Imperial College London
Department of Computing

Ph.D. Thesis Corrections

Robust Statistical Deformable Models

Author: Epameinondas Antonakos
Supervisor: Dr. Stefanos Zafeiriou
Examiners: Dr. Stefan Leutenegger, Prof. Lourdes Agapito

I would like to thank the examiners for the fruitful discussion and constructive feedback on my Ph.D. work. This document summarises the corrections made to the thesis. Each correction is also highlighted in the thesis document using red colour.

General comments

Comment 1

> *"Please be more careful with the use of both proof (page 91) and optimal (pages 84, 100). optimal has to be put into context explaining the cost function and underlying assumptions; and proof should only ever be used when there is a formal proof of a theorem."*

I removed the use of "proof" in page 93 (91 in the old version of the thesis) and made sure that the word is not misused anywhere else. I also fixed all occurrences of the word "optimal" (pages 45, 50, 51, 86, 102, 104, 106, 122).

Comment 2

> *"Please be consistent with $\boldsymbol{E}$ and $\boldsymbol{I}$ notation for identity matrices."*

The symbol $\mathbf{E}$ is now consistently used for denoting an identity matrix. The changes were made in Equations 4.18, 4.21, 4.33, 4.35, 6.3, 6.17, and pages 49, 85, 102, 103.

Chapter 1: Introduction

Comment 1

> *"Please provide some more context around newer trends, specifically 3D generative models as well as deep learning along with some reasoning about why you have decided not to pursue these directions."*

I added a new paragraph at the end of Section 1.1 (pages 4, 5). The paragraph is the following:

As explained above, the work presented in this Ph.D. thesis aims to solve the problem of landmark localization by exploring generative and discriminative 2D Deformable Models. Nevertheless, there has been significant research effort on directions that approach the problem in different ways. Specifically, these are the most important current trends and the reasons why they are not within the scope of this thesis:

- 3D facial shape estimation from monocular images is the main alternative to 2D Deformable Models. The predominant lines of research include 3D Morphable Model (3DMM) [4, 5, 6, 7, 8, 26] and Shape-from-Shading (SfS) [2, 11, 17, 31, 36]. 3DMM is a generative statistical model of the 3D shape and texture of a deformable object. The biggest advantage of 3DMMs is the fact that dense 3D shape modeling provides a more natural and accurate representation of the human face that overpasses the limitations and ambiguities of 2D sparse landmarks (*e.g.*, the semantic meaning of the 2D landmarks around the jaw is ambiguous and inconsistent over the head pose variation [30]). However, capturing 3D facial data is a tedious task that also requires specialised acquisition devices that cannot operate under unconstrained conditions. As a result, there only exist small databases with limited variance that capture a few hundred faces under laboratory conditions [26, 4] and are not suitable neither for "in-the-wild" applications, nor for training discriminative methodologies. These are the main reasons why 3D Deformable Models are not within the scope of this thesis. Nevertheless, during the last year, 3D Deformable Models have re-attracted increased interest thanks to the development of the first powerful 3D models trained on thousands of subjects [8, 7], as well as the organization of the first challenges on the task [16].

- Deep Learning, and more importantly, Convolutional Neural Networks (CNNs) have become the most popular trend in Computer Vision and have significantly contributed in improving the performance of various tasks such as image classification [20, 33, 34, 15], generic object detection [12, 28], semantic segmentation [12, 23, 9, 13] and instance segmentation [27, 14]. The progress witnessed over the last decade is highly related to the spatial accuracy that CNNs were able to achieve over time, starting from boxes, moving to coarse instance regions until reaching accurate pixel-level labelling. As a result, it was not until recently that CNNs were able to perform tasks with accurate spatial localization, such as body pose estimation [35, 41] and facial landmark localization [29, 32, 43, 37, 19, 13]. However, despite the fact that facial databases include reasonably large numbers of "in-the-wild" annotated images for the generative or discriminative methodologies of this thesis, they are not large enough in order to train CNNs. As a matter of fact, LFPW [3] and HELEN [21], which are the largest facial databases annotated with 2D landmark points, consist of 1035 and 2330 images, respectively. This is orders of magnitude less than the size of ImageNet [10] ($\sim 15M$), MegaFace [18] ($1M$), WIDER [40] ($\sim 400k$) or Microsoft COCO [22] ($330k$) that are commonly used for other tasks. Finally, it is worth mentioning that the research community has been actively attempting to increase the size of annotated data during the last few months [42], which will benefit Deep Learning approaches and potentially further improve face alignment accuracy.

Chapter 2: Literature Review

Comment 1

> *"Page 18: methodologies that that employ: please correct."*

Fixed.

Chapter 3: Basic Definitions and Notation

Comment 1

> *"Equation (3.12): Is there a reason for the order of variables to be the inverse of the shape model? If not, please make it consistent."*

Fixed. The model notations are now consistent.

Chapter 4: Feature-based Lucas-Kanade and Active Appearance Models

Comment 1

> *"Please clarify that no image pyramid was used in this approach."*

A paragraph is added in Section 4.5 (page 55) to clarify this. Specifically:

"Note that commonly LK and AAMs fitting is performed using an image pyramid with progressively increasing the number of shape and appearance parameters as the image resolution increases [1, 24, 25, 39]. However, in the following experiments of this chapter, the image pyramid is not employed in order to facilitate and simplify the comparisons. Using multiple fitting scales would make it difficult to derive any conclusions about the various features and approaches, such as the representation power, number of appearance and shape eigenvectors, convergence rate, etc. Nevertheless, a multi-level pyramid fitting framework is employed in the rest of this thesis, as also explained in individual Chapters 5, 6 and 7."

