MULTI-AGENT DEEP COLLABORATION LEARNING FOR FACE ALIGNMENT UNDER DIFFERENT PERSPECTIVES

Congcong Zhu, Suping Wu, Zhenhua Yu, Xing Wang, Hao Liu*

School of Information Engineering, Ningxia University, Yinchuan, 750021, China

ABSTRACT

In this paper, we propose a multi-agent deep collaboration learning method (MADCL) for simultaneously detecting 2D facial landmarks and 3D facial landmarks projected from 3D to 2D, which aims at distinguishing the ambiguity caused by different perspectives. Above two facial annotations, there are a large number of public semantic areas and some very important private semantic areas. Our single agent captures and memorizes private features for iterations and multiple agents collaborate to learn public features. To achieve this, we design a collaboration learning mechanism to capture, memorize and share semantic information for enhancing the feature representation. Moreover, the input of traditional cascade regression methods is cropped directly from the raw facial image via the shape-indexed manner, which leads that the poor initial shapes likely bring about the predicted results getting worse and worse. We introduce the Markov decision process (MDP) to reason a better position of the initial shape by a reward function that reflects the shape quality. Authentic experimental results indicate that our MADCL consistently outperforms most state-of-the-art methods on two widely-evaluated challenging datasets.

Index Terms— Face alignment, reinforcement learning, deep neural network, multi-task learning, biometrics.

1. INTRODUCTION

In the last decades, there were a number of classic and effective approaches for face alignment [1–4]. For example Cootes *et al*. [3] used the appearance model to reconstruct the face and estimate the shape. However, these reconstructed approaches cannot capture facial details in complex scenes, such as large head poses and occlusion. [4–7] addressed the face alignment as a cascaded regression process, which refine the initial shape to the final shape in a coarse-to-fine manner. Tzimiropoulos *et al*. [8] employed linear regressors are not powerful enough to exploit the complex and nonlinear relationship between the face data and facial shapes. In these cases, features employed are hand-crafted, which requires strong prior knowledge by hand. After the application of CNNs in this field, works such as [1, 9–13] achieved breakthroughs of face alignment, which learn discriminative features directly from pixels. Deng *et al*. [10] proposed heat map based approach leads to the loss of the continuity and integrity of the face shape in the case of occlusion and facial mutilation.

With the large pose issues taken into consideration, 3D face fitting methods [2, 14] have been considered, which aims to fit a 3D morphable model (3DMM) [2, 14–16] to a 2D image. This model requires complex 2D mapping of point cloud data, which requires a lot of computing resources and a large amount of 3D point cloud data as strong prior knowledge. Bulat *et al*. [9] used the similarity of 2D landmarks and 3D landmarks projected on 2D images to generate 3D landmarks, which heavily relies on the quality of 2D facial landmarks.

In this paper we propose a multi-agent deep collaboration learning method (MADCL) for robust joint 2D and 3D face alignment, Fig.1 shows the advantages of our approach

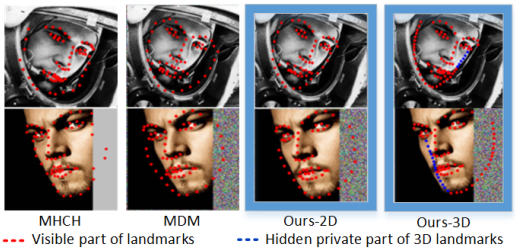

Fig. 1. Our proposed MADCL versus conventional face alignment methods. The MHCH uses the Hourglass Network based on heat map, which leads to the loss of the continuity and integrity of the facial landmarks in the case of occlusion and facial mutilation. The MDM is plagued by initialization problems and falls into local optimization in a large pose. Our approach mines more semantic information through collaboration learning and introduces initialization adjustment policy, so that agents can better reason the undamaged facial shapes and the hidden self-points. For example, the facial contour lines of 3D landmarks are shown above.

*indicates corresponding author($pswuu@nxu.edu.cn$). This work was supported in part by the National Science Foundation of China under Grant 61662059 and in part by the Research and Innovation Foundation of First-class Universities in Western China under Grant ZKZD2017005.

