MULTI-DROPOUT REGRESSION FOR WIDE-ANGLE LANDMARK LOCALIZATION

Gee-Sern Hsu, Cheng-Hua Hsieh

National Taiwan University of Science and Technology, Taipei, Taiwan

ABSTRACT

We propose the Multi-Dropout Regression Network (MDRN) for real-time facial landmark localization across extreme poses. Different from most landmark localization methods only work for -45° to 45° in yaw, the proposed MDRN works for the full coverage of -90° to 90°. It employs the Single Shot Multibox Detector (SSD) [1] as a preprocessor for fast and accurate face detection. Given an SSD detected face, the MDRN locates the landmarks. The MDRN is composed of 2 double-layer convolution blocks, 1 triple-layer convolution block and 3 fully-connected layers. Unlike most networks with only one dropout layer connected to the last convolution layer, in the MDRN each convolution block is followed by a max-pooling layer and a dropout layer ahead of connecting to the next processing layer. Experiments reveal that multiple dropouts better stabilize the regression and improve the accuracy of landmark localization. To locate the landmarks on profile faces and other extreme poses, the MDRN is trained on an augmented database composed of imagery of synthesized poses. A comparison study shows that the proposed solution delivers a comparable performance to the state of the art for wide-angle landmark localization.

Index Terms— Face Alignment, Deep Learning, Convolutional Neural Network

1. INTRODUCTION

Fully automatic facial landmark localization is split into two phases, the first is face detection and the second is the landmark localization on the detected faces. Most studies on landmark localization and face alignment only focus on the second phase, assuming that the locations of faces can be obtained by a face detector. The popular Viola-Jones detector is used in [2, 3, 4], and the Tree Structured Model (TSM) is used in [5]. However, the Viola-Jones detector cannot handle faces with large rotation, expression variation and unbalanced illumination conditions; and the TSM detector is too slow to handle practical applications. Another big issue with these and many other landmark localization approaches is that they can only handle up to $45°$ in yaw, while many applications, e.g., cross-pose recognition, require landmark localization across wide angles, i.e., up to $90°$ in yaw.

Thanks to Taiwan Ministry of Science and Technology (MOST) for funding.

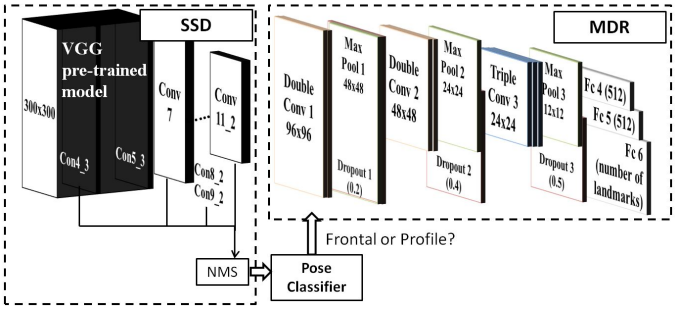

Fig. 1. The architecture of MDRN combined SSD, pose classifier and Multi-Dropout Regressor.

Several contemporary methods achieve high accuracy in limited pose range. The Supervised Descent Method (SDM) [2] learns a sequence of descent directions that minimizes the mean of the cost functions sampled at different points. The Regressing Local Binary Features (RLBF) in [3] explores a better learning based *locality* principle and achieves a better accuracy than the SDM. The locality principle is to learn the discriminant characteristics of local binary features. Although both SDM and RLBF attain high accuracy with fast speed, they can only handle poses within 45° and cannot deal with wide-angle landmarking.

