Survey on Face Tracking with Deep Learning

Vinay Balasubramanian[1] and Jilliam Diaz Barros[2]

[1] v_balasubr18@cs.uni-kl.de
[2] jilliam_maria.diaz_barros@dfki.de

Abstract. Face tracking is the underlying task for many applications such as face recognition, expression analysis and 3D face modeling. Detecting facial landmarks in wild(unconstrained) conditions still remains a challenging task. In this paper, we review different face tracking architectures and their performance in challenging conditions. We focus on deep-learning based methods that exploit the temporal information across frames, i.e video-based methods. Recent developments include using an encoder-decoder network, recurrent network, deep reinforcement learning, and two-stream network. This paper aims to compare those approaches in terms of accuracy, the dataset(s) used for training, evaluation metrics and robustness to large head poses and occlusions.

Keywords: Face tracking, Facial landmarks, Deep Learning, Reinforcement Learning, Temporal information

1 Introduction

Face tracking is a computer vision task of tracking the face across all frames of a video. It can be done by tracking a bounding box around the face across frames. Another way is to track specific landmarks around the face. Facial landmarks are localized around facial components such as eyes, ears, nose, mouth and jawline. Face tracking technology plays an important role in computer vision applications such as *Face recognition* [6], *Expression recognition* [3] and *Face modeling* [13]. This is a challenging problem as the videos are usually captured in unconstrained conditions. They may have illumination inconsistencies, large head poses, blurriness, occlusions, etc.

There are various approaches to this problem. Some of them are image-based methods, where the models are trained on still frames and the detection also happens independently at each frame. Other methods are video-based and use an incremental-learning technique to exploit the temporal connection between successive frames. Figure 1 shows a generic high-level architecture of a video-based landmark detection pipeline. The main idea is that the frames of the video are given in a sequential manner to the model and the temporal connection between the frames is utilized to track facial landmarks.

The rest of this paper is organized as follows: Section 2 lists publicly available datasets for face tracking. Section 3 presents some of the state-of-the-art face tracking approaches that use deep learning. In section 4 we compare these approaches. Section 5 concludes our paper.

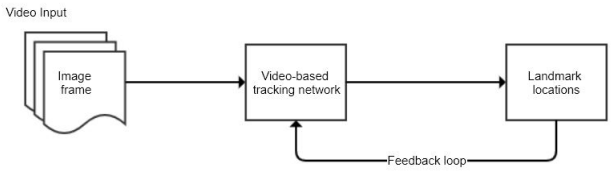

Fig. 1: Generic architecture of video based methods. Landmarks detected in the current frame are used as an initialization for the next frame

2 Datasets

In this section, we list the datasets commonly used for landmark-based face tracking. These datasets are publicly available for research purposes. Datasets can be categorized into constrained datasets and unconstrained datasets (in the wild). Table 1 shows various image-based and video datasets. Most methods use the 300-VW(300 face Videos in the Wild) [18] and the TF(Talking Faces) [1] datasets to evaluate their models and compete in the 300-VW challenge.

Table 1: Datasets used for facial landmark detection and tracking

Dataset	**Description**	**Video/ Image**	**Contains**	**Points**	**Wild**
AFLW [11]	Annotated Facial Landmarks in the Wild	Image	Around 25k annotated face images.	21	Yes
COFW [5]	Caltech Occluded Faces in the Wild	Image	1007 occluded face images.	29	Yes
HELEN [12]	HELEN facial feature dataset	Image	2000 training and 330 test images. Includes additional annotations.	194	Yes
IBUG [2]	IBUG dataset	Image	135 images with difficult poses and expressions.	68	Yes
LFPW [4]	Labeled Face Parts in the Wild	Image	1432 images.	29	Yes
LFW [10]	Labeled Faces in the Wild	Image	13,233 images of 5749 people.	10	Yes
3D Menpo [20]	The 3D Menpo database	Image & Video	Around 12k images and 280k video frames, with 2D and 3D landmarks.	84	Yes
300-W [17]	300 Faces-In-The-Wild	Image	600 images in the wild	68	Yes
FM [16]	Face Movies	Video	2150 images of 6 videos.	68	Yes
RWMB [19]	Real-World Motion Blur	Video	20 videos with motion blur.	68/98	Yes
TF [1]	Talking Face	Video	5000 frames of a person engaged in a conversation.	68	No
300VW [18]	300 videos in the wild	Video	114 videos with 218,595 frames.	68	Yes

3 Face Tracking Approaches

In this section, we describe some of the state-of-the-art approaches for video-based facial landmark tracking. Deep learning methods, in general, use CNN and RNN to detect landmarks.

