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 face tracking methods and their performance in challenging conditions. This paper focuses on those methods that use deep learning and exploit the temporal information, i.e video-based methods. Recent developments include using an encoder-decoder network, recurrent network, deep reinforcement learning, two-stream network, etc This paper aims to compare various approaches in terms of accuracy, computational efficiency, 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 specific landmarks around the face across all frames in a given video. Face Tracking technology plays an important role in computer vision applications such as *Face analysis*, *Person Identification*, *Activity recognition*, *Expression analysis*, *Face modeling* etc. This is a challenging problem as the videos may not be captured in constrained conditions and may have illumination inconsistencies, large head poses, 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. Some other methods are video-based that use an incremental-learning technique by exploiting the temporal connection between successive frames. This survey focuses on video-based methods. Figure 1 shows a generic high-level architecture of a video-based landmark detection pipeline.
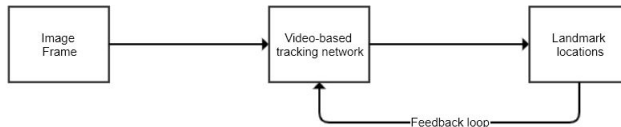

Fig. 1: Generic architecture of video based methods. Landmarks detected in the current frame are used as an initialization for the next frame

2 Datasets

The following table shows a multitude of datasets to train and evaluate face tracking models:

Table 1: Datasets

Dataset	Description	Contains	Wild?	Publicly available?
300VW	300 videos in the wild	114 videos with 218,595 frames with 68 landmarks per frame	Yes	Yes
AFLW	Annotated Facial Landmarks in the Wild	Around 25k annotated face images with 21 landmarks per image	Yes	Yes
LFW	Labeled Faces in the Wild	13,233 images of 5749 people detected and centered by Viola Jones face detector	Yes	Yes
Helen	Helen facial feature dataset	2000 training and 330 test images with accurate consistent annotations of primary facial components	Yes	Yes
LFPW	Labeled Face Parts in the Wild	1432 images with 29 landmarks on each image	Yes	Yes
TF	Talking Face	500 images of 5 videos with 68 landmarks on each image	No	Yes
FM	Face Movies	2150 images of 6 videos with 68 landmarks on each image	Yes	Yes
SynHead	Synthetic dataset	510,960 frames of 70 head motion tracks that include large face pose variations	No	Yes
BIWI	Biwi kinect head pose database	24 videos with over 15k frames of 20 people	Yes	Yes
COFW	Caltech Occluded Faces in the Wild	1007 images with 29 manually annotated landmarks on each image	Yes	Yes

3 Face Tracking Approaches

3.1 Recurrent Encoder-Decoder Network for Video-based Face Alignment (2017) [3]

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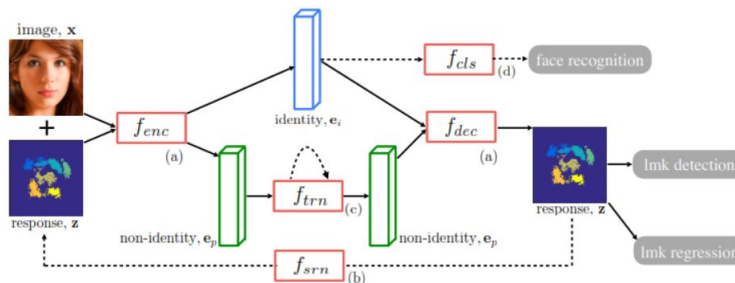
Source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas.
RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face
Alignment

Fig. 2: Overview of REDNet pipeline

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 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

This model 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 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.

3.2 Dynamic Facial Analysis using Recurrent Neural Networks (2017) [1]

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.

The model 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 randomly split 300-VW dataset.

