Convolutional aggregation of local evidence for large pose face alignment

Adrian Bulat
adrian.bulat@nottingham.ac.uk
Georgios Tzimiropoulos
yorgos.tzimiropoulos@nottingham.ac.uk

Computer Vision Laboratory
University of Nottingham
Nottingham, UK

Abstract

Methods for unconstrained face alignment must satisfy two requirements: they must not rely on accurate initialisation/face detection and they should perform equally well for the whole spectrum of facial poses. To the best of our knowledge, there are no methods meeting these requirements to satisfactory extent, and in this paper, we propose Convolutional Aggregation of Local Evidence (CALE), a Convolutional Neural Network (CNN) architecture particularly designed for addressing both of them. In particular, to remove the requirement for accurate face detection, our system firstly performs facial part detection, providing confidence scores for the location of each of the facial landmarks (local evidence). Next, these score maps along with early CNN features are aggregated by our system through joint regression in order to refine the landmarks' location. Besides playing the role of a graphical model, CNN regression is a key feature of our system, guiding the network to rely on context for predicting the location of occluded landmarks, typically encountered in very large poses. The whole system is trained end-to-end with intermediate supervision. When applied to AFLW-PIFA, the most challenging human face alignment test set to date, our method provides more than 50% gain in localisation accuracy when compared to other recently published methods for large pose face alignment. Going beyond human faces, we also demonstrate that CALE is effective in dealing with very large changes in shape and appearance, typically encountered in animal faces.

1 Introduction

Face alignment refers to the problem of localising a set of fiducial points on the human face. Being a long-standing problem in Computer Vision research, a multitude of approaches with various degrees of success have been proposed so far to solve it. With the advent of cascaded regression [10] and its application to face alignment [7, 27, 28, 34], state-of-the-art is now considered to have reached a satisfactory level of performance for frontal faces including faces with difficult illumination, expression and occlusion. Yet, the problem of face alignment under very large pose variation (including alignment of profile faces) has received little attention so far. This paper proposes a CNN architecture that copes well for the case of (a) inaccurate initialisation/face detection, and (b) severe self-occlusions, and hence it is particularly suitable for arbitrary pose face alignment.

Recently, regression has been the standard approach to face alignment. Because learning a direct mapping from image features to landmark locations might be hard, most approaches

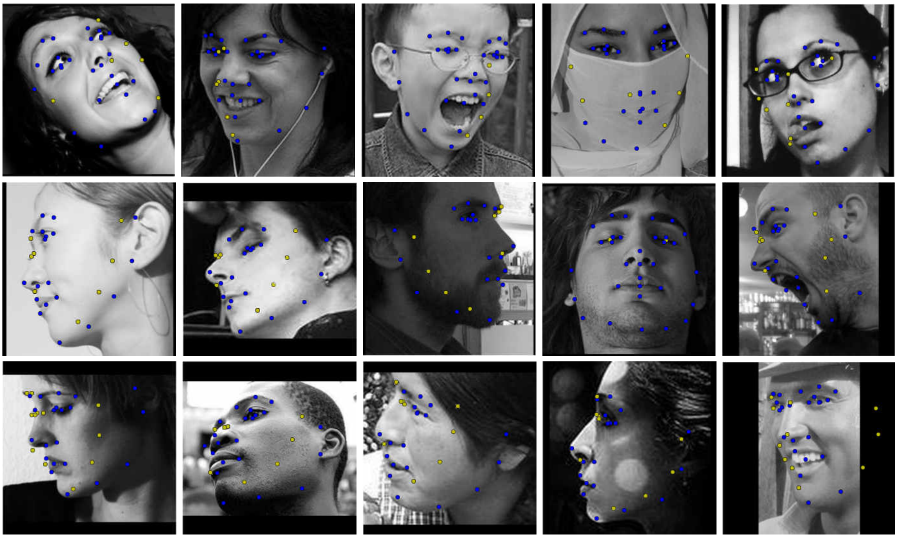


Figure 1: Qualitative fitting results produced by CALE on AFLW-PIFA test set. Observe that our method copes well for both occlusions and difficult poses. Blue/Yellow points indicate visible/invisible landmarks. All the keypoints are detected from a **3D perspective**, so the non-visible (yellow) points are actually accurately localised for the majority of cases.

