International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 12, December 2015, PP 128-133
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Facial Landmark Detection using Ensemble of Cascaded Regressions

Martin Penev[1*], Ognian Boumbarov[2]

[1] *Faculty of Telecommunications, Technical University, Sofia, Bulgaria*
[2] *Faculty of Telecommunications, Technical University, Sofia, Bulgaria*

ABSTRACT

This paper presents an ensemble of regressions approach for estimation of the positions of facial landmarks in frontal and near-frontal face images. Our approach learns three different cascades of regressors and fuses their predictions into one final precise estimate using Gradient Boosting Regression Trees (GBRT). The cascaded model starts from an approximate estimate of the landmark positions and iteratively improves it. Each weak regressor is making a prediction based on Histograms of Oriented Gradients (HOG) features and performs linear SVM regression using ε-insensitive loss function. The GBRT regression produces the final result using a feature set consisting of the output prediction of each cascade together with extracted HOG features. Our approach achieves state-of-the-art results when tested on current challenging datasets.

Keywords: Facial Landmark Localization, Facial Features, HOG, Linear Regression, Ensembling of Regression Models

INTRODUCTION

The search for improved algorithms for facial landmark detection is still a subject to active research. Current state-of-the-art methods have already achieved impressive results, but there are still problems due to variability in face shapes, and many limiting factors such as head pose, variations in orientation and background, facial expressions, different lighting conditions and partial occlusions. The most typical applications where the precise face alignment plays a significant role are the analysis of expressions and emotions, face recognition, eye tracking, gender, and age estimation.

The facial landmarks are selected to have discriminatory characteristics, or may serve as a distinguishing mark on the face. Commonly used keypoints are corners of the eyes, the tip of the nose, corners of the nostrils, corners of the mouth, the end points of the arcs of the eyebrows, the outline of the ear, chin (Figure 1).

RELATED WORK

Facial landmark localization approaches can be divided into three broad categories: (1) Active Appearance Model -based method, (2) cascaded regression method, and (3) detection based method [17].

The Active Appearance Model (AAM) [9] is considered as the most classic method. It searches for shape parameters through minimizing the residual between the face appearance and a face template.

The method does not generalize well and is sensitive to initialization.

Cascaded regression method estimates the face shape through a cascade of regressors. It starts from a raw initial guess of landmark positions and learns regressors that iteratively map shape-dependent features into shape increment. Examples of cascaded regression algorithms include the approach by Cristinacce and Cootes [10] which employs boosted regression for facial landmarks alignment. Xiong and De la Torre [15] propose a Supervised Descent Method (SDM) which learns descent directions and does linear mapping on non-linear SIFT features. Cao et al. [5] use pixel differences as features and implement nonlinear boosted regression. Burgos-Artizzu et al. [3] build a cascaded regression model with occlusion detection and voting strategy to cope with severe occlusion. Regression forest

Address for correspondence:

martin@tu-sofia.bg

voting for accurate shape fitting was proposed by Cootes et al. [8]. Random forests [2] and random ferns [13] are frequently used in recent research papers as the regression algorithms.

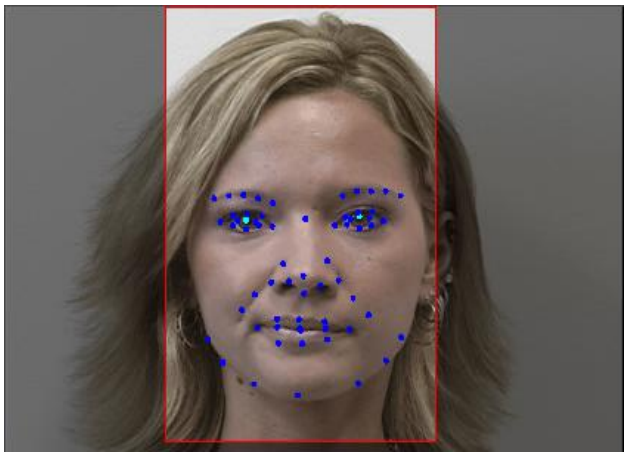

Figure1. *Example of a face image with annotated facial landmarks. The red outline indicates the face detector result*

Detection based approaches detect object parts independently and then estimate the shape directly from the detections [6] or through flexible part models [4].

Sun et al. [14] recently achieved state-of-the-art results by using three-level deep convolutional network trained for facial landmark detection. The convolutional network detects approximate locations of the landmarks in the lower cascade and refines the estimations in higher cascade.

