A CNN Cascade for Landmark Guided Semantic Part Segmentation

Aaron S. Jackson, Michel Valstar, Georgios Tzimiropoulos

School of Computer Science, The University of Nottingham, Nottingham, UK
{aaron.jackson, michel.valstar, yorgos.tzimiropoulos}@nottingham.ac.uk

Abstract. This paper proposes a CNN cascade for semantic part segmentation guided by pose-specific information encoded in terms of a set of landmarks (or keypoints). There is large amount of prior work on each of these tasks separately, yet, to the best of our knowledge, this is the first time in literature that the interplay between pose estimation and semantic part segmentation is investigated. To address this limitation of prior work, in this paper, we propose a CNN cascade of tasks that firstly performs landmark localisation and then uses this information as input for guiding semantic part segmentation. We applied our architecture to the problem of facial part segmentation and report large performance improvement over the standard unguided network on the most challenging face datasets. Testing code and models will be published online at http://cs.nott.ac.uk/~psxasj/.

Keywords: pose estimation, landmark localisation, semantic part segmentation, faces

1 Introduction

Pose estimation refers to the task of localising a set of landmarks (or keypoints) on objects of interest like faces [1], the human body [2] or even birds [3]. Locating these landmarks help establish correspondences between two or more different instances of the same object class which in turn has been proven useful for fined-grained recognition tasks like face and activity recognition. Part segmentation is a special case of semantic image segmentation which is the task of assigning an object class label to each pixel in the image. In part segmentation, the assigned label corresponds to the part of the object that this pixel belongs to. In this paper, we investigate whether pose estimation can guide contemporary CNN architectures for semantic part segmentation. This seems to be natural yet to the best of our knowledge this is the first paper that addresses this problem. To this end, we propose a Convolutional Neural Network (CNN) cascade for landmark guided part segmentation and report large performance improvement over a standard CNN for semantic segmentation that was trained without guidance.

Although the ideas and methods presented in this paper can probably be applied to any structured deformable object (e.g. faces, human body, cars, birds), we will confine ourselves to human faces. The main reason for this is the lack

of annotated datasets. To the best of our knowledge, there are no datasets providing pixel-level annotation of parts and landmarks at the same time. While this is also true for the case of human faces, one can come up with pixel-level annotation of facial parts by just appropriately connecting a pseudo-dense set of facial landmarks for which many datasets and a very large number of annotated facial images exist, see for example [4]. Note that during testing we do not assume knowledge of the landmarks' location, and what we actually show is that a two-step process in which a CNN firstly predicts the landmarks and then uses this information to segment the face largely outperforms a CNN that was trained to directly perform facial part segmentation.

Fig. 1. Example faces and their corresponding output from the CNN cascade.

1.1   Main contributions

In summary, this paper addresses the following research questions:

1. Is a CNN for facial part segmentation needed at all? One might argue that by just predicting the facial landmarks and then connecting them in the same way as we created the part labels, we could get high quality facial part segmentation thus completely by-passing the part segmentation task. Our first result in this paper is that indeed the latter method slightly outperforms a CNN trained for facial part segmentation (without guidance though).
2. Can facial landmarks be used for guiding facial part segmentation, thus reversing the result mentioned above? Indeed, we show that the proposed CNN cascade for landmark guided facial part segmentation largely outperforms both methods mentioned above without even requiring very accurate localisation of the landmarks. Some example output can be seen in Fig 1.

2   Related work

This section reviews related work on semantic segmentation, facial landmark localisation (also known as alignment) and facial part segmentation.

Face Alignment State-of-the-art techniques in face alignment are based on the so-called cascaded regression [5]. Given a facial image, such methods estimate the landmarks' location by applying a sequence of regressors usually learnt from SIFT [6] or other hand-crafted features. The regressors are learnt in a cascaded manner such that the input to regressor k is the estimate of the landmarks' location provided by regressor $k-1$, see also [7–11]. The first component in the proposed CNN cascade is a CNN landmark detector based on VGG-16 [12] converted to a fully convolutional network [13]. Although the main contribution of our paper is not to propose a method for landmark localisation, our CNN landmark localisation method performs comparably with all aforementioned methods. One advantage of our method over cascaded regression approaches is that it is not sensitive to initialisation and hence it does not rely on accurate face detection.

