A Deep Learning Approach to Joint Face Detection and Segmentation

Khoa Luu, Chenchen Zhu, Chandrasekhar Bhagavatula, T. Hoang Ngan Le, and Marios Savvides

Abstract Robust face detection and facial segmentation are crucial pre-processing steps to support facial recognition, expression analysis, pose estimation, building of 3D facial models, etc. In previous approaches, the process of face detection and facial segmentation are usually implemented as sequential, mostly separated modules. In these methods, face detection algorithms are usually first implemented so that facial regions can be located in given images. Segmentation algorithms are then carried out to find the facial boundaries and other facial features, such as the eyebrows, eyes, nose, mouth, etc. However, both of these tasks are challenging due to numerous variations of face images in the wild, e.g. facial expressions, illumination variations, occlusions, resolution, etc. In this chapter, we present a novel approach to detect human faces and segment facial features from given images simultaneously. Our proposed approach performs accurate facial feature segmentation and demonstrates its effectiveness on images from two challenging face databases, i.e. Multiple Biometric Grand Challenge (MBGC) and Labeled Faces in the Wild (LFW).

1 Introduction

The problem of facial segmentation has been intensely studied for decades with the aim of ensuring generalization of an algorithm to unseen images. However, a robust solution has not yet been developed due to a number of challenging properties of the problem. For example, facial pose varies in captured images, illumination changes over time, different cameras have different outputs for an identical object, movement of objects cause blurring of colors, skin tones vary dramatically across individuals and ethnicities, and some background objects' color is similar to human facial skin. In this chapter, we present a novel approach based on deep learning framework to simultaneously detect human faces and segment facial features from

K. Luu (✉) • C. Zhu • C. Bhagavatula • T.H.N. Le • M. Savvides
CyLab Biometrics Center, Department of Electrical & Computer Engineering,
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: kluu@andrew.cmu.edu; chenchez@andrew.cmu.edu; cbhagava@andrew.cmu.edu;
thihoanl@andrew.cmu.edu; msavvid@ri.cmu.edu

M. Kawulok et al. (eds.), *Advances in Face Detection and Facial Image Analysis*,
DOI 10.1007/978-3-319-25958-1_1

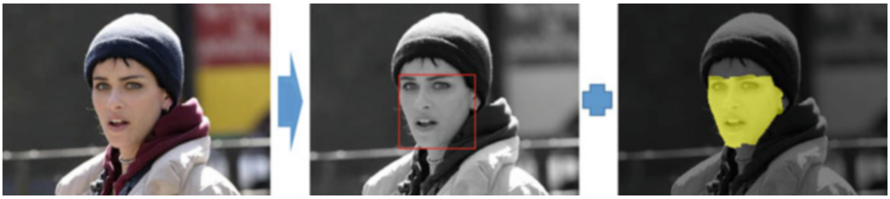

Fig. 1 Joint face detection and segmentation using our proposed algorithm on an example form the LFW database

given digital images. Instead of focusing on general object detection in natural images [1], our method aims at building a robust system for face detection and facial feature segmentation. Our proposed method takes the advantages of the Multiscale Combinatorial Grouping (MCG) algorithm [2] and the deep learning Caffe framework [3] to robustly detect human faces and locate facial features. One-class Support Vector Machines (OCSVM) [4] are developed in the later steps to verify human facial regions. Finally, the region refinement technique [1] and the Modified Active Shape Models (MASM) [5] method are used in the post-processing steps to refine the segmented regions and cluster facial features resulting in bounding boxes and segmentations as shown in Fig. 1.

Compared to FaceCut [6], an automatic facial feature extraction method, our proposed approach contains several critical improvements. Firstly, instead of using an off-the-shelf commercial face detection engine as in [6], our method robustly performs face detection and facial boundary segmentation at the same time. Secondly, instead of using a color-based GrowCut approach [7] that is unable to deal with grayscale and illuminated images, our method robustly works with both grayscale and color images captured in various resolutions. In addition, since our method is implemented using a deep learning framework, it is able to learn features from both human faces and non-facial background regions. Therefore, our method is able to robustly detect human faces and segment facial boundaries in various challenging conditions. It also is able to deal with facial images in various poses, expressions, and occlusions. Our approach is evaluated on the NIST Multiple Biometric Grand Challenge (MBGC) 2008 still face challenge database [8] to evaluate its ability to deal with varying illuminations and slight pose variation as well as on the challenging Labeled Faces in the Wild (LFW) database [9].

