Advances in Engineering Research, volume 138

2nd International Conference on Machinery, Electronics and Control Simulation (MECS 2017)

Summary on Facial Landmark Detection

Jinghao Wen[1, a]

[1]School of Optoelectronic Information, University of Electronic Science and Technology of China

Chengdu, China

[a]578115435@qq.com

Abstract: Facial landmark detection has important applications in many aspects such as facial recognition, expression recognition, facial attributes analysis, and so on. It compares the detected images with images in dataset to find matched faces, which achieves the identification purpose. Because of its wide use in business, security, identification and other aspects, it gains more and more attention. State-of-art researches on deep convolutional neutral network (DCNN) separate DCNN into DCNN-I(Inner) and DCNN-C(Contour) to get more accurate detection. In HeHOP estimation, performed on local areas projected relative to the head orientation and position, a binary feature extraction approach based on depth data is developed to estimate the head orientation directly by global linear regression. An energy based partitioning of the input domain with a direct control of the final variance perpartition is proposed. A pose-indexed based multi-view (PIMV) face alignment method that obtains a more accurate prediction is proposed in. Researches also proposed a method to handle both images with severe occlusion and images with large head poses. In this paper, we will summarize some state-of-art methods on FLD and analysis their advantages and disadvantages. Based on this, we will have a discussion on some feasible research areas.

Keywords: Facial Landmark Detection, DCNN, HeHOP, partition input domain, PIMV.

1. Introduction

As a vital step of diverse face related applications, facial landmark detection (FLD) has received much attention in the computer vision field. It refers to the localization of the fiducial points on facial images and it is essential for many facial analysis tasks such as face recognition, expression recognition, and facial attributes analysis. In recent years, new methods are proposed to deal with tasks in different smaller aspects among this.

In this paper, we will summarize 5 methods and discuss on them. Baddar *et al.* [1] proposed a method on deep convolutional neutral network (DCNN) based on DCNN with facial contour and facial component constraints. To get more accurate detection, It fantastically separates DCNN into DCNN-I(Inner) and DCNN-C(Contour) to detect them . Novel learning strategies are proposed for both DCNN-I and DCNN-C and the results are jointly combined. Schwarz *et al.* [2] proposed a method on a different area, head pose estimation. A binary feature extraction approach based on depth data is developed to estimate the head orientation directly by global linear regression. Then, other state-of-the-art methods for head pose estimation are also used in the paper for further detection. Romero *et al.* [3] explores an energy based partitioning of the input domain with a direct control of the final variance per partition. Before this, an assumption that controlling the energy per partition can be used to modulate the complexity of the problem to make it suitable for different classifiers is proposed. Qi *et al.* [4] proposed a pose-indexed based multi-view (PIMV) face alignment method that obtains a more accurate prediction which is also robust to pose variations as well as partial occlusions in real world face alignment tasks. Wu *et al.* [5] regard the landmarks on the self-occluded facial parts as occluded points, where the face

itself is the occluder and consider images with large head poses as special cases of images with occlusion and treat them similarly. Based on this, they proposed a method to handle both images with severe occlusion and images with large head poses.

The method of DCNN captures facial component variations, and fine-tunes the corresponding hyper parameters independently. It analyses different layers of the data for further detection. The method of HeHOP combines the advantages of both methods and improves accuracy and speed in head pose detection. Similar to the method of DCNN, the method of Romero *et al.* partitions input domain into smaller domains that can be solved by one simple classifier. Jointly the results are estimated for further details. Based on cascaded pose regression (CPR) framework and its weaknesses, the multi-view face alignment method is based on a pose-indexed shape searching space which is established by a series of pose-shape pairs. The method of Wu *et al.* iteratively predicts the occlusions and location of the landmark. Based on the advantages and disadvantages of 5 papers, we will discuss on some feasible research areas.

