A Facial Features Detector Integrating Holistic Facial Information and Part-based Model

Eslam Mostafa[1,2] Asem A. Ali[3] Ahmed Shalaby[4] Aly Farag[1]
[1]CVIP Lab, University of Louisville, Louisville, KY 40292, USA.
[2]Electrical Engineering Department, Alexandria University, Alexandria, Egypt.
[3]Electrical Engineering Department, Assiut University, Assiut 71515, Egypt.
[4]Kentucky Imaging Technology (KIT), Louisville, KY 40245, USA.

Abstract

We propose a facial landmarks detector, in which a part-based model is incorporated with holistic face information. In the part-based model, the face is modeled by the appearance of different face parts and their geometric relation. The appearance is described by pixel normalized difference descriptor. This descriptor is the lowest computational complexity as compared with existing state-of-the-art while it has a similar accuracy. On the other hand, to model the geometric relation between the face parts, the complex Bingham distribution is adapted. This is because the complex Bingham distribution has a symmetric property so it is invariant to rotation, scale, and translation. After that the global information is incorporated with the local part-based model using a regression model. The regression model estimates the displacement to the final face shape model. The the proposed detector is evaluated on two datasets. Experimental results show that it outperforms the state-of-the-art approaches in detecting facial landmarks accurately.

1. Introduction

Face understanding is considered one of the most important topics in computer vision field since the face is a rich source of biometrics in social interaction. Facial landmarks extraction is the corner stone in the success of different face analysis and understanding biometrics applications. Facial features, also known as facial landmarks or facial fiducial points, have a semantic meaning. Facial features are mainly located around facial components such as eyes, mouth, nose, and chin. Facial feature points detection (FFPD) refers to a supervised or a semi-supervised process using abundant manually labeled images. FFPD usually starts from a rectangular bounding box, which implies the location of a face and is returned by a face detector e.g., [1, 2]. This bounding box can be employed to initialize the positions of facial features.

Facial features can be classified into three types: Points label parts of the face with an application-dependent significance, such as the center of an eye or sharp corners of a boundary. Points label application-independent elements, such as the highest point on the face in a particular orientation, or the highest point along the bridge of the nose (the curvature extrema). Finally, points are interpolated from the previous two types such as points along the chin. According to various application scenarios, different numbers (e.g., 17, 29, 68) of facial feature points are labeled. Whatever the number of the points is, these points should cover several frequently-used areas: eyes, nose, and mouth. These areas carry the most important information for both discriminative and generative purposes. Generally speaking, more points indicate richer information, although it is more time-consuming to detect all the points.

Detecting the facial features is a challenging problem due to both the rigid (scale, rotation, and translation) and the non-rigid (such as facial expression variation) face deformations. Existing methods of facial features detection can be broadly grouped into three categories: Constrained Local Model (CLM)-based methods, active appearance model-based methods, and regression-based methods.

Active Appearance Model (AAM)-based methods model the appearance variation from a holistic perspective. In the training phase of these algorithms, principal component analysis (PCA) is applied to a set of labeled faces (manually annotated faces) to model the intrinsic variation in the shape and the texture. This results in a parameterized model (eigenshapes and eigenfaces) that can represent large variations in the shape and the texture with a small set of parameters. The AAM algorithm aims to find the model's coefficients that minimize the difference between the texture as sampled from a test image and the texture that is synthesized by the model. However, the coefficients of the

model are defined over a high dimensional space, making it is impossible to find its global maximum. Many trials have been done as an improvement and extension for AAM by cootes [3]. In their survey paper, Gao et al. [4] discussed the recent developments on AAM.

In the regression-based methods, the shape is directly estimated from the appearance without learning any shape or appearance models. Regression-based approach learns a regression function that maps the image appearance (features) to the target output (shape). Zhou and Comaniciu [5] proposed a shape regression method based on boosting [6, 7]. Their method proceeds in two stages: First, the rigid parameters are found by casting the problem as an object detection problem, which is solved by a boosting-based regression method. Second, a regularized regression function is learned from perturbed training examples to predict the non-rigid shape. Haar-like features are fed to the non-rigid shape regressors. Cao et al. [23] proposed a two-level cascaded learning framework based on boosted regression [9]. Unlike the method in [5], which learns the regression map for each individual facial feature, their method directly learns a vectorial map that combines all landmarks. The main drawbacks of the regression methods are the sensitivity to initialization and the need of huge amount of memory compared with the CLM-based methods.