The rest of this chapter is organized as follows. In Sect. 2, we review prior work on graph cuts based approaches to image segmentation and face detection methods. In Sect. 3, we present our proposed approach to simultaneously detect human faces and facial boundaries from given input images. Section 4 presents experimental results obtained using our proposed approach on two challenging databases, MBGC still face challenge database and the LFW database. Finally, our conclusions on this work are presented in Sect. 5.

2 Related Work

In this section, we describe prior graph cuts-based image segmentation algorithms and previous face detection approaches.

2.1 Image Segmentation

Recent development in algorithms such as graph cuts [10], GrabCut [11], GrowCut [7] and their improvements have shown accurate segmentation results in natural images. However, when dealing with facial images, it is hard to determine the weak boundary between the face and the neck regions by using conventional graph cuts based algorithms. This can be seen in Fig. 2, which shows an incorrectly segmented face obtained using the classical GrowCut algorithm. It is also to be noted that GrowCut requires the manually marking of points in the foreground and background and that it can't separately segment facial components such as eyes, nose and mouth.

In the graph cuts algorithm [10], input images are treated as graphs and their pixels as nodes of the graphs. In order to segment an object, max-flow/min-cut algorithms are used. For the rest of this chapter we refer to this original approach to image segmentation as the GraphCuts method. The GrabCut [11] algorithm was later introduced to improve the GraphCuts method by using an iterative segmentation scheme and applying graph cuts at intermediate steps. The input provided by the user consists of a rectangular box around the object to be segmented. The segmentation process is employed using the color statistical information inside and outside the box. The image graph is re-weighted and the new segmentation in each step is refined by using graph cuts.

Grady [12] presented a new approach to image segmentation using random walks. This approach also requires initial seed labels from the user but allows the use of more than two labels. A random walker starting at each unlabeled pixel first reach a pre-labeled pixel and its analytical probabilities are calculated. Then, image segmentation is carried out by assigning each unlabeled pixel the label for which the greatest probability is calculated.

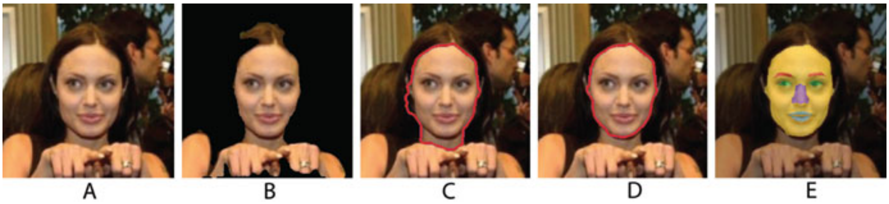

Fig. 2 Comparisons of facial segmentation algorithms: (**a**) original image, (**b**) face segmentation using color based information, (**c**) face segmentation results obtained using GrowCut, (**d**) automatic segmentation of facial features using our modified GrowCut algorithm and statistical skin information, (**e**) facial features segmentation using our proposed algorithm

2.2 Face Detection

Face detection has been a well studied area of computer vision. One of the first well performing approaches to the problem was the Viola-Jones face detector [13]. It was capable of performing real time face detection using a cascade of boosted simple Haar classifiers. The concepts of boosting and using simple features has been the basis for many different approaches [14] since the Viola-Jones face detector. These early detectors tended to work well on frontal face images but not very well on faces in different poses. As time has passed, many of these methods have been able to deal with off-angle face detection by utilizing multiple models for the various poses of the face. This increases the model size but does afford more practical uses of the methods. Some approaches have moved away from the idea of simple features but continued to use the boosted learning framework. Li and Zhang [15] used SURF cascades for general object detection but also showed good results on face detection.