Semantic Segmentation Thanks to its ability to integrate information from multiple CNN layers and its end-to-end training, the Fully Convolutional Network (FCN) of [13] has become the standard basic component for all contemporary semantic segmentation algorithms. The architecture of FCN is shown in Fig. 2. One of the limitations of the FCN is that prediction is performed in low-resolution, hence a number of methods have been recently proposed to compensate for this by usually applying a Conditional Random Field (CRF) on top of the FCN output. The work of [14] firstly upsamples the predicted scores using bilinear interpolation and then refines the output by applying a dense CRF. The method of [15] performs recurrent end-to-end training of the FCN and the dense CRF. Finally, the work in [16] employs learnt deconvolution layers, as opposed to fixing the parameters with an interpolation filter (as in FCN). These filters learn to reconstruct the object's shape, instead of just classifying each pixel. Although any of these methods could be incorporated within the proposed CNN cascade, for simplicity, we used the VGG-FCN [12]. Note that all the aforementioned methods perform unguided semantic segmentation, as opposed to the proposed landmark-guided segmentation which incorporates information about the pose of the object during both training and testing. To encode pose specific information we augment the input to our segmentation network with a multi-channel confidence map representation using Gaussians centred at the predicted landmarks' location, inspired by the human pose estimation method of [17]. Note that [17] is iterative an idea that could be also applied to our method, but currently we have not observed performance improvement by doing so.

Part Segmentation There have been also a few works that extend semantic segmentation to part segmentation with perhaps the most well-known being the Shape Boltzman Machine [18, 19]. This work has been recently extended to incorporate CNN refined by CRF features (as in [14]) in [20]. Note that this work aims to refine the CNN output by applying a Restricted Boltzmann Machine on top of it and does not make use of pose information as provided by landmarks. In contrast, we propose an enhanced CNN architecture which is landmark-guided, can be trained end-to-end and yields large performance improvement without the need of further refinement.

Face Segmentation One of the first face segmentation methods prior to deep learning is known as LabelFaces [21] which is based on patch classification and further refinement via a hierarchical face model. Another hierarchical approach to face segmentation based on Restricted Boltzmann Machines was proposed in [22]. More recently, a multi-objective CNN has been shown to perform well for the task of face segmentation in [23]. The method is based on a CRF the unary and pairwise potentials of which are learnt via a CNN. Softmax loss is used for the segmentation masks, and a logistic loss is used to learn the edges. Additionally, the network makes use of a non-parametric segmentation prior which is obtained as follows: first facial landmarks on the test image are detected and then all training images with most similar shapes are used to calculate an average segmentation mask. This mask is finally used to augment RGB. This segmentation mask might be blurry, does not encode pose information and results in little performance improvement.
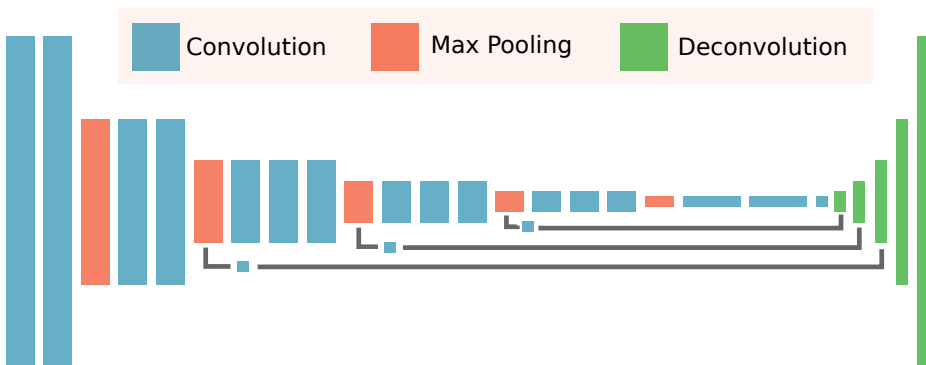

Fig. 2. Overview of the Fully Convolutional Network [13], low level information providing refinement are reintroduced into the network during deconvolution.

3   Datasets

There are a few datasets which provide annotations of pixel-level parts [24–26] but to the best of our knowledge there are no datasets containing both part and landmark annotations. Hence, in our paper we rely on datasets for facial landmarking. These datasets provide a pseudo-dense set of landmarks. Segmentation masks are constructed by joining the groundtruth landmarks together to fully enclose each facial component. The eyebrows are generated by a spline with a fixed width relative to the normalised face size, to cover the entire eyebrow. The selected classes are background, skin, eyebrows, eyes, nose, upper lip, inner mouth and lower lip. While this results in straight edges between landmarks, the network can learn a mean boundary for each class. The output from the network will be actually smoother than the groundtruth.

