International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

Facial Landmark Localization – A Literature Survey

Dhananjay Rathod[Å*], Vinay A, Shylaja SS[Å] and S Natarajan[Å]

[Å]Department of Information Science and Engineering, PES Institute of Technology, Bangalore, Karnataka, India

Accepted 25 May 2014, Available online 01 June2014, **Vol.4, No.3 (June 2014)**

Abstract

Automatic facial point detection plays arguably the most important role as an intermediary step for many subsequent face processing operations that ranges from biometric recognition to the understanding of mental states. Distinguishable characteristic of points on face is known as facial landmarks eg. eye corner, mouth corner etc. Though it is conceptually very simple to locate landmarks, the computer vision problem has proven extremely challenging due to compound factors such as scale, pose, expression, occlusion, illumination and inherent face variability. The purpose of this survey is to give an overview of facial landmarks localization techniques and their progress over last 7-8 years.

Keywords: *Facial landmarks, localization, detection, face tracking, face recognition*

1. Introduction

A landmark is a recognizable natural or man-made feature used for navigation feature that stands out from its near environment and is often visible from long distances. Facial landmarks is defined as the detection and localization of certain keypoints points on the face which have an impact on subsequent task focused on the face, like animation, face recognition, gaze detection, face tracking, expression recognition, gesture understanding etc. Facial landmark are a prominent feature that can play a discriminative role or can serve as anchor points on a face graph.

Facial landmarks are the nose tip, eyes corners, chin, mouth corners, nostril corners, eyebrow arcs, ear lobes etc. For ease of analysis most landmark detection algorithm prefers an entire facial semantic region, such as the whole region of a mouth, the region of the nose, eyes, eyebrows, cheek or chin. The facial landmarks are classify in two groups, primary and secondary , or fiducial and ancillary. This distinction is based on reliability of image features detection techniques. For example, the corners of the mouth, of the eyes, the nose tips and eyebrows can detected relatively easily by using low level image features, e.g. SIFT, HOG. The directly detected landmarks are referred as fiducial. The fiducial group of landmarks and they play a more prominent role in facial identity and face tracking. The search for secondary landmarks is guided by primary landmarks. The secondary landmarks are chin, cheek contours, eyebrow and lips midpoints, non-extremity points, nostrils. It takes more prominent role in facial expression.

2. Application

This section list out the applications where facial landmarks plays a prominent role:

The facial landmarks plays a prominent role in application like face recognition, face animation, facial expression analysis, feature based face recognition, 3D face reconstruction, face tracking, registration, head gesture understanding, verification, surveillance system, lip reading, image editing, sign language interpretation, facial expression transfer, to identify facial attributes and much more. Below we give in more details how these applications are dependent on landmarks:

- Face recognition: Face recognition first locates the eye region and then extract holistic features from the windows centered on various regions of interest. The coordinates of located landmarks give rise to a number of geometric properties as angle and distance between facial components.
- Expression understanding: Facial expression forms a visual channel for nonverbal messages and emotions. They also in supporting role while spoken in communication. The coordinates of landmarks and spatial configuration provides a way to analyze facial expression and to describe head gestures.
- Face registration: Face registration is the most important factor affecting face recognition performance. Landmark points are used to establish point to point correspondence for face registration. It is also be useful to build 3D face models from multiple images or stereo or sequence of video.
- Face tracking: Most face tracking algorithms works on tracked landmark sequences. Face graph model is fitted to 60-80 facial landmarks. Face tracking is realized by letting the model graph to evolve according to facial components, geometrical relations between facial components and face shape parameter. The advantageous of landmark based tracking is that

*Corresponding author: **Dhananjay Rathod**

both facial deformations and the head motion are estimated jointly.

3. Challenges of Landmark Localization:

Despite the conceptual simplicity of facial landmarks detection, In computer vision there are some challenges. The emerging applications like surveillance system, gesture recognition requires that landmark localization algorithms should run in real time parallel with the computational power of an embedded system, such as intelligent cameras. Such type of application requires a more robust algorithms against a confounding factors such as illumination effects, expression and out of plane pose. There are four main challenges in localizing facial landmarks are as follows:

- Variability: Landmark appearances differ due to extrinsic factors such as partial occlusion, pose, illumination, camera resolution and expression, also due to intrinsic factors such as face variability between individuals. Facial landmarks can sometimes be only partially observed due to hand movements or self-occlusion due to extensive head rotations or occlusions of hair. Also facial landmark detections are difficult because of illumination artifacts and facial expressions. A facial landmark localization algorithm that delivers the target points in a time in an efficient manner and works well across all intrinsic variations of faces has not yet been feasible.
- Accuracy and number of landmarks require: Based on the intended application the number of landmarks and its accuracy varies. For example, In face recognition or in face detection tasks, primary landmarks like two mouth corner, four eyes corner and nose tips may be adequate. On the other hand, higher level tasks face animation or facial expression understanding require greater number of landmarks e.g. from 20-30 to 60 - 80 with higher accuracy. Fiducial landmarks are need to be determine with more accuracy because they often guide the search of secondary landmarks.
- Lack of globally accepted and error free dataset: Most of the dataset provides annotations with different markups and accuracy of their fiducial point is questionable. The accuracy of landmark localization algorithm is largely depend on the data set used for training. Each algorithm uses different dataset to train and evaluate performance so it is difficult to compare algorithms.
- Acquisition conditions: Acquisition conditions, such as resolution, background clutter, illumination can affect the landmark localization performance. The landmark localizers trained in one database have usually inferior performance when tested on another database.

4. Facial Landmark Localization Techniques

4.1 Component based Deformable Mode

(Yuchi Huang, et al,2007) presented component based deformable model for generalized face alignment, they used a novel bi-stage statistical framework to account for both local and global shape characteristics. It uses separate Gaussian models for shape components instead of using statistical analysis on the entire shape which preserve more detailed local shape deformations. Each model of components used the Markov Network for search strategy. They used Gaussian Process Latent Variable Model to give control of full range shape variations, hence it makes out better description of the nonlinear interrelationships over shape components. This approach allows system to preserve the full range low frequency shape variations, and also the high frequency local deformations caused by exaggerated expression.

Database: YALE FACE DATABASE B
Key point detected: 79
Image / video: Images only
Output faces:

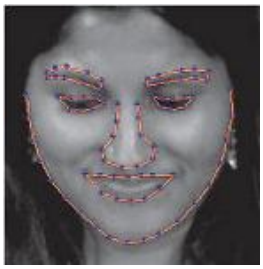

Fig.1 Output of component based deformable model

4.2. Cascade Deformable Shape model:

(Xiang, et al, 2013) presented a two-stage cascaded deformable shape model to effectively and efficiently localize facial landmarks with large head pose variations. For face detection, they proposed a group sparse learning method to automatically select the most salient facial landmarks. 3D face shape model detects pose free facial landmark initialization. The deformation executes in two stages, the first step uses mean-shift local search with constrained local model to rapidly approach the global optimum. Second step, uses component-wise active contours to discriminatively refine the subtle shape variation. It improves performance over CLM and multi-ASM in face landmark detection and tracking.

Database: MultiPIE, AR, LFPW,LFW and AFW ,Talking Face Video
Key point detected: 65
Image / video: Images and video both
Output faces:

Fig.2 Output of cascade deformable shape model

4.3. Constrained Local Neural Field model (CLNF)

(Baltrusaitis, et al, 2013) presented the Constrained Local Neural Field model for facial landmark detection. This model uses probabilistic patch expert (landmark detector) that learn non-linear and spatial relationships between the input pixels and the probability of a landmark being aligned and Non-uniform Regularized Landmark Mean-Shift (NRLM)optimization technique, which takes the account the reliabilities of each patch expert leading to better accuracy.

Database: LFPW and Helen
Key point detected: 65
Image / video: Images only
Output faces

Fig.3 Output of constrained local neural field model

4.4. Deep Convolutional Network Cascade

(Yi Sun, et al, 2013) presented three level deep convolitonal networks cascade model. The outputs of multiple networks at each level are fused to give robust and accurate estimation. At first level, convolutional networks extracts high level features over whole face region. It has two advantages first, the texture context information over the entire face is utilized to locate keypoint. Second, the network are trained to predict all the keypoints simultaneously, geometric constraints among keypoints are encoded .Therefore, method avoids local minimum caused by ambiguity and data corruption. In next two level the initially predicted keypoints are finely tune to achieve high accuracy. This method is not good for locating facial landmarks in large number .

Database: LFW,LFPW and BioID
Key point detected: 5
Image / video: Images only
Output faces:

Fig.4 Output of deep convolutional network cascade

4.5. Coarse-to-fine Convolutional Network Cascade

(Zhou, et al, 2013) presented four level convolutional network cascade, which tackles the problem in coarse-to-fine manner. Each network level refine a subset of facial landmarks generated in previous network levels and predicts explicit geometric constraints like the position and rotation angles of specific facial component to rectify inputs of the current network level. The first level estimate bounding boxes for inner points and contour points separately. Second level, predicts an initial estimation of the positions for the inner points which are refined in third level for each component. The last layer improves the predication of mouth and eyes by taking rotated image patch as new input. Two levels of separate networks are used for contour points. This method is good for locating facial landmarks in huge numbers (above 50).

Database: 300-W (300 faces in wild)
Key point detected: 68
Image / video: Images Only
Output faces:

Fig.5 Output of coarse-to-fine convolutional network cascade model

4.6. Explicit Shape Regression

(Xudong Cao, et al, 2012) presented very efficient and highly accurate Explicit Shape Regression approach. This model directly learns a vectorial regression function to infer the whole facial shape and explicitly minimize localization error in training data. The inhereent shape encoded into the regressor in learning framework applied from coarse to fine to during testing. This approach uses two level boosted regression, a correlation based feature and shape-indexed features selection method hence regression is more effective. It shows highly accurate and efficient results. Regression process is extremely fast in test 15 ms for 87 landmarks shape.

Database: BioID, LFPW and LFW87
Key point detected: 5,29 or 87
Image / video: Images Only
Output faces:

Fig.6 Output of explicit shape regression model

4.7. Constrained local models (CLM)

(Cristinacce, et al, 2008)This model is developed on Active Appearance Model (AAM). CLM differs from AAM because it is not a generative model for the whole face, instead it produces landmark templates iteratively and use a shape constrained search technique. The position vectors of the landmarks templates are estimated using Bayesian formulation. The posterior distribution in Bayesian formula incorporates both the image information via template matching scores and the statistical shape information. Therefore, positions of new landmarks are predicted in the joint shape model and the light of the image, and then templates are updated by sampling from the training images.

Database: BioID and XM2VTS
Key point detected: 22
Image / video: Images Only
Output faces

Fig.7 Output of constrain local model

4.8. Parameterized Kernel Principal Component Analysis (PKPCA)

(Torre, et al, 2008) presented the parameterized kernel principle component analysis (PKPCA) model by extending KPCA to incorporate geometric transformation into formulation and applying gradient descent algorithm for fast alignment. This model differ from PCA because it can model non-linear structure in data in variant to rigid or non-rigid deformations. It does not required manually labeled training data.

Database: CMU Multi-PIE
Key point detected: 46
Image / video: Images Only
Output faces

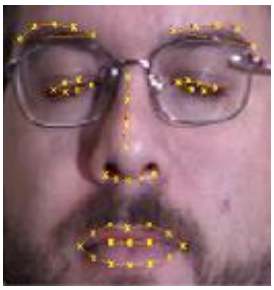

Fig.8 Output of PKPCA model

4.9. Multiple Kernel Learning

(Rapp, et al, 2011) presented multiple kernel, it uses two patches on the face. One covers the eye region and the other covering the mouth region. In testing, pixels of a respective regions should be part of a target landmark, using the multi-resolution windows (progressively smaller nested windows) texture data is extracted that captures information from global to local view range. Every resolution level, information is passed into a different kernel and the convex combination of these kernels, instead of concatenating pyramidal information. For each resolution level there is dedicated kernel which forms a multi-kernel SVM. The multi-kernel SVM is trained using center-surround architecture and the surround windows forming the negative examples. After the discovery of landmark points, without having any spatial relationship among them initially. To reach the plausible shapes, point distribution models are invoked. The point distribution models are particularly focuses to the mouth and to the eye-eyebrow pairs. The shape alternatives are evaluated using Gaussian mixture models (GMMs), so that the point combinations that possess the highest sum of SVM scores and that fit best to the learned models are selected.

Database: Cohn-Kanade and Pose, Illumination and Expression (PIE)
Key point detected: 17
Image / video: Images only
Output faces:

Fig.9 Output of multiple kernel learning

4.10. Semi-Supervised learning

(Tong, et al, 2012) address to the often imperfect and tedious task of manual landmark labeling, and to overcome this suggest a scheme to partly automate land-marking. In their method, a small percentage (e.g., 3%) of faces need to be hand labeled, while most the faces are automatically marked. This is done by propagating the few exemplars landmarking information to the whole set. On the minimization of the pair wise pixel differences resulting in two error terms, the learning is based. The penalty in one term makes the warping of each un-marked image toward all other un-marked images, irrespective of the content they become more alike. While penalty in the other term controls the warping of un-marked images toward marked images. The warping function itself can be a piecewise affine warp to model a non-rigid transformation or a global affine warp for the whole face.

Database: Notre Dame, Caltech 101and FERET

Key point detected: 33
Image / video: Images only
Output faces:

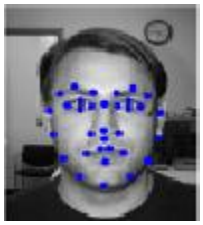

Fig.10 Output of semi supervised learning

4.11. Bayesian Approach

(Belhumeur, et al, 2013) presented a fully Bayesian approach to find landmark positions from local evidences. An interesting aspect about their work is that the local detector outputs are collected from a cohort of exemplars (sample faces with annotated landmarks), which provide non-parametrically the global model information. The local detector consists of a sliding window whose size is proportional to IOD and which collects SIFT features. In the next stage, the global detector models the configurational information of the ensemble of primary points. The joint probability of the location on 'n' landmarks and the vector of their local detectors outputs is maximized. This method surpasses in accuracy the performance of the manual landmarking in most of the 29 landmarks considered.

Database: BioID and LFPW
Key point detected: 29
Image / video: Images only
Output faces

Fig.11 Output of Bayesian approach

4.12. Conditional Regression Forests

(Dantone, et al, 2012) proposed pose-dependent landmark localization scheme that is achieved by conditional random forests. Conditional regression forests learn several conditional probabilities over the parameter space, and deals with facial variations in appearance and shape, while regression forests try to learn the probability over the parameter space from all face images in the training set. The head pose is quantized into five segments of i) right profile ii) right iii) front iv) left and v) left profile faces and specific random forests are trained. Both texture and 2D displacement vectors that are defined from the centroid of each patch to the remaining ones described the local properties of a patch. Texture is described by normalized gray values in order to cope with illumination changes in addition to Gabor filter responses. Training of conditional random forests is very similar to random forests, except that the probability of assigning a patch to a class is conditioned on the given head pose. This method locates landmarks in a query image at real-time speed.

Database: LFW
Key point detected: 10
Image / video: Images Only
Output faces:

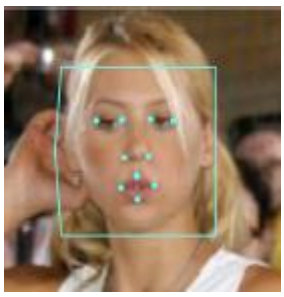

Fig.12 Output of conditional regression forests

4.13. Multi View

(Zhanpeng Zhang, et al, 2014) presented a real time Multi-View facial landmark localization in RGB-D images. This system is able to estimate 3D head pose and 2D landmark localization. The model extract random local binary patterns of different scales, and estimate the facial parameters with hierarchical regression techniques. At first, 3D face positions and rotations are estimated via a random regression forest. Afterwards, 3D pose is refined by fusing the estimation from the RGB observation. The depth channel and RGB channel are used at different stages, the depth input is fed to the random forest for face detection and pose estimation at the beginning stage, the RGB input is fed to gradient boosted decision trees (GBDT) for head pose and hierarchical facial landmark location regression when the face is available. The pose estimation results from the depth and RGB inputs are weighted and combined to improve the precision.

Database: BIWI Kinect Head Pose
Key point detected: 13
Image / video: Images only
Output faces:

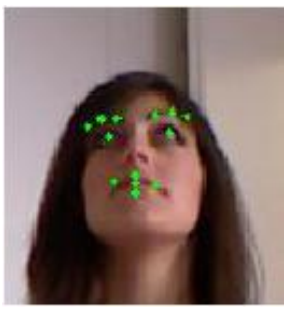

Fig.13 Output of multi-view model

4.14. Combinatorial Search and Shape Regression

(Sukno, et al, 2012) have managed the combinatorial problem using RANSAC algorithm. First, they find out the reliable features using spin images as features and the missing are regressed using the multivariate Gaussian model encompassing all 3D landmark coordinates. The correct landmarks are sort out from the multitude of candidate points, all combinations of four points are used and RANSAC is used as the basis of the feature matching procedure. For the missing landmarks the median of the closest candidates is considered. The PCA instrumented shape fitting term for the detecting landmarks is has been used in the cost function consist of a part accounting for the reconstruction error, the other part accounts for the distance from the inferred landmarks to their closest candidates.

Database: FastSCAN
Key point detected: 8
Image / video: Images and video both
Output faces

Fig.14 Output of combinatorial search and shape regression

4.15. Tree-structured models

(Xiangxin Zhu, et al, 2012) address the three linked problems of face detection, face pose estimation and facial landmarks localization. This algorithm is shape driven and local and global information are merged right from beginning. Since pose is part of estimation, the algorithm practically works as a multi-view algorithm. Multi-view implemented by considering several (30 to 60) local patches that are connected as a tree which collectively describe the landmark related region of the face.

Database: CMU MultiPIE and AFW
Key point detected: 61
Image / video: Images only
Output faces:

Fig.15 Output of tree structured model

4.16. Supervised Descent Method

(Xuehan Xiong, et al, 2013) presented a Supervised Descent Method (SDM) for localizing facial landmarks. At first, it takes an image with manually labeled landmarks. Then it ran through training images to give initial configuration of landmarks. It uses SIFT function to extract initial landmarks. In training, SDM tries to minimize difference between manually labeled landmarks and initially located landmarks by SIFT (Δx). This method does not learn any shape or appearance model in advance from training data. The SDM learns a series of descent direction and re-scaling factors to produce a sequence of updates.SDM directly learns descent direction from training data by learning a linear regression between, (Δx) and difference of SIFT value of manually and extracted landmarks ($\Delta \Phi$). In testing, based on descent direction and re-scaling factors learn in training SDM estimates landmarks. SDM learns descent direction without computing Jacobian nor Hessian matrix, which are computationally expensive. SDM is fast and accurate as compare to shape models.

Database: LFPW and LFW-A&C, RU-FACS
Key point detected: 49
Image / video: Images and video both
Output faces

Fig.16 Output of Supervised Descent Method

5. Performance Metrics

For evaluating landmark localization performance one can define two different metrics-

a) Ground truth based localization error.
b) Task-oriented performance.

The straightforward way to estimate landmark localization performance is to validate with the manually annotated ground-truths. In task oriented performance, impact of the landmarking accuracy is measured on performance scores of a task. The localization performance can be expressed in terms of the normalized root mean square error (NRMSE), it will be computed per landmark or normalized mean square error (NMSE) can be averaged over all the landmarks and generates global precision figure. Inter Ocular Distance (IOD) which defines the distance between two eyes center is used for normalizing landmarks. Performance measure can be made independent of the camera zoom factor by dividing normalizing landmark localization errors with IOD.

Conclusions

This paper presents some of the recent survey of landmark localization techniques with a brief insight and results obtained by the authors. We observed that few of them work effectively on 2D images and few on 3D images and video.

References

Yuchi Huang, Qingshan Liu, Metaxas, D (2007), A Component Based Deformable Model for Generalized Face Alignment, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1-8.

Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, Metaxas, D.N (2013), Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model, *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 1944-1951.

Baltrusaitis, T., Robinson, P., Morency, L.-P(2013), Constrained Local Neural Fields for robust facial landmark detection in the wild, *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*,pp.354-361.

Yi Sun, Xiaogang Wang, Xiaoou Tang (2013), Deep Convolutional Network Cascade for Facial Point Detection, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3476-3483.

Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin(2013), Extensive Facial Landmark Localization with Coarse-to-fine Convolutional Network Cascade, *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pp. 386-391.

Xudong Cao, Yichen Wei, Fang Wen, Jian Sun (2012), Face alignment by Explicit Shape Regression, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2887-2894.

David Cristinacce, Tim Cootes (2008), Automatic feature localisation with constrained local models, *Pattern Recognition*, 41, pp. 3054-3067.

De la Torre, F., Minh Hoai Nguyen (2008), Parameterized Kernel Principal Component Analysis: Theory and applications to supervised and unsupervised image alignment, *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1-8.

Rapp, V., Senechal, T., Bailly, K., Prevost, L (2011), Multiple kernel learning SVM and statistical validation for facial landmark detection, *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 265-271.

Y Tong, X Liu, FW Wheelerb, PH Tub, *(*2012*)* Semi-supervised facial landmark annotation. *Comput. Vis. Image Understand*. 116*(*8*)*, pp. 922–935.

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N. (2013), Localizing Parts of Faces Using a Consensus of Exemplars, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35, pp. 2930-2940.

Dantone, M., Gall, J., Fanelli, G., Van Gool, L (2012), Real-time Facial Feature Detection using Conditional Regression Forests, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2578-2585.

Zhang, Z., Zhang, W., Liu, J., Tang, X. (2014), Multi-view Facial Landmark Localization in RGB-D Images via Hierarchical Regression with Binary Patterns, *Circuits and Systems for Video Technology, IEEE Transactions on*, pp, pp.1.

Federico M. Sukno, John L. Waddington, and Paul F. Whelan, (2012), 3D Facial Landmark Localization Using Combinatorial Search and Shape Regression, *ECCV'12 Proceedings of the 12th international conference on Computer Vision*, volume -1, pp. 32-41.

Xiangxin Zhu, Ramanan, D (2012), Face detection, pose estimation, and landmark localization in the wild, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2879-2886.

Xuehan Xiong , de la Torre, F.(2013), Supervised Descent Method and Its Applications to Face Alignment, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 532-539.