The Cascaded Deformable Shape Model (CDSM) [6] selects the most salient facial landmarks by using a group sparse learning approach. It can detect wide-angle landmarks, but it takes > 5 secs to locate the landmarks on a face in the AFW. The Regressive Tree Structured Model (RTSM) is composed of a coarse TSM (c-TSM) and a refined TSM (r-TSM). The c-TSM is defined with fewer parts on a low-resolution image, and the r-TSM is defined with more parts on a high-resolution image. The c-TSM performs as a fast but coarse face detector that searches through the whole image for facial candidates. The facial candidates are then processed by the r-TSM for removing the false positives and locating landmarks on the true positives. It takes the RTSM 0.7 sec to locate the landmarks on a face in the AFW. 3DDFA [7] combines a cascaded CNN regressor and 3DMM, and formulate the problem as 3DMM fitting problem. The dense 3D shape allows a design with pose-invariant appearance features for CNN learning.

We propose the Multi-Dropout Regression Network (MDRN) which employs the Single Shot Multibox Detec-

Fig. 2. Samples from the test set of the WIDER FACE [9] with detection bounding boxes given by the proposed SSD face detector. Following the defined partition [9], the WIDER FACE is split into training set and validation set with 16,106 images (199k faces) and test set with 16,097 images (194k faces).

tor (SSD) as a preprocessor for fast face detection and the core multi-dropout network for wide-angle landmark localization. In our framework, the SSD face detection preprocessor define the ROIs for training data, and the landmarks are trained based on these ROIs. Therefore the preprocessor must be included when applying our framework for localizing landmarks. In the following, we first present the MDRN architecture in Sec.2, and the experimental evaluation in Sec.3, and then a conclusion in Sec.4.

2. MULTI-DROPOUT REGRESSION NETWORK

The architecture of the MDRN is shown in Fig.1. It is composed of an SSD face detector, a pose classifier and the core multi-dropout regressor.

2.1. Single Shot Multibox Face Detector

The Single Shot Multibox Detector (SSD) was proposed for real-time object detection. Unlike the Faster R-CNN (Region-based Convolutional Neural Network) [8] which needs to integrate the Region Proposal Network (RPN) for generating possible object regions, the SSD uses the anchor boxes of difference scale and aspect ratios over multiscale feature maps for locating object locations. SSD achieve 77.2 mAP on VOC2007 and 46 FPS with Titan X, faster and higher accuracy than Faster R-CNN which acquire 73.2 mAP and 7 FPS. Fig.1 shows the SSD detector in the MDRN. The loss function considered in the SSD takes the following form:

$$E(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (1)$$

We train the SSD using one of the latest large face databases, WIDER FACE [9]. The 393,703 labeled faces in the WIDER FACE illustrate large variant conditions. Fig.2 shows a few samples from the test partition of the database

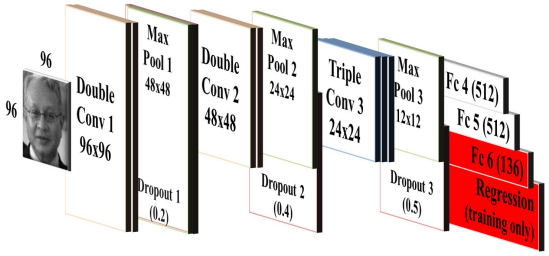

Fig. 3. The proposed Multiple Dropout Regressor (MDR) network where the red layer denotes the 136D output layer for locating 68 landmarks.

with the detection bounding boxes given by the SSD. Details of the experiments are reported in Sec. 3.

2.2. Pose Classifier

The pose classifier is built on the VGG network [10], and trained on the 300W Large Pose (300W-LP) database [7] for determining the pose of an input face. The input face is classified to either frontal ($\leq 45°$) or profile ($> 45°$). The 300W-LP is an augmented version of the 300W database. The augmentation explores the face profiling technique, which reverses the approach for face frontalization [11, 12], and generates multiple poses for each face in the 300W database. The core part of face profiling involves the 3D Morphable Model (3DMM) fitting following the Multi-Features Framework (MFF) on the face region, and 3D meshing on the background region [11]. All augmented data are landmark annotated, and thus constitute an appropriate database for training our cross-pose landmark regressor networks.