3.1 Dynamic Facial Analysis using Recurrent Neural Networks (2017) [7]

This approach uses RNN for head pose estimation and facial landmark tracking. It proposes RNN as an alternative approach that performs better than previous video-based approaches for dynamic facial analysis which use Kalman filters or particle filters. The method is inspired by the fact that RNNs and Bayesian filters are operationally very similar, although Bayesian filters need problem-specific hand-tuning. Given sufficient data, an RNN can be trained to do the same task and avoid problem-specific tracker engineering. The authors create a synthetic dataset **SynHead** to cater to the need for large training data.

The approach employs FC-RNN to exploit the generalization from a pre-trained CNN and consists of CNN layers followed by recurrent layers as dense layers. Figure 2 shows the proposed architecture for head pose estimation and tracking. The CNN and RNN are trained together end-to-end. The head pose is estimated in terms of pitch, yaw and roll angles. The network is a modified VGG16 with an extra fully connected layer with 1024 neurons and the output layer consists of 3 neurons for the pitch, yaw and roll angles. For facial landmark detection, the same network is used with the only difference that the output layer contains 136 neurons corresponding to the locations of the 68 landmarks. RNN makes the model robust to occlusions and large head poses. The weights of the RNN are fixed once the training is complete, and the model cannot adapt dynamically to new data.

Fig. 2: Proposed end-to-end CNN RNN network. Source: [7]

3.2 Two Stream Transformer Networks (2017) [14]

This approach proposes a two-stream deep learning method that decomposes the video input to spatial and temporal streams. The spatial stream aims to capture appearance information from still frames and it is trained to transform image pixels to landmark positions directly on still frames and then to refine the current facial shape based on the previous shape. On the other hand, the temporal stream aims to capture temporal consistency information across successive frames.

Figure 3 shows the proposed TSTN architecture. The temporal stream consists of an encoder-decoder module. The encoder is trained to encode the spatial information as active appearance codes that capture the whole face changes across frames in the temporal dimension. The decoder remaps the learned codes to the original face input size. The temporal consistency information for each landmark is used to improve alignment accuracy. It also consists of a two-layer RNN in between the encoder-decoder module. The first layer captures spatial-temporal appearance features whereas the second layer memorizes the temporal information across frames. Facial landmarks are determined by a weighted fusion of both spatial and temporal streams. The landmark positions are refined simultaneously in both the streams. The weights for the fusion of both the streams are not learnable and has to be set manually. The authors conducted experiments for different weights(100,0), (0,100), (80,20), (20,80), (50,50) and achieved the best performance for equal weight of 50% for both the streams

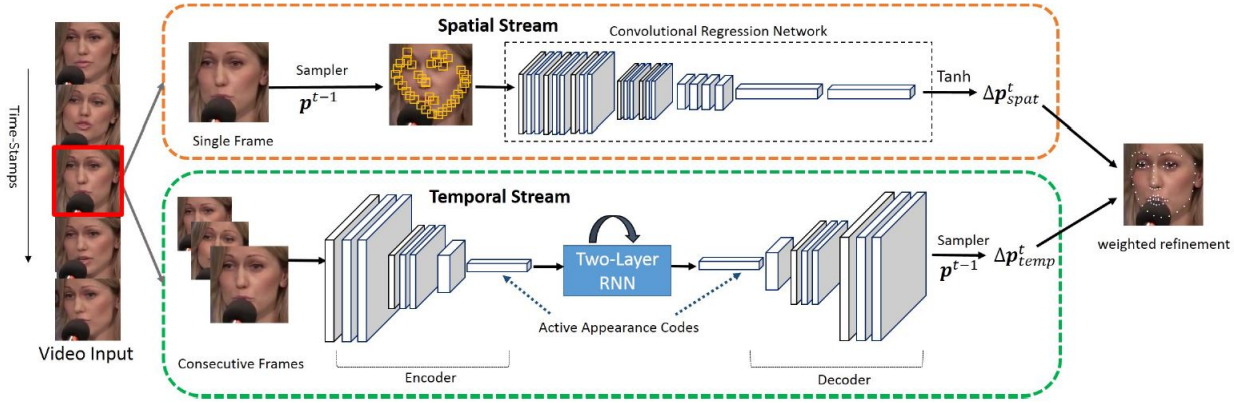

Fig. 3: TSTN pipeline. Source: [14]