More recent work on face detection has tended to focus on using different models such as a Deformable Parts Model (DPM) [16, 17]. Zhu and Ramanan's work was an interesting approach to the problem of face detection in that they combined the problems of face detection, pose estimation, and facial landmarking into one framework. By utilizing all three aspects in one framework, they were able to outperform the state-of-the-art at the time on real world images. Yu et al. [18] extended this work by incorporating group sparsity in learning which landmarks are the most salient for face detection as well as incorporating 3D models of the landmarks in order to deal with pose. Chen et al. [19] have combined ideas from both of these approaches by utilizing a cascade detection framework while synchronously localizing features on the face for alignment of the detectors. Similarly, Ghiasi and Fowlkes [20] have been able to use hierarchical DPMs not only to achieve good face detection in the presence of occlusion but also landmark localization. However, Mathias et al. [21] were able to show that both DPM models and rigid template detectors similar to the Viola-Jones detector have a lot of potential that has not been adequately explored. By retraining these models with appropriately controlled training data, they were able to create face detectors that perform similarly to other, more complex state-of-the-art face detectors.

All of these approaches to face detection were based on selecting a feature extractor beforehand. However, there has been work done in using a Convolutional Neural Networks (CNNs) to learn which features are used to detect faces [22]. Neural Networks have been around for a long time but have been experiencing a resurgence in popularity due to hardware improvements and new techniques resulting in the capability to train these networks on large amounts of training data.

Our approach is to use a CNN to extract features useful for detecting faces as well with one key difference. In all of the prior approaches, the face is detected as a bounding box around the face. Since faces are not a rigid, rectangular object, a bounding box will necessarily include more than the face. Ideally, a detector would be able to give a pixel level segmentation of the face which could be used to ensure only features from the face are extracted for any problems that use face detection. Some of the works comes close to determining a tighter bound on the face detection by finding landmarks on the face including a facial boundary. Another work similar to our approach presented in [1]. The difference is that our method aims at working on robust face detection and facial feature segmentation including subdividing the face into smaller semantic regions. In addition, since there is only one class in our defined system, i.e. the face region, instead of multiple classes, the One-Class Support Vector Machines (OCSVM) [4, 23] are a good fit in our problem. Our approach uses CNNs to determine a pixel level segmentation of the faces and use these to generate a bounding box for measuring performance against other methods.

3 Deep Learning Approach to Face Detection and Facial Segmentation

In this section, we present our proposed deep learning approach to detect human faces and segment features simultaneously. We first review the CNNs and its very fast implementation in the GPU-based Caffe framework [3]. Then, we will present our proposed approach in the second part of this section.

3.1 Deep Learning Framework

Convolutional Neural Networks are biologically inspired variants of multilayer perceptrons. The CNN method and its extensions, i.e. LeNet-5 [24], HMAX [25], etc., simulate the animal visual cortex system that contains a complex arrangement of cells sensitive to receptive fields. In this model, designed filters are treated as human visual cells in order to explore spatially local correlations in natural images. It efficiently presents the sparse connectivity and the shared weights since kernel filters are replicated over the entire image with the same parameters in each layer. The pooling step, a form of down-sampling, plays a key role in CNNs. Max-pooling is a popular pooling method for object detection and classification since max-pooling reduces computation for upper layers by eliminating non-maximal values and provides a small amount of translation invariance in each level.

Although CNNs can explore deep features, they are computationally very expensive. The algorithm runs faster when implemented in a Graphics Processing Unit (GPU). The Caffe framework [3] is a rapid deep learning implementation using CUDA C++ for GPU computation. It also supports bindings to Python/Numpy and MATLAB. It can be used as an off-the-shelf deployment of state-of-the-art models. In our proposed system, the AlexNet [26] network is implemented in Caffe and is employed to extract features from candidate regions. Given a candidate region, two 4096-dimensional feature vectors are extracted by feeding two types of region images into the AlexNet and the two vectors are concatenated to form the final feature vector. The AlexNet has 5 convolution layers and 2 fully connected layers. All the activation functions are Rectified Linear Units(ReLU). Max pooling with a stride of 2 is applied at the first, second and fifth convolution layers. Two types of region images are cropped box and region foreground. They are both warped to a fixed 227 × 227 pixel size.

3.2 Our Proposed Approach

In our proposed method, the MCG algorithm is first employed to extract superpixel based regions. In each region, there are two kinds of features calculated and used as inputs into the GPU-based Caffe framework presented in Sect. 3.1. The OCSVM is then applied to verify facial locations. Finally, the post-processing steps, i.e. region refinement and MASM, are employed in the final steps.