2.3. Multi-Dropout Regressor

The proposed Multi-Dropout Regressor (MDR) is composed of a feature extraction block and a regression layer. The feature extraction block aims to locate the landmarks inside a facial region regardless of variations caused by illumination, expression, pose and other parameters. The regression layer defines the landmarks to be learned. The training is split into two phases. In Phase 1, we train the MDR using the Multi-Task Facial Landmark (MTFL) dataset [13] for 5-landmark localization with 5 location outputs at the regression layer. The MTFL contains 10,000 face images with 5 landmarks, including two pupils, nose and two corners of the mouth. In Phase 2, we *fine tune* the last fully-connected layer of the feature extraction block and change the dimension of the regression layer so that the MDR can locate the desired number of landmarks, which is 68 in our case. Fine-tune refers to a process that freezes some of the weights of the network and only allows part of the weights to change. We only allow the coefficients of the three fully-connected layers to change in Phase 2, on the 300-W training set. Fig.3 shows our fine-tuning network.

As the boundaries of the SSD detected faces are close to physical facial boundaries, the receptive field in Phase 2 is enlarged to encompass the cross-boundary features. Experiments show that this enlargement improves the accuracy of the landmarks. The mean error can be reduced for 2% in 300-W.

We train the 5-landmark MDR in Phase 1 with within-face landmarks without considering the face-boundary landmarks. The within-face landmarks are located in the facial region and thus can better capture the local variations within the face. Compared with the face-boundary landmarks that capture the local variations across the facial boundaries, the within-face landmarks make the network better learn the characteristics within faces. Experiments show that the MDR can be difficult to converge when including the face-boundary landmarks in Phase 1. Euclidean distance is chosen as the loss function for linear regression.

Dropout prevents overfitting and offers an efficient way to approximately combine exponentially many different network architectures. The multiple dropout is proposed in this study to stabilize the network and refrain it from overfitting. The dropout rate is determined by running a sequence of different rates and comparing the performance. The selected dropout rates are 0.2, 0.4 and 0.5 for the first double-convolution layer, the second double-convolution layer and the trple-convolution layer, respectively, as shown in Fig.3.

For locating the landmarks on faces of extreme poses, we also adopt the above 2-phase training. In Phase 1, we train the MDR on the 300W-LP for locating the 5 landmarks, including eyes corner, corner of mouth, corner of eyebrow and tip of the nose. In Phase 2, we fine tune the last fully-connected layer of the feature extraction block and change the dimension of the regression layer so that the MDR can locate the desired number of landmarks, which is 39 in our profile case. We only allow the coefficients of the fully-connected layers to change in Phase 2, on the 300W-LP dataset.

3. PERFORMANEC EVALUATION

Two face detection algorithm, Faster R-CNN and SSD, are considered in our experiments. However, the latter gives AP 90.64 and 95.7 on the AFW and, and the former gives 94.45 and 95.28, with runtime speed 15 FPS (Faster R-CNN built on the ZF network [14]) v.s. 35 FPS (SSD300). Finally, we select SSD for our face detector because they achieve similar performance, however, SSD is two times faster than Faster R-CNN.

To emphasize the performance for handling in-the-wild conditions, we choose the AFW [15] and the PASCAL Face [16] for evaluating face detection, and the AFW and 300W for evaluating landmark localization. The PASCAL Face is used only for studying face detection as its samples are not landmark annotated. Although the 300W dataset is generally accepted as a good benchmark for assessing landmark

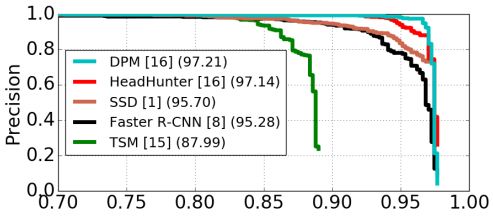

Fig. 4. Precision-recall rates on the AFW dataset with Average Precision (AP) in the parentheses. TSM is the p-1050 model [15]. The DPM and HeadHunter are both from Mathias et al. [16]

