INFORMATIK

2D Image Processing & Augmented Reality
Winter Semester 2019/2020
Survey on Face Tracking with Deep Learning

Vinay Balasubramanian

v_balasubr18@cs.uni-kl.de

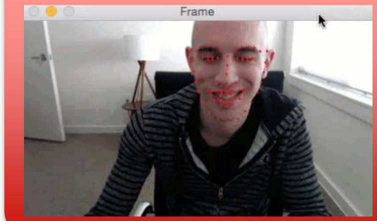
Supervisor: Jilliam Diaz Barros

Outline

- Face Tracking – Quick recap
- Datasets
- Approaches
- Comparison
- Conclusion and Questions

Quick recap

Face Tracking



Tracking a face across all frames of a video.

Can be bounding-box based tracking or landmark based (track specific number of keypoints around facial components)

Applications - Face analysis, Person identification, Activity recognition, Expression analysis, Face modeling etc.

Challenges - Videos can be captured in unconstrained conditions.

May have illumination variations, large head poses, occlusions, etc.

Image source: Google

Datasets

- AFLW
- COFW
- Helen
- IBUG
- LFPW
- LFW
- 3D Menpo
- 300-W
- BIWI
- FM
- RWMB
- SynHead
- TF
- 300-VW

Method

- Video – Sequence of frames with temporal connection
- Sequence data? (Use RNN)

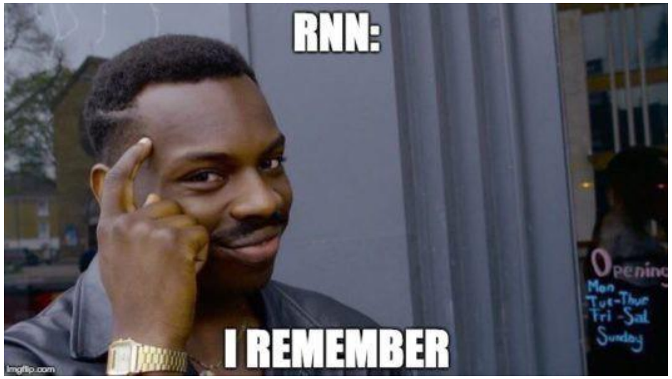