Multiscale Combinatorial Grouping [2] is considered as one of the state-of-the-art approaches for bottom-up hierarchical image segmentation and object candidate generation with both high accuracy and low computational time compared against other methods. MCG computes a segmentation hierarchy at multiple image resolutions, which are then fused into a single multiscale hierarchy. Then candidates are produced by combinatorially grouping regions from all the single scale hierarchies and from the multiscale hierarchy. The candidates are ranked based on simple features such as size and location, shape and contour strength.

Our proposed approach first employs MCG on the input image $\mathbf{X}$ to extract N face candidate regions $\mathbf{x}_i, i = 1..N$. Then, features are extracted from two types of representations of the regions, i.e. the bounding box of the region with only the foreground $\mathbf{x}_i^F$ and the ones with background $\mathbf{x}_i^B$ as shown in Fig. 3. The first representation $\mathbf{x}_i^F$, i.e. the segmented region, is used to learn the facial information. Meanwhile, the second type $\mathbf{x}_i^B$, i.e. the bounding boxes, aims at learning relative information between face and their background information. In this way, the learning features include both human faces and common backgrounds. This approach helps our proposed system robust against various challenging conditions as presented in Sect. 1.

We fuse the two representations as inputs into the CNN to extract deep features $\mathbf{x}_i^D$. In the later step, One-class Support Vector Machines are developed to verify human facial regions. Finally, the region refinement technique and the Modified Active Shape Models are used in the post-processing steps to refine the segmented regions and cluster facial features.

3.2.1 One-Class Support Vector Machines (OCSVM)

Given a set of N feature vectors $\{\mathbf{x}_1^D, \dots, \mathbf{x}_N^D\}$, where $\mathbf{x}_i^D \in \mathbb{R}^d$ are features extracted using deep learning, the OCSVM algorithm finds a set of parameters $\{\mathbf{x}_0, r\}$, where $\mathbf{x}_0 \in \mathbb{R}^d$ denotes the center position and $r \in \mathbb{R}$ is the radius. They describe the optimal sphere that contains the training data points while allowing some slack for errors. This process can be employed by solving the following minimization problem:

$$\Lambda(\mathbf{x}_0, r, \epsilon) = r^2 + C\sum_{i=1}^{N}\epsilon_i$$
$$s.t., \qquad \|\mathbf{x}_i^D - \mathbf{x}_0\|^2 \leq r^2 + \epsilon_i, \forall i, \epsilon_i \geq 0 \tag{1}$$

where $C \in \mathbb{R}^n$ denotes the tradeoff between the volume of the description and the errors, and ϵ_i represents the slack variables to describe data points outside the hyper-sphere. The quadratic programming problem in Eq. (1) can be mathematically solved by using Lagrange Multiplies and setting partial derivatives to zeros as follows:

$$\begin{aligned} L(\mathbf{x}_0, r, \epsilon, \alpha, \gamma) = r^2 + C\sum_{i=1}^{N}\epsilon_i \\ -\sum_i \alpha_i\{r^2 + \epsilon_i - (\mathbf{x}_i \cdot \mathbf{x}_i - 2\mathbf{x}_0 \cdot \mathbf{x}_i + \mathbf{x}_0 \cdot \mathbf{x}_0)\} \\ -\sum_i \gamma_i\epsilon_i \end{aligned} \tag{2}$$

L has to be minimized with respect to $\mathbf{x}_0, \mathbf{r}$ and ϵ, and maximized with respect to α, γ.

The distance from a point $\mathbf{z}$ to center of hyper-sphere $\mathbf{x}_0$ is defined as:

$$\|\mathbf{z} - \mathbf{x}_0\|^2 = (\mathbf{z} \cdot \mathbf{z}) - 2\sum_i \alpha_i(\mathbf{z} - \mathbf{x}_i) + \sum_{i,j} \alpha_i\alpha_j(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{3}$$

The decision rule in our One-class Support Vector Machines approach is computed using the regular ℓ_2-norm distance between the center $\mathbf{x}_0$ and a new point $\mathbf{z}$ as follows:

$$f(\mathbf{z}) = \begin{cases} 1, & \text{if}\|\mathbf{z} - \mathbf{x}_0\|^2 \leq r^2 \\ -1, & \text{otherwise} \end{cases} \tag{4}$$

In Eq. (4), when the computed ℓ_2-norm distance between $\mathbf{x}_0$ and $\mathbf{z}$ is smaller or equal to the radius r, the given point $\mathbf{z}$ will be considered as the target.