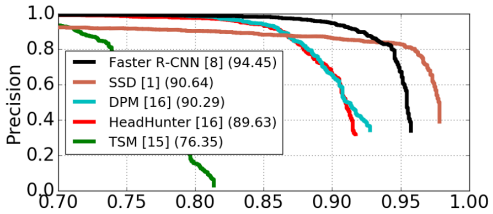

Fig. 5. Precision-recall rates on the PASCAL FACE dataset with Average Precision (AP) in the parentheses.

localization, it does not contain samples with poses large than 45° in yaw, which restrains its effectiveness for evaluating the landmark localization on poses beyond that range.The AFW offers 468 in-the-wild faces with wide-angle poses.

The face detection performance of the SSD, in terms of the precision-recall rates, on the AFW and the PASCAL Face are shown in Fig.4 and 5, together with the performances of three state-of-the-art methods, the TSM [15], Faster R-CNN, the DPM and HeadHunter [16]. We follow the evaluation protocol proposed in the development of the DPM and HeadHunter detectors when making Fig.4 and 5 for showing performances in precision and recall. The performance comparison on the PASCAL Face is shown in Fig 5 The SSD face detector outperforms the TSM with a clear margin, and it is comparable to the state of the art, Faster R-CNN, the DPM and HeadHunter, with AP 90.64% on the AFW and 95.70% on PASCAL Face.

When calculating the landmark localization error, we adopt a common metric [17, 3] that normalizes the location error to the (horizontal) distance between the eyes for poses with yaw angle $< 45^o$, or to the (vertical) distance between the eye and the mouth for poses beyond that range. This normalized location error is averaged over all landmarks and images in a dataset, and represented in terms of percentage.

Table 1 shows the normalized landmark localization error of the proposed approach, along with the performances of other state-of-the-art approaches. The codes of these state-of-the-art approaches are mostly released by the authors, except the CDSM [6], which we do not have the codes and obtain its performance directly from their paper (with result on AFW only). For the first four approaches that we have codes, the

Table 1. Landmark accuracy and localization-only runtime speeds. The location error is normalized to either inter-pupil distance ($\leq 45^o$) or eye-to-mouth ($> 45^o$) distance, in terms of %.$(\cdot)$ indicates the time including face detection and landmark localization. * indicates running on Matlab and @ means running on GPU.

	AFW	300-W			Runtime/Face	Runtime/Face
Method	$> 45°$	Common	Challenge	Full Set	in MPIE (ms)	in AFW (ms)
SDM [2]	-	5.57	15.40	7.50	25	29
RLBF [3]	-	4.95	11.98	**6.32**	**10**	**11**
DRMF [5]	-	6.65	19.79	9.22	0.9k*	1.1k*
TSM [15]	7.11	8.22	18.33	10.20	(8.8k), (25.6k)*	(24.9k), (71.2k)*
CDSM [6]	6.62	-	-	-	-	(5.8k)*
3DDFA [7]	-	6.15	10.59	7.01	25@	25@
3DDFA+SDM	-	5.53	9.56	**6.31**	50@	54@
Proposed	**6.17**	5.23	11.52	**6.42**	**10@**	**10@**

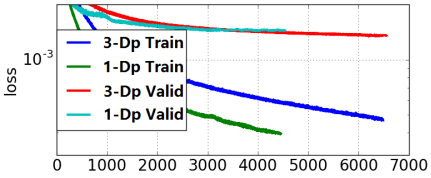

Fig. 6. Different convergences with different number of dropout layers. 3-Dp refers to three dropout layers and 1-Dp refers to one dropout layer.

errors and the runtime speeds in the table are based on our tests on the same platform. The DRMF codes cease to run when presented a face with yaw angle $> 45^o$, so we leave the corresponding grids empty. Since the 300W dataset can be split into a Common subset and a Challenging subset [3], we report the performances of the two subsets and of the overall full set. The Runtime/Face is the average time needed for landmark localization on each face. As some codes are in Matlab and some are in C/C++, we tag a star "*" to those in Matlab, and we have both types of codes for the proposed approach and TSM. Note that the TSM and the CDSM combine face detection and landmark localization in the model, we put their runtime in a parenthesis $(\cdot)$. The performances shown in Table 1 can be summarized as follows.

