A 3D Approach to Facial Landmarks: Detection, Refinement, and Tracking

Jan Čech, Vojtěch Franc, Jiří Matas
Center for Machine Perception, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University in Prague
{cechj, xfrancv, matas}@cmp.felk.cvut.cz

***Abstract*—A real-time algorithm for accurate localization of facial landmarks in a single monocular image is proposed. The algorithm is formulated as an optimization problem, in which the sum of responses of local classifiers is maximized with respect to the camera pose by fitting a generic (not a person-specific) 3D model. The algorithm simultaneously estimates a head position and orientation and detects the facial landmarks in the image. Despite being local, we show that the basin of attraction is large to the extent it can be initialized by a scanning window face detector. Other experiments on standard datasets demonstrate that the proposed algorithm outperforms a state-of-the-art landmark detector especially for non-frontal face images, and that it is capable of reliable and stable tracking for large set of viewing angles.**

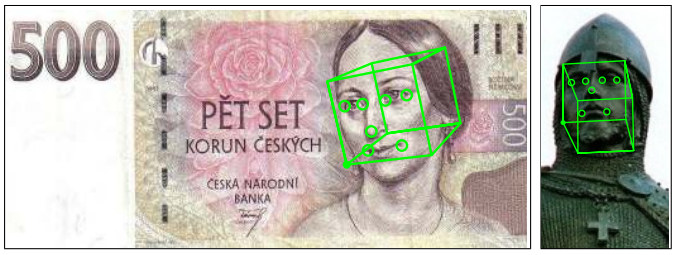

Fig. 1. The proposed method jointly estimates the position of seven facial landmarks in the image and the head pose (position and orientation) with respect to the camera frame. The landmarks are shown as circles and the pose is visualized by projecting a virtual 3D cube around the head into the image. The method is robust, the classifier generalizes even to images of an artistic engraving or a bronze statue.

I. Introduction

Facial landmarks refer to characteristic points on a face like the corners of the mouth, the corners of the eyes or the tip of the nose. Detection of facial landmarks in images of faces is an important step in most face image interpretation tasks.

Most existing facial landmark detectors simultaneously model local appearance around the landmarks and their geometrical configuration. The local appearance is described either by generative models (e.g. [1]) or by discriminatively trained detectors (e.g. [2]). The geometrical structure of the landmarks is usually modeled by a Point Distribution Model [3], describing the landmark positions of a face in canonical pose, and by a subsequent transformation from the canonical pose into 2D image coordinates. Both PDM of 2D shapes (e.g. [1]) and 3D shapes have been proposed (e.g [4]). Fitting the shape models into image requires optimization of a highly non-convex fitness function typically carried out as local gradient search sensitive to the initial estimate. The problem with local optima is mitigated either by re-initializing the optimization, or by using global but expensive optimization methods (e.g. [5]) or by simplifying the shape model. A prominent example of a simplified 2D shape prior is the Pictorial Structure Model (e.g. [6]) representing the shape by a pair-wise energy function whose global optimum can be found efficiently by dynamic programming. Excellent results of PSM based facial landmark detectors have been demonstrated e.g. in [7], [8]. On the other hand, the PSM detectors can describe only a limited range of face poses and thus a multi-view detector must be composed of several PSMs (e.g. [7]). Currently, it is not fully understood how the PSMs fitted by a global method compare to the genuine 3D shape models fitted by local methods.

In this paper we show that a robust and sufficiently precise landmark detector is obtained by fitting the simplest possible 3D shape model into the image using a full projective transformation. The method fits the 6D pose of the mean face, obtained by the process defined in Sec. II-B, and outputs 2D landmark positions together with 3D position and orientation of the face. In addition, we propose a novel method for discriminative learning of the local detectors used to guide the fitting of the 6D pose. We learn a scoring function whose value decreases approximately linearly with the Euclidean distance from the true landmark position. This method produces unimodal peaks around the true landmark positions which helps to make the basin of attraction sufficiently large. The closest to our approach is the work of [2], who employ a 3D shape model, that has an extra degree of freedom compared to our method. Other differences include the assumed camera model, weak perspective in [2] vs. full perspective, and the local detector learning, standard AdaBoost [2] vs. the novel learning method. Contributions of the paper are:

1) We show that modeling shape by a 3D mean face (not a person specific model) and a fully projective transform is enough to obtain precise landmark positions and the head pose estimation.
2) We propose a novel method for learning local landmark detectors which produces nicely behaving score functions with a large basin of attraction.
3) We provide a thorough comparison of the proposed method with a state-of-the-art implementation of PSM based detector [8]. The compared methods use exactly the same local detectors and differ just in the used shape prior. We show that both models provide comparable accuracy on near-frontal images but 3D shape model consistently wins on profile faces.