In the constrained local model-based methods, the local texture and the shape prior models are the main components. For the texture model, the local texture around a given facial feature is modeled i.e., the pixels intensity in a small region around the feature point. While for the shape model, the relationship among facial features are modeled. Both models are learned from labeled exemplar images (manually labeled images).

The texture-based features detection can be formulated either as a regression or a classification problem. For the regression problem, the displacement vector from an initial point to the actual feature point is estimated. For the classification problem, a sliding window runs through the image to determine if each pixel is a feature or a non-feature. The sliding window searching approach is the standard approach for object detection. This approach requires two steps object representation and classification. Instead of directly using the pixel intensity as a descriptor, the texture model can be constructed using different descriptors such as Haar-like [10], local binary pattern (LBP) [11], Gabor [12], scale-invariant feature transform (SIFT) [13]. The seminal work of viola and Jones [1] is considered the corner stone for many development in the area of object detection. The object is represented by Haar-like features and adaboost is used for classification and feature selection. Recently, many researchers [19] used the histogram of gradient orientation for an object representation and Support Vector Machine (SVM) for classification. The histogram of gradient orientation is invariant to illumination and affine transformation. However, the main drawback of using this representation over Haar-like features with adaboost is the execution time.

Texture-based detectors are imperfect for many reasons. One of them is the visual obstructions (e.g., hair, glasses, hands, etc.), which can greatly affect the results. Another reason is that the detection of each facial feature is independent from the others and it ignores the relation among these facial feature points. To overcome these disadvantages, constraints related to the relative positions of the facial features can be established using a shape model. The shape model either is used to filter the output of the texture model or both models are combined together into a single formula.

Cristinacce et al. [14] modeled the relative positions of facial features by pairwise reinforcement of feature responses and modeled the texture around facial features using PCA as in Active Shape Model (ASM). Valstar et al. [15] modeled the shape using Markov Random Field (MRF) and the texture using Haar-like features with a boosting classifier. These two approaches use a shape model to filter the output of the texture model. They use a single distribution for the shape model, but this is not suitable for modeling a wide range of poses. Felzenszwalb et al. [16] modeled the relation between facial features by a graph tree where the relation between each two nodes is a gaussian distribution and the texture is modeled using an iconic representation. To handle different poses, Everingham et al. [17] extended the relation between facial feature points from a single Gaussian distribution into a mixture of Gaussian trees. Also, they used Haar-like features with boosting instead of the iconic representation to represent the texture around facial feature points. Zhu et al. [18] built on [17], but they combined the texture and shape models into a single formula and used Histogram of Oriented Gradients (HOG) features [19] to represent the texture around each facial feature points. Belhumeur et al. [24] used a non-parametric approach for shape modeling. They used information from their large collection of diverse labeled exemplars and represented the texture around each facial feature point using SIFT features. They used SVM [21] to classify each pixel as a candidate facial feature or not. Their algorithm takes 1 sec to detect a facial feature.

Unlike conventional state-of-the-art approaches, which use the shape model to filter the results of texture-based detectors, in the proposed work, the facial features detection problem is formulated as minimizing an energy function that simultaneously incorporates information from both texture and shape models. However, the parametric shape model have a drawback since it penalizes shapes that are far from the mean shape. Therefore, we propose adding another stage to refine the output, which corresponds to the minimum energy, by using a regression model that esti-

mates the displacement to the final face shape model. The regression model is based on a global texture model to give complementary information with the local texture model, which is used in the previous stage. Therefore, the proposed facial features detector combines the advantages of the part-based face model and the holistic face model. The face is modeled using a part-based model where the texture around facial points is modeled using the pixel difference descriptor and complex Bingham distributions are used to model features' relative positions (the shape). The output of the first stage is refined by a regression model that is build using non-parametric global information. We built on Mostafa and Farag work [22], however, there are two main differences: (a) The appearance is described by a pixel normalized difference descriptor. This descriptor is three times faster than other state-of-the-art texture descriptors, as shown in experimental results. (b) Incorporate non parametric holistic information to proposed face model for achieving few pixels accuracy.