2. Discussions

2.1 Discussions on DCNN and highly efficient orientation and position estimation

In this method, the landmarks are detected as a linear regression problem, whose loss function can be written as

$$E = \frac{1}{2}\sum_{i=1}^{N} ||y_i - f(x_i; W)||^2, \quad (1)$$

where y_i is a vector of all landmarks coordinates on facial components, x_i represents an input image and f(.) is a function of x_i parameterized by the learned hyper parameter W. The relationship between landmarks on facial components can be characterized by inter-facial component relationship and intra-facial component relationship. In the inter-facial component relationship, landmarks on each facial component are constraint by the location of the corresponding facial component. In the intra-facial component relationship, facial component variations can affect other facial component landmarks. They fork shared lower layers into branches at the higher layers. The lower layers are joint while the higher ones are separated, which results in an improved FLD that is more robust to occlusion and local variation. In the forward propagation pass, the feature x^l at layer *L* are obtained in function σ(.) parameterized by $W^l$to the previous layer features x^{l-1}. In higher layers, it is similar. Each layer can transform into similar replica in lower layers. The higher layers loss from k-th branch is independently back propagating by

$$\varepsilon^{lk} = \frac{\partial E^k}{\partial W^k} = (y_i - (W^{lk})^T x_i)x_i^T, \quad (2)$$

which updates the filters corresponding to each facial component. The errors of each layer are described by$\varepsilon^{l-1} = (W^l)^T\varepsilon^l \frac{\partial\sigma(u^l)}{\partial u^l}$; the partial derivative $\frac{\partial\sigma(u^l)}{\partial u^l}$ is the gradient of the l-th layer activation function. Then, each layer is decomposed into lower layers to detect more details. For pre-training DCNN-C, a network similar is utilized and trained with all facial landmarks in this method. The distance between the estimated landmarks and the ground truths normalized by the inter-ocular distance, known as mean error, is defined

$$\text{error} = \frac{1}{N}\frac{\frac{1}{M}\sum_{j=1}^{M} p_{i,j} - g_{i,j}|_2}{|l_i - r_i|_2}, \quad (3)$$

where M is the number of landmarks, **p** is the ground truth and **l** and **r** are the positions of both eyes.

In the highly efficient orientation and position estimation work, regression-based approach for facial landmark estimation from in several aspects to obtain head orientations and positions from depth maps

is proposed, which uses similar methods in details. To estimate the 3D head location and rotation from a single depth scan, a stage-by-stage approach is used while each stage provides an update towards the correct head orientation and position.

The optimal global linear regression:

$$\min_{W^t} \sum_{i=1}^{N} \left| \left| \Delta\hat{\theta}_i^t - W^t \Phi_i^t \right| \right|_2^2 + \lambda ||W^t||_2^2. \quad (4)$$

The ground-truth residuals:

$$\Delta\hat{\theta}_i^t = \begin{pmatrix} \Delta\hat{l}_i^t \\ \Delta\hat{\theta}_i^t \end{pmatrix} = \begin{pmatrix} \hat{l}_i - l_i^{t-1} \\ Q(\hat{R}_i (R_i^{t-1})^{-1}) \end{pmatrix}. \quad (5)$$

The head orientation and position of each frame:

$$\theta_i^t = f\left(\theta_i^{t-1}, \Delta\theta_i^t\right) = \begin{pmatrix} l_i^{t-1} + \Delta l_i \\ Q(\Delta R_i^t R_i^{t-1}) \end{pmatrix}. \quad (6)$$

$$\Delta\theta_i^t = \begin{pmatrix} \Delta l_i^t \\ \Delta\theta_i^t \end{pmatrix} = W^t \Phi_i^t. \quad (7)$$

For each stage, the features are estimated locally and jointly. With the formulas, an update of the head position and orientation towards the current head orientation in each stage is estimated and shown.

2.2 Discussions on partitioning the input domain for classification

1) Direction of projection: The strategy for partitioning the input domain constructs a binary tree using a recursive procedure. Minimizing the energy in each partition, as measured by its variance, can reduce the complexity of the input domain for each classifier.