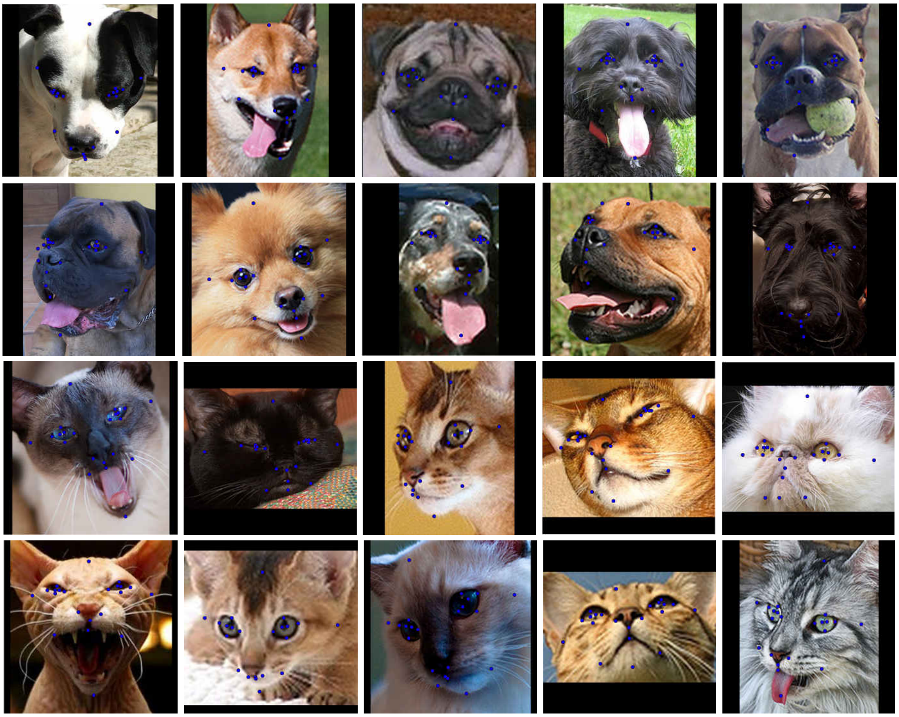


Figure 2: Qualitative results produced by CALE on our Cats&Dogs dataset.

learn a cascade of regressors, applied in a progressive manner by initialising each regressor with the output of the previous one in the cascade. Such methods have been shown to produce remarkably accurate results on a number of face datasets with significant expression, illumination change and to some extent occlusion like LFPW[4], Helen [17] and 300-W[21]. Yet, it is well-known that such methods (a) are sensitive to initialisation (see for example [29]), and (b) that their performance deteriorates for large pose datasets (e.g. AFLW-PIFA [13, 16] and AFW [36]) especially when there is a significant number of self-occluded landmarks or when there are large rotations (both out-of-plane and in-plane) and, in general, unfamiliar poses. Due to poor visibility caused by self-occlusion and due to the large number of unfamiliar poses, it is unclear whether cascaded regression methods can learn a mapping from a large number of occluded/non-visible parts to landmark coordinates.

To address the aforementioned limitations of prior work, in this paper we propose a CNN architecture for large pose face alignment which we call Convolutional Aggregation of Local Evidence (CALE). CALE by-passes the requirement for accurate face detection by firstly using a CNN detector to perform facial landmark detection, providing at the same time confidence scores for the location of each of the facial landmarks (local evidence). Next, CALE aggregates the local evidence for each facial landmark through joint CNN regression of the confidence scores, in order to refine the landmarks' location. Besides playing the role of a graphical model, CNN regression is a key feature of our system, guiding the network to rely on context for predicting the location of occluded landmarks, typically encountered in very large poses. The proposed architecture is very simple and can be trained end-to-end with intermediate supervision. We show that our system achieves large performance improvement on AFLW-PIFA, which is, to the best of our knowledge, by far the most difficult test set for face alignment to date.