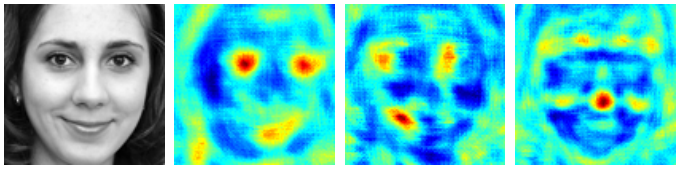

Fig. 2. An example of color-coded coded classifier scores computed inside the face-detector bounding-box (normalized to 100×100 pixels): an inner left eye corner, a left mouth corner, a nose tip.

II. Facial landmarks and a head pose

The problem addressed includes: (1) localization of landmarks in the image, and (2) estimation of the head pose, i.e., a position and orientation with respect to the camera coordinate system. Our solutions uses local classifiers. Each landmark classifier takes the image and for a query pixel returns a score proportional to how likely the landmark occurrence centred at the pixel is, see Sec. II-A. Having the landmarks detected in the image, and their 3D model, it is possible to estimate the pose of the model, as described in Sec. II-B. Finally, we show that these two problems (the landmark detection and head pose estimation) are coupled and can be solved as a single optimization problem. This is described in Sec. II-C.

A. Local landmark classifiers

Let us define a score function $c_i(\mathbf{x}, I)$ which evaluates the likelihood of the i-th landmark being at position $\mathbf{x}$ in the image I, i.e. the most likely position is $\hat{\mathbf{x}}_i = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_i} c_i(\mathbf{x}, I)$ where $\mathcal{X}_i$ denotes the searched positions. We consider a linearly parametrized score $c_i(\mathbf{x}, I; w_i) = \langle \Psi(\mathbf{x}, I), w_i \rangle$ where $\langle \cdot, \cdot \rangle$ stands for a dot product, $\Psi(\mathbf{x}, I) \in \Re^n$ denotes a feature descriptor applied on a patch cropped from the image I around the position $\mathbf{x}$ and $w_i \in \Re^n$ is a weight vector associated with the i-th landmark. We construct the feature descriptor $\Psi(\mathbf{x}, I)$ by concatenating the Local Binary Patterns (256 valued code assigned to patch 3×3) computed in all positions of the cropped patch normalized to size 20×20, 10×10 and 5×5 pixels, respectively. By this process we obtain $256(18^2 + 8^2 + 3^2)$-dimensional sparse ($18^2 + 8^2 + 3^2$ non-zero elements) binary feature descriptor whose values are to some extent invariant against a scale and lighting conditions. The side of the cropped squared patch is 0.3 of the bounding box side returned by the face detector.

The proposed method uses these score functions to guide the search for the most likely configuration of the 3D face pose. It is common to learn the score functions by two-class classification methods, like the Support Vector Machines or AdaBoost, learning the score that best separates example patches collected at the true positions from the patches sampled around the true position. These methods do not take the distance from the true landmark position explicitly into account. In turn, there is no guarantee that the learned score will form unimodal peaks around the true positions. In this paper we propose a different approach which learns the score function such that its value decreases at least linearly with the Euclidean distance measured from the truth landmark position. To this end, we define the loss $\ell_i(\mathbf{x}, I, w_i) = \max_{\mathbf{x}' \in \mathcal{X}_i} \left(\|\mathbf{x}' - \mathbf{x}\| + \langle \Psi(\mathbf{x}', I), w_i \rangle - \langle \Psi(\mathbf{x}, I), w_i \rangle \right)$ where $\mathbf{x}$ denotes the true position of the i-th landmark in the image I. It is seen that the value of $\ell_i(\mathbf{x}, I, w_i)$ upper bounds the Euclidean distance between the true position $\mathbf{x}$ and the position with maximal score, i.e. $\hat{\mathbf{x}}_i = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_i} c_i(\mathbf{x}, I)$.

Given a training set $\{(I^1, \mathbf{x}_i^1), \ldots, (I^m, \mathbf{x}_i^m)\}$ containing pairs $(I^j, \mathbf{x}_i^j)$ of image I^j and the ground truth positions $\mathbf{x}_i^j$ of the i-th landmark, we learn the parameters w_i of the score function $c_i(\mathbf{x}; w_i)$ by solving

$$w_i = \operatorname*{argmin}_{w \in \Re^n} \left(\frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{j=1}^{m} \ell_i(\mathbf{x}_i^j, I^j, w) \right), \qquad (1)$$

where $\lambda > 0$ is a positive constant penalizing large weighs in order to prevent over-fitting and its optimal value is tuned on a separate validation set. The problem (1) can be seen as an instance of the Structured Output Support Vector Machines with margin-rescaling loss [9]. The formulation (1) translates learning of the i-th local detector to an unconstrained convex optimization problem which can be solver efficiently, e.g. we use the cutting plane algorithm [10].

B. From landmarks to 6D head pose

Once the landmarks are detected in the image, it is possible to estimate the head position and orientation. However, additionally a 3D model of the landmark configuration, and certain parameters of the camera projection are also needed.