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) [2]

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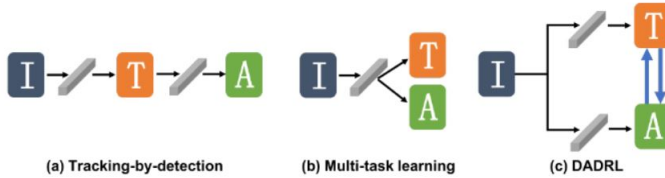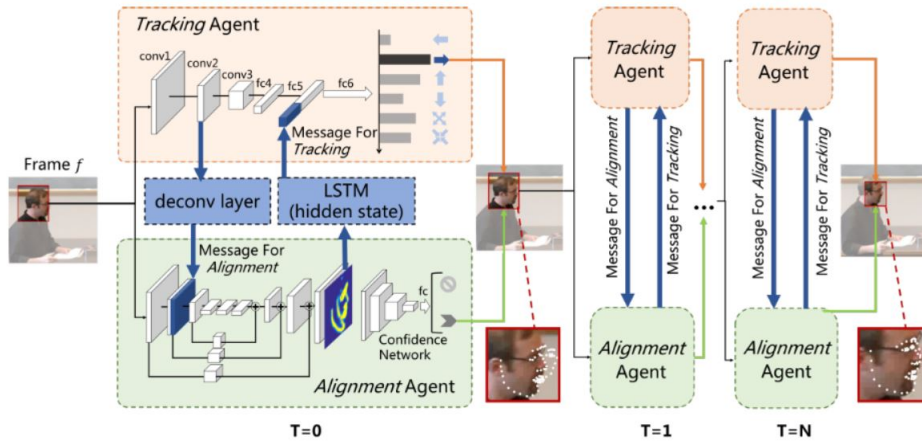
Source: Minghao Guo, Jiwen Lu, Jie Zhou. Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking

Fig. 4: Strategies for deformable face tracking

If I_k is the kth frame, B_k is the bounding box for the kth frame and V_k is the vector of L landmarks, then by probabilistic duality -

$$104 \quad p\left(B_k|I_k\right)p\left(V_k|B_k, I_k\right) = p\left(V_k|I_k\right)p\left(B_k|V_k, I_k\right)$$

105

106   The learning objectives of bounding box tracking and landmark detection are treated as two con-
107 ditional probabilities and the dependency between these two tasks are formulated as two marginal
108 distributions. Since the ground-truth marginal distributions are not available, communication chan-
109 nels between the agents are used as alternatives.

Source: Minghao Guo, Jiwen Lu, Jie Zhou. Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking
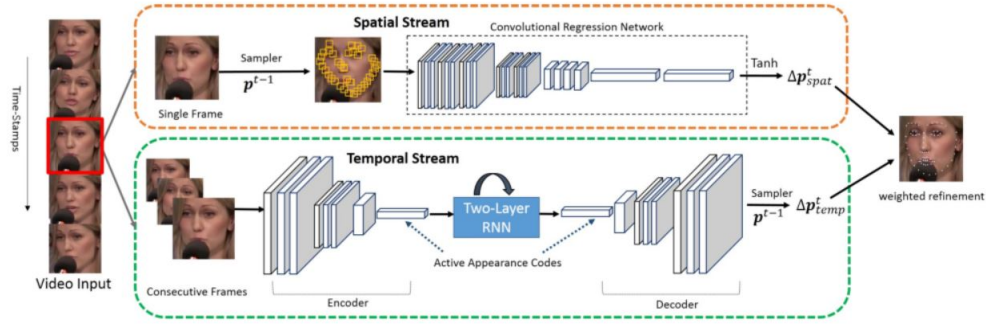
Fig. 5: DADRL architecture

110 ## 3.4   Two Stream Transformer Networks (2017)

111 This approach aims to capture both spatial as well as temporal information. It proposes a two-
112 stream deep learning method for video-based face alignment. The spatial stream aims to capture
113 information on still frames. It is trained to transform image pixels to landmark positions directly on
114 still frames. The temporal stream aims to capture temporal consistency information across successive
115 frames. It is trained to encode all facial changes in the temporal dimension. It is followed by an RNN
116 to model the sequential information over consecutive frames. Facial landmarks are determined by a
117 weighted fusion of both spatial and temporal streams. Figure 5 shows the proposed architecture.

118 ## 3.5   Face Alignment Recurrent Network (2017)

119 This recurrent regression approach for landmark detection uses the LSTM model to exploit both
120 spatial and temporal information. The predicted landmark location is used as a basis for estimation
121 in the next stage (spatial) and used as a basis for estimation in the next frame (temporal).

6

Fig. 6: TSTN pipeline

4 Performance Comparison

5 Discussion

6 Conclusion

References

1. Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. *Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network.* 2017.
2. Minghao Guo, Jiwen Lu*, and Jie Zhou. *Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking.* 2018.
3. Xi Peng, Rogerio S. Feris, Xiaoyu Wang, and Dimitris N. Metaxas. *RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment.* 2017.