3.3 Face Alignment Recurrent Network (2017) [9]

Previous state-of-the-art regression-based approaches start with an initial shape estimation and iteratively estimate the facial shape at successive stages by estimating an increment from the previous estimation. This paper proposes to improve the cascade shape regression by using LSTM and Region Convolutional Neural Network (RCNN). The LSTM model exploits both spatial and temporal information for landmark detection in images and videos in uncontrolled conditions. The predicted landmark location is used as a basis for estimation in the next stage and frame in spatial and temporal dimensions. The process continues recurrently until the face shape is finalized. The proposed model is for facial landmark detection in images under uncontrolled conditions and not for facial landmark tracking.

Figure 4 shows the training architecture of Face Alignment Recurrent Network(FARN). The face image, initial face shape, and ground truth shape are given as input to the network. The image is passed through several convolutional and max-pooling layers to obtain a feature map. The initial face shape contains facial landmarks. Region of Interest (ROI) pooling is applied around each landmark to obtain ROI pooling features. The ROI pooling features are concatenated and given to a fully connected layer followed by an LSTM layer. The network outputs the predicted shape increment over the initial face shape. The initial face shape is summed over the predicted shape increment to obtain updated initial face shape. This process continues recurrently for T stages.

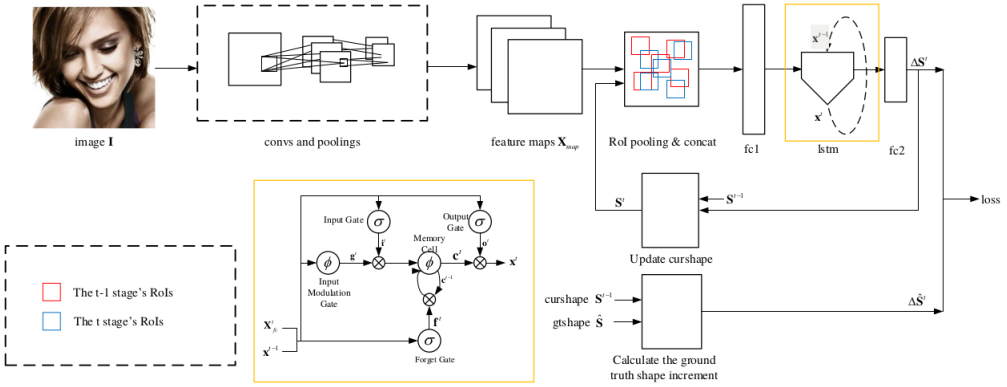

Fig. 4: FARN training architecture. Source: [9]

3.4 Recurrent Encoder-Decoder Network for Video-based Face Alignment (2018) [15]

This method leverages temporal information to predict facial landmarks in each frame and uses recurrent learning at both spatial and temporal dimensions. At the temporal level, the features are separated into *temporal-variant* features such as pose and expression, and *temporal-invariant* features such as facial identity. Recurrent learning is only applied to the temporal-variant features. This feature disentangling has shown to achieve better generalization and more accurate results. Figure 5 shows the pipeline of recurrent encoder-decoder network (REDNet).

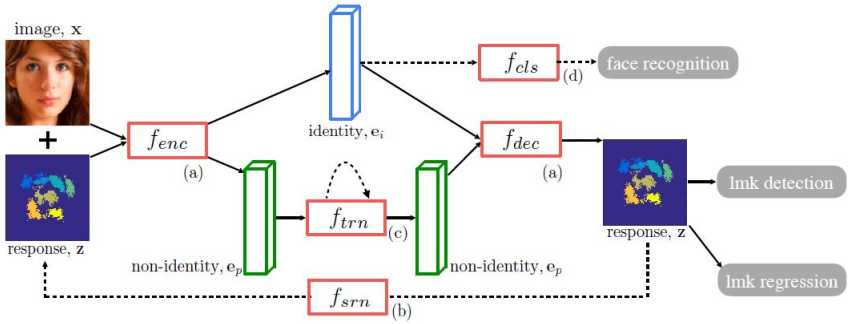

Fig. 5: Overview of REDNet pipeline. Source: [15]

The network consists of 4 modules:

(1) **Encoder-Decoder**:- The encoder encodes features from a single video frame into an intermediate low dimensional representation by performing a sequence of convolutions, pooling and batch normalization. The decoder upsamples the low dimensional representation and transforms it into a response map that contains facial landmarks.

