol (sec IV-B), the competing methods (sec IV-C) and, finally, we report the achieved results (sec IV-D).

TABLE I

THE DISCRETIZATION OF THE VIEWING ANGLE (YAW).

Viewing angle names ($\phi \in \Phi$)				
$-$profile	$-$half-profile	frontal	half-profile	profile
Viewing angle ranges				
$(-110^\circ, -60^\circ >$	$(-60^\circ, -15^\circ >$	$(-15^\circ, 15^\circ)$	$< 15^\circ, 60^\circ)$	$< 60^\circ, 110^\circ)$
Number of landmarks detected in ϕ				
13	19	21	19	13

A. Implementation details of the proposed detector

The proposed detector is described in Section III. Here we summarize the implementation details of the particular instance of the multi-view landmark detector which was used in our experiments.

We discretized the viewing angle (yaw angle) as follows $\Phi = \{-\text{profile}, -\text{half-profile}, \text{frontal}, \text{half-profile}, \text{profile}\}$. For each view, we detect a different number of landmarks since the actual number of visible landmarks varies due to the self-occlusions. The precise ranges of the yaw angle defining the views $\phi \in \Phi$ together with the number of landmarks to be detected are listed in Table I.

The graphs $G_\phi = (V, E)$ for individual views are depicted in Figure 4. Each node is denoted by its ID and name of the corresponding landmark. We show only pictures of the positive viewing angles, since $-$half-profile and $-$profile are just mirrored versions of half-profile and profile.

The internal settings of individual detectors are as follows. The normalized frame is set to 60×60 pixels, enlargement factor of the face box provided by our in-house face detector is 1.5 in both width and height. We use the Sparse LBP features for the appearance models (2) of each landmark with a template window of size 9×9 pixels for all landmarks except of the root (tip of the nose), which has a bigger template 15×15 pixels. For the deformation cost (3) we use a separable quadratic function of the displacement vector as defined in (4).

B. Datasets and the evaluation protocol

In this section we describe the evaluation protocol. We use the AFLW [17] database for both training and evaluation and Multi-PIE [15] database just for evaluation. We divided the AFLW database into training (approx. 80%) and testing (20%) parts. Since we need to tune the regularized parameter λ, we further split the training part into a subset (approx. 80%) used to learn the joint parameter vector $\boldsymbol{w}$ and a part (approx. 20%) used to tune λ.

We found some inconsistency both in the viewing angle and the landmark positions in the annotation of the AFLW dataset. We used the 3D landmark detector [29] initialized from the original manual annotation to make the annotation consistent. However, the imprecision in the viewing angle remained in some examples large. Since the proposed model is sensitive to the yaw angle more than to landmark positions by definition, we manually selected just approximately 700 examples per view, removing the badly annotated examples. In the end, we use just $3,398$ examples to train the detector. This is not optimal, since we use a very high-dimensional features ($\dim(\boldsymbol{w}) = 1,335,360$) and our detector would certainly benefit from using more training examples. Despite this limitation we obtained a very good results as shown below.

Once the joint parameter vector $\boldsymbol{w}$ is learned, we evaluate the detector on the testing examples. For the evaluation we use the following approach. When the detector's estimate of the viewing angle matches the ground truth annotation, then we compute the average localization error by

$$L_{\text{mean}} \quad = \quad \kappa(\boldsymbol{s}) \frac{1}{|V|} \sum_{j=1}^{|V|} \|\boldsymbol{s}_j - \boldsymbol{s}_j^*\| \tag{9}$$

otherwise the localization error is set to infinity.

To remove the dependency on the particular face detector in the comparison with other methods, we cropped the images around the face bounding boxes found by our in-house face detector and enlarged them by 30% in both width and height.

In the case that the number of landmarks provided by the competing method is different than in our setting (i.e. 21 landmarks), we select the maximally overlapping set of the landmarks and in the error evaluation we consider only the matching subset.

C. Competing Methods

This section summarizes all methods compared with the proposed detector.

