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. In this paper, we review different face tracking architectures and their performance in challenging conditions. We focus on deep learning based methods which 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, robustness to large head poses and occlusions, etc

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 may involve tracking specific landmarks around the face, or tracking a bounding box around the face across frames. Face Tracking technology plays an important role in computer vision applications such as *Face recognition* [6], *Expression recognition* [3] and *Face modeling* [15]. This is a challenging problem as the videos may not be captured in constrained conditions and 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. We make a comparison between the different architectures, datasets used for training and testing, evaluation metrics and robustness to challenging conditions. Figure 1 shows a generic high-level architecture of a video-based landmark detection pipeline.

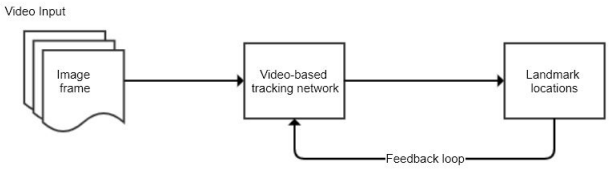


Fig. 1: Generic architecture of video based methods. Landmarks detected in the current frame are used as an initialization for the next frame

2 Datasets

In this section, we list the datasets commonly used for landmark-based face tracking. There are many publicly available datasets for face tracking. Datasets can be categorized into constrained

datasets and unconstrained datasets (in the wild). Table 1 shows various image-based and video datasets.

Table 1: Datasets

Dataset	Description	Contains	Wild?
300VW [7, 19, 21]	300 videos in the wild	114 videos with 218,595 frames with 68 landmarks per frame	Yes
AFLW [13]	Annotated Facial Landmarks in the Wild	Around 25k annotated face images with 21 landmarks per image	Yes
LFW [12]	Labeled Faces in the Wild	13,233 images of 5749 people detected and centered by Viola Jones face detector	Yes
Helen [14]	Helen facial feature dataset	2000 training and 330 test images with 194 landmarks and accurate annotations of primary facial components	Yes
LFPW [4]	Labeled Face Parts in the Wild	1432 images with 29 landmarks on each image	Yes
TF [1]	Talking Face	5000 frames of a person engaged in a conversation with 68 landmarks in each frame on each image	No
FM [18]	Face Movies	2150 images of 6 videos with 68 landmarks on each image	Yes
SynHead [9]	Synthetic dataset	510,960 frames of 70 head motion tracks that include large face pose variations	No
BIWI [8]	Biwi kinect head pose database	24 videos with over 15k frames of 20 people	Yes
COFW [5]	Caltech Occluded Faces in the Wild	1007 occluded face images with 29 manually annotated landmarks	Yes
IBUG [2]	IBUG dataset	135 images with difficult poses and expressions	Yes
RWMB [20]	Real-World Motion Blur	10000 faces with 98 landmarks including occlusion, blur, illumination changes etc.	Yes

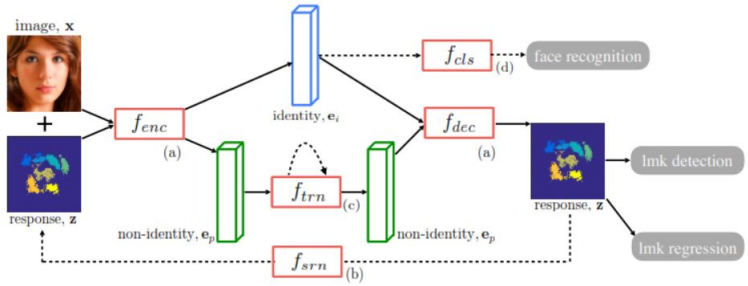
3 Face Tracking Approaches

In this section, we look at 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 Recurrent Encoder-Decoder Network for Video-based Face Alignment (2016) [17]

This method leverages temporal information to predict facial landmarks in each frame. It 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 and recurrent learning is applied to the temporal-variant features. This feature disentangling has shown to achieve better generalization and more accurate results. Figure 2 shows the pipeline of recurrent encoder-decoder network

The network consists of 4 modules -



Source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas. RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment

Fig. 2: Overview of REDNet pipeline

(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 an LSTM unit consisting of 256 hidden neurons. Trained using T successive frames. Detection and regression tasks are performed frame by frame. The prediction loss is calculated at each time step.