Similar to Support Vector Machines (SVMs), OCSVM can be presented in a kernel domain by substituting the inner products of $\mathbf{x}_i^D$ and $\mathbf{x}_j^D$, where $i,j = 1..N$, with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ to enhance the learning space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\mathbf{x}_i - \mathbf{x}_j)^2/s^2 \tag{5}$$

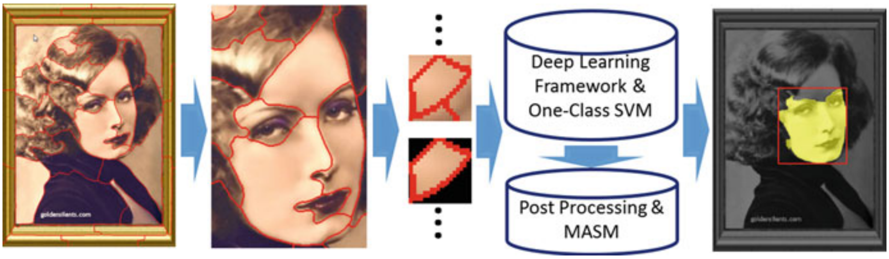

Fig. 3 Our proposed framework for face detection and facial segmentation simultaneously

where s denotes the standard deviation. Then, given a region feature $\mathbf{z} \in \mathbb{R}^d$, its distance to center $\mathbf{x}_0$ can be computed as:

$$\|\mathbf{z} - \mathbf{x}_0\|^2 = K(\mathbf{z}, \mathbf{z}) - 2 \sum_i \alpha_i K(\mathbf{z}, \mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

Our work aims at finding facial region detection as an one-class pattern problem. Therefore, OCSVM is employed to model deep learning features. The learned facial features boundary is more accurate when adding some negative samples during training steps as presented in [4]. Figure 3 illustrates our proposed approach to joint face detection and facial feature segmentation.

3.2.2 Post-processing Steps

Non-maximum suppression (NMS) [1] is employed on the scored candidates to refine segmented regions. Then, the Modified Active Shape Model (MASM) technique is used to refine the locations of facial feature components in the final step. The deep learning features extracted from CNNs are categorized to produce category-specific coarse mask predictions to refine the remaining candidates. Combining this mask with the original region candidates provides a further boost.

Though there are many approaches to landmark localization on face images, many of them use a 68 point landmarking scheme. Some well known methods such as Supervised Descent Method (SDM) by Xiong and De la Torre [27] and the work by Yu et al. [18] are such examples. However, with the 68 point scheme defined by these methods, segmenting certain regions of the face becomes much more difficult. Particularly the nose region as the landmarks do not give any estimate as to the boundary of the nose. For this reason, we have chosen to use the Modified Active Shape Model [5] method for the task of automatic facial landmark localization. Even though MASM may not perform as well at localization of some points, the additional points in the landmarking scheme are a necessity to the final semantic segmentation of the face region. However, MASM is a PCA based approach and does not always accurately localize landmarks in faces with shapes radically different from those exhibited by faces in the training set. Our proposed method can overcome the shortcomings of ASM based approaches in order to ensure that the facial region and individual facial features are segmented accurately as possible. Instead of using commercial face detection engines, our approach robustly finds facial locations that are used as inputs for MASMs. More importantly, the segmented facial boundaries extracted using the OCSVMs and the deep features are used to refine the facial feature regions.

4 Our Experimental Results

Our proposed models are trained on images of the CMU Multi-PIE database [28]. The OCSVM is trained on images drawn from various facial angles so that the face detection and segmentation is robust against off-angle faces. The MASM component is trained from images drawn from the frontal view set of the database.