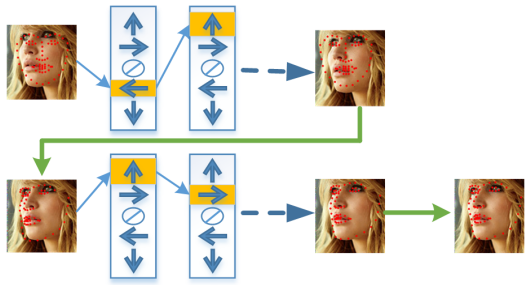

Fig. 2. Our initialization strategy. We defined five actions as the output of ActionNet: up, right, stop, left, down. The initial shape is adjusted to the better initialization position in a limited number of actions before each iteration.

compared with other approaches. We carefully design a communication mechanism to capture more context information. In addition, we solve the initialization sensitive problem of traditional cascade regression algorithm to some extent, Fig.2 illustrates our initialization strategy.

The main contributions of our work are summarized as follows:

1) We model face alignment under different perspectives as a multi-task learning framework. To achieve this, we design a cooperative learning for better interaction among multiple tasks. As result, our method learns and memorizes private features by a single agent, while multiple agents cooperate to capture and memorize public features.

2) Compared with conventional face alignment methods, we carefully design a initialization strategy based on the MDP. Following the RDN [17] our initialization strategy learns a set of actions from the reward function to adjust the initial shape of every iteration to the reasonable location for robust cascade regression process.

2. MULTI-AGENT COLLABORATION MODEL

In our method, we propose a multi-agent sharing feature information model. We introduce a learning mechanism to exchange semantic information between 2D-agent and 3D-agent, which was designed to adapt large pose and self-occlusion estimation, and it can get plausible both 2D and 3D alignment performance in unconstrained environments. In Fig.3, we illustrate the application of the MADCL for the task of joint 2D-3D face alignment.

2.1. Collaboration Learning

In the face alignment network, the motivation of this network is learning to extract appearance features mainly including information of eyes, mouth, eyebrows, nose, and outline. Hence, the 2D-agent pays more attention to the salient semantic information of the face and the 3D-agent learns to capture more context information to reason the self-occlusion points.

As can be seen from the prediction results in Fig.1, the 2D and 3D facial annotations have lots of areas with the same semantic information. This means that a large number of public features can be learned from these regions under the shape-indexed sampler conditions. Moreover, The biggest difference between two annotations lie in the labeling self-occlusion points. The 2D facial landmarks utilize the facial contour points to replace the self-occlusion points, while the 3D facial landmarks directly display these points. These different semantic areas provide the private feature.

As described in Fig.3, we introduce long short-term memory (LSTM) network [18] as a communication channel between the 2D-agent and the 3D-agent, which learns and memorizes the public features of 2D and 3D facial landmarks. With the change of shape-indexed raw patches in the iteration process, the public features of the the 2D and 3D facial landmarks are also updated. The LSTM network has a forget gate and an input gate, which captures consistent public information in time sequence more accurately. In particular, a single agent utilizes a recurrent neural network (RNN) [19] to learn and memorize the consistency information in the iterative process, and this information is private and cannot degrade other agents.

Let $P_t = [p_1, p_2, ...p_L] \in R^{2\times L}$ denote the predicted shape vector with L points at t-th iteration, where p_i represents the coordinates for i-th landmark. Moreover, let $P^* = [p_1^*, ..., p_L^*]$ denote the groundtruth.

Mathematically, our face alignment agent optimizes the following objective function:

$$\min J = \sum_{t=1}^{T} \left\| \Delta P_t - \left(P^* - P_{t-1}^I\right) \right\|_2^2, \tag{1}$$

where T, ΔP_t, P^* and P^{t-1} denote the number of iteration, the facial shape residual, the groundtruth and the adjusted results of previous iteration, respectively.

2.2. Initialization Strategy

Our architecture defines two same agents in the 2D alignment network and the 3D alignment network to interact within the MDP with a trajectory of states and actions, a state transition function, a reward function to evaluate shape quality and a policy network to seek a sequence of plausible actions.