Consider a training dataset S where each sample $s \in S$ has a feature vector $x_s \in R^{d*1}$. The set of samples in node n, denoted as $C_n \subseteq S$ into two new nodes n and n , is divided with the sets of samples $C_{n+1} \subset C_n$ and $C_n + 2 \subset C_n$ respectively, such that $C_{n+1} \cup C_{n+2} = C_n$ and $C_{n+1} \cap C_{n+2} = \emptyset$.The variance of the node samples along a unit vector $\hat{u}$ is $\sum_{s \in C_n} ((x_u - \bar{x}_n)^T \hat{u})^2$, where $\bar{x}_u$ is the mean value of the samples in the node. For the zero-mean data matrix, the node data matrix is normalized row-wise when using features with different ranges.

$$\bar{x}_n = X_n - \bar{x}_n{}^T \in R^{dx|C_n|}, \quad (8)$$

$$X_n = \left[x_1, x_2, \ldots, x_{|s|}\right] \in R^{dx|S|}, \quad (9)$$

$$\max_{\hat{u}} \frac{\hat{u}^T \sum \hat{u}}{\hat{u}^T \hat{u}} \quad \text{s.t.} \quad \hat{u}^T \hat{u} = 1, \quad (10)$$

$$p_n = \bar{X}_n^T u_1 \in R^{|C_n| x 1}, \quad (11)$$

$$C_{n+1} = \{s \in C_n : p_n, s > \tau_n\} \quad (12)$$

$$C_{n+1} = \{s \in C_n : p_n, s \leq \tau_n\} \quad (13)$$

2). Projection threshold: For selecting the splitting threshold τ_n, the projection threshold is set as the projection of the mean value of the data on u_1. Since the data is centred prior to the projection, the projection threshold $\tau_n = 0$ $\forall n$. The sorted projections $q_{n=sort\,(p_n)}$ are searched to find the sample division that yields the minimum combined variance. The projection threshold is selected as the midpoint between the two samples that encompass the optimal division.

$$R = \text{argmin}_i \, \text{var}\left(q_{n,(1:i)}\right) + \text{var}\left(q_{n,(i+1:end)}\right) \quad (14)$$

The projection threshold $\tau_n = \frac{q_{n,r} + q_{n,(r+1)}}{2}$ is selected as the midpoint between the two samples that encompass the optimal division. It guarantees that the combined variance of the partitions along the direction of projection is minimized. A better node division tends to achieve a comparable performance with fewer nodes as seen in the experimental results.

3). Characteristics of the partitioning method: The energy based partitioning method has the following characteristics:(1) The size of the tree can be defined based on a variance threshold instead of the number of nodes, as opposed to algorithms such as k-means, which require the number of partitions to be known a priori. This allows partitioning to be done based on the expected energy in a partition.(2) The method does not require an initialization as with clustering methods such as k-means and GMMs. Therefore the produced results are deterministic.(3) Finding the corresponding partition for a new sample by traversing the node tree has a maximum complexity of O (depth Of Tree $*$ d).

2.3 Discussions on pose-indexed based multi-view method for face alignment

In shape cascaded regression methods, starting with an initial shapeS_0, shape S is progressively refined by updating the previous estimated shape stage-by-stage.

$$S_t = S_{t-1} + \Phi_t(I, S_{t-1}), \quad (15)$$

1). Multi-view Generative Model: Formally, face images from source images and are obtained normalized to the same size. A face image is decomposed into a non-overlapping grid of patches. The kernel of this model is to build a dictionary D= $\{p_1, p_2, \dots, p_D\}$ which is composed of different image patches with pose cases, where $p_d (d \in \{1 \dots D\})$ represents the image patch in the dictionary and D represents the length of the dictionary. For the patches in the dictionary, a probability distribution of the true pose over all patches is composed. Then, a best matched pose of the input image with Bayes rule is obtained by using this probability distribution. Based on the posterior probability of true pose, the most similar pose parameter α for an input image can be estimated by using Bayes rule $P_{(\alpha|U,\ W)} = \frac{P(u_m, \hat{d}|\alpha, W_{m\bullet})p(\alpha)}{P(U)}$, where U=$[u_1 \dots u_M]$ indicates the input image which consists of M non-overlapping patches and u_M represents the $m^{th} (m \in \{1 \dots M\})$ path in this image. In short, in the estimation process, the position of the most similar patch with the input patch in the dictionary through likelihood estimation is calculated. Then this position is combined with the weight parameters W to estimate the face pose α.