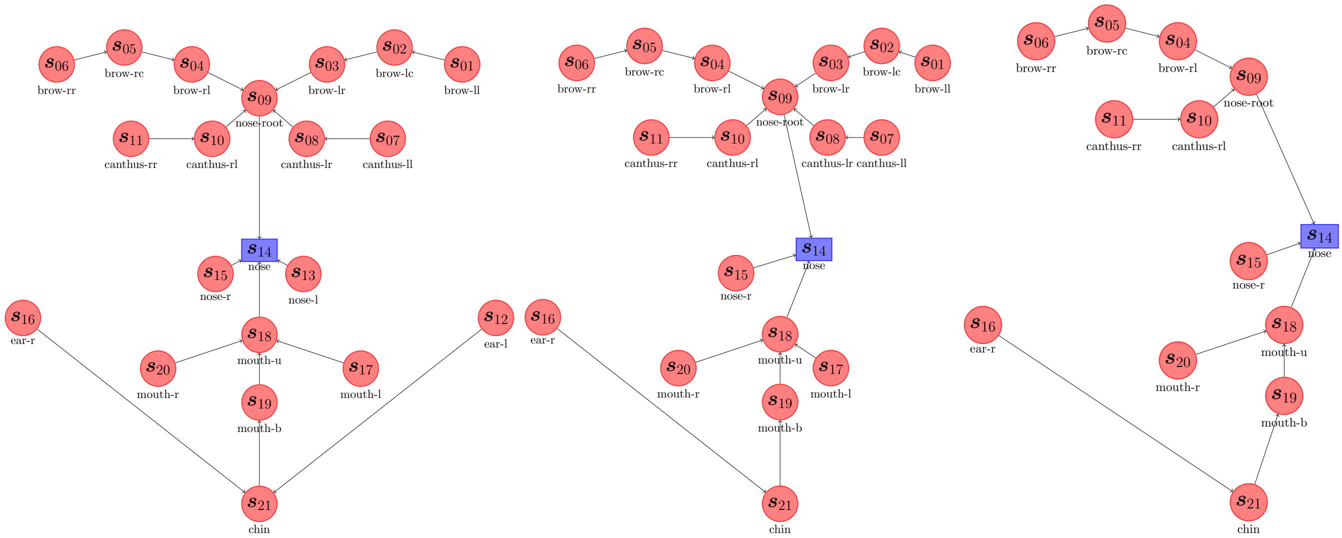
1) Baseline — Independent Detectors: To show the benefits of the multi-view detector, we use the following baseline approach. We create a multi-view detector, where for each $\phi \in \Phi$ we use an independently trained single-view DPM detector. The particular single-view detector is selected based on the response of the face-detector which provides a rough estimate of the viewing angle ϕ. The individual detectors have the same settings as the proposed multi-view detector and they are trained as described in [27].

2) Detector of Zhu & Ramanan [34] : We use the code provided by the authors with the fully shared model "p99". Even though this is supposed to be the fastest model available, the detection speed on the cropped AFLW images was very slow (tens of second per cropped image). The drop of accuracy by using the fully shared model was not reported to be dramatic enough, so we have not tried the independent model, which would require much more time to run.

3) CHEHRA [3]: We use the implementation of recent state-of-the-art facial landmark detector provided by the authors. The detector uses a discriminative 3D facial deformable shape model fitted to a 2D image by a cascade of linear regressors. The detector was trained on the 300W dataset [21].

D. Results

This section summarizes the achieved results. To have just a single number comparison of the detector accuracy, we introduce statistics $E5$ and $E10$ defined as the percentage of the testing examples with the average localization error (9) not higher than 5% and 10%, respectively. Table II shows



(a) frontal (21 landmarks) (b) half-profile (19 landmarks) (c) profile (13 landmarks)

Fig. 4. Graph representation of individual detectors for positive viewing angles. The nodes are represented by the red circles, edges by black lines connecting them. Note that all graphs are in fact trees. The root node is represented by the blue square. Note that the core of all the graphs is the same and just the self-occluded landmarks with its incident edges are removed in non-frontal cases.

the achieved results on both AFLW and Multi-PIE databases expressed in terms of the $E5$ and $E10$. Note that Zhu & Ramanan detector uses a part of the Multi-PIE database for training hence these results are positively biased.

For more detailed evaluation we report the cumulative histogram of the average localization error (9). In Figure 5 we show the cumulative histogram evaluated on the testing subset of the AFLW 5(a) and Multi-PIE 5(b). In Figure 6 the cumulative histograms are shown for a subset of non-frontal faces only. By non-frontal, we mean all faces with the ground-truth viewing angle outside the $(-15°, 15°)$ interval. It is seen that the proposed detector outperforms all competing methods on the AFLW dataset. The results on the Multi-PIE dataset show that our detector achieves smaller localization error, but has a slightly higher yaw misclassification rate (2.36%) compared to the detector of Zhu & Ramanan which was, in contrast to our detector, trained on this database.

In Table III, we present the timing evaluation of the proposed multi-view detector, the baseline method and the detector of Zhu & Ramanan. The baseline method is fastest, but it uses the estimate of the viewing angle provided by the face detector. Common face detectors typically do not provide such information in which case the proposed detector has a clear advantage.