Our second contribution in this paper is an investigation of CALE's alignment performance beyond human faces and, in particular, on animal faces. As animal faces exhibit a much larger degree of variability in shape and appearance as well as in pose and expression, animal face alignment is a much more difficult problem which, to the best of our knowledge, has never been systematically explored in the past by the Computer Vision community. Although drawing a direct comparison is not possible, our results, both quantitative and qualitative (see Figs 1 and 2), show that CALE's performance on animal faces is not far from that on human faces.

2 Related Work

This section reviews related work on face alignment.

2D face alignment. State-of-the-art in 2D face alignment are techniques based on cascaded regression, see for example [7, 27, 28, 34]. Most commonly, such methods rely on hand-crafted features, are sensitive to face detection initialisation [29], might require a cascade with many steps, and most notably have been shown to work well mainly for frontal datasets like LFPW[4], Helen [17] and 300-W[21] in which most of the landmarks are visible. On the contrary, our method does not rely on accurate face detection, uses a single regression step and can cope well with arbitrary poses and severe self-occlusion. Notably, the idea of aggregating local evidence for facial landmark localisation has been explored within methods based on the so-called Constrained Local Model (CLM) [1, 2, 8, 22]. Note that all CLM-based methods use hand-crafted features and have been shown to be largely outperformed by cascaded regression methods. On the contrary, we show that our method, which can be seen as a deep version of the CLM, largely outperforms all prior work on large pose face alignment.

Large pose face alignment. State-of-the-art methods for large pose face alignment include techniques that attempt to perform face alignment by fitting a 3D Morphable Model (3DMM) to a 2D image [13, 14, 37]. The work in [13] aligns faces using a sparse 3D point distribution model the parameters of which along with the projection matrix are estimated by cascaded regression. Notably, [13] introduces AFLW-PIFA, the most challenging, to the best of our knowledge, dataset for large pose unconstrained face alignment. The work in [14] extends [13] by fitting a dense 3DMM using a cascade of CNNs. A similar approach to [14] has been also proposed in [37]. Besides 3DMM-based approaches, the work in [37] performs large pose 2D face alignment based on compositional cascaded learning, a novel way to perform model averaging within cascaded regression. Despite its elegant formulation, [37] completely avoids regressing non-visible landmarks and suffers from many of the problems common in all cascaded regression techniques (please see above). Compared to [13, 14, 37], our system by-passes the burden of fitting a 3D model and compared to [37], our method avoids the limitations of cascaded regression. On AFLW-PIFA, our system reduces the error reported in [13, 14, 37] by more than 50% ([37] does not report performance on this dataset).

CNNs for face alignment. CNNs have been applied to the problem of face alignment, only recently. One of the very first attempts that uses a simple CNN to regress landmark locations on the face image was proposed in [24]. The work in [32] proposes to combine facial landmark localisation with attribute classification through multi-task learning. One limitation of both methods is that they can detect 5 landmarks only. Very recent work includes [14, 37] mentioned above and [26] which extends [28] within recurrent neural networks. Our work largely outperforms [14] on AFLW-PIFA, while [26] has been applied to frontal face alignment only, not reporting performance on AFLW or other large pose face alignment datasets (e.g. AFLW-PIFA [13, 16] and AFW [36]).

CNN regression. Recently, methods based on CNNs have been shown to produce state-of-the-art results for many Computer Vision tasks like image recognition [23], object detection [11] and semantic image segmentation [18]. In the context of landmark localisation, it is natural to formulate the problem as a regression one in which CNN features are regressed in order to provide a joint prediction of the landmarks, see for example recent works on human pose estimation [3, 5, 20, 25]. The idea of joint regression of part detection scoremaps for localisation has been explored in [5], however in the context of human pose estimation.