This process is illustrated in Fig. 3.

Fig. 3. Example groundtruth segmentation mask produced from the groundtruth landmarks.

For our experiments we used the 68-point landmark annotations provided by the 300W challenge [27]. In particular the training sets of LFPW [28], Helen [29], AFW [30] and iBUG [27] are all used for training while the 300W test set (600 images) is used for testing. Both training and test sets contain very challenging images in terms of appearance, pose, expression and occlusion.

This collection of images undergoes some pre-processing before they are used to train the network. The faces are normalised to be of equal size and cropped with some noise added to the position of the bounding box. Not all images are the same size, but their height is fixed at 350 pixels. With probability $p = 0.5$, a randomly sized black rectangle, large enough to occlude an entire component is layered over the input image. This assists the network in learning a robustness to partial occlusion.

4   Method

We propose a CNN cascade (shown in Fig. 4 and listed in Table 1) which performs landmark localisation followed by facial part segmentation. Our cascade was based on the VGG-FCN [12, 13] using Caffe [31] and consists of two main components:

1. Firstly, an FCN is trained to detect facial landmarks using Sigmoid Cross Entropy Loss.
2. Secondly, inspired by the human pose estimation method of [17], the detected 68 landmarks are encoded as 68 separate channels each of which contains a 2D Gaussian centred at the corresponding landmark's location. The 68 channels are then stacked along with the original image and passed into our segmentation network. This is a second FCN trained for facial part segmentation using as input the stacked representation of 2D Gaussians and image, and a standard Softmax loss.
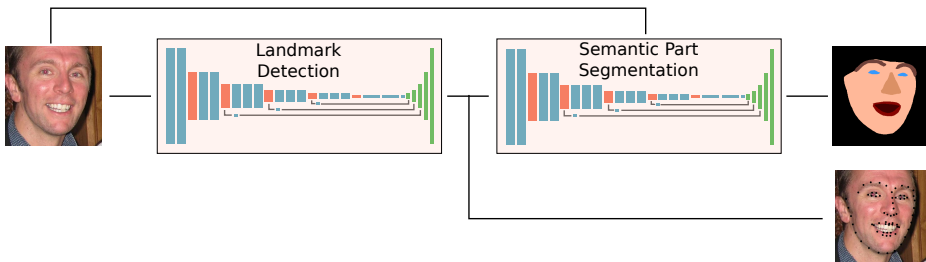
Fig. 4. The proposed architecture, comprising of two separate Fully Convolutional Networks. The first performs Landmark Detection, the output of which is encoded as multichannel representation which is then passed into the Semantic Part Segmentation network.

Overall we encode pose specific information by augmenting the input to our segmentation network with a multi-channel confidence map representation using Gaussians centred at the predicted landmarks' location. Hence, our FCN for semantic segmentation is trained to produce high quality, refined semantic masks by incorporating low level information with globally aware information. Each of the aforementioned components is now discussed in more detail:

Facial Landmark Detection The training procedure for landmark detection is similar to training FCN for part segmentation. Landmarks are encoded as 2D Gaussians centred at the provided landmarks' location. Each landmark is allocated its own channel to prevent overlapping with other landmarks and allow the network to more easily distinguish between each point. The main difference with part segmentation is the loss function. Sigmoid Cross Entropy Loss [3] was chosen to regress the likelihood of a pixel containing a point. More concretely, given our groundtruth Gaussians $\hat{p}$ and predicted Gaussians p, each of equal dimensions $N \times W \times H$, we can define the Sigmoid Cross Entropy loss l as follows:

$$l = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{W} \sum_{j=1}^{H} [p_{i,j}^n \log(\hat{p}_{i,j}^n) + (1 - p_{i,j}^n) \log(1 - \hat{p}_{i,j}^n)].$$

The loss was scaled by $1e^{-5}$ and a learning rate of 0.0001 was used. The network was trained in steps as previously described, for approximately 400,000 iterations, until convergence.