METHOD

Review of the Cascaded Regression Model

Face shape is represented as a vector of landmark locations $S = (x_1, x_2, \cdots, x_n) \in R^{2n}$,

where n is the number of landmarks. $x_i \in R^2$ is the 2D coordinates of the i-th facial landmark. Given a face image I, the cascaded model estimates the shape starting from an initial estimate S^0 and progressively refines the shape by a cascade of T regressors, $r^{1\cdots T}$. Each regressor predicts an update vector ΔS, which is added to the current shape estimate:

$$S^t = S^{t-1} + r^t(I, S^{t-1}) \quad (1)$$

The final estimated shape can be written as:

$$S^T = S^0 + \sum_{t=1}^{T} r^t(I, S^{t-1}) \quad (2)$$

Each regressor r^t predicts the shape increments based on features, extracted from I and indexed relative to the current shape estimate S^t. This introduces some geometric invariance and makes the feature computation process more robust. Given N training samples $\{(I_i, S_i)\}_{i=1}^{N}$, during training each regression function r^t is learnt by minimizing the mean square error between the current estimated shape and the ground-truth shape:

$$r^k = \arg min_r \sum_{i=1}^{N} \left\| S_i - \left(S_i^{k-1} + r\left(I_i, S_i^{k-1}\right) \right) \right\|_2^2 \quad (3)$$

Proposed Algorithm

First, we detect the face region in the images using OpenCV's Viola-Jones face detector. The training face boxes are perturbed by a certain horizontal and vertical translation and an optional scaling, and the new samples are added to the training dataset to create a more robust model. Furthermore, as faces are symmetric, we mirror the images and double the size of the training data. All the face images are scaled to 128x128 pixels size. The ground-truth coordinates of the landmark points are modified accordingly.

We use an ensemble of three different cascades of linear regressing functions (Figure 2) and fuse their prediction results using Gradient Boosting Regression Trees (GBRT). The regressors of the various cascades are trained from a different random subset of the training samples.

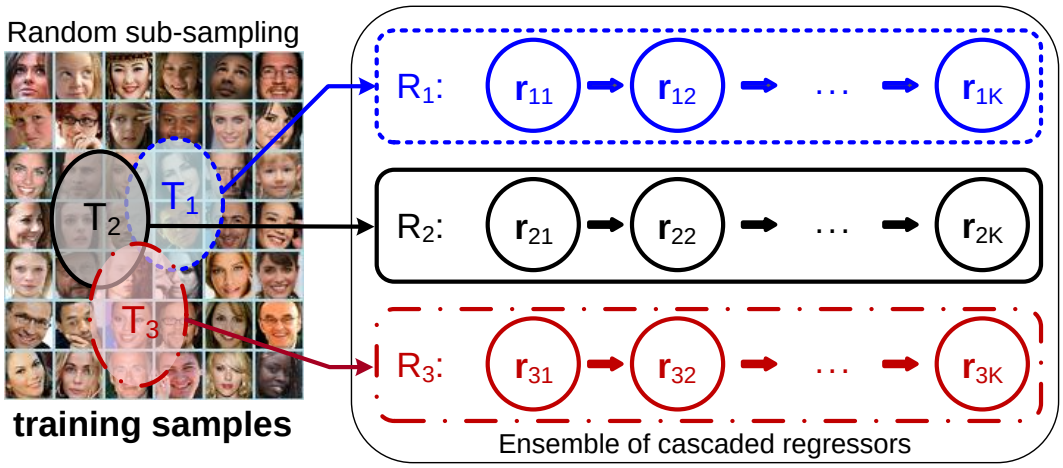

Figure2. *Ensemble of three cascades of regressors*

The regression functions at each stage of the cascade extract local features from the estimated landmark positions in the previous cascade level and concatenate them into one feature vector. Yan *et al.* [16] have compared different local image descriptors (HOG, SIFT, Gabor, and LBP) and found that the HOG descriptor worked best. Our implementation uses Histograms of Oriented Gradients (HOG) features for the weak regressors. We address the regression problem by using L2-regularized L2-loss linear support vector machine. We use LIBLINEAR open source library [11] in our linear regression implementation. Starting from the mean face shape, which is calculated from the ground-truth data of the training dataset, each regressor refines the landmarks position estimate (Figure 3). We limit the number of regressors in each cascade to five. By generating different random subsets of training data and varying the hyper-parameters of the support vector regressors (the epsilon in loss function, tolerance, regularization factors), we trained 100 different cascades. We evaluated their prediction using the validating dataset and kept ten models with the least median error. Then we calculated Pearson's correlation measures of the ten regressors and selected three of them that were less correlated to use in the ensemble.