Our approach is tested on 500 random images from the MBGC database to evaluate its robustness to varying illumination and in-plane rotation of faces. As can be seen from Table 1, which shows the mean and standard deviation of the point to point landmark fitting errors (for 79 landmarks) obtained by MASM, FaceCut and our proposed approach across all MBGC test images when compared to manually annotated images (ground truths), the results using our approach in improved fitting accuracy of the facial boundary landmarks. Our approach is also tested on images from the LFW database and was again able to accurately segment facial features in these challenging everyday images. Figure 4 shows sample detection and segmentation results using our approach on images from LFW databases respectively and demonstrates the effectiveness of the algorithm.

5 Conclusions

This paper has presented a novel approach to automatically detect and segment human facial features. The method combines the positive features of the deep learning method and the Modified Active Shape Model to ensure highly accurate

Table 1 Comparison of the mean (in pixels) and standard deviation (in pixels) of the fitting error across landmarks produced by MASM and our FaceCut approach on images from the MBGC database

Algorithms	Mean (all landmarks)	Std. dev. (all landmarks)	Mean (facial boundary landmarks)	Std. dev. (facial boundary landmarks)
MASM [5]	9.65	11.66	14.85	12.21
FaceCut [6]	9.56	11.51	14.46	11.59
Our approach	9.47	11.20	14.24	11.12

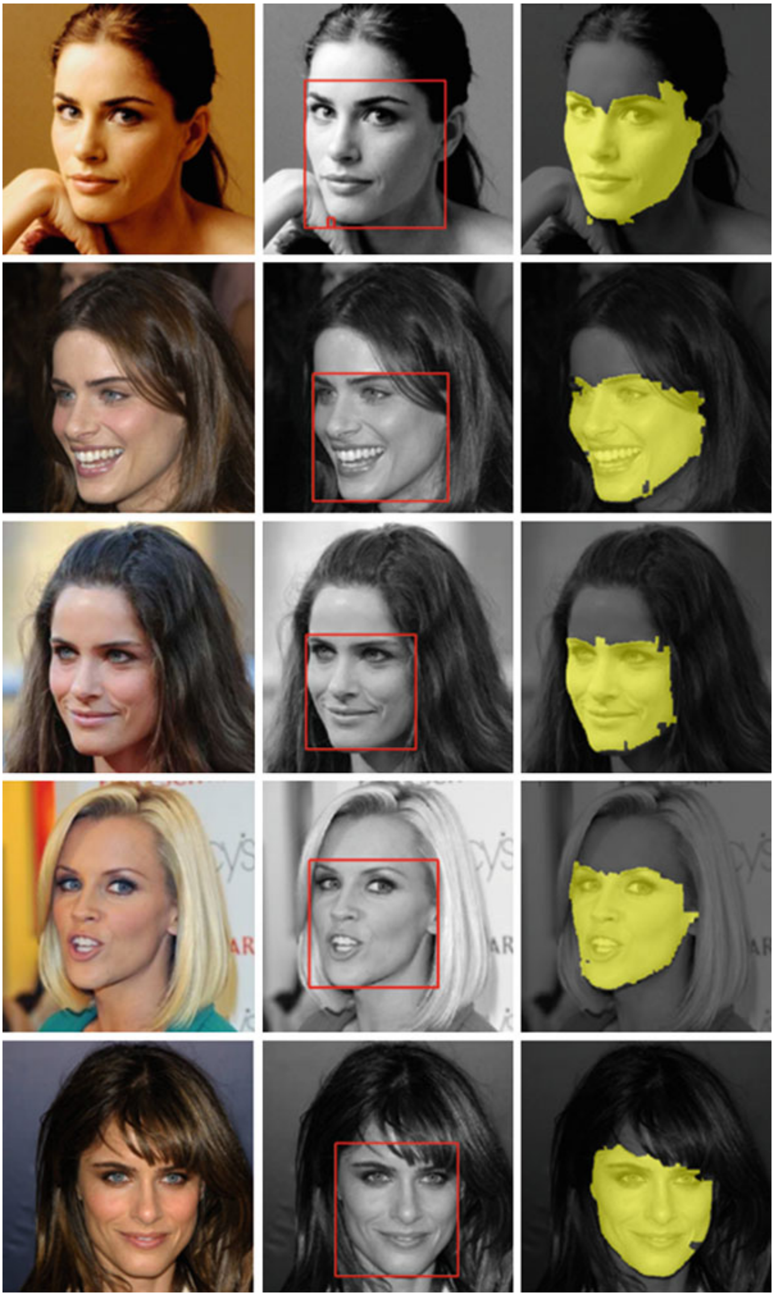

Fig. 4 The results using our proposed approach on the LFW database: the *first column*: original image, the *second column*: face detection results, the *third column*: facial features segmented

and completely automatic segmentation of facial features and the facial boundary. The effectiveness of detection and segmentation using our approach is demonstrated on unseen test images from the challenging MBGC and LFW databases.