Guided Facial Part Segmentation To train our guided FCN part segmentation network we followed [13]. Softmax Loss was also used. If N is the number of outputs (in our case, classes), $p_{i,j}$ is the predicted output for pixel (i, j), and n is the true label for pixel (i, j), then the Softmax loss l can be defined as:

$$l = \frac{-1}{N} \sum_{i=1}^{W} \sum_{j=1}^{H} \log(p_{i,j}^n).$$

We firstly trained an unguided FCN for facial part segmentation following [13]. Initially, the network was trained as 32 stride, where no information from the lower layers is used to refine the output. This followed by introducing information from pool4, followed by pool3. A learning rate of 0.0001 was chosen, and a momentum of 0.9. The network was trained for approximately 300,000 iterations until convergence.

Then, our guided FCN was initialised from the weights of the unguided one, by expanding the first layer to accommodate the additional 68 input channels. As mentioned earlier, each channel contains a 2D Gaussian centred at the corresponding landmark's location. A key aspect of our cascade is how the landmarks' location is determined during training. We cannot use the groundtruth landmark locations nor the prediction of our facial landmark detection network on our training set as those will be significantly more accurate than those observed during testing. Hence, we applied our facial landmark detection network on our validation set and recorded the landmark localisation error. We used this error to create a multivariate Gaussian noise model that was added to the groundtruth landmark locations of our training set. This way our guided segmentation network was initialised with much more realistic input in terms of landmarks' location. Furthermore, the same learning rate of 0.0001 was used. For the first 10,000 iterations, training was disabled on all layers except for the first. This allowed the network to warm up slightly, and prevent the parameters in other layers from getting destroyed by a high loss.

Table 1. The VGG-FCN [12, 13] architecture employed by our landmark detection and semantic part segmentation network.

Layer Name	Kernel	Stride	Outputs	Layer Name	Kernel	Stride	Outputs
conv1_1	3×3	1×1	64	pool4	2×2	2×2	–
conv1_2	3×3	1×1	64	conv5_1	3×3	1×1	512
pool1	2×2	2×2	–	conv5_2	3×3	1×1	512
conv2_1	3×3	1×1	128	conv5_3	3×3	1×1	512
conv2_2	3×3	1×1	128	pool5	2×2	2×2	–
pool2	2×2	2×2	–	fc6_conv	7×7	1×1	4096
conv3_1	3×3	1×1	256	fc7_conv	1×1	1×1	4096
conv3_2	3×3	1×1	256	fc8_conv	1×1	1×1	68 or 7
conv3_3	3×3	1×1	256	deconv_32	4×4	2×2	68 or 7
pool3	2×2	2×2	–	score_pool4	1×1	1×1	68 or 7
conv4_1	3×3	1×1	512	deconv_16	4×4	2×2	68 or 7
conv4_2	3×3	1×1	512	score_pool3	1×1	1×1	68 or 7
conv4_3	3×3	1×1	512	deconv_8	16×16	8×8	68 or 7

5   Experiments

5.1   Overview of Results

In all experiments we used the training and test sets detailed in Section 3. As a performance measure, we used the familiar intersection over union measure [13]. We report a comparison between the performance of four different methods of interest:

1. The first method is the VGG-FCN trained for facial part segmentation. We call this method **Unguided**.
2. The second method is the part segmentation result obtained by joining the landmarks obtained from VGG-FCN trained for facial landmark detection. We call this method **Connected Landmarks**.
3. The third method is the proposed landmark guided part segmentation network where the input is the groundtruth landmarks' location. We call this method **Guided by Groundtruth**.
4. Finally, the fourth method is the proposed landmark guided part segmentation network when input is detected landmarks' location. We call this method **Guided by Detected**.

The first two methods are the baselines in our experiments while the third one provides an upper bound in performance. The fourth method is the proposed CNN cascade.

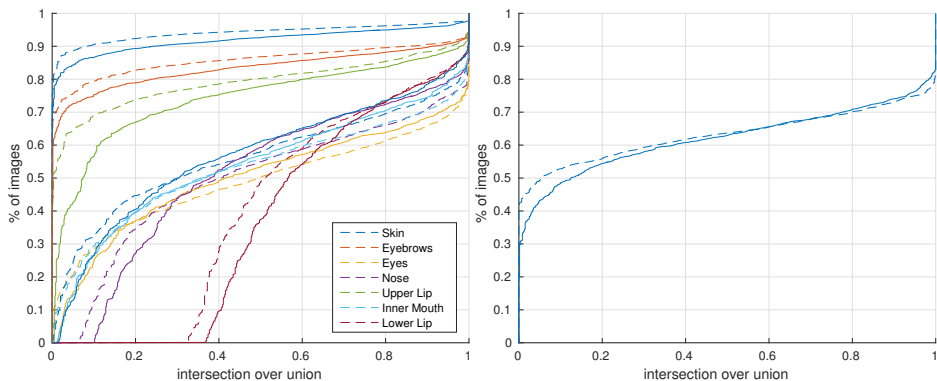### 5.2   Unguided Facial Part Segmentation

Fig. 5. Comparison of Unguided (—) and Connected Landmarks (- -). Per-class averages shown on the right.