Fig.6 shows the the loss convergence during training and validation for different numbers of dropouts. The validation loss with one dropout (1-Dp Valid) stops to decrease as epoch exceeds 3,000, but the associated training loss (1-Dp Train) keeps going down. This is a clear sign of overfitting. However, the validation loss with three dropouts (3-Dp Valid) decreases along with the associated training loss.

The proposed approach performs the best for handling the faces $> 45^o$ in yaw, as the RLBF [3], SDM [2] and DRMF [5] cannot handle this pose range. This capacity makes the proposed approach one of the most appropriate landmark localization approaches for wide-angle landmark localization. Fig.7 shows samples with landmarks located by the best three algorithms, the proposed approach, RLBF and SDM in our test.

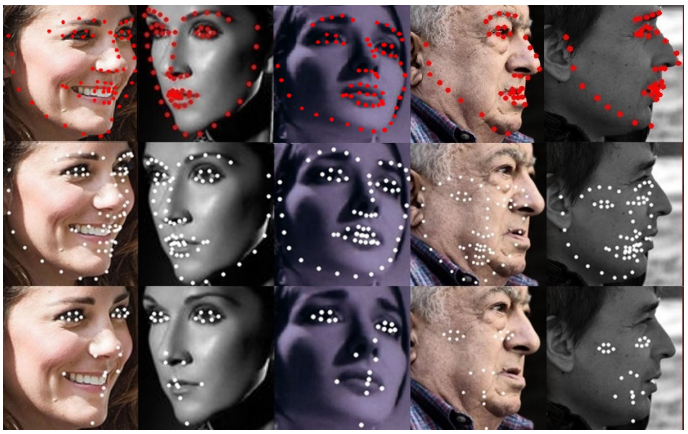

Fig. 7. Comparison of different landmark localization approaches. The first row is obtained by the proposed approach, the second by RLBF [3] and the bottom by SDM [2]. RLBF and SDM fail to handle profile or nearly profile faces, but the proposed approach works well for all poses.

4. CONCLUSION

Although the TSM-based approaches can locate wide-angle landmarks, they suffer from sluggish processing speeds as their core models integrate face detection and landmark localization. The proposed approach exploits the SSD, one of the latest and fastest object detectors, for face detection, and combines this face detector with a pose classifier and the proposed Multi-Dropout Regressor (MDR). The MDR is experimentally proven stable to train and able to yield high accuracy in localization. Experiments show that the proposed approach can be one of the best solutions for the alignment of faces with large rotation.

In addition to cross-pose recognition, other cross-pose facial analyses, such as age and expression identification, also attract general attention in recent years. Wide-angle face alignment can be a vital component in such analyses. We wish this study could motivate more research in this regard.

5. REFERENCES

[1] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[2] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[3] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014, pp. 1685–1692.

[4] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *IJCV*, vol. 91, no. 2, pp. 200–215, Sep 2011.

[5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," *CVPR*, 2013.

[6] 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 *ICCV*, 2013.

[7] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.

[8] 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 (ANIPS)*, 2015, pp. 91–99.

[9] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learning Representations (ICLR)*, 2015.

[11] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *CVPR*, 2015, pp. 787–796.

[12] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[13] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*. Springer, 2014, pp. 94–108.

[14] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[15] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[16] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *ECCV*, 2014.

[17] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011, pp. 545–552.