We denote a vector $\mathbf{v} = (v_1, v_2, v_3)^T$, its homogeneous extension is denoted as $\tilde{\mathbf{v}} = (v_1, v_2, v_3, 1)^T$, and the operator which transforms a vector to Euclidean representation as $[\mathbf{v}]_E = (v_1/v_3, v_2/v_3)^T$. A perspective camera $\mathbf{K}[\mathbf{R}(\mathbf{\Phi}) \mid \mathbf{t}]$, with intrinsic matrix $\mathbf{K}$, rotation matrix $\mathbf{R}$ parametrized by $\mathbf{\Phi} = (\alpha, \beta, \gamma)^T$ which are roll, pitch, yaw angles respectively and the translation $\mathbf{t} = (t_x, t_y, t_z)^T$ projects a 3D point $\mathbf{X}_i = (X_i, Y_i, Z_i)^T$ into a 2D image image point

$$\mathbf{x}_i = (x_i, y_i)^T = \left[\mathbf{K}[\mathbf{R}(\mathbf{\Phi}) \mid \mathbf{t}] \tilde{\mathbf{X}}_i \right]_E. \qquad (2)$$

Now assume, the camera intrinsic matrix $\mathbf{K}$ is known and a 3D model of the landmarks $\mathbf{X}_i$ as well as the 2D landmark points $\mathbf{s}_i = (s_x^i, s_y^i)^T$ detected in the image are given. The position $\mathbf{t}$ and orientation $\mathbf{\Phi}$ of the 3D model with respect to the camera, i.e., the 6D pose, is then calculated based on N correspondences between the 3D model points and 2D image points $\mathbf{X}_i \leftrightarrow \mathbf{s}_i$. Then,

$$\{\mathbf{\Phi}^*, \mathbf{t}^*\} = \arg\min_{\mathbf{\Phi}, t} \sum_{i=1}^{N} \left\| \mathbf{s}_i - \left[\mathbf{K}[\mathbf{R}(\mathbf{\Phi}) \mid \mathbf{t}] \tilde{\mathbf{X}}_i \right]_E \right\|_2 \qquad (3)$$

minimizes the sum of geometric re-projection errors $\|\mathbf{s}_i - \mathbf{x}_i\|_2$. Problem (3) is known as the Perspective n-Point problem (PnP) [11], [12]. A minimal number of points determining the camera pose is 3. The P3P is one of the minimal problems in computer vision. It has an algebraic solution which leads to a set of polynomial equation with several real solutions. In cases the set of correspondences may contain outliers, the minimal solution is repeatedly used in a RANSAC scheme [13]. After the outliers are removed, the problem (3) is iteratively solved by Levenberg-Marquardt optimization starting from the solution having the maximum support in RANSAC. This is a strategy we follow in case of estimating the pose from given landmarks.

The algorithm requires 3D positions of landmarks $\mathbf{X}_i$, which are usually not available precisely, since a 3D shape of the face is difficult to obtain from a single image. Nevertheless, similarly to [14] we use an average model of the 3D configuration of the landmarks computed over a dataset of subjects. The model was built from the MultiPie dataset [15]. This is a multi-view dataset consisting of images of 250 subjects captured by synchronized cameras around each subject. Using a standard structure from motion [16], we first reconstruct full calibration of all cameras. Then having a manual annotation of facial landmarks in the images, we triangulate their 3D positions. Each such 3D landmark model is normalized to canonical coordinates: The model is scaled so that the distance between the eye centres was equal to 1. The model is translated, such that the zero point was at the centre of gravity of all reconstructed landmarks. Finally the model is rotated, such that the horizontal direction coincided with the direction from the left to the right eye centre and the vertical direction coincided with the direction from the centre of the mouth corners to the to the centre between the eyes. The 3D models in canonical coordinates of all subjects are therefore registered. Then, the average model $\mathbf{X}_i$ of N landmarks is the average out of 100 subjects. This 3D model is used in all our experiments.

The method assumes the knowledge of camera calibration $\mathbf{K}$ which is usually not available. Nevertheless, a reasonable guess on the camera intrinsic parameters can be typically made. In all experiments, we placed the principal point in the centre of the image and the focal length equal to the maximum of the image width and image height (sensor width in pixels). The same guess on the camera calibration is made in the popular Bundler [16] when the calibration is not available. This parameter is not critical, we show in many experiments that the choice we made leads to a good precision in landmark detection and accuracy of the orientation. Of course, the focal length influences the translation. The estimated face position is either closer or farther then the estimated value if the true camera has a different focal length.

In [14], [2], an affine camera model is used, which has the advantage of a simple solution of the model parameters. We use a full perspective camera. We aim to use the algorithm in a situation when the affine camera is not a suitable model, as e.g. a laptop webcam. A person may be very close to the camera, with a wide field of view. The differences of depth of the facial landmarks are no longer negligible to the distance to the face centre. This is exactly the situation when the affine camera approximation is poor as explained in [17], p. 169.