3 Convolutional Aggregation of Local Evidence

In our system (CALE), a CNN detector is firstly trained to detect the individual facial landmarks thus by-passing the requirement for accurate face detection. At the same time, the CNN detector provides confidence scores for the location of each of the facial landmarks (local evidence). Next, CALE aggregates the local evidence for each facial landmark through joint CNN regression of the confidence scores stacked with high-resolution CNN features, in order to refine the landmarks' location. The CNN detector and regressor are described in detail in the following subsections. The proposed architecture is illustrated in Fig. 3.

CNN detection. One of the main issues with almost all prior techniques on face alignment is face detection initialisation. It is well-known that face alignment methods are sensitive to how accurate the face detection algorithm is, with faces in difficult poses being

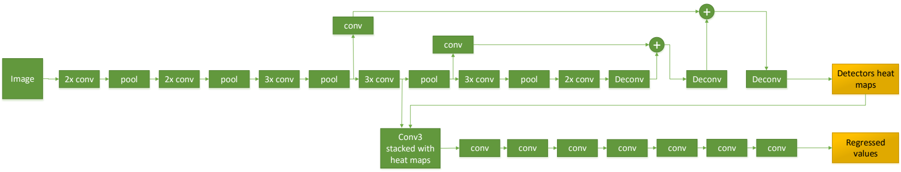


Figure 3: Proposed architecture for Convolutional Aggregation of Local Evidence (CALE).

usually detected with less accuracy. A second important, but not so well-discussed, issue is that typically face alignment methods are tight with a specific face detector, with alignment accuracy rapidly deteriorating if a different face detector (than the one that the face alignment algorithm was trained on) is used. Notably, face alignment methods are usually tight to both the statistics of the face detector and the definition of the face region that the detector was trained on.

To overcome the strong dependency on the face detector, CALE firstly performs detection of the individual facial landmarks, expecting as input a grayscale image, scaled and cropped based on a bounding box, used only for obtaining an rough estimate of the face size and location. See Figs 1 and 2. Notably, the CNN landmark detectors of CALE are trained jointly as follows: the detectors are based on a fully convolutional VGG [23], also making use of earlier level CNN features to increase spatial resolution, as proposed in [18]. To avoid the problem of neighbouring landmarks overlapping with each other, we used the training procedure for detecting landmarks described in [31]: each landmark is encoded as a binary image, with the values being within a specific radius around each landmark's location set to 1 (otherwise they are set to 0). Hence, the output of this network is a set of N channels, one for each landmark. A radius of 10 pixels was found to work well for a face size of 175 pixels. Finally, the facial landmark detectors are trained jointly using the pixelwise sigmoid cross entropy loss [31].

CNN regressor. The CNN regressor of CALE aims to play the role of a graphical model, enforcing additional shape constraints necessary for enhancing accuracy and robustness to occlusion [5, 20]. To this end, the N landmark heatmaps produced by the CNN detector are stacked with high resolution CNN features produced from `conv3_3` (see Fig. 3) and then are fed to CALE's CNN regressor. The CNN regressor has seven convolutional layers, the first for of which have a kernel size varying from 7 to 17, ensuring a sufficiently large receptive field. The last three layers have a kernel size equal to 1, and are the equivalent of the fully connected layers [18]. Finally, the regressor has N output channels, one for each landmark. As in [20], we represent each landmark with a Gaussian (with standard deviation of 9 pixels) centred at the landmark's ground truth location. Finally, the CNN regressor is trained to regress the location of all landmarks jointly using the L2 loss [20].