To establish a baseline, an unguided fully convolutional network was firstly trained. This was done as described in the FCN paper [13] and Section 4. Some

visual results can be seen in Fig. 8. Additionally, a second baseline was obtained by simply connecting the landmarks of our facial landmark detection network also described in Section 4. The performance of both baselines can be seen in Fig. 5. We may observe that connecting the landmarks appears to offer slightly better performance than FCN for part segmentation alone. Nevertheless, we need to emphasise that the groundtruth masks were obtained by connecting the landmarks and hence there is some bias towards the connecting the landmarks approach.

5.3 Guided Facial Part Segmentation with Groundtruth
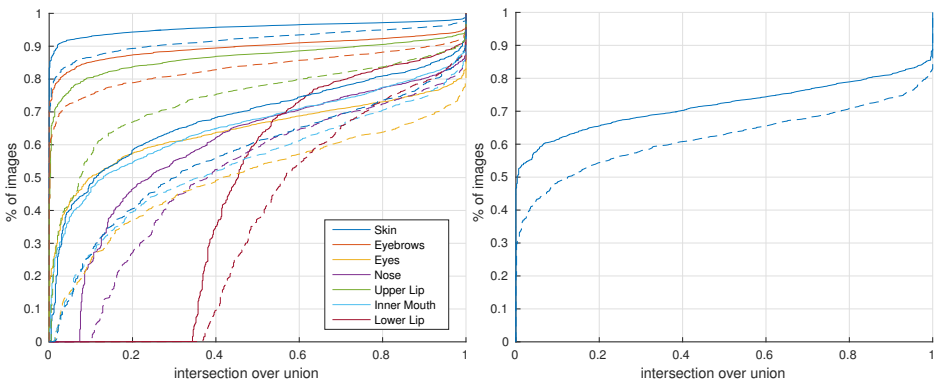
Fig. 6. Comparison of guided by groundtruth (——) and unguided (- -) facial part segmentation. Per-class averages shown on the right.

To establish an upper bounds to our performance, a fully convolutional network was trained to accept guidance from groundtruth landmarks. As described in Section 4, the guidance is provided in the form of landmarks encoded as 2D Gaussians. The performance difference between unguided and groundtruth guided part segmentation can be seen in Fig. 6. As we may observe the difference in performance between the two methods is huge. These results are not surprising given that the groundtruth semantic masks are generated from the landmarks guiding the network. Furthermore, landmark detection offers an advantage because, in the case of faces, there can only be one tip of the nose, and one left side of the mouth. Giving some information to the network about where it is likely to be located can offer a significant advantage. Our next experiment shows that this is still the case when detected landmarks are used instead of groundtruth landmarks.
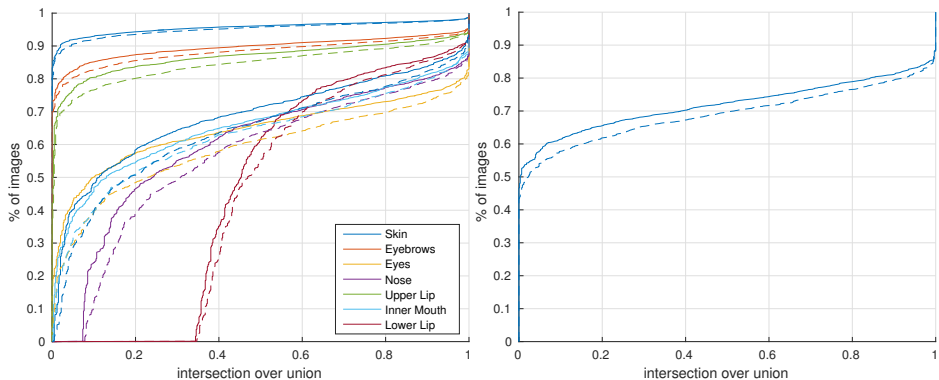
Fig. 7. Comparison of guidance from groundtruth landmarks (—) and guidance from detected landmarks (- -). Per-class averages shown on the right.