(2) **Spatial recurrent learning**:- The purpose is to find the exact location of landmarks in a coarse-to-fine manner by iteratively providing the previous prediction as feedback along with the video frame. This is carried out in 2 steps - *Landmark Detection* and *Landmark Regression*. Landmark detection step locates 7 major facial components whereas landmark regression step refines predicted locations of all 68 landmark positions

(3) **Temporal recurrent learning**:- This is proposed to model the temporal-variant factors such as pose and expression. The temporal variations in the temporal-invariant factors (non-identity code) are modeled using a Long Short Term Memory(LSTM) unit consisting of 256 hidden neurons. Trained using T successive frames. Detection and regression tasks are performed frame by frame.

(4) **Supervised identity disentangling**:- Complete identity and non-identity factor disentangling cannot be guaranteed. More supervised information is needed to achieve better separation of the features. This module applies identity constraint to the identity code(temporal-invariant factors) to further separate identity code from the non-identity code(temporal-variant factors). Face recognition is applied to the identity code to classify the people present in the frames. This is shown to yield better generalization and better test accuracy

This model may have high computational complexity due to the spatial recurrent learning block and the usage of VGG16 for the encoder and the decoder.

3.5 Dual-Agent Deep Reinforcement Learning (2018) [8]

This approach exploits the fact that bounding box tracking and landmark detection are dependent. The accuracy of the detected facial landmarks depends on how good the bounding box is. Figure 6 shows different strategies for deformable face tracking including the proposed DADRL (Dual-Agent Deep Learning) architecture. This framework is designed for simultaneous bounding box tracking and landmark detection in an interactive manner and uses reinforcement learning to learn to make adaptive decisions during face tracking. The architecture consists of a *Tracking agent*, an *Alignment agent* and *communication channels* between the agents, as shown in Figure 7. The two agents are trained simultaneously to learn two conditional distributions. The message channels are trained using deep Q-learning algorithm.

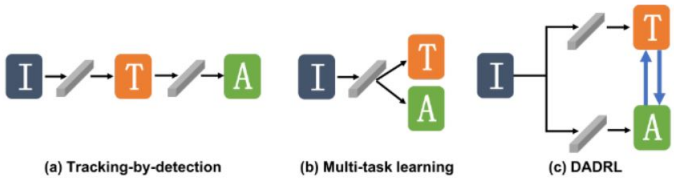

Fig. 6: Strategies for deformable face tracking. Source: [8]

If I_k is the k^{th} frame, B_k is the bounding box for the k^{th} frame and V_k is the vector of L landmarks, then the probabilistic duality is given by:

$$p\left(B_k|I_k\right)p\left(V_k|B_k,I_k\right) = p\left(V_k|I_k\right)p\left(B_k|V_k,I_k\right) \tag{1}$$

The learning objectives of bounding box tracking and landmark detection are treated as two conditional probabilities and the dependency between these two tasks is formulated as two marginal distributions. Since the ground-truth marginal distributions are not available, communication channels between the agents are used as alternatives to satisfy the probabilistic duality. For each frame, the terminal state of the previous frame is used for initializing the current state. The two agents decide a sequence of actions based on the observed state and exchanged messages, to adjust the