Action: In our method, we let the adjusting movements $a \in \{up, right, stop, left, down\}$ over a continuous space as the MDP action, which means an offset to refine the initial landmarks as input of the following iterations.

State: The state is defined as the set of facial appearance features observed based on current shape, which is locally cropped directly from the raw facial image via shape-indexed manner. This appearance features serve as the input of the

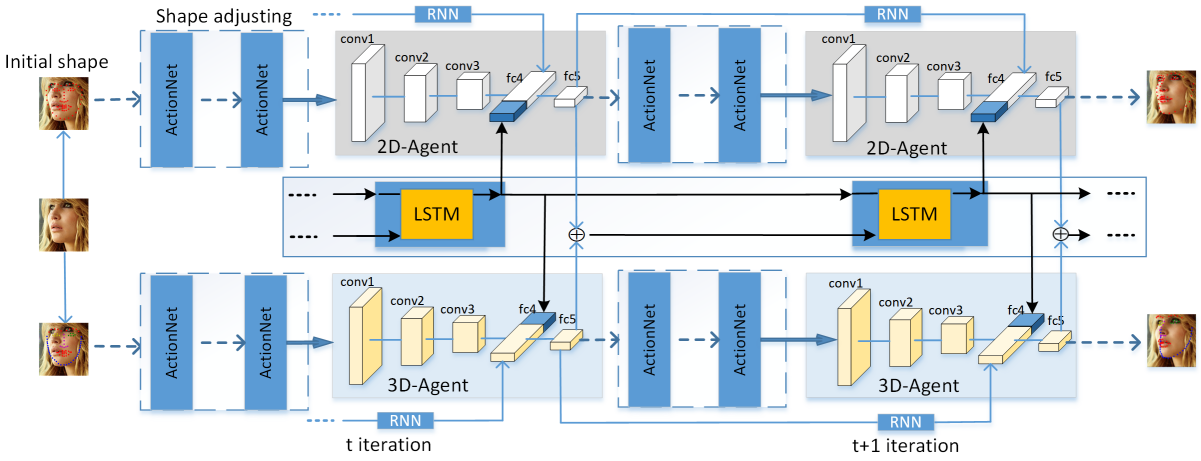

Fig. 3. Illustration of the proposed collaboration learning execution in our MADCL. Specifically, Our MACDL adjusts shape between each iteration to avoid the failure to capture meaningful semantic information caused by the poor initialization of shape. In order to mine more context information, one single agent uses an RNN to learn and memorize private features between iterations and the output of multiple agents are concatenated together as a input of the LSTM to capture and memorize public features. where $\oplus$ denotes concatenate operation and these dotted lines denote repeated operations that are omitted, respectively.

ActionNet for predicting an action to adjust the location of the initial shape.

State Transitions: We define a MDP state transition, which includes two transition processes: the shape change caused by the action and the features information change caused by the shape change. For example, at the i-th adjustment, the shape is adjusted by selected action as $\boldsymbol{p}_t^{i+1} = \boldsymbol{p}_t^i + \boldsymbol{a}^i$. Simultaneously, the observed features information change as $\boldsymbol{s}^{i+1} = o\left(\boldsymbol{I}, \boldsymbol{p}_t^{i+1}\right)$.

Reward: The reward function r^i reflects the landmark detection quality improvements. It measures the misalignment descent and is defined as follow:

$$r^i = \begin{cases} e^i - e^{i+1}, & \text{if} \quad a \in \{up, down, left, right\}, \\ +\eta, & \text{if} \quad a = stop \quad \text{and} \quad e^0 - e^i \geq 0, \\ -\eta, & \text{otherwise.} \end{cases}$$

where e and η denote the normalized point-to-point distance and the empirical value. Note that for the stop action, we use a different reward value because it leads to a terminate state.

Learning Stage: The reinforcement learning stage aims to train parameters of ActionNet. The ActionNet performs five actions and outputs the corresponding $Q\left(s, a\right)$. Based on $Q\left(s, a\right)$, the agent will choose the action that is associated with the highest reward. $Q\left(s, a\right)$ iteratively updates using the Bellman as follows:

$$Q\left(s, a\right) = r + \gamma \text{max} Q\left(s', a'\right), \tag{2}$$

where γ and $\text{max}Q\left(s', a'\right)$ denote the discount factor and the future maximum benefit.