(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 to further separate identity code from the non-identity code. 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

3.2 Dynamic Facial Analysis using Recurrent Neural Networks (2017) [9]

This approach uses RNN for joint estimation and face 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, inspired by the fact that RNNs and Bayesian filters are operationally very similar. 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 head pose is estimated in terms of pitch, yaw and roll angles 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. It consists of CNN layers followed by recurrent layers as dense layers. RNN is more robust to occlusions and large head poses. Figure 3 shows the

proposed end-to-end network for joint estimation and tracking. The CNN and RNN are trained together end-to-end. 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.

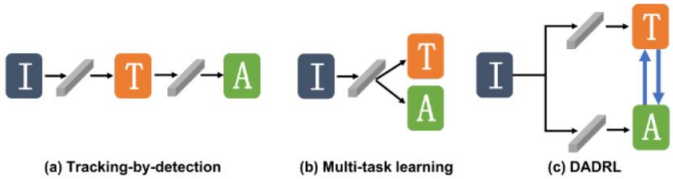
For each frame, the mean Euclidean distance of the 68 landmarks normalized by the diagonal distance of the ground truth box is computed. The metrics used for evaluation are *area under the curve* which is the area under the cumulative error distribution curve, and *failure rate* which is the percentage of images whose errors are larger than a given threshold.

Source: Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz. Dynamic Facial Analysis: From Bayesian Filters to Recurrent Neural Network

Fig. 3: Proposed end-to-end CNN RNN network

3.3 Dual-Agent Deep Reinforcement Learning (2018) [10]

This approach exploits the fact that bounding box tracking and landmark detection are dependent. The accuracy of facial landmarks detected depends on how good the bounding box is. Figure 4 shows different strategies for deformable face tracking. This paper proposes DADRL (Dual-Agent Deep Learning) framework for simultaneous bounding box tracking and landmark detection in an interactive manner. It uses reinforcement learning to learn to make adaptive decisions during face tracking. The architecture consists of a *Tracking agent* and an *Alignment agent* and *communication channels* between the agents. The two agents are trained simultaneously to learn two conditional distributions. Figure 5 shows the proposed architecture. The message channels are trained using deep Q-learning algorithm

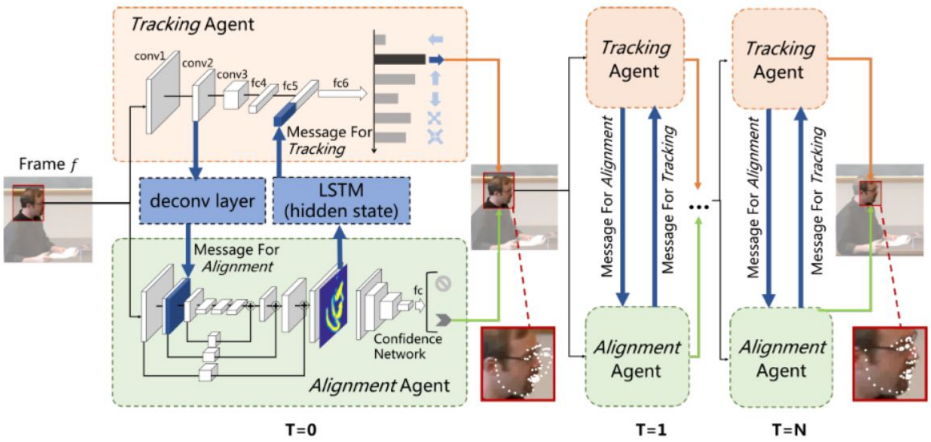


Source: Minghao Guo, Jiwen Lu, Jie Zhou. Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking

Fig. 4: Strategies for deformable face tracking

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 by probabilistic duality -

$$p\left(B_k|I_k\right)p\left(V_k|B_k,I_k\right) = p\left(V_k|I_k\right)p\left(B_k|V_k,I_k\right)$$



Source: Minghao Guo, Jiwen Lu, Jie Zhou. Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking

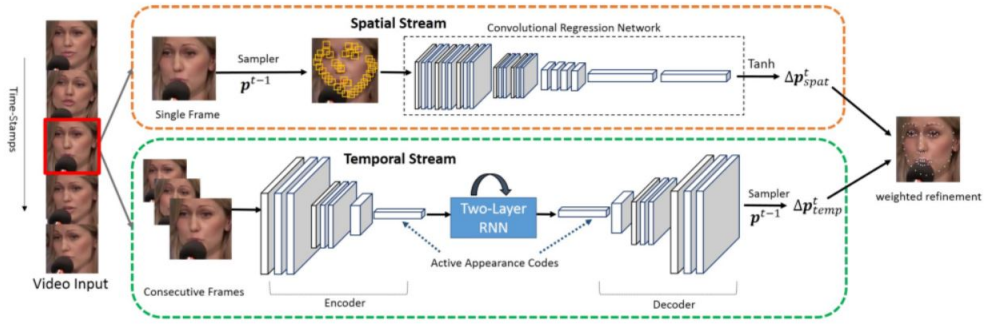
Fig. 5: DADRL architecture

The learning objectives of bounding box tracking and landmark detection are treated as two conditional probabilities and the dependency between these two tasks are 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 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.