Training. We trained our CNN landmark detectors by fine-tuning from a VGG-16 network that was previously trained on ImageNet [9]. We followed a training procedure similar to the one described in [18] by firstly, performing a "network surgery" which converts VGG-16 to a fully convolutional network. We firstly trained the 32-stride model with a learning rate of $1e-7$ for 10 epochs. Because the 32-stride version of the network does not provide enough resolution, we went all the way down to 8-stride. The detectors were trained under this setting for 20 epochs (25 for the Cats&Dogs dataset) with a learning rate of $1e-8$. Then, we gradually reduced the learning rate twice, down to $1e-10$. All the new learned layers were initialised with zeros. In order to avoid early divergence, we froze the learning for all CNN detector layers and set temporary the learning rate to 0, training only the CNN regressor. We trained the sub-network for 30 epochs with a learning rate of $1e-6$. After 20 epochs, we lowered it to $1e-7$ and continued the training until convergence was reached. The entire network (CNN detector and CNN regressor) was then trained jointly, in an end-to-end fashion for 5 more epochs. All the new layers added were initialised with a random Gaussian distribution with standard deviation of 0.01.

Regarding data augmentation, we applied image flipping and scale jittering (0.8-1.2). Because the images provided in the AFLW-PIFA dataset were grayscale, the human face alignment model was trained with grayscale images, while the one for animals using colour images.

All models were trained and tested using Caffe[12] on a single Titan X GPU. The models and the code will be published on our page.

4 Results

We firstly report results on the most challenging and large scale dataset for large pose human face alignment, namely AFLW-PIFA [13], illustrating that CALE reduces the error achieved by state-of-the-art methods [35, 37] by more than 50%. Then, we report results on our Cats&Dogs dataset, illustrating, for the first time, that a face alignment method is capable of achieving similar performance on both animal and human faces.

4.1 Human faces

We have opted not to report results on LFPW[4], Helen [17] and 300-W[21] which are all frontal datasets containing a small portion of test images and are currently being considered as saturated [35, 37]. Instead we report performance on AFLW-PIFA which is by far the most challenging dataset for large pose face alignment [13]. In particular, the authors of [13] created a subset of AFLW [16] that has a balanced distribution of yaw angles (from -90 degrees to 90 degrees) including 3901 images for training and a large number of 1299 for testing. Notably, besides the existing 21 key points, this subset contains 13 new landmarks, making the total number of annotated keypoints equal to 34. All the images are annotated from a 3D perspective which makes the landmark location prediction even more difficult, making AFLW-PIFA the most challenging dataset for face alignment. We report results on the original 21 point annotations [13] as well as on the new ones, based on 34 points [14].

The evaluation metric used for AFLW-PIFA subset is the Normalized Mean Error (NME), which is the average of the normalized (by the face size as defined in [14]) estimation error of the visible landmarks:

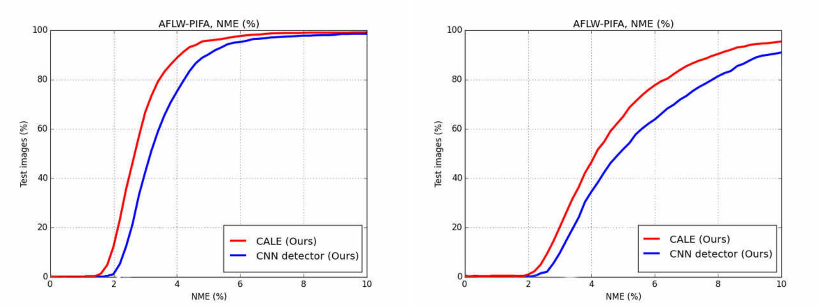
$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_i|v_i|_1}\sum_{j}^{N_k} v_i(j)\left\|\widetilde{L}_i(:,j)-L_i(:,j)\right\|, \tag{1}$$

where N is the total number of faces, N_k the number of keypoints and v_i the corresponding visibility label for the image I_i. For each image, the error is normalized by f_i, which for ALFW-PIFA is the square root of the face size calculated from the bounding box as in [14].

