# 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 technology plays an important role in Computer Vision applications such as *Face analysis*, *Person Identification*, *Activity recognition*, *Sentiment analysis* etc. There are many challenges for face tracking such as unconstrained conditions like large pose variations, illumination changes, occlusion etc. Image-based methods predict facial landmarks on still frames whereas video-based methods detect facial landmarks through a sequence of images. This paper focuses on those methods that exploit the temporal information, i.e video-based methods. Recent developments include facial tracking using encoder-decoder approach, recurrent learning, deep reinforcement learning, two stream network etc. This paper aims to compare various approaches in terms of accuracy and computational efficiency.

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

## 1 Introduction

### 1.1 Recurrent Encoder-Decoder Network for Video-based Face Alignment (2017)

This method leverages temporal information to predict facial landmarks in each frame. It uses recurrent learning at both spatial and temporal dimensions. The features are separated into *temporal-variant* features such as pose and expression, and *temporal-invariant* features such as facial identity in order to get better generalization and more accurate results. Temporal recurrent learning is applied to temporal-variant features.

The network consists of 4 modules -

(1) **Encoder-Decoder**:-

The encoder encodes features from a single video frame into a low dimensional representation. The decoder upsamples the low dimensional representation.

(2) **Spatial recurrent learning**:-

The purpose is to find the exact location of landmarks in a coarse-to-fine manner by iteratively providing previous prediction as feedback along with the video frame. This is carried out in 2 steps - Landmark detection and landmark regression

(3) **Temporal recurrent learning**:-

Trained using $T$ successive frames. Detection and regression tasks are performed frame by frame. This is to model the temporal-variant factors such as pose and expression.

(4) **Supervised identity disentangling**:-

Separating the features into temporal-variant and temporal-invariant is not completely guaranteed. More supervised information is needed to achieve better separation of the features leading to better generalization and increased accuracy.

### 1.2 Dynamic Facial Analysis using Recurrent Neural Networks (2017)

This approach improvises on previous 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. The model uses a synthetic dataset **SynHead** for training to estimate head pose. 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.

### 1.3 Dual Agent Deep Reinforcement Learning (2018)

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. The two tasks are modeled in a probabilistic manner by following a Bayesian model. The architecture consists of a *Tracking agent* and an *Alignment agent* and *communication channels* between the agents.

### 1.4 Two Stream Transformer Networks (2017)

This approach aims to capture both spatial as well as temporal information. It proposes a two-stream deep learning method for video-based face alignment. Spatial stream aims to capture information on still frames. It is trained to transform image pixels to landmark positions directly on still frames. Temporal stream aims to capture temporal consistency information across successive frames. It is trained to encode all facial changes in the temporal dimension. It is followed by a RNN to model the sequential information over consecutive frames. Facial landmarks are determined by a weighted fusion of both spatial and temporal streams.

### 1.5 Face Alignment Recurrent Network (2017)

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

## 2 My approach

[**?**] test citations

## 3 Experiments

## 4 Conclusion