In order to enable ActionNet to accurately predict $Q\left(s, a\right)$, we minimize following loss to update all parameters:

$$L = \mathbb{E}\left[Q\left(s^i, a^i\right) - \left(r^i + \gamma \text{max} Q\left(s^{i+1}, a^{i+1}\right)\right)\right]^2 \tag{3}$$

both agents use the same optimization process, so that each agent can capture common context information to share.

3. EXPERIMENTS

We presented the wildly used benchmarking datasets, evaluation protocols and evaluation settings. In particular, our model used 3148 images shared by 300-W [20] and 300-W part of the 3D Menpo static [21, 22]. Specifically, we used the outer-eye-corner distance as the normalizing factor and evaluated our method by using the standard normalised landmarks mean error and the cumulative errors distribution (CED) curve.

3.1. Datasets

We evaluated our MADCL approach and compared with existing state-of-the-art methods on the 300-W and the 300-W part of the 3D Menpo static face alignment datasets.

These annotated face images were collected from completely unconstrained conditions, which exhibits large variations in pose, expression, illumination, etc. We utilized the training sets of LFPW(2000), HELEN(811) and AFW(337) to train our model. Then we evaluated our method on the 224-image LFPW testing set, the 330-image HELEN testing

Table 1. Comparisons of averaged errors of our proposed MADCL with the state-of-the-arts on the 2D 300-W (68-lm).

Methods	Challenging	Common	Full
SDM [4]	15.40	5.57	8.35
ESR [5]	17.00	5.28	7.58
LBF [23]	11.98	4.95	6.32
CFSS [6]	9.98	4.73	5.76
PIFA [14]	9.88	5.43	6.30
TCDCN [24]	8.60	4.80	5.54
3DDFA [2]	9.60	4.70	5.98
R-DSSD [25]	8.60	4.80	5.54
MDM [1]	8.87	3.74	4.78
TSR [11]	7.56	4.36	4.99
SBR [26]	8.14	3.39	4.36
MADCL(w/o CM)	6.89	3.53	4.19
MADCL	**6.75**	**3.46**	**4.11**

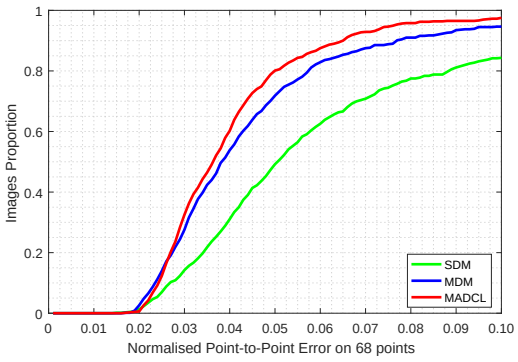

Fig. 4. CED curves of our MADCL compared to the state-of-the-arts on the 2D 300-W fullset.

set as well as the 135-image IBUG. We also investigated our approach by following another wildly used evaluation setting: using the LFPW and the HELEN testing set as the Commonset (554), the 135-image IBUG dataset as the Challengingset (135), and the union of them as the Fullset (689).

3.2. Results and Analysis

We compared our approach against the state-of-the-art methods of 2D face alignment on 300-W and 300-W part of the 3D Menpo static. Fig.1 shows the experiment results of our approach compared with the MHCH and the MDM.

2D face alignment: The results are shown in Table 1, Fig.1 and Fig.4. We see that our MADCL consistently obtains higher performance than the state-of-the-arts, and even outperforms previous methods by a large margin since our MADCL applies initialization strategy to address the local optimal problem of conventional cascade methods. Moreover, we observe that our method performs best among all of the

Table 2. Comparisons of averaged errors of our proposed MADCL with the state-of-the-arts on the 3D 300-W (84-lm).