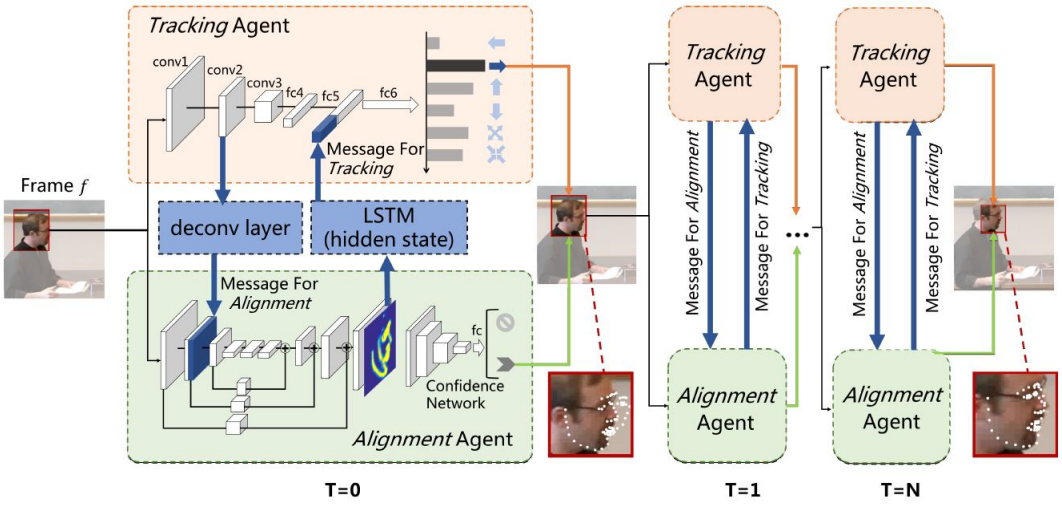

Fig. 7: DADRL architecture. Source: [8]

bounding box and regress facial landmarks simultaneously. The messages sent from the tracking agent to the alignment agent are encoded by a deconvolution layer. It provides additional textural information to the alignment agent to improve its robustness. The messages from the alignment agent to the tracking agent are encoded by an LSTM unit. It provides 3D pose information to the tracking agent to improve bounding box tracking.

4 Performance Comparison

In this section we compare the above methods starting from the datasets used for training and testing, evaluation metrics, evaluation on common dataset and robustness to challenges.

Dynamic Facial Analysis [7] is trained using the created SynHead dataset with the L2 loss function and tested on the BIWI dataset for head-pose estimation. It is then fine-tuned using training data from the BIWI dataset. For landmark detection, the corresponding model is trained and tested using a randomly split 300-VW dataset. For each frame, the mean Euclidean distance of the 68 landmarks is computed. It is normalized by the diagonal distance of the ground truth box. The metrics used for evaluation are *area under the curve*(AUC) and *failure rate*(FR). AUC is the area under the cumulative error distribution curve. FR is the percentage of images whose errors are larger than a given threshold. The proposed RNN-based architecture is more robust to large head poses and occlusions compared to per-frame estimation.

TSTN [14] is trained using the 300-VW training set. The pre-trained spatial stream network is finetuned beforehand. The model is evaluated on the testing sets of Talking Face(TF) and 300-VW datasets. Normalized Root-Mean-Square-Error and cumulative error distribution plots are used for evaluating the model. The temporal stream consisting of the encoder-decoder module and the 2-layer RNN provides consistency over time. The spatial stream which provides complementary appearance information from still frames, achieves robustness to large head poses, occlusions and variations of expressions.

FARN [9] is trained on the training partition consisting of training sets of COFW, LFPW, Helen and the entire AFW with 3148 images in total. The testing partition contains 3 parts - the common subset, the challenging subset, and the full set. The common subset consists of testing set of LFPW and Helen with 554 images in total. The challenging subset consists of the IBUG dataset which contains additional annotations for 135 images in difficult poses and expressions. The full set consists of both the common subset and the challenging subset with 689 images. The model

is evaluated using point-to-point Root-Mean-Square-Error between the face shape and the ground truth annotations. The end-to-end trained model runs extremely fast (18ms) with robustness to large head poses and occlusions.

REDNet [15] is trained on both image and video datasets with different configurations for different datasets. The training happens in 3 steps. In the first step, the network without the temporal recurrent learning and supervised identity disentangling modules is pre-trained using the image datasets AFLW, Helen and LFPW. In the second step, supervised identity disentangling is included and trained with other modules using the image-based LFW dataset. In the third step, the temporal recurrent learning module is included and the entire model is fine-tuned using the video dataset 300-VW. Inter-ocular distance is used to normalize the Root-Mean-Square-Error(RMSE). The coarse-to-fine strategy in the two-step spatial recurrent learning(landmark detection and landmark regression) makes the model robust to large head pose and partial occlusions.

DADRL [8] is trained in two stages. The **first stage** is the *supervised learning stage* in which the two agents are trained separately. All training data from 300 faces in the wild challenge (300-W image dataset) is used to train the alignment agent. The 300-VW training set is used to train the tracking agent. The communicated messages are set to zero in this stage. The **second stage** is the *reinforcement learning stage* in which the whole network is trained with the 300-VW training set. The model is evaluated on the test set of 300-VW. For evaluation, Normalized Root-Mean-Square-Error and cumulative error distribution plots are used. The communication channels between the agents has shown to provide robustness to occlusions. The authors also propose a DADRL-3D which is trained on the 3D Menpo dataset [20]. It is more robust to large pose as it is trained on 3D data.