C. Joint estimation of landmarks and the head pose

The pose estimation algorithm presented in the previous subsection has satisfactory performance when the landmarks are provided accurately. However, since they are detected by a separate algorithm that does not enforce their locations to be a projection of a 3D model, the method is sensitive to their fluctuations. Moreover, the 3D model itself provides an excellent prior on the configuration of the landmarks in the image. Therefore we formulate the tasks of detection of landmarks and the head pose estimation as a single optimization problem

$$\{\mathbf{\Phi}^*, \mathbf{t}^*\} = \arg\max_{\mathbf{\Phi},t} \sum_{i=1}^{N} \mathrm{c}_i\Big(\big[\mathbf{K}[\mathbf{R}(\mathbf{\Phi}) \mid \mathbf{t}]\tilde{\mathbf{X}}_i\big]_E\Big), \quad (4)$$

where $\mathrm{c}_i(\mathbf{x}_i)$ stands for a classifier response of landmark i, see Sec. II-A, located at image position $\mathbf{x}_i$. Notice that this problem is very similar to problem (3), but instead of optimizing the re-projection error, we propose to optimize the sum of responses of individual landmark classifiers with respect to the camera pose.

Landmarks $\mathbf{x}_i$ are found by projecting the 3D model $\mathbf{X}_i$ into the image by the camera at the optimum rotation and translation $\{\mathbf{\Phi}^*, \mathbf{t}^*\}$ according to eq. (2).

The maximum of (4) is found iteratively by gradient descent. The gradient of the criterion has a special structure. To simplify the notation, let us collect all parameters into $\mathbf{\Theta} = \{\mathbf{\Phi}, \mathbf{t}\}$, and denote the projection of i-th model point into the image as $\mathrm{p}_i(\mathbf{\Theta}) = \big[\mathbf{K}[\mathbf{R}(\mathbf{\Phi}) \mid \mathbf{t}]\tilde{\mathbf{X}}_i\big]_E = (x_i, y_i)^T$. Then the criterion in (4) becomes $F(\mathbf{\Theta}) = \sum_{i=1}^{N} \mathrm{c}_i\big(\mathrm{p}_i(\mathbf{\Theta})\big)$.

The gradient with respect to the parameters is

$$\frac{\partial F(\mathbf{\Theta})}{\partial \mathbf{\Theta}} = \sum_{i=1}^{N} \frac{\partial \mathrm{c}_i\big(\mathrm{p}_i(\mathbf{\Theta})\big)}{\partial \mathbf{\Theta}} = \sum_{i=1}^{N} \frac{\partial \mathrm{c}_i\big(\mathrm{p}_i(\mathbf{\Theta})\big)}{\partial \mathrm{p}_i(\mathbf{\Theta})} \frac{\partial \mathrm{p}_i(\mathbf{\Theta})}{\partial \mathbf{\Theta}}, \quad (5)$$

where

$$\frac{\partial \mathrm{c}_i\big(\mathrm{p}_i(\mathbf{\Theta})\big)}{\partial \mathrm{p}_i(\mathbf{\Theta})} = \left[\frac{\partial \mathbf{c}_i(x_i,y_i)}{\partial x_i}, \quad \frac{\partial \mathbf{c}_i(x_i,y_i)}{\partial y_i}\right] = {\mathbf{J_c}}^i\big(\mathrm{p}_i(\mathbf{\Theta})\big), \quad (6)$$

$$\frac{\partial \mathrm{p}_i(\mathbf{\Theta})}{\partial \mathbf{\Theta}} = \begin{bmatrix} \frac{\partial x_i(\mathbf{\Theta})}{\partial \theta_1}, & \ldots, & \frac{\partial x_i(\mathbf{\Theta})}{\partial \theta_6} \\ \frac{\partial y_i(\mathbf{\Theta})}{\partial \theta_1}, & \ldots, & \frac{\partial y_i(\mathbf{\Theta})}{\partial \theta_6} \end{bmatrix} = {\mathbf{J_p}}^i(\mathbf{\Theta}). \quad (7)$$

The gradient is therefore a sum of products of two matrices. Matrix ${\mathbf{J_p}}^i(\mathbf{\Theta})$ is the 2×6 Jacobian matrix. The derivatives are rather complex due to a non-linear nature of the mapping $\mathrm{p}_i(\mathbf{\Theta})$, but they are computed analytically. Matrix ${\mathbf{J_c}}^i(x_i, y_i)$ of size 1×2 is spatial gradient of the classifier response. The derivatives are computed using a symmetric Gaussian kernel of size σ and finite differences. The scale σ influences how far the gradient "sees", nevertheless too large σ may smooth the responses too much and mislead the optimization. Moreover, a classifier has to be evaluated in a window of $4\sigma \times 4\sigma$ pixels around the target pixel, which can be expensive for large σ. Empirically we found that $\sigma = 3$ works well and this value is used in all our experiments.