V. Conclusions

We have proposed a multi-view facial landmark detector based on the DPM whose parameters are learned from examples by the Structured Output SVM algorithm. The experimental evaluation shows that the proposed detector outperforms the current state-of-the-art methods in terms of the detection accuracy on very challenging "in the wild" images from the AFLW dataset. The detector runs in real time on a standard PC.

The open-source implementation of the proposed detector and the re-annotated AFLW [17] database can be downloaded from the following link:
`http://cmp.felk.cvut.cz/~uricamic/clandmark/`

TABLE II
THE LOCALIZATION ERROR ON THE AFLW AND MULTI-PIE.

AFLW

	E10 [%]	Yaw mis-classifications [%]
proposed	39.59	25.02
baseline	33.32	38.81
Zhu & Ramanan	7.89	57.18
CHEHRA	39.03	—

Multi-PIE

	E5 [%]	E10 [%]	Yaw mis-classifications [%]
proposed	74.95	82.61	15.39
baseline	47.1	53.41	44.66
Zhu & Ramanan	70.99	86.76	13.03

TABLE III
THE AVERAGE TIME IN SECONDS SPENT ON SINGLE FACE DETECTION.

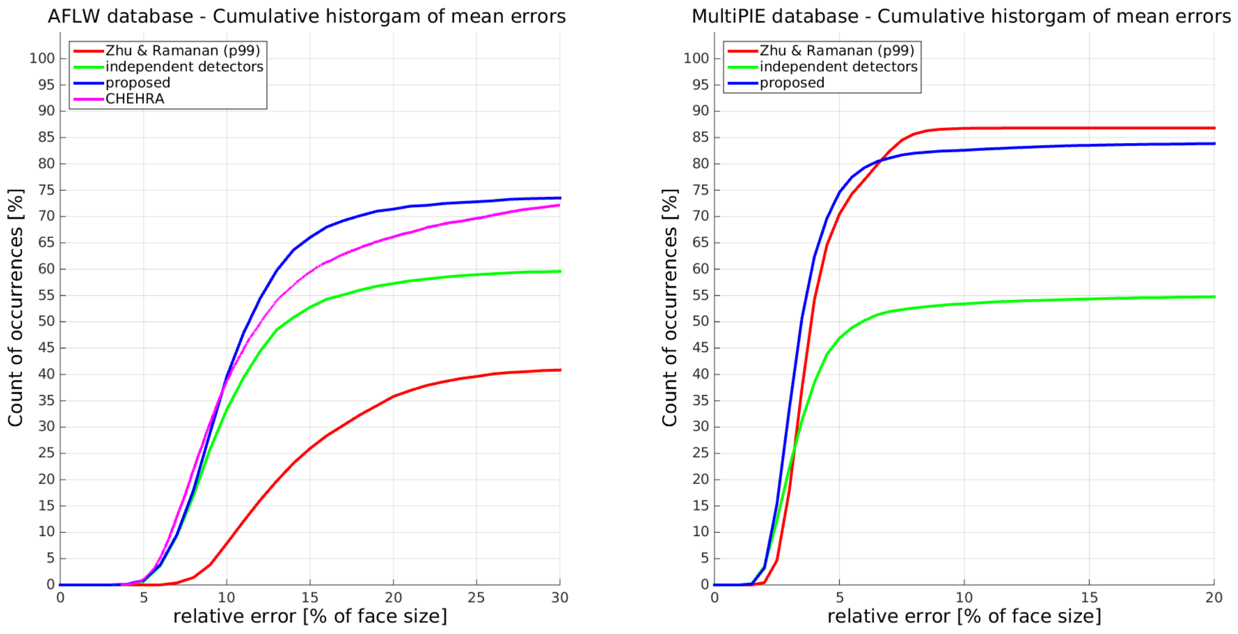
	proposed time [s]	baseline time [s]	Zhu & Ramanan time [s]
AFLW	0.011	0.0027	30
Multi-PIE	0.011	0.0027	9.9

References

[1] *Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, September 4-7, 2006*, 2006.