5.4 Guided Facial Part Segmentation with Detected Landmarks

With our upper bound and baselines defined, we can now see how much of an improvement we can achieve by guiding the network with our detected landmarks. The output of the landmark detection network is passed into the part segmentation network along with the original input image. We acknowledge that the performance of our landmark detector is far from groundtruth. We measure the performance as mean point to point Euclidean distance normalised by the outer interocular Euclidean distance, as in [27]. This results in an error of 0.0479. However, we show that the performance of the segmentation is improved significantly. The results of facial part segmentation guided by the detected landmarks, compared to the network guided by groundtruth landmarks can be seen in Fig 7. Our main result is that performance of the guided by detected network is very close to the that of the guided by groundtruth illustrating that in practice accurate landmark localisation is not really required to guide segmentation. Some visual results can be seen in Fig. 8. Also, performance over all components for all methods is given in Fig. 9.

6   Conclusion

In this paper we proposed a CNN architecture to improve the performance of part segmentation by task delegation. In doing so, we provided both landmark localisation and semantic part segmentation on human faces. However, our method should be applicable to our objects as well. This is the focus of our ongoing work. We are also looking into how the segmentation masks can be further used to improve landmark localisation accuracy, thus leading to a recurrent architecture. Future work may also compare the performance of this method with a multitask architecture.
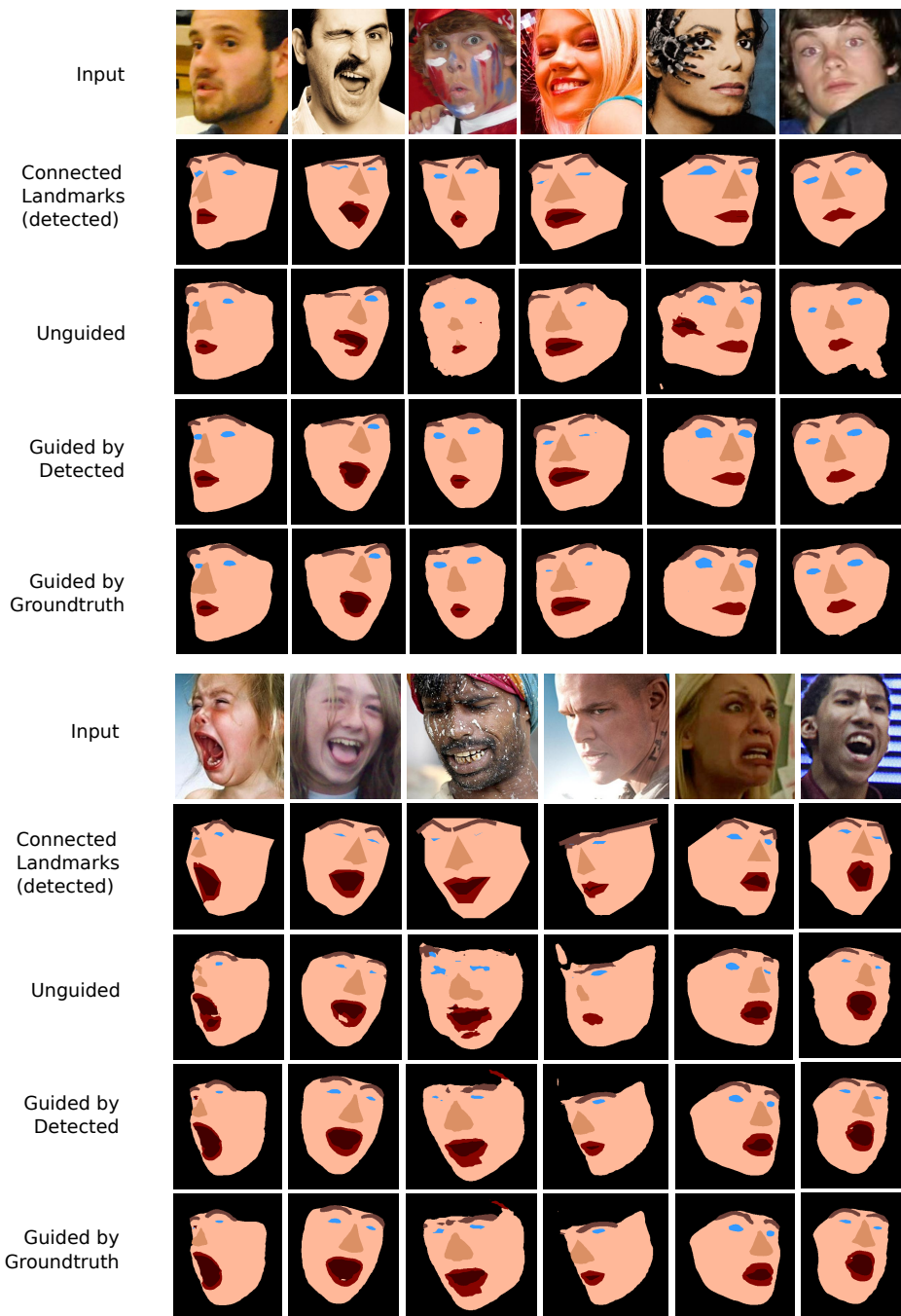
Fig. 8. Some visual results showing where the unguided network begins to fail, and where the guidance begins to pay off. Observe how visually close the results of the guided by groundtruth landmarks and the guided by detected landmarks networks are.
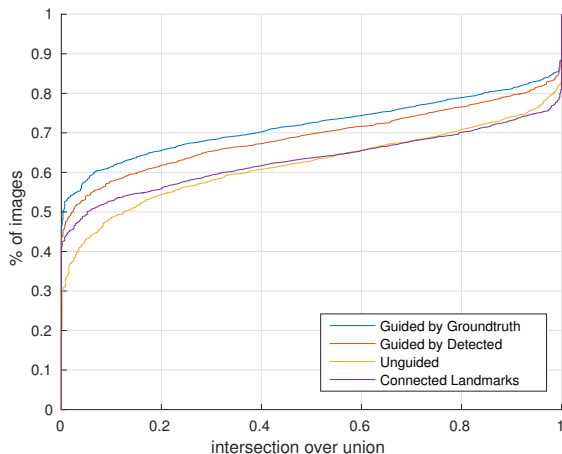
Fig. 9. Average performance of the four tested methods over all facial components: part segmentation guided by groundtruth landmarks, part segmentation guided by detected landmarks, unguided part segmentation, part segmentation by joining up the detected landmarks.