The initialization $\{\mathbf{\Phi}_0, \mathbf{t}_0\}$ is required. Nevertheless, the algorithm often converges to a correct solution even when the initial solution was far away from the correct one. This ability allows the algorithm to be initialized by a face detector which provides raw estimate on the face position and a dominant head orientation. Of course, the algorithm can be initialized by any landmark detector, including [8]. In this case, the initial solution is computed by (3).

The single 3D landmark model is used for all subjects across various facial expressions. The algorithm fits this model to the data in fact. This is neither a problem even for atypical faces (e.g. children) nor facial expressions, since the variation of canonically normalized faces is surprisingly small, the set of landmarks we are using is rather rigid, and the classifier is tolerant to possible small displacements, see Fig. 2.

The proposed algorithm has a very low computational complexity. The proposed algorithm has the data model of

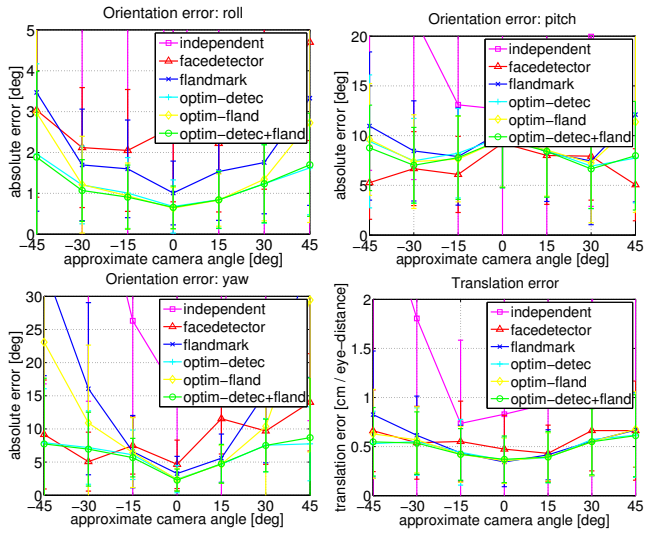

Fig. 3. Camera pose error statistics on the Multipie dataset.

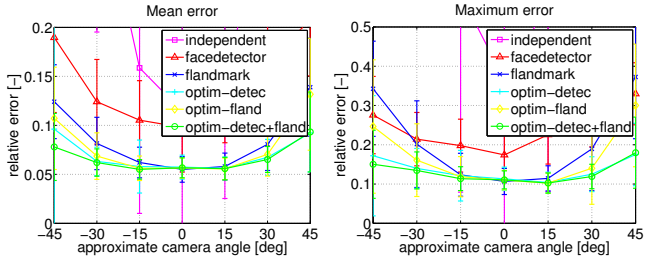

Fig. 4. Relative landmark displacement statistics on the Multipie dataset.

the same kind as Flandmark algorithm [8]. Though Flandmark returns a global optimum (of a different criterion), it evaluates the classifier exhaustively in a large area around each landmark and then a discrete optimization problem is solved by running dynamic programming on the tree. However, optimization (4) requires only 5–30 iterations. Evaluating the criterion as well as the gradient is cheap.

III. Experiments

We demonstrate that the proposed method is robust and accurate in both landmark detection and head pose estimation, and that it can be easily used for tracking. We compare with several baseline methods:

1) **independent**. This is a very naïve baseline, where the landmarks are found as locations of the maximum response of individual classifiers over the entire bounding box of the face detector. The pose is found by the PnP method, see Sec. II-B.
2) **face-detector**. This baseline is based on the face-detector only. We use a commercial detector[1] based on Waldboost [18] which is able to detect non-frontal faces. Besides the bounding-box, it also returns a raw estimate of yaw angle γ. This gives us an estimate of head orientation $\mathbf{\Phi} = (0, 0, \gamma)$. A position (and size) of the bounding-box gives an initial estimate of head position $\mathbf{t}$. Landmarks are found by projecting the 3D model into the image by (2).

[1]Eyedea recognition, Ltd. http://www.eyedea.cz/

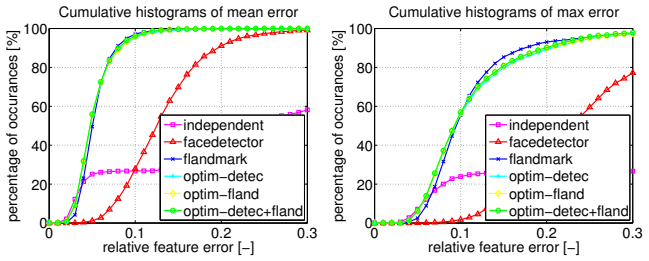

Fig. 5. Results on the LFW dataset - cumulative histograms of a relative displacement error.

