INFORMATIK

augmented VISION DFKI

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

2D Image Processing & Augmented Reality
3rd Semester
Survey on Face Tracking with Deep Learning

Vinay Balasubramanian

v_balasubr18@cs.uni-kl.de

Supervisor: Jilliam Diaz Barros

Outline

- Face tracking is a computer vision task that involves tracking a specific number of landmarks on the face detected across all frames of a video.
- Applications include Face analysis, Person identification, Activity recognition, Sentiment analysis, Face modeling etc.
- It is a challenging problem as the videos are captured in unconstrained conditions which may include illumination variations, large head poses, occlusions.

Outline

- There are two basic approaches – Image-based and Video-based.
- Image-based methods operate only on still frames.
- Video-based methods make use of temporal information to predict facial landmarks in each frame.

Recurrent Encoder-Decoder Network for Video-based Face Alignment

- **Temporal-variant features such as pose and expression are separated from Temportal-invariant features such as facial identity.**
- **Employs recurrent learning at both spatial and temporal dimensions.**
- **The network consists of 4 modules - Encoder-Decoder, Spatial recurrent learning, Temporal recurrent learning and Supervised identity disentangling.**

Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network

- Improvises on previous approaches for dynamic facial analysis that use Kalman/Particle filters.
- CNN layers followed by recurrent layers as dense layers.
- Uses FC-RNN to exploit generalization from a pre-trained CNN.

Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking

- **Exploits the fact that bounding box tracking and landmark detection tasks are dependent. The accuracy of the latter depends on how good the former is.**
- **The two tasks are modeled in a probabilistic manner by following a Bayesian model.**
- **The architecture consists of a *Tracking agent, Alignment agent* and *Communication channels* between the agents.**

Two-stream Transformer Networks for Video-based Face Alignment

- Two stream deep learning method to capture spatial as well as temporal information.
- Spatial stream captures information on still frames.
- Temporal stream captures temporal consistency information across successive frames. It is followed by a RNN to model the sequential information over consecutive frames.
- Finally, facial landmarks are determined by a weighted fusion of spatial and temporal streams.

Face Alignment Recurrent Neural Network

- Recurrent regression based approach
- Uses LSTM to exploit both spatial and temporal information.
- Spatial – The predicted landmark location is used as basis for estimation in the next stage.
- Temporal – The predicted landmark location is used as basis for estimation in the next frame.

Comparison metrics

- **Following are some of the metrics used to compare the above methods in this survey :–**
  - Dataset used for training (In the wild vs constrained)
  - Evaluation metrics
  - Number of landmarks tracked
  - Kind of landmarks retrieved (2D or 3D)
  - Robustness to large pose variations, illumination changes

augmented VISION DFKI

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Thank You