Methods	Challenging	Common	Full
CFSS [6]	11.64	5.61	6.79
PIFA [14]	10.41	5.66	6.59
3DDFA [2]	10.20	4.63	5.72
MDM [1]	9.30	4.02	5.13
MHCH [10]	8.39	3.94	4.81
MADCL(w/o CM)	7.31	3.62	4.34
MADCL	**7.14**	**3.55**	**4.25**

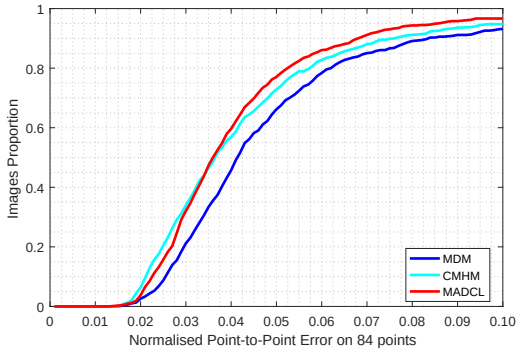

Fig. 5. CED curves of our MADCL compared to the state-of-the-arts on the 3D 300-W fullset.

state-of-the-art methods on the Challenging subset. This fully shows that our MADCL approach overcomes the problem of large poses that the conventional cascade regression methods have not solved the problem.

3D face alignment: In this subsection, we compared our MADCL against the state-of-the-art methods on 300-W part of 3D Menpo static. The results are shown in Table 2 , Fig.1 and Fig.5. The 3D landmarks have a large number of self-occlusion points under the condition of large poses. We see that our MADCL consistently obtains higher performance than the state-of-the-arts. Note that our approach significantly outperforms previous methods by a large margin on the Challenging subset, which reflects the effectiveness of handling self-occlusion point reasoning under large head rotations and extreme poses.

4. CONCLUSION

In this paper, we have proposed a multi-agent deep collaboration learning method (MADCL) for joint face alignment. We design a collaboration learning mechanism to capture, memorize and share semantic information. The experimental results have demonstrated the robust results of our approach. How to use MDP to make multiple agents better collaboration learning will be a desirable future work.

5. REFERENCES

[1] 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, pp. 4177–4187.

[2] Shi HL et al., "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.

[3] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active appearance models," *TPAMI*, vol. 23, no. 6, pp. 681–685, 2001.

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

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[6] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015, pp. 4998–5006.

[7] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *ECCV*, 2014, pp. 1–16.

[8] Georgios Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *CVPR*, 2015, pp. 3659–3667.

[9] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017, pp. 1021–1030.

[10] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou, "Cascade multi-view hourglass model for robust 3d face alignment," in *FG*, 2018, pp. 399–403.

[11] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *CVPR*, 2017, pp. 3691–3700.

[12] Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao, "Tracking via robust multi-task multi-view joint sparse representation," in *ICCV*, 2013, pp. 649–656.

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," *arXiv:1803.07835*, 2018.

[14] Amin Jourabloo and Xiaoming Liu, "Pose-invariant face alignment via cnn-based dense 3d model fitting," *IJCV*, vol. 124, no. 2, pp. 187–203, 2017.

[15] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3d morphable model," *TPAMI*, vol. 25, no. 9, pp. 1063–1074, 2003.

[16] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu, "Joint face alignment and 3d face reconstruction," in *ECCV*. Springer, 2016, pp. 545–560.

[17] Hao Liu, Jiwen Lu, Minghao Guo, Suping Wu, and Jie Zhou, "Learning reasoning-decision networks for robust face alignment," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[18] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[19] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *ISCA*, 2010.

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCVW*, 2013.

[21] Stefanos Zafeiriou, Grigorios Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis, "The 3d menpo facial landmark tracking challenge," 2018.

[22] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, Stefanos Zafeiriou, et al., "3d face morphable models "in-the-wild"," in *CVPR*, 2017.

[23] 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.

[24] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for face alignment with auxiliary attributes," *TPAMI*, vol. 38, no. 5, pp. 918–930, 2016.

[25] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou, "Learning deep sharable and structural detectors for face alignment," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1666–1678, 2017.

[26] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *CVPR*, 2018, pp. 360–368.