Firstly, we compare the performance of our CNN detector alone with that of the overall CNN architecture (CALE). We opted to report performance on both occluded and visible points. The results on AFLW-PIFA are given in Table 1 and Fig. 4. We observe that although

	21 points (vis.)	21 points	34 points (vis)	34 points
CNN detector	3.32	5.53	3.63	5.96
CALE	**2.63**	**4.38**	**2.96**	**4.97**

Table 1: Performance analysis of CALE on AFLW-PIFA using NME (%). Results are reported on both 21 and 34 points. Results marked with `(vis)` are calculated on visible points only, while the rest are calculated on both occluded and visible landmarks.



a) Evaluation on visible points only b) Evaluation on both invisible and visible

Figure 4: NME-based (%) comparison between CNN detector and CALE on AFLW-PIFA on 34 points.

the CNN detector alone performs very well, CALE largely outperforms it achieving very high alignment accuracy. The performance improvement offered by CALE is even greater on the occluded points, verifying the usefulness of the CNN regressor for the difficult poses and occlusions of AFLW-PIFA.

Next, we compare the performance of our method with that of currently considered state-of-the-art methods for large pose face alignment, also including the very recent works of [14] and [35]. Tables 2 and 3 summarise our results on AFLW-PIFA on both 21 and 34 points for the visible points only. From Table 2, we observe that CALE largely outperforms all other methods by a remarkable more than 50%, reducing the error of the second best performing method [35] to more than half. Similarly, from Table 3, we observe that the improvement over the second best performing method approaches 37%. Note that prior work reports on visible points, only. To the best of our knowledge we are the first to report results on non-visible landmarks too, please see Table 1. Remarkably, the performance of CALE when evaluated on all points - both visible and occluded (see Table 1) surpasses the performance of all existing methods when these are evaluated on visible points only (see Tables 2 and 3). Fitting results from AFLW-PIFA can be seen in Fig. 1.

CDM [30]	CFSS [33]	ERT [15]	SDM [28]	PIFA [13]	CCL [35]	Ours
8.59	6.75	7.03	6.96	6.52	5.81	**2.63**

Table 2: NME-based (%) comparison on AFLW-PIFA on 21 points (visible landmarks only). The results for CFSS, ERT and SDM are taken from [35].

Evaluation	PIFA [13]	RCPR [6]	PAWF [14]	Ours
AFLW-PIFA	8.04	6.26	4.72	**2.96**

Table 3: NME-based (%) comparison on AFLW-PIFA evaluated on 34 points (visible landmarks only). The results for PIFA, RCPR and PAWF are taken from [14].

4.2 Animal faces

While human face alignment is a well-studied problem, the problem of animal face alignment, to the best of our knowledge, has never been systematically explored in the past by the Computer Vision community. As animal faces exhibit a much larger degree of variability in shape and appearance as well as in pose and expression, animal face alignment is considered a much more difficult problem. Cats and dogs, the two species chosen here, are the most popular companion animals, worldwide and of enormous societal and economic importance. Motivated by our results on human face alignment, we investigate CALE's performance on cat and dog face alignment. Although drawing a direct comparison is not possible, our results, both quantitative and qualitative (see Figs 1 and 2), show that CALE's performance on animal faces is not far from that on human faces.

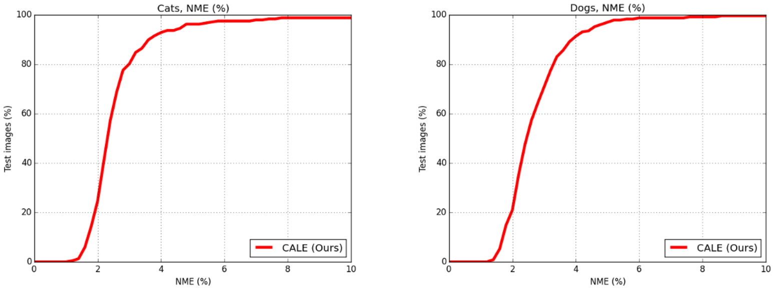
Our Cats&Dogs dataset is a subset of the Oxford-IIIT-Pet dataset [19] which contains a rich variety of cats/dogs breeds, making the dataset particularly challenging. Our dataset contains 1511 images of cats and 1514 of dogs. For both animals, we kept 250 images for testing and used the rest for training. We used 22 landmarks similarly defined for both species (see 2). To measure performance, we used the same metric as the one used for AFLW-PIFA.