Acknowledgements

References

1. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. TPAMI **23**(6) (2001) 681–685
2. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
3. Zhang, N., Shelhamer, E., Gao, Y., Darrell, T.: Fine-grained pose prediction, normalization, and recognition. arXiv preprint arXiv:1511.07063 (2015)
4. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: CVPR-W. (2013)
5. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR. (2010)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
7. E. Snchez-Lozano, B. Martinez, G.T., Valstar, M.: Cascaded continuous regression for real-time incremental face tracking. In: European Conference on Computer Vision. (2016)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR. (2012)
9. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR. (2013)
10. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: CVPR. (2015)
11. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: CVPR. (2015)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
14. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR. (2015)
15. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: CVPR. (2015)
16. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1520–1528
17. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: CVPR. (2016)
18. Eslami, S., Williams, C.: A generative model for parts-based object segmentation. In: NIPS. (2012)
19. Eslami, S.A., Heess, N., Williams, C.K., Winn, J.: The shape boltzmann machine: a strong model of object shape. IJCV **107**(2) (2014) 155–176
20. Tsogkas, S., Kokkinos, I., Papandreou, G., Vedaldi, A.: Deep learning for semantic part segmentation with high-level guidance. arXiv preprint arXiv:1505.02438 (2015)
21. Warrell, J., Prince, S.J.: Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE (2009) 2481–2484

22. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2480–2487
23. Liu, S., Yang, J., Huang, C., Yang, M.H.: Multi-objective convolutional learning for face labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3451–3459
24. Bo, Y., Fowlkes, C.C.: Shape-based pedestrian parsing. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 2265–2272
25. Kae, A., Sohn, K., Lee, H., Learned-Miller, E.: Augmenting crfs with boltzmann machine shape priors for image labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2019–2026
26. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR. (2014)
27. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: International Conference on Computer Vision, (ICCV-W), 300 Faces in-the-Wild Challenge (300-W), Sydney, Australia, 2013, IEEE (2013)
28. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR. (2011)
29. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: ECCV. (2012)
30. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark estimation in the wild. In: CVPR. (2012)
31. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)