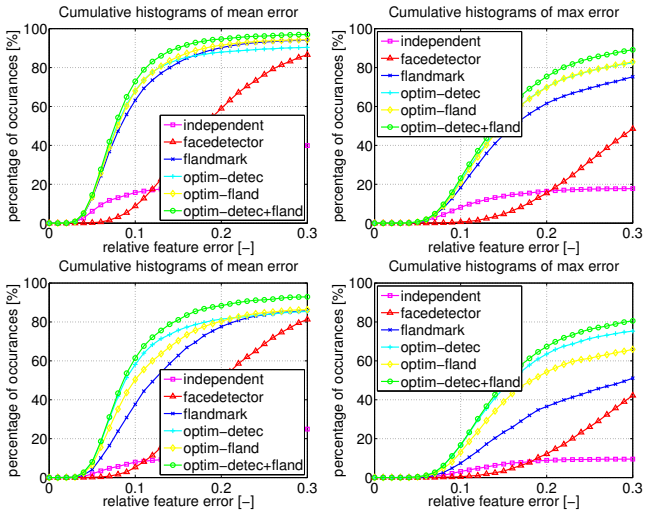

Fig. 6. Results on the AFLW dataset - cumulative histograms of a relative displacement error. The upper row are results on the entire dataset, the lower lower row are results on the subset of non-frontal faces.

3) **flandmark** [8]. This algorithm is a recent award-winning algorithm for facial landmark detection. The head pose is found by the PnP method, see Sec. II-B.
4) **optim-detec**. This is a proposed method solving (4) initialized by $\{\mathbf{\Phi}, \mathbf{t}\}$ obtained by the face-dector as described above. The method does not depend on flandmark, it is its alternative.
5) **optim-fland**. The same as above, but it is initialized by the pose from landmarks detected by flandmark.
6) **optim-detec+fland**. This is a method which combines **optim-detec** and **optim-fland**. It is initialized by both the face detector and flandmark and the final solution is selected that has a better value of the criterion in (4).

Standard error statistics to evaluate the landmark precision were used. Given the ground-truth locations of landmarks in the image (obtained by manual annotation) $\mathbf{x}_i^{gt}$ and landmarks found by an algorithm $\mathbf{x}_i$, we define the mean and maximum relative displacement errors: $\bar{e} = \frac{1}{\kappa N} \sum_{i=1}^{N} \left\| \mathbf{x}_i^{gt} - \mathbf{x}_i \right\|_2$, $e_{\max} = \frac{1}{\kappa} \max_i \left\| \mathbf{x}_i^{gt} - \mathbf{x}_i \right\|_2$, where normalization κ is the length of the facial mid-line in pixels (a distance between the centre of the eye-centres and the centre between mouth corners). In case that not all the ground-truth landmarks are annotated, this value is estimated from the size of the ground-truth bounding box $a \times a$ as $\kappa = 0.3749a$, where

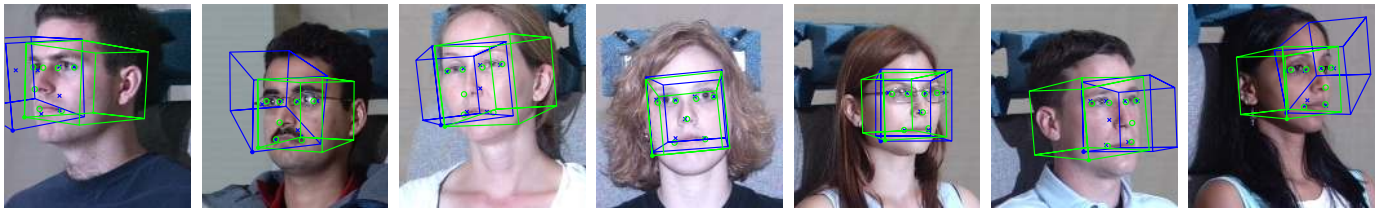

Fig. 7. Samples from the MultiPie dataset. The proposed **optim-detec+fland** (green, circles) and **flandmark** (blue, crosses).

the coefficient is found from images with a full annotation.

All algorithms always output 7 landmarks: outer and inner eye corners, mouth corners, and the nose tip. The tested methods were evaluated in three standard datasets: MultiPie [15], LFW [19], and AFLW [14].

A. The MultiPie dataset

The MultiPie dataset [15] contains images of 250 subjects synchronously captured by cameras displaced by about 15 degrees around each subject, see Fig. 7. The advantage of this dataset is that the multi-camera setup provides a ground-truth for the head orientation and position. As already discussed in Sec. II-B, a subject specific 3D landmark model is obtained by triangulating the ground-truth landmark annotation using full camera calibration [16] and normalization into the canonical coordinates. Then for an image, having the ground-truth landmark annotations $\mathbf{x}_i^{gt}$ in it, we found the ground-truth pose $\{\mathbf{\Phi}_{gt}, \mathbf{t}_{gt}\}$ by PnP using the *subject specific* 3D model and the truth cameras. To increase accuracy of the ground-truth pose, we use full set of 21 landmarks which were annotated manually. This set includes extra landmarks as ears, chin, eyebrows, etc.