Fig. 5 and Table 4 summarise our results on 22 points. As we may observe, CALE literally produces the same fitting accuracy for both species. Next, we attempt to make a comparison between CALE's performance on human and animal face alignment using 9 commonly defined points (2 on the corners of each eye, 1 on the nose, 3 on the upper mouth and 1 on the jaw). Note that direct comparison is by no means straightforward as although our Cats&Dogs dataset has "similar" training and testing sets for both species, AFLW-PIFA is very different, including more images for training and testing and very large pose variation. Fig. 6 shows the obtained results. We may observe that CALE produces literally the same performance for humans and cats, while the performance on dogs is inferior. This performance deterioration is mainly due to the upper mouth and jaw landmarks which are more noisy for dogs. Note however that when evaluation is done on all points (see Fig. 5 and Table 4), this gap in performance diminishes illustrating that the difference in performance shown in Fig. 6 is magnified by the not so large number of landmarks used.

Evaluation

Evaluation	Ours
Cats&Dogs (Cats subset)	**2.72**
Cats&Dogs (Dogs subset)	**2.71**

Table 4: NME-based (%) performance on Cats&Dogs on 22 points.



a) Cats. b) Dogs.

Figure 5: NME-based (%) performance on Cats&Dogs on 22 points.

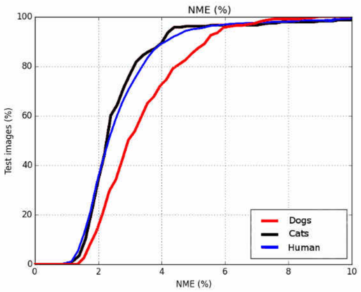


Figure 6: NME-based (%) comparison between human and animal faces on 9 commonly defined points (2 on the corners of each eye, 1 on the nose, 3 on the upper mouth and 1 on the jaw).

5 Conclusions

We proposed Convolutional Aggregation of Local Evidence, a very simple CNN architecture for large pose face alignment. We showed that such an approach is particularly suitable for the case of large amount of self-occlusion typical in profile faces and unfamiliar poses. The proposed architecture is very simple and was shown to achieve large performance improvements on the most difficult datasets for large pose face alignment, for both human and animal faces.

6 Acknowledgment

This work was supported in part by the EPSRC project EP/M02153X/1 Facial Deformable Models of Animals.

References

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

[2] Akshay Asthana, Stefanos Zafeiriou, Georgios Tzimiropoulos, Shiyang Cheng, and Maja Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE TPAMI*, 37(6):1312–1320, 2015.

[3] Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab. Robust optimization for deep regression. In *ICCV*, 2015.

[4] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[5] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.

[6] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[8] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, 2010.

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *ICCV*, 2015.

[14] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016.

[15] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[16] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV-W*, 2011.

[17] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*, 2012.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[19] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

[20] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

[21] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR-W*, 2013.

[22] J.M. Saragih, S. Lucey, and J.F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[25] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[26] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[27] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015.

[28] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[29] Heng Yang, Xuhui Jia, Chen Change Loy, and Peter Robinson. An empirical study of recent face alignment methods. *arXiv preprint arXiv:1511.05049*, 2015.

[30] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *CVPR*, 2013.

[31] Ning Zhang, Evan Shelhamer, Yang Gao, and Trevor Darrell. Fine-grained pose prediction, normalization, and recognition. *arXiv preprint arXiv:1511.07063*, 2015.

[32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*. 2014.

[33] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[34] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[35] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016.

[36] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark estimation in the wild. In *CVPR*, 2012.

[37] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. *CVPR*, 2016.