REDNet [15], Dynamic Facial Analysis [7], TSTN [14] and FARN [9] provide testing results on challenging category of 300-VW test set for 68 landmarks. REDNet [15] and TSTN [14](7 landmarks) provide results on Talking Face dataset [1]. REDNet [15] provides results for both 68 and 7 landmarks in both datasets. Table 2 reports the RMSE of the compared methods on 300-VW and TF [1] datasets. DADRL [8] reports normalized RMSE for a few videos with heavy occlusions from the most challenging Category 3 of 300-VW. It has the lowest error among the others on the 300-VW dataset for 68 landmarks. REDNet [15] reports RMSE for Category 3 of 300-VW dataset and TF dataset for both 7 and 68 landmarks. TSTN [14] shows the best performance on the controlled TF dataset for 7 landmarks. Although FARN [9] does not have the lowest error, it performs in real time (18ms). REDNet [15] takes around 40ms to process an image. DADRL-3D is the only method that tracks 3D landmarks.

Table 2: Evaluation on 300-VW and TF test sets

	300-VW		**TF**	
Method	**RMSE(68 landmarks)**	**RMSE(7 landmarks)**	**RMSE(68 landmarks)**	**RMSE(7 landmarks)**
Dynamic Facial Analysis [7]	6.16	-	-	-
TSTN [14]	5.59	-	-	**2.13**
FARN [9]	5.90	-	-	-
REDNet [15]	5.15	**5.29**	**2.77**	2.89
DADRL [8]	**3.09**	-	-	-

5 Conclusion

In this paper, we have reviewed some of the state-of-the-art deep learning methods for video-based face alignment. All of these methods avoid hand-engineering by using neural networks. All these methods use RNN in common to model temporal information. Dynamic facial analysis [7] uses CNN followed by RNN to exploit temporal information. TSTN [14] uses spatial and temporal streams to capture appearance information on still frames and temporal information across frames. FARN [9] uses LSTM to detect facial landmarks. REDNet [15] uses an encoder-decoder network to separate

temporal-variant and temporal-invariant factors and applies recurrent learning to the temporal-variant factors. DADRL [8] exploits the dependency between bounding box generation and facial landmark tracking by treating the two tasks as agents and using deep Q-learning to maximize the accuracy in facial landmarks over several iterations. Although the recent methods have shown robustness to large head poses and occlusions, face tracking under difficult illumination is still a challenge.

References

1. Fgnet: Talking face video. *Tech. rep.*, 2004.
2. Christos Sagonas a, Epameinondas Antonakosa, Georgios Tzimiropoulosb, Stefanos Zafeirioua, and Maja Pantic. 300 faces in-the-wild challenge: database and results, 2016.
3. Jeremy Bailenson, Emmanuel (Manos) Pontikakis, Iris Mauss, James Gross, Maria Jabon, Cendri Hutcherson, Clifford Nass, and Oliver John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, 66, 2008.
4. Peter Belhumeur, David Jacobs, David Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35:2930–40, 12 2013.
5. X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *2013 IEEE International Conference on Computer Vision*, Dec 2013.
6. Paola Campadelli, Raffaella Lanzarotti, and Chiara Savazzi. A feature-based face recognition system. In *12th International Conference on Image Analysis and Processing*. IEEE, 2003.
7. Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1548–1557, 2017.
8. Minghao Guo, Jiwen Lu, and Jie Zhou. Dual-agent deep reinforcement learning for deformable face tracking. In *Proceedings of the European Conference on Computer Vision*, 2018.
9. Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, and Yihong Gong. Face alignment recurrent network. *Pattern Recognition*, 74:448–458, 2018.
10. Gary Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Tech. rep.*, 10 2008.
11. 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. pages 2144–2151, 11 2011.
12. Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.
13. Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016.
14. Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 08 2017.
15. Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. Red-net: A recurrent encoder–decoder network for video-based face alignment. *International Journal of Computer Vision*, 2018.
16. Xi Peng, Shaoting Zhang, Yu Yang, and Dimitris N Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *Proceedings of the IEEE international conference on computer vision*, 2015.
17. Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.
18. Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.
19. Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuochang Ye, and Chen Qian. Fab: A robust facial landmark detection framework for motion-blurred videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5462–5471, 2019.
20. Stefanos Zafeiriou, Grigorios G Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2503–2511, 2017.