2. The Proposed Face Model

In the first stage, the facial image is modeled using a local parametric model to detect N facial features. The model is based on pixel normalized difference texture with a SVM classifier as a local texture detector that is combined with a mixture of complex Bingham distributions as parametric shape model.

2.1. Local Texture Detector

The texture-based features detection is formulated as a classification problem where a sliding window runs through a sub-image and a SVM classifier determines if each pixel is a feature or a non-feature. The texture around each facial feature is represented by the difference in values of random pixels in the neighbourhood of this feature. To make this descriptor illumination invariant, the difference is normalized by the average intensity. This descriptor is the lowest computational complexity compared with existing state-of-the-art while it has a similar accuracy.

To detect a facial point i, the sliding window-based classifier scans its corresponding search area. Then the score $S(D_{z_i})$, which measures if the pixel at position z is the feature i, is calculated as follows.

$$S(D_{z_i}) = \sum_{t=1}^{r} \alpha_{t_i} \mathfrak{z}_i \mathfrak{S}_{t_i}, \tag{1}$$

where α_{t_i} is the weight of each support vector t for the feature i, $\mathfrak{z}_\mathfrak{i}$ is the extracted texture, which is the random pixel normalized difference, and $\mathfrak{S}_{t_i}$ are the support vectors [21].

For a perfect texture-based detector, the response map, which represents the score at each pixel in the search area, of the classifier is homogenous as the probability of the pixel being a feature is high at the true position and decreases smoothly going away from this position. Therefore, the output of the SVM classifier should be regularized to handle false positives in the classification step. The classifier is regularized with a variance normalization factor by dividing the output probability of the classifier with the standard deviation $\sigma_{\aleph(z)}$ of the output probability among the neighborhood $\aleph(z)$. Then, the texture-based probability $P(D_{z_i})$ of the position z to be the feature i can be written as

$$P(D_{z_i}) = \frac{K}{\sigma_{\aleph(z_i)}} S(D_{z_i}), \tag{2}$$

where K is a normalization constant.

Candidate positions of a feature are the peaks in the corresponding response map and they are estimated using the non-maximal suppression technique. Figure 1 shows an example of the response map for the tip of the nose as well as its candidate positions. The candidates are illustrated for subjects with different poses. Also, Fig. 2 shows a subset of positive samples for the tip of the nose. This figure shows the high variation of the appearance within the positive samples.

In the training stage, the face detection box is resized to 50×50 pixels. The patch size around a given facial feature position has been empirically determined to be 13×13 pixels for optimum running time and accuracy. Positive samples are chosen at the corresponding manually annotated locations. Whereas negative samples are chosen away from the corresponding annotated locations by at least 10 pixels.

2.2. Combining Texture and Shape Model

We formulate this problem as hidden variables (positions of facial features $Z = [z_1, z_2 ... z_N]$) are estimated based on observable variables (image gray level I). This problem can be represented as a Bayesian framework of Maximum-A-Posteriori (MAP) estimation. The probability model of the input image and the facial feature positions is given by the joint distribution $P(I, Z) = P(I|Z)P(Z)$, where $P(I|Z)$ is the conditional distribution of the original image given the facial feature positions i.e., the texture-based model. $P(Z)$ is the distribution of the facial feature positions i.e., the shape prior model. The MAP estimate of facial feature positions given the image is expressed as

$$Z^* = \arg\max P(I|Z)P(Z). \tag{3}$$

Since each facial feature has its sliding window classifier scanning its corresponding search area, the output of each facial point texture detector can be considered independent from the others. Therefore, $P(I|Z)$, which represents the probability of similarity between the texture of the face to off-line model given the facial feature vector, is the overall