2). Pose-indexed Shape Searching Space: The mean shape of every set as the corresponding shape for each pose are taken and normalized by Procrustes Analysis. A shape space base on a series of mean shapes with pose labels ranging from -60° to 60° is established. 121 yaw angles to mark candidate shapes that the yaw angle is smaller than 60° are used in this paper. To decrease the errors caused by the estimated poses, two strategies are introduced to improve the robustness of the method: (1)The average shape of each face shape set induced by poses are taken as the pose-indexed shape to reduce the estimated error caused by the pose estimation. (2) Each yaw angle from -60° to 60° corresponding to the whole 121 shapes are estimated for instead of merging multiple pose angles for one shape.

3). Shape Regression with Local Binary Features: Taking advantages of the cheap computation and high speed come from the sparse binary features, the local binary features (LBF) are determined to be combined with our shape initialization method. Starting from a more accurate shape provided by PIMV, the random forest as repressors is used to minimize the error between the estimated shape and the ground truth at each iteration during the training phase.

2.4 discussions on robust facial landmark detection under significant head pose and occlusion

1). The general framework: For one specific image with binary land-mark occlusion vector $\mathbf{c} \in [0,1]^{D_t}$,$p^d$ measures the probability that a landmark is visible (c_d=1).

Algorithm 1: The general framework
Initialize the landmark locations x_0 using the mean face;
Assume all the landmarks are visible $p^0 = 1$
for t=1,2,…,T or convergence **do**

Update the landmark visibility probabilities given the images, the current landmark locations and the occlusion pattern *Loss (.).*

$$f_t: \mathbf{I}, x^{t-1}, Loss(.) \rightarrow \Delta \mathbf{p}^t$$

$\mathbf{p}^t=\mathbf{p}^{t-1}+\Delta \mathbf{p}^t$

Update the landmark locations given the images, the current landmark locations, and the landmark visibility probabilities.

g_t:$\mathbf{I}, x^{t-1}, \mathbf{p}^t \rightarrow \Delta \mathbf{x}^t$

$\mathbf{x}^t=\mathbf{x}^{t-1}+\Delta \mathbf{x}^t$

end

Output the estimated landmark locations given the images, the current landmark locations $x^T$and the binary occlusion vector based on the predicted visibility probabilities p^T.

The general framework of the proposed robust cascade regression method is shown in Algorithm 1. When updating the visibility probabilities, denoted asf_t, a constrained supervised regression model is introduced to predict the landmark visibility probability update Δp_t based on the image, the current landmark locations $x_{t-1}$and the occlusion pattern embedded in a loss function *Loss (.)*.When updating the landmark locations, a regression function g_t is used to predict the landmark location updateΔx_t based on the image, the current landmark locations x_{t-1}, and the visibility probabilities p_t.

2).Update the landmark visibility probability: The landmark visibility probability and landmark occlusion are difficult to predict. Landmark visibility probabilities based on the appearance and shape information from all points are proposed to be updated and the learned explicit occlusion pattern is also proposed to be used as a constraint.

(1). Landmark visibility prediction model: Landmark visibility prediction depends on the local appearance, the current shape, and the occlusion pattern. To encode the appearance information, SIFT features of the local patches centered at the current landmark locations are used, denoted as $\Phi(\mathbf{I},\mathbf{x}^{t-1}) \in R^{D_l D_f}$, while$D_f$=128 is the dimension of features. To encode the shape information, we calculate the differences of x, y coordinates for pairwise landmarks to get the shape features denoted as$\varphi(\mathbf{x}^{t-1}) \in R^{D_l(D_l-}$. Combined together, a concatenated feature vector is generated, denoted as$\Psi(I,x^{t-1})=[\Phi(\mathbf{I},\mathbf{x}^{t-1}); \varphi(\mathbf{x}^{t-1})]$.