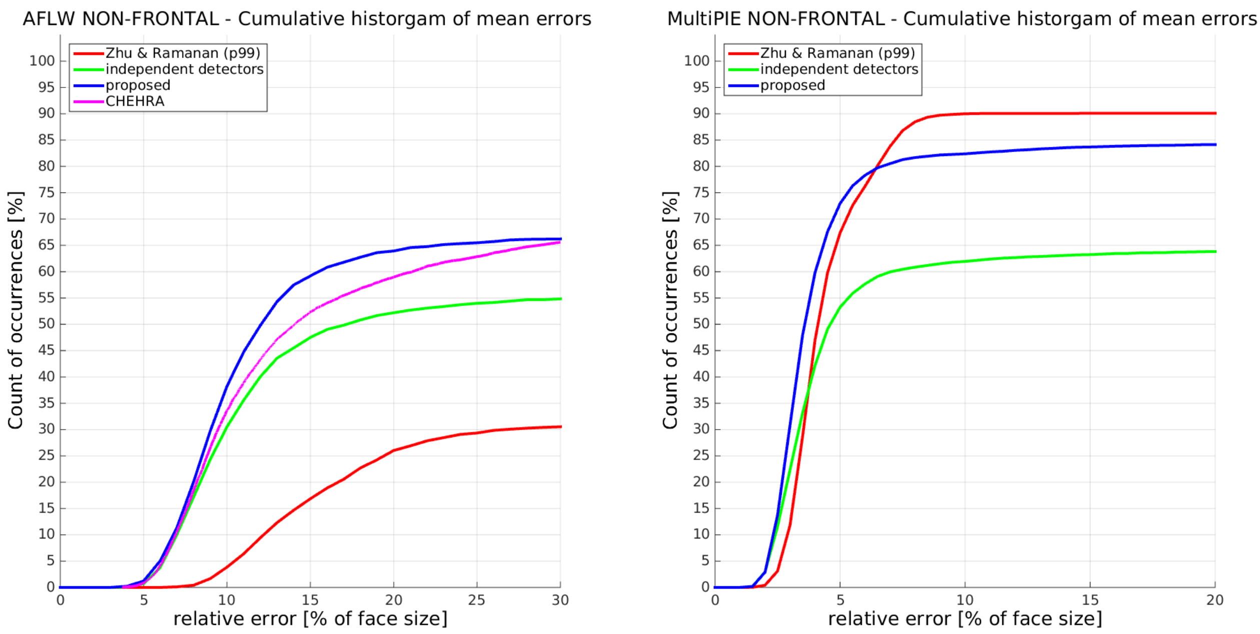
[2] A. Ali Salah, H. c. Akakin, L. Akarun, and B. Sankur. Robust facial landmarking for registration. *Annales des Télécommunications*, 62(1-2):83–108, 2007.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. June 2014.



(a) (b)

Fig. 5. Cumulative histograms of the average localization error measured on the testing subset of the AFLW 5(a) and Multi-PIE 5(b) datasets.



(a) (b)

Fig. 6. Cumulative histograms of the average localization error measured on the testing subset of non-frontal faces from the AFLW 5(a) and Multi-PIE5(b) datasets.

[4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552. IEEE, 2011.

[5] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. pages 1063–1074, 2003.

[7] T. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image*

Understanding, 61(1):38–59, 1995.
[9] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC* [1], pages 929–938.
[10] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy" – automatic naming of characters in tv video. In *BMVC* [1], pages 899–908.
[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2009.
[12] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Internatinal Journal of Computer Vision*, 61:55–79, January 2005.
[13] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 8(19):415–428, 2012.
[14] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.
[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807 – 813, 2010. Best of Automatic Face and Gesture Recognition 2008.
[16] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kälviäinen, and J. Matas. Feature-based affine-invariant localization of faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(9):1490–1495, 2005.
[17] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.
[18] D. Lee, J. Chung, and C. D. Yoo. Joint estimation of pose and face landmark. In *Proceedings of the 12th Asian Conference on Computer Vision*, 2014.
[19] A. M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):748–763, 2002.
[20] S. Milborrow and F. Nicolls. Active Shape Models with SIFT Descriptors and MARS. *VISAPP*, 2014.
[21] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2013.
[22] E. Sangineto. Pose and expression independent facial landmark localization using dense-surf and the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):624–638, 2013.
[23] P. Setthawong and V. Vanijja. Modified deformable parts model for side profile facial feature detection. In B. Papasratorn, N. Charoenkitkarn, V. Vanijja, and V. Chongsuphajaisiddhi, editors, *Advances in Information Technology*, volume 409 of *Communications in Computer and Information Science*, pages 212–220. Springer International Publishing, 2013.
[24] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *arXiv preprint arXiv:1212.1824*, 2012.
[25] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[26] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
[27] M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, pages 547–556, February 2012.
[28] M. F. Valstar, B. Martínez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, pages 2729–2736. IEEE, 2010.
[29] J. Čech, V. Franc, and J. Matas. A 3D Approach to Facial Landmarks: Detection, Refinement, and Tracking. In *ICPR*, 2014.
[30] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *In SMC05*, pages 1692–1698, 2005.
[31] L. Williams. Pyramidal parametrics. In *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '83, pages 1–11, New York, NY, USA, 1983. ACM.
[32] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539. IEEE, 2013.
[33] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *14th IEEE International Conference on Computer Vision*, 2013.
[34] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.