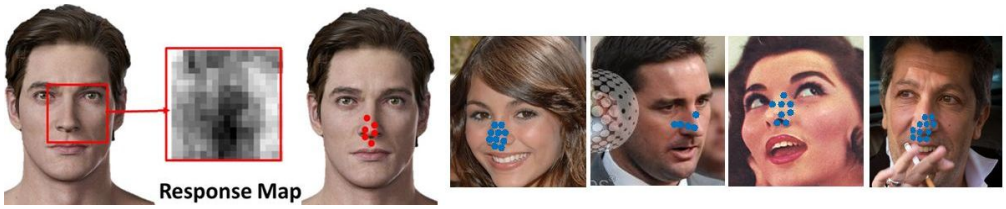

Figure 1. The response map of the tip of the nose and examples of its candidates for different subjects.

probability of N facial features based only on the texture-based detector and is given by

$$P(I|Z) = \prod_{i=1}^{N} P(D_{z_i}). \tag{4}$$

A mixture of complex Bingham distributions is chosen to represent the shape model i.e., the distribution of the facial feature positions $P(Z)$. The complex Bingham distribution is more robust in modeling the joint probability of the location of facial features than existing models. Existing models need a preprocessing step before using the shape prior to filter out scale, translation, and rotation using least-square approaches (e.g., Procrustes analysis), which can introduce errors to the system due to noise and outliers. Since the probability distribution function (PDF) of a complex Bingham has a symmetric property, there is no need to filter out rotation. Scale and translation can be easily removed by a simple mean and normalization step [22].

Therefore, the MAP estimate of the features can be formulated as an energy minimization of the function $E(Z)$:

$$E(Z) = -\frac{HZ^*AHZ}{\| HZ \|^2} - \sum_{i=1}^{N} \log P(D_{Z_i}), \tag{5}$$

where A is a $(N-1) \times (N-1)$ complex Bingham parameters matrix and H is the Helmert matrix. Note that this energy function is non-linear so it is not amenable to gradient descent-type algorithms. It is solved by a stochastic approach, which is simulated annealing.

3. Non-parametric global information for detection refinement

Random fern regression is used to find the displacement from the positions of the detected features Z^* that minimize the energy function Eq. 5 to more accurate positions with few pixels accuracy. The regression model learns the relation between the appearance around these detected points Z^* and their displacement from the ground truth positions. Since this relation is very complex, a single regression model is not sufficient.

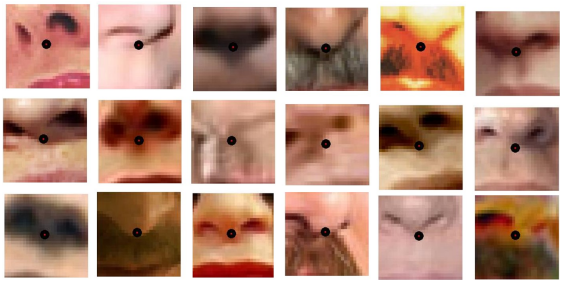

Figure 2. Illustration of intrinsic variation in the appearance of the nose tip

Therefore, a boosted regression model is used. Where $T = 500$ random ferns regressors ($\Im^1$,$\Im^2$,...,$\Im^T$) are combined in an additive manner. Given the face image I and the detected facial feature points Z^*, each random fern computes a shape increment δZ from the appearance descriptor around these points and updates the detected facial feature points in a cascaded manner:

$$Z^t = Z^{t-1} + \Im^t(I, Z^{t-1}), \qquad t = 1, 2, ..., T \tag{6}$$

where Z^t is the positions of facial feature points that are generated by the random fern regressor t, while $Z^0 = Z^*$. Each fern is learned by minimizing the sum of the alignment error in the training set. The alignment error is the difference between the detected positions for facial feature points and the corresponding ground truth positions. In the training stage, the regression function $\Im$, is learned by minimizing the alignment error as follows.