$\text{minimize}_{\Delta \mathbf{p}^t} ||\Delta \mathbf{p}^t - T\hat{}t\Psi(I, x\hat{}(t-1))||_2^2+\lambda E_{p^t}[\text{Loss}(\mathbf{c})]$

Subject to $\mathbf{p}^t=\mathbf{p}^{t-1}+\Delta \mathbf{p}^t(0 \leq \mathbf{p}^t \leq 1)$

$E_{p^t}[\text{Loss}(\mathbf{c})]=\sum_{k=1}^{2D_t} \text{Loss}(\mathbf{c_k}) P(\mathbf{c_k}; \mathbf{p}^t)$

$P(C; P) =\prod_{d=1}^{D_t} {p_d}^{c_d} (1-p_d)^{1-c_d}$,

p^t refers to the landmark visibility probabilities for the next iteration.

(2). Learning the landmark visibility prediction model:Parameters:

$$\theta=\{W_1,b_1,W_2,b_2\}, \quad (16)$$

Reconstruction errors are minimized:

$$\theta^* = \text{argmin} \sum ||c_i - \sigma(W_2\sigma(W_1 c_i + b_1) + b_2||_2^2, \quad (17)$$

$$T^{t^*} = \text{argmin}_{T^t} \sum_i ||\Delta p_i^t - T^t \Psi(I_i, x_i^{t-1})||_2^2, \quad (18)$$

$$[\text{Loss}(c; \theta) = ||c - \sigma(W_2\sigma(W_1 c_i + b_1) + b_2||_2^2, \quad (19)$$

(3). Inference with the landmark visibility prediction model:Monte Carlo approximation is used to calculate the second term over the samples:$E_{p^t}[Loss(\boldsymbol{c})] \approx \frac{\text{const}}{K}\sum_{k=1}^{K} Loss(\widetilde{\boldsymbol{c_k}})\, P^t(\widetilde{\boldsymbol{c_k}})$(K=5000 in our experiments).Denote the gradient of the objective function, w.r.t. Δp_t as δ

$$\Delta=2(\Delta p^t - T^t\Psi(I_i, x^{t-1})) + \lambda\frac{\text{const}}{K}\sum_{k=1}^{K} \text{Loss}(\widetilde{C_k})\frac{\partial p^t \widetilde{C_k}}{\partial \Delta p^t}, \quad (20)$$

$$p^t = p^{t-1} + \Delta p^t, \text{in the range}[0,1]^{D_1}. \quad (21)$$

3). Update the landmark locations: The location update vector is predicted with linear regression function as bellow:$\Delta \mathbf{x}^t = R^t[\sqrt{p_i^t} \circ \Psi(I, x^{t-1})]$, "○" denotes the block-wise product between the square of the landmark visibility probabilities and the appearance features from corresponding point. Parameter learning can be formulated as a weighted least squares problem with closed form solution:

$$R^t = \text{argmin}_{R^t} \sum_i ||\Delta x_i^t - R^t[\sqrt{p_i^t} \circ \Psi(I_i, x_i^{t-1})]||^2_{\text{diag}(W_i)}. (22)$$

Three kinds of databases are used in the paper. The first kind of databases contain general "in-the-wild" images collected from the internet with near-frontal head poses and limited occlusion. The second kind of database contains "in-the-wild" images with severe occlusion. The third kind of database contains images with large head poses. The error as the distance between detected landmarks and the ground truth landmarks normalized by the inter-ocular distance are calculated. The average error across all available annotated landmarks from the testing databases is also calculated. Auto encoder with one hidden layer is used and 20 and 25 hidden nodes for experiments are set.

3. Advantages and feasible improvements in research areas.

3.1 Results and advantages

The method of DCNN shows a much lower mean error rate in almost all aspects that it separates DCNN into DCNN-I and DCNN-C to detect both in deeper layers. In this case, more details on facial components are detected. And joint together, the deep neutral convolutional network of facial contour is set up. Based on this, detection on facial contour is furthered. It captures facial component variations, and fine-tunes the corresponding hyper parameters independently.