Random forest regression is a very popular technique for its efficiency to handle nonlinear regression problems. We use the XGBoost library [7], which is a parallel implementation of the gradient boosting tree classifier and have demonstrated excellent performance. The final results of the estimated landmark positions are obtained by GBRT regression using a feature set consisting of the output prediction of each cascade together with HOG features, extracted from the face image. Those HOG features are calculated at the points of the mean predicted shape from the three cascades.

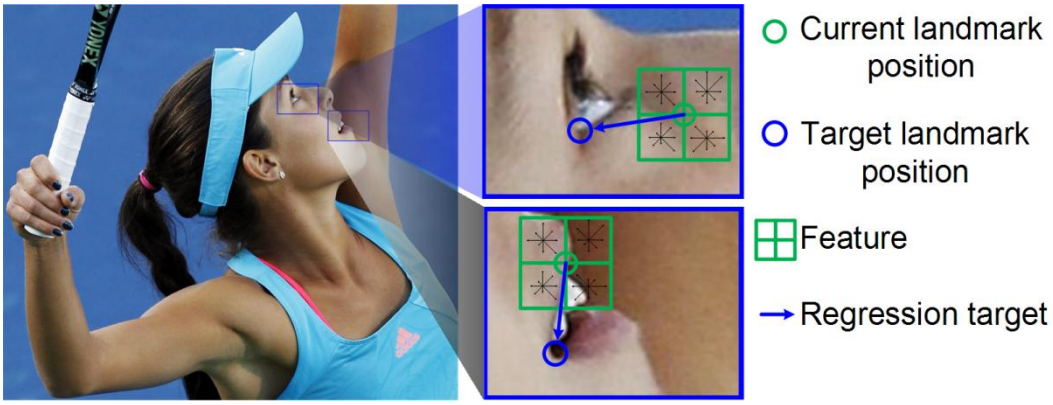

Figure3. *Weak regression illustration*

EXPERIMENTS

In our experiments on facial feature detection we used two datasets with annotated faces – the LFPW dataset [1] and the Helen dataset [12].

LFPW: The Labeled Face Parts in-the-wild (LFPW) dataset consists of 1,287 images collected from the internet. The images contain faces with large variations of facial expressions, illumination, head pose, and occlusions.

HELEN: The Helen dataset contains 2,330 annotated images downloaded from flickr.com website. The face images are of a high resolution, and the provided annotations are very detailed.

We split the LFPW dataset into two parts – one for training and the other for validation. The Helen dataset was used only for testing of the results.

To evaluate the accuracy of our method, we used as error measure the point-to-point Euclidean distance, normalized by the distance between the outer corners of the eyes. Facial landmark detection performance was assessed on the 68 landmark points markup scheme of Figure 4. Some images with detected landmarks are shown in Figure 5. Finally, the cumulative error rates were calculated for the Helen dataset (Figure 6).