References

1. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in *European Conference on Computer Vision (ECCV)* (2014)
2. P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2014)
3. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, *Caffe: Convolutional Architecture for Fast Feature Embedding* (2014). arXiv preprint arXiv:1408.5093
4. D.M.J. Tax, One-class classification, Ph.D. thesis, Delft University of Technology, June 2001
5. K. Seshadri, M. Savvides, Robust modified active shape model for automatic facial landmark annotation of frontal faces, in *Proceedings of the* 3*rd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2009), pp. 319–326
6. K. Luu, T.H.N. Le, K. Seshadri, M. Savvides, Facecut - a robust approach for facial feature segmentation, in *ICIP* (2012), pp. 1841–1848
7. V. Vezhnevets, V. Konouchine, "GrowCut" - interactive multi-label N-D image segmentation by cellular automata, in *Proceedings of the International Conference on Computer Graphics and Vision (GRAPHICON)* (2005)
8. P.J. Phillips, P.J. Flynn, J.R. Beveridge, W.T. Scrugs, A.J. O'Toole, D. Bolme, K.W. Bowyer, B.A. Draper, G.H. Givens, Y.M. Lui, H. Sahibzada, Joseph A. Scallan III, S. Weimer, Overview of the multiple biometrics grand challenge, in *Proceedings of the* 3^{rd} *IAPR/IEEE International Conference on Biometrics* (2009), pp. 705–714
9. P.N Belhumeur, D.W. Jacobs, D Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, in *Computer Vision and Pattern Recognition (CVPR)* (2011), pp. 545–552
10. Y.Y. Boykov, M.P. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2001), pp. 105–112
11. C. Rother, V. Kolmogorov, A. Blake, "GrabCut": interactive foreground extraction using iterated graph cuts, in *Proceedings of ACM Transactions on Graphics (SIGGRAPH)* (2004), pp. 309–314
12. L. Grady, Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **28**(11), 1768–1783 (2006)
13. P. Viola, M. Jones, Robust real-time face detection. Int. J. Comput. Vis. **57**, 137–154 (2004)
14. C. Zhang, Z. Zhang, A survey of recent advances in face detection. Tech. Rep. MSR-TR-2010-66, June 2010
15. J. Li, Y. Zhang, Learning surf cascade for fast and accurate object detection, in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 3468–3475
16. X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), pp. 2879–2886
17. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)

18. X. Yu, J. Huang, S. Zhang, W. Yan, D.N. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in *2013 IEEE International Conference on Computer Vision (ICCV)* (2013), pp. 1944–1951
19. D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, in *Computer Vision ECCV 2014*, eds. by D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars. Lecture Notes in Computer Science, vol. 8694 (Springer International, New York, 2014), pp. 109–122
20. G. Ghiasi, C. Fowlkes, Occlusion coherence: localizing occluded faces with a hierarchical deformable part model, in *CVPR* (2014)
21. M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in *Computer Vision ECCV 2014*, eds. by D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars. Lecture Notes in Computer Science, vol. 8692 (Springer International, New York, 2014), pp. 720–735
22. C. Garcia, M. Delakis, Convolutional face finder: a neural architecture for fast and robust face detection. IEEE Trans. Pattern Anal. Mach. Intell. **26**(11), 1408–1423 (2004)
23. H. Jin, Q. Lin, H. Lu, X. Tong, Face detection using one-class SVM in color images, in *Proceedings of the International Conference on Signal Processing* (2004)
24. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
25. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 411–426 (2007)
26. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *NIPS* (2012), pp. 1097–1105
27. X. Xiong, F. de la Torre, Supervised descent method and its applications to face alignment, in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 532–539
28. R. Gross, I. Matthews, J.F. Cohn, T. Kanade, S. Baker, Multi-PIE, in *Proceedings of the* 8^{th} *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008)* (2008), pp. 1–8