The orientation error in roll, pitch, yaw angles was measured (see Fig. 4 in [14] for a definition) in the range from -45 to +45 degrees of the yaw angle. The results are shown in Fig. 3 as mean absolute errors over 100 subjects. The plots have error bars of standard deviations. We can see, that **independent** algorithm has poor performance, **face-detector** only estimate the yaw and since the other angles are close to zero, its performance is fair. The **flandmark** algorithm performs well, but quickly deteriorates with larger angles of yaw, its prior model does not capture well non-frontal faces. The proposed methods **optim-detec**, **optim-fland**, **optim-detec+fland** are better in all experiments. The optimization always improves results over the initialization. Several examples when the proposed method outperforms **flandmark** are shown in Fig. 7. The error slightly increases for larger angles as well since the individual landmark classifiers were trained on near-frontal images. The error is the most significant for the pitch angle. This is probably caused by the mean 3D model. We observed that the largest variation in subject-specific models is in the length of the nose. The discrepancy between the true and used 3D model has an impact in the pitch angle. The position error is normalized by the eye-distance. Assuming the eye-distance is 10 cm, for **optim-detec+fland**, the precision is 1.5 cm for a subject distant about 100 cm from the camera.

Accuracy of the estimated landmarks $\bar{e}, e_{\max}$ as a function of the yaw angle is also measured. The average statistics over 100 subjects are shown in Fig. 4. Notice that **optim-detec+fland** has the average $\bar{e} < 0.1$ for all tested angles.

	LFW	AFLW	AFLW (non-frontal)
independent	26.75	15.15	7.42
face-detector	23.40	7.02	4.21
flandmark [8]	**96.23**	59.92	34.32
optim-detec	94.48	64.93	54.66
optim-fland	94.85	64.31	47.27
optim-detec+fland	95.37	**69.29**	**57.80**

TABLE I. PERCENTAGE OF IMAGES HAVING AN AVERAGE LANDMARK DISPLACEMENT $\bar{e} \leq 0.1$.

B. The LFW and the AFLW datasets

The LFW dataset [19] contains typically near-frontal face images. We tested on a random split of 2.7k images. The AFLW dataset [14] is a large dataset of various images downloaded from Flickr. These images, see Fig. 8, seem to be "wilder" compared to LFW. A large range of viewing angles in roll, pitch and yaw, the level of varying facial expressions, the level of occlusions, varying illumination, varying quality of the images in the sense of focus or motion blur, certain level of post-processing and artistic effects sometimes present make this dataset particularly challenging. The difficulty is reflected in the false negative rate of the face-detector, which was not outstanding even with multi-view well-trained commercial detector. It missed about 40% of the annotated faces, which were excluded from the evaluation. Furthermore, we decided to exclude too small faces (smaller than 150 px), which seems to have fairly imprecise manual annotation. This results in a set of 14.2k faces. Additionally, we selected another subset of non-frontal images. This subset is selected as those images having either of roll, pitch, yaw angles grater than 25 degrees. This subset of 2.8k faces we denote AFLW non-frontal.

Relative displacement errors $(\bar{e}, e_{\max})$ were measured. The results as cumulative histograms are presented in Fig. 5 and Fig. 6. For LFW, the results of **flandmark** and all proposed methods are almost identical. For much more difficult AFLW dataset, performance for all tested algorithms is lower, however all proposed methods outperform **flandmark**. This is especially significant for AFLW non-frontal. These observations are summarized in Tab. I, which shows a percentage of faces having an average landmark displacement $\bar{e} \leq 0.1$. As shown in [8], this level of the average displacement is considered an acceptable solution and takes a limited precision of the manual ground-truth annotation into account. For AFLW **optim-detec** and **optim-fland** have almost equal results despite they are initialized very differently, **optim-detec+fland** outperforms **flandmark** by almost 10%. For AFLW non-frontal, **optim-detec+fland** is better by 23% than **flandmark**.

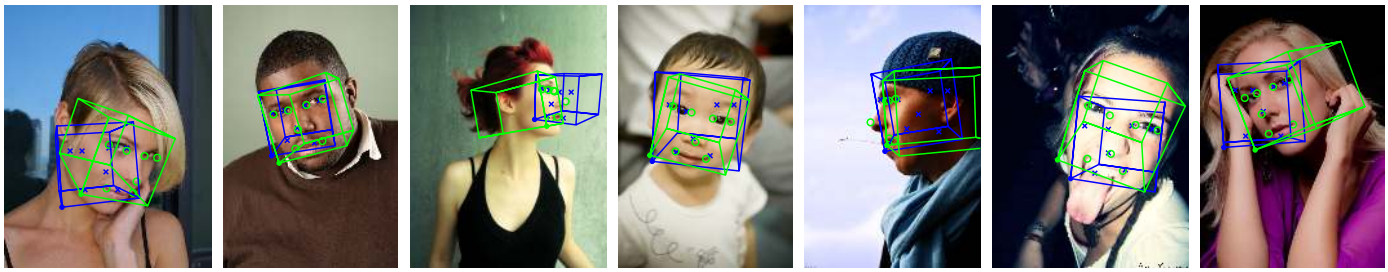

Fig. 8. Selected results on the AFLW dataset. The proposed **optim-detec+fland** (green, circles) and **flandmark** (blue, crosses).