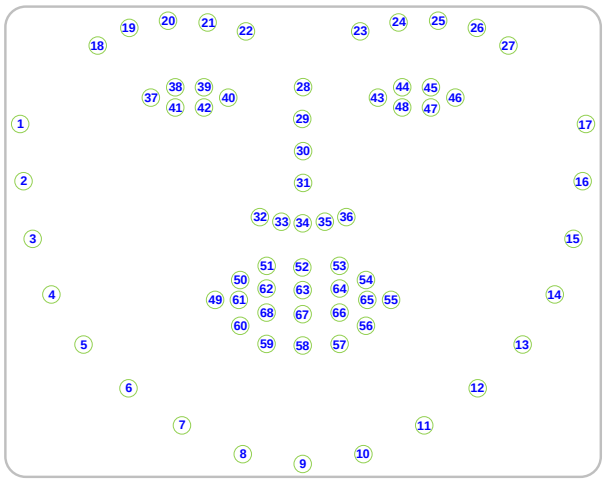

Figure4. *The 68 points markup used for our annotations*

Figure5. *Results on some images from Helen dataset*

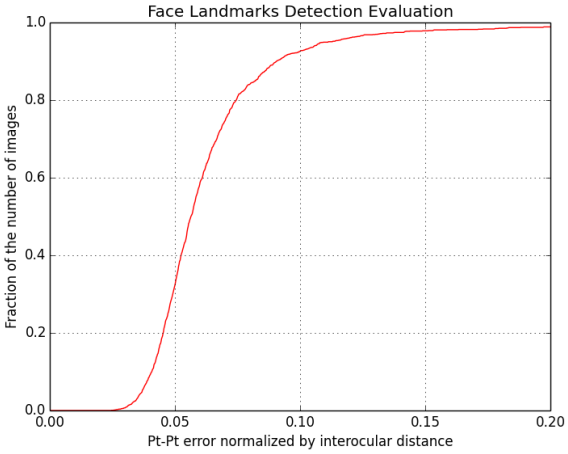

Figure6. *Cumulative error rates for Helen dataset*

CONCLUSION

In this paper, we proposed a method for facial landmarks detection. The algorithm is very accurate and is suitable for applications that do not require real-time analysis. From the conducted experiments on Helen databases, we achieved detection performance compared with the current state-of-art methods. The method can be extended to future research with other nonlinear features to make it more robust against head pose variations.

ACKNOWLEDGEMENT

This work was supported by contract DFNI I02/1 for the research project: "Intelligent man-machine interface for assistive medical systems in improving the independent living of motor disabled users" of the Bulgarian Research Fund of the Ministry of Education and Science".

REFERENCES

[1] Belhumeur, P. N., D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 545-552.

[2] Breiman, L., "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001/10/01 2001.

[3] Burgos-Artizzu, X. P., P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," Proceedings of the IEEE International Conference on Computer Vision, pp. 1513–1520, 2013.

[4] Burl, M., M. Weber, and P. Perona, "A probabilistic approach to object recognition using local photometry and global geometry," in Computer Vision — ECCV'98. vol. 1407, H. Burkhardt and B. Neumann, Eds., ed: Springer Berlin Heidelberg, 1998, pp. 628-641.

[5] Cao, X., Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," International Journal of Computer Vision, vol. 107, pp. 177-190, 2014/04/01 2014.

[6] Cevikalp, H., B. Triggs, and V. Franc, "Face and landmark detection by using cascade of classifiers," in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, 2013, pp. 1-7.

[7] Chen, T. and T. He, "xgboost: eXtreme gradient boosting," 2015.

[8] Cootes, T., M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in Computer Vision – ECCV 2012. vol. 7578, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 278-291.

[9] Cootes, T. F., G. J. Edwards, and C. J. Taylor, "Active appearance models," in Computer Vision — ECCV'98. vol. 1407, H. Burkhardt and B. Neumann, Eds., ed: Springer Berlin Heidelberg, 1998, pp. 484-498.

[10] Cristinacce, D. and T. F. Cootes, "Boosted regression active shape models," Proceedings of the British Machine Vision Conference 2007, pp. 79.1–79.10, 2007.

[11] Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," J. Mach. Learn. Res., vol. 9, pp. 1871-1874, 2008.

[12] Le, V., J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," in Computer Vision – ECCV 2012. vol. 7574, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 679-692.

[13] Ozuysal, M., M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, pp. 448-461, 2010.

[14] Sun, Y., X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," presented at the Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[15] Xuehan, X. and F. de la Torre, "Supervised descent method and its applications to face alignment," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 532-539.

[16] Yan, J., Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, 2013, pp. 392-396.

[17] Zhu, S., C. Li, C. C. Loy, and X. Tang, "Transferring landmark annotations for cross-dataset face alignment," CoRR, vol. abs/1409.0602, 2014.

AUTHOR'S BIOGRAPHY

Martin Penev, received his MSc Degree from the Technical University of Sofia in 2000. Currently, he is a Ph.D. student in the Department of Radio Communications and Video Technologies. His area of research is digital image processing and pattern recognition. *martin@tu-sofia.bg*

Ognian Boumbarov, received his MSc Degree from the Technical University of Sofia in 1973. He obtained a Ph.D. Degree in 1985. Since 1999, he is Assoc. Professor in Sound and Image Processing Laboratory, Department of Radio Communications and Video Technologies, TU-Sofia. His area of research is audio and video systems, signal and image processing, pattern recognition, neural networks and multimodal biometric analysis and identification. *olb@tu-sofia.bg*