Among the HeHOP method shows a trade-off between runtime and accuracy by varying the number of stages. The group of HeHOP (Highly efficient Head Orientation and Position) fast works much better than the others in general. Increasing the number of trees results in little improvement of the success rate, while increasing the number of stages improves the success rate. However, the runtime rises much faster by increasing the tree depth than by increasing the number of trees. It combines the advantages of both methods and improves accuracy and speed in head pose detection. Using the regression-based stage-by-stage algorithm, binary features in each layer are detected from deep information of local area. Thus, the blocking is reduced while the computational efficiency and robustness are obtained. In this case, on one hand, ways of extracting binary features based on deep data is proposed, which works well on head orientation and pose detection. On the other hand, using quaternion, the head orientation can be estimated directly from overall linear regression. Compared with state-of-the-art methods, the algorithm shows no missing frames while running more than 3 times as fast than existing methods.

In the situation of one pixel-wise partitioning and image-wise partitioning, partitioning improved the classification results for the Avg. Label method and the Linear SVM. In comparison to using only a pixel-wise partitioning, the combination of both partitioning approaches reduced the amount of false negatives, such as misclassified sidewalk pixels which a purely pixel- based classifier could not differentiate from the road. Two methods of partitioning energy based on detection are proposed. Energy is directly controlled in each partition as limiting the domain energy reduces the complexity of the problem. It improves the function of SVM and classifiers during partitions. With maximum variance in each partition defined, the number of partitions is not required as a prier. With assumptions that control the energy in each partition, the complexity is reduced. Therefore, it can be adequate to different classifiers.

In the multi-view face alignment method, shape index is used as initialization form of iteration, which solves problems based on differential initializations. With various poses, it works better than other

methods during blocking and lighting. Although the method of initialization is rough, it shows more robustness and accuracy than using the average shape. During face alignment, the multi-angle models are used as estimating shape for initialization. Then the binary features are taken in a high speed as well as low cost, while the regression is used to predict the pose increment until coverage.

In the situation of robust facial landmark detection under significant head pose and occlusion, a unified and robust framework for cascading regression is proposed, which iteratively predicts and handles the occlusions and location of the landmark. To improve robustness, the visibility of the landmark is gradually updated. For the congestion estimation, the supervised regression method is used instead of the direct estimation of the binary blocking vector. Besides, the blocking pattern is explicitly added as a constraint to improve the performance of prediction blocking. For landmark detection, the visibility of landmarks, the local appearance, and the local shape are updated iteratively. For the occlusion of the landmark detection, based on the possibility of different visibilities, different points of different processing methods, a clear prediction of the shape features.

3.2 Feasible improvements in research areas

To solve complex classification tasks better and get more details, we can partition input based on Class label distribution, in the way of data cluster. The algorithm in the situation of robust facial landmark detection under significant head pose and occlusion can be improved in the following aspects, real time tracking detection or directly improve its ability to solve more complex situation in the actual situation. In each method, we can estimate data using the regression-based stage-by-stage algorithm to improve accuracy and speed. DCNN method can be used in pose and multi-angle-based detection to separate input and analyze data.

4. Conclusion

In this paper, we conclude several state-of-art methods on facial landmark detection. We analyze the advantages and disadvantages and give our opinions on them. Then, we propose some feasible improvements in research areas.

5. References

[1] W. J. Baddar, J.Son, D. H. Kim, S.T.Kim, and Y. M.Ro, “A deep facial landmarks detection with facial contour and facial components constraint,” Image and Video Systems Lab , School of Electrical Engineering, KAIST,Daejeon,305-701, Republic of Korea

[2] A. Schwarz, Z. Lin, and R. Stiefelhagen, “Highly efficient orientation and position estimation,”Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology,Robert Bosch GmbH., fanke.schwarz, zhuang.ling@bosch.com , rainer.stiefelhagen@kit.edu

[3] A. R. Romero, S. Jayawardena,M.Cox and P.V. K.Borges, “Partitioning the input domain for classification,” Autonomous Systems Laboratory, CSIRO, Australia,Autonomous Systems Lab, ETH Zurich

[4] H. Qi, Q. Zhao, X.Wang, M. Pei, “Pose-indexed based multi-view method for face alignment,” Beijing Key Lab of Intelligent Information Technology,School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

[5] Y. Wu, Q.Ji, “Robust facial landmark detection under significant head pose and occlusion,” ECSE Department, Rensselaer Polytechnic Institute,110 8th street, Troy, NY, USA,{wuy9,jiq}@rpi.ed.