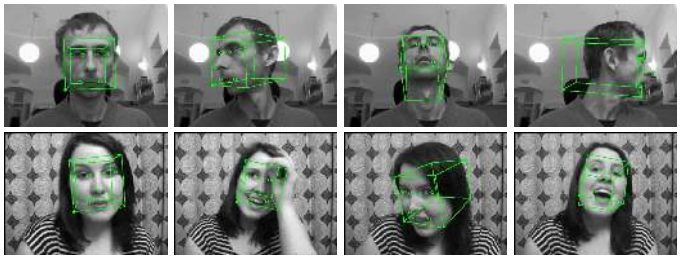

Fig. 9. Tracking. See the robustness against a wide range of angles, fast motions, partial occlusions, and facial expressions in supplementary videos http://cmp.felk.cvut.cz/~cechj/icpr2014/ .

Considering the challenges in the AFLW dataset, and that the proposed method is not designed to be working in profile views (or views where not all landmarks are visible) since local classifiers are trained in near-frontal images and since occluded landmarks corrupt the optimization (4).

C. Tracking

It is natural to use the proposed local optimization method in a tracking task. Pose $\{\mathbf{\Phi}, \mathbf{t}\}$ of the first frame of the sequence is found by **optim-detec+fland**. The optimization (4) of subsequent frames is initialized by the solution of the previous frame. Despite the difficulty of the tested sequences, see Fig. 9, including large angle ranges, fast motions, partial occlusion, speech and facial expressions, the algorithm does not loose a track in about 0.6k and 1.7k frames respectively. This qualitative result confirms the robustness and stability of the proposed optimization scheme.

IV. Conclusion

We have presented a real-time local optimization based method which has good results in landmark detection and refinement, head pose estimation and tracking. We tested on three standard datasets. Under a comparable condition, with similar data model, the algorithm outperforms algorithm [8], a recent award winning algorithm, by a significant margin in a difficult AFLW [14] dataset (by 10% in the entire set and by 23% in the non-frontal subset). This success is based on a simple, but satisfactory 3D modelling (without using a person specific model) employing a perspective camera in conjunction with a novel learning of local classifiers.

Acknowledgement. The first and the third author were supported by the Czech Science Foundation Project GACR P103/12/G084. The second author was supported by the project ERC-CZ LL1303.

References

[1] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. BMVC*, 2006, pp. 929–938.

[2] L. Ang, D. Yangzhou, W. Tao, L. Jianguo, L. Eric Q, Z. Yimin, and Z. Yong, "Fast facial landmark detection using cascade classifiers and a simple 3D model," in *Proc. of ICIP*, 2011.

[3] T. Cootes and C. Taylor, "Active shape models – smart snakes," in *Proc. BMVC*, 2006, pp. 929–938.

[4] G. Lie and K. Takeo, "3D alignment of face in a single image," in *Proc. CVPR*, 2006.

[5] B. Amberg and T. Vetter, "Optimal landmark detection using shape models and branch and bound," in *ICCV*, 2011, pp. 455–462.

[6] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Internatinal Journal of Computer Vision*, vol. 61, pp. 55–79, January 2005.

[7] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[8] M. Uřičář, V. Franc, and V. Hlaváč, "Detector of facial landmarks learned by the structured output SVM," in *Proc. VISAPP*, 2012, pp. 547–556.

[9] 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*, vol. 6, pp. 1453–1484, 2005.

[10] C. H. Teo, S. Vishwanthan, A. J. Smola, and Q. V. Le, "Bundle methods for regularized risk minimization," *Journal of Machine Learning Research*, vol. 11, pp. 311–365, 2010.

[11] J. A. Grunert, "Das pothenotische problem in erweiterter gestalt nebst über seine anwendungen in der geodäisie," *Grunerts Archiv für Mathematik und Physik, Band 1*, pp. 238–248, 1841.

[12] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *IJCV*, vol. 81, no. 2, pp. 155–166, 2009.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] 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.

[15] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[16] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *IJCV*, vol. 80, no. 2, pp. 189–210, 2008.

[17] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge: Cambridge University, 2003.

[18] J. Šochman and J. Matas, "Waldboost - learning for time constrained sequential detection," in *Proc. CVPR*, 2005, pp. 150–157.

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