Comment 2

> *"Please give some details how you implemented solving the optimisation problem and how this relates to timings."*

A paragraph is added in Section 4.5.2 (pages 62-64) which gives more details on the implementation and explains how it affects the timings. Specifically:

"The AAM fitting used in these experiments is implemented in Matlab using the Moore-Penrose pseudoinverse, which, despite the fact that it ensures robustness, it is computationally expensive. Additionally, as mentioned before, the fitting is not performed using an image pyramid. These two factors make the fitting procedure reported in Tab. 4.2 slower than expected. However, note that the aim of these experiments is to make a fair comparison of the computational complexity between the different feature types. It is not in the scope of this work to provide an optimized implementation of AAMs or features. Faster AAM optimization can be achieved with the framework proposed in [25, 38]. One could also use GPU or parallel programming to achieve faster performance and eliminate the cost difference between various features and also between the two composition scenarios of $\mathcal{F}$ and $\mathcal{W}$. Finally, by applying a multi-scale fitting using an image pyramid greatly speeds up the fitting procedure, since convergence is achieved in less iterations, as shown in Chapter 5 (Sec. 5.3) and Chapter 7."

Chapter 6: Automatic Construction of Deformable Models

Comment 1

> *"Figure 6.2: Please clarify that the figure was taken from Stefanos paper."*

Fixed.

Comment 2

> *"You claim that the IGO features are better separating the PCA enabling the somewhat magical convergence of the automatic construction of the model. In our discussion, however, we found that it works just as well with SIFT. Please clarify, as otherwise the claim is misleading."*

IGO features have been shown to suppress outliers at the very last components keeping the principal components clean [39, **?**]. This is what it is also explained in Sec. 6.2.1 and Fig. 6.2.

My point during our discussion was that other powerful features would work as well (e.g. HOG or SIFT) because of the statistics of the data that we employ for the experiments shown in the chapter. Specifically, the employed datasets (LFPW and HELEN) do not include many faces with extreme poses lots of nearly frontal images that

The method in this chapter assumes that we apply a face detector with a very small (almost zero) false positive rate, which is relatively easy to achieve for the employed datasets (LFPW, HELEN). Additionally, the employed databases do not include lots of extreme poses and most of the images are nearly frontal. This means that the data on which the method is applied are relatively clean, which does not

Chapter 7: Adaptive Cascaded Regression

Comment 1

> *"Page 117: estimate of the shape parameters pk) that: remove)."*

Fixed.

Comment 2

> *"Section 7.2.3. How do you set the lambdas? Please explain."*

This was very briefly explained in the "Implementation Details" paragraph of the experiments Section 7.3 (page 124). I removed it from there and I created a new paragraph in pages 122, 123 that explains how these parameters are fine-tuned, so that it is more clear and easier to find. Specifically, the paragraph is the one below:

"$\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ is a set of weights that control the linear combination between the regression-based descent directions and the Gauss-Newton descent directions. They are treated as a set of hyperparameters that are fine-tuned prior to fitting. Intuitively, given the properties of regression and Gauss-Newton descent directions explained above and shown in Fig. 7.1, we expect the regression-based descent directions to dominate the optimization on the first few iterations, as they are able to move towards the correct direction with steps of large magnitude. Then, the Gauss-Newton descent steps are necessary in order to converge to an accurate local minimum. The hyperparameters λ_k are fine-tuned by running extensive cross-validation experiments that perform grid search using the mean point-to-point error normalized with the interocular distance as evaluation criterion."

Comment 3

> *"Why does [157] not appear in the graphs of the evaluation? It seems there is a wrong citation. Please correct."*

The citations in the legends where wrong. They are now fixed for Figures 7.5, 7.6 and 7.10.

Conclusion

Comment 1

> *"with the gradient descent directions from Gauss-Newton optimization: strictly speaking this is not gradient descent, since Gauss-Newton is a second order method... Please adjust."*

Fixed in pages 115, 120, 121, 131 and 136.

References

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision (IJCV)*, 56(3):221–255, 2004.

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(8):1670–1687, 2015.

[3] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[5] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(9):1063–1074, 2003.

[6] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models ”in-the-wild”. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[7] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision (IJCV)*, 2017.

[8] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, 2016.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[11] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding (CVIU)*, 109(1):22–43, 2008.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[13] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 511–520. Springer, 2016.

[17] Ira Kemelmacher-Shlizerman. Internet based morphable model. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3256–3263, 2013.

[18] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016.

[19] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. *arXiv preprint arXiv:1706.01789*, 2017.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[21] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 679–692. Springer, 2012.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[24] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60(2):135–164, 2004.

[25] George Papandreou and Petros Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008.

[26] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 296–301. IEEE, 2009.

[27] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1990–1998, 2015.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[29] Samuel Rivera and Aleix M Martinez. Learning deformable shape manifolds. *Pattern Recognition*, 45(4):1792–1801, 2012.

[30] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localisation "In-The-Wild"*, 47:3–18, 2016.

[31] Patrick Snape and Stefanos Zafeiriou. Kernel-pca analysis of surface normals for shape-from-shading. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1059–1066, 2014.

[32] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483. IEEE, 2013.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[35] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1799–1807, 2014.

[36] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face normals in-the-wild using fully convolutional networks. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[37] George Trigeorgis, Patrick Snape, Mihalis Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016. IEEE.

[38] Georgios Tzimiropoulos and Maja Pantic. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013.

[39] Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. Robust and efficient parametric face alignment. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[40] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[41] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3073–3082, 2016.

[42] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of IEEE Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2017.

[43] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(5):918–930, 2016.