3.4 Two Stream Transformer Networks (2017) [16]

This approach aims to capture both spatial information on still frames as well as temporal information across frames. It 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. 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. The temporal stream aims to capture temporal consistency information across successive frames. It 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. Figure 6 shows the proposed architecture. The landmark positions are refined simultaneously in both the streams.



Source: Hao Liu, Senior Member, IEEE, Jianjiang Feng, Member, IEEE, Jie Zhou, Senior Member, IEEE. Two-Stream Transformer Networks for Video-based Face Alignment

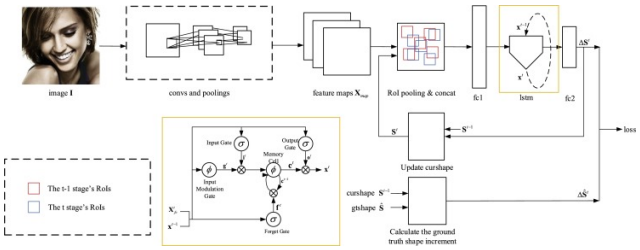
Fig. 6: TSTN pipeline

3.5 Face Alignment Recurrent Network (2017) [11]

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 (spatial) and used as a basis for estimation in the next frame (temporal). The process continues recurrently until the face shape is finalized. Figure 7 shows the training architecture of Face Alignment Recurrent Network. The face image, initial face shape, and ground truth shape is 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 in a recurrent manner for T stages.

4 Performance Comparison

[17] 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



Source: Hao Liu, Senior Member, IEEE, Jianjiang Feng, Member, IEEE, Jie Zhou, Senior Member, IEEE. Two-Stream Transformer Networks for Video-based Face Alignment

Fig. 7: FARN training architecture

distance is used to normalize the Root Mean Square Error(RMSE). [9] is trained using the created SynHead dataset with L2 loss function and tested on the BIWI dataset. 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. [10] 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. [16] 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. [11] is trained on the training partition consisting of training sets of 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.

[17], [9], [16] and [11] provide testing results on challenging category of 300-VW test set for 68 landmarks. [17] and [16] provide results on Talking Face dataset [1]. [17] provides results for both 68 and 7 landmarks in both datasets. [10] uses normalized point-to-point error for evaluation and hence cannot be compared with the other methods. Table 2 reports the RMSE of the compared methods on 300-VW and TF [1] datasets.

5 Conclusion and Discussion

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. [17] seem to perform the best on 300-VW dataset for both 7 and 68 landmarks. [16] has the least error on TF dataset for 7 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)**
REDNet	**5.15**	**5.29**	**2.77**	2.89
Dynamic Facial Analysis	6.16			
DADRL				
TSTN	5.52			**2.13**
FARN	5.49			

References

1. Fgnet: Talking face video. 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*, 66:303–317, 05 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*, pages 1513–1520, Dec 2013.
6. Paola Campadelli, Raffaella Lanzarotti, and C. Savazzi. A feature-based face recognition system. pages 68– 73, 10 2003.
7. Grigorios Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. 12 2015.
8. Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101, 02 2013.
9. Jinwei Gu, Xiaodong Yang, Shalini Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. pages 1531–1540, 07 2017.
10. Minghao Guo, Jiwen Lu, and Jie Zhou. *Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, pages 783–799. 09 2018.
11. Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, and Yihong Gong. Face alignment recurrent network. *Pattern Recognition*, 74, 09 2017.
12. 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.
13. 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.
14. Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas Huang. Interactive facial feature localization. 10 2012.
15. Feng Liu, Qijun Zhao, xiaoming Liu, and Dan Zeng. Joint face alignment and 3d face reconstruction with application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2018.
16. 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.
17. Xi Peng, Rogerio Feris, Xiaoyu Wang, and Dimitris Metaxas. A recurrent encoder-decoder network for sequential face alignment. 08 2016.
18. Xi Peng, Shaoting Zhang, Yang Yu, and Dimitris Metaxas. Piefa: Personalized incremental and ensemble face alignment. 12 2015.

19. Jie Shen, Stefanos Zafeiriou, Grigorios Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. 12 2015.
20. 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, 2019.
21. Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. pages 3659–3667, 06 2015.