$$\Im^t = \arg\min \|\hat{Z} - (Z^{t-1} + \Im^t(I, Z^{t-1})\|, \tag{7}$$

where $\hat{Z}$ is the ground truth positions of the facial feature points, which are manually annotated. For each regressor t, the holistic appearance of the facial image is represented by approximately 1000 normalized differences of pixels that are randomly chosen around the shape Z^{t-1}. To construct a good fern, we downsample these descriptors such that each descriptor in the fern is highly discriminative to the regression target and the fern's descriptors are complementary when they are composed. To achieve this, we use correlation-based feature selection method [23]:

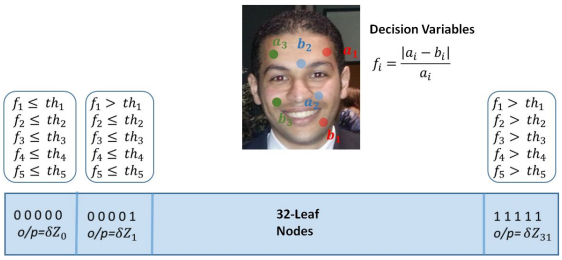

Figure 3. Illustration of regression outputs using pixel difference-based texture in the random tree regression.

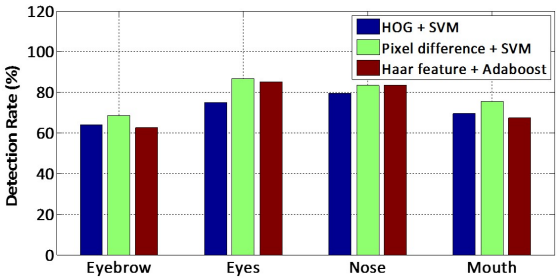

Figure 4. A detection rate comparison between three different appearance descriptors: HOG, pixel difference, and Haar-like.

1. Project the regression target, difference between positions of the current facial feature points and the ground truth, to a random direction to produce a scalar.

2. Among 1000 descriptors, select a descriptor with the highest correlation to the scalar.

3. Repeat steps (1) and (2) 5 times to obtain 5 descriptors.

4. Construct a fern by 5 descriptors with random thresholds.

The regression function in each fern is estimated by dividing the training data into 2^5 bins based on the selected 5 pixel difference descriptors. Each bin is associated with regression output δZ^t that minimizes the alignment error of the training samples falling into this bin. Figure 3 shows an illustration for a testing fern using the pixel difference-based texture.

4. Experiments

The first experiment is conducted to evaluate the effect of using the pixel difference descriptor. Figure 4 shows a comparison between different descriptors: Haar-like, HOG, and pixel difference descriptors with respect to the detection accuracy. The histogram of orientated gradient and pixel difference descriptors show a similar detection accuracy. The main differences are: the size of the window that describes the facial point appearance in the HOG descriptor is half the size of the window in the pixel difference descriptor. Larger window size captures more global information, which may be needed, however, the execution time will increase dramatically in the case of the HOG descriptor. On the other hand, the execution time of the pixel difference descriptor-based approach is not a function in the window size.

The performance of the proposed facial features detector is evaluated on BIO-ID dataset and Labeled Face Parts in the Wild (LFPW) dataset [24]. The number of facial features is chosen to be 68 points. The facial feature points detector is evaluated using the cumulative distribution of the relative error. The relative error is the distance between the

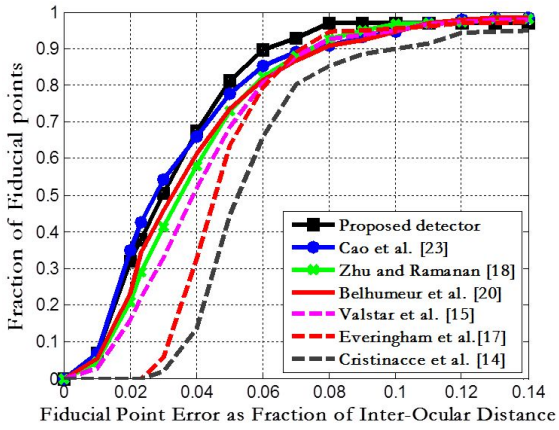

Figure 5. A comparison of the cumulative error distributions measured on BIO-ID dataset.

detected facial feature point and the corresponding manually annotated point (ground truth) divided by the ground truth distance between the two eyes. At every point in the curve, the x-axis shows the relative error, and the y-axis is the percentage of facial feature points that have relative error less than or equal the value of x-axis.

The majority of the research about facial features detection in the literature reported their results on the BIO-ID database. Therefore, it is included here as a testing dataset. The BIO-ID dataset contains 1521 images for 23 distinct subjects. Each image shows a near frontal view of a face in a controlled indoor environment without illumination and occlusion problems. On the other hand, Belhumeur et al. [24] released LFPW as a challenging uncontrolled dataset. It consists of 1432 faces from images collected from the web. The dataset contains different challenges: pose, existence of shadow, presence of occlusion objects as sunglasses or subject's hand, existence of in-plane rotation, and blurred images.

Figure 5 shows the cumulative error distributions for the proposed detector compared to those reported by [25], [18], [24], [15], [17], and [14] on BIO-ID datasets, respectively.

Figure 6. Samples of results of the proposed facial feature detector on Labeled Faces Parts in the Wild (LFPW) dataset.

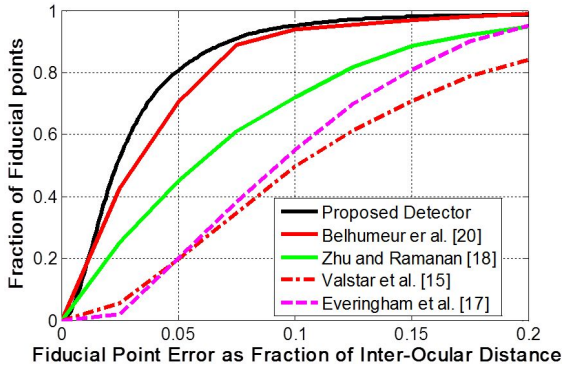

Figure 7. A comparison of the cumulative error distributions measured on LFPW dataset.

The proposed detector and detectors in [23], [18], [24], [15], and [17] have comparable performance on BIO-ID, since this database includes near frontal facial images, which are captured in controlled indoor environments with no illumination and occlusion problems. For the LFPW dataset, Fig. 6 shows samples of the results of the proposed facial features detector on this dataset. Figure 7 shows the cumulative error distributions for the proposed detector compared to the state-of-the-art approaches on this dataset. It is worth mentioning that for such experiment Cao et al. [23] and Burgos et al. [26] reported their performance using mean error as percentage of interocular distance instead cumulative error distribution. Therefore, their results are not included in Fig. 7. The proposed detector and the detector in [24] show a similar performance. This performance is the highest accuracy as compared with other approaches. However, the Belhumeur et al. detector [24] takes long time since

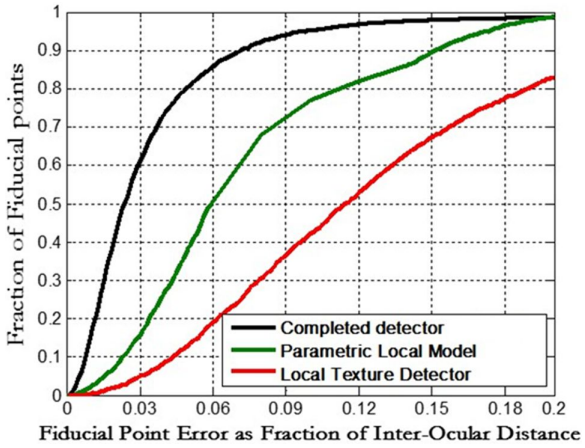

Figure 8. Effects of each component in the proposed approach: local texture detector only, local texture detector with shape constraint, and the full proposed approach.

it is based on the SIFT representation, which is extensive feature in extraction. Also, it needs extra memory as compared with most of existing algorithm, since it needs to save shapes instead of the parameters of the shape model. While the proposed shape model is a mix between the parametric and the non-parametric shape model. The parametric shape model with the texture local detector does not need a lot of memory. While, the regression random ferns need more memory but it is still less than others.

Finally, three experiments are conducted to investigate the performance of the approach's components. Figure 8 shows the results of these experiments. The first experiment highlights the accuracy of the local texture detector where the pixel difference descriptor with support vector machine is used to find the best candidate for each facial feature point without any shape constraint. The second experiment highlights the enhancement in the accuracy after integrating the parametric shape model. In this experiment, the facial feature points are detected based on optimized energy function 5 that combines the complex Bingham distributions with the texture model. The last one highlights the enhancement in the accuracy due to the fine tuning stage, which is based on the ferns regression model.

5. Conclusion

In this work, we proposed a facial landmarks detector. The proposed detector combines a part-based model and holistic face information. In the part-based model, the facial features detection problem was formulated as an energy minimization function that incorporates information from both texture and shape models. The texture around facial feature point was represented by the normalized difference in values of random pixels in the neighbourhood of this fa-

cial feature point. This descriptor is faster than other state-of-the-art texture descriptors. A mixture of complex Bingham distributions was chosen to represent the shape model. Finally, the global information was incorporated with the local part-based model using a regression model. The proposed detector results outperform the state-of-the-art in detecting facial landmarks.

References

[1] P. Viola, and M. Jones, "Robust Real-time Object Detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137-154, May 2001.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, "Cascade object detection with deformable part models," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2010, pp. 2241-2248.

[3] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 23, no. 6, pp. 681-685, Jun. 2001.

[4] X. Gao, Y. Su, X. Li, D. Tao,"A review of active appearance models," in *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 40, no. 2, pp. 145-158, 2010.

[5] S. Zhou, D. Comaniciu, "Shape regression machine," in *Proceedings of the 20th International Conference on Information Processing in Medical Imaging*, 2007, pp. 13-25.

[6] Y. Freund, R. Schapire, "decision-theoretic generalization of on-line learning and an application to boosting," in *Journal of Computer and System Science*, vol.55, pp. 119-139, 1997.

[7] J. Friedman, T. Hastie, R. Tibshiani, "Additive logistic regression: a statistical view of boosting," in *The Annals of Statistics*, vol. 38, no. 2, pp. 337-374, 2000.

[8] X. Cao, Y. Wei, F. Wen,J. Sun, "Face alignment by explicit shape regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2887-2894.

[9] N. Duffy, "Boosting methods for regression," in *Machine Learning*, vol 47, pp. 153-200, 2002.

[10] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proceedings of International Conference in Computer Vision*, 1998, pp. 555-562.

[11] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[12] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Optical Society of America, Journal, A: Optics and ImageScience*, vol. 2, pp. 1160-1169, 1985

[13] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91-110, Nov. 2004.

[14] D. Cristinacce, T. Cootes, and I. Scott, "A Multi-Stage Approach to Facial Feature Detection," in *Proceedings of the British Machine Vision Conference*, 2004, pp. 231-240.

[15] M. Valstar and B Martinez and X Binefa and M Pantic, "Facial Point Detection using Boosted Regression and Graph Models," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2010, pp. 2729-2736.

[16] F. Felzenszwalb, and P. Huttenlocher, "Pictorial Structures for Object Recognition," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55-79, Jan 2005.

[17] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy - Automatic Naming of Characters in TV Video," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 92.1-92.10.

[18] X. Zhu, and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879-2886.

[19] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

[20] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2011, pp. 545-552.

[21] C. Chang, and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[22] E. Mostafa, and A. Farag, "Complex Bingham Distribution for Facial Feature Detection," in *Proceedings of European Conference on Computer Vision Workshops*, 2012, pp. 330-339.

[23] X. Cao, Y. Wei, F. Wen,J. Sun, "Face alignment by explicit shape regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2887-2894.

[24] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2011, pp. 545-552.

[25] Q. Cao, Y. Ying, P. Li,"Similarity Metric Learning for Face Recognition",*IEEE International Conference on Computer Vision (ICCV)*, 2013.

[26] X. Burgos, P. Perona, P. Dollr, " Robust face landmark estimation under occlusion", in *Proceedings of International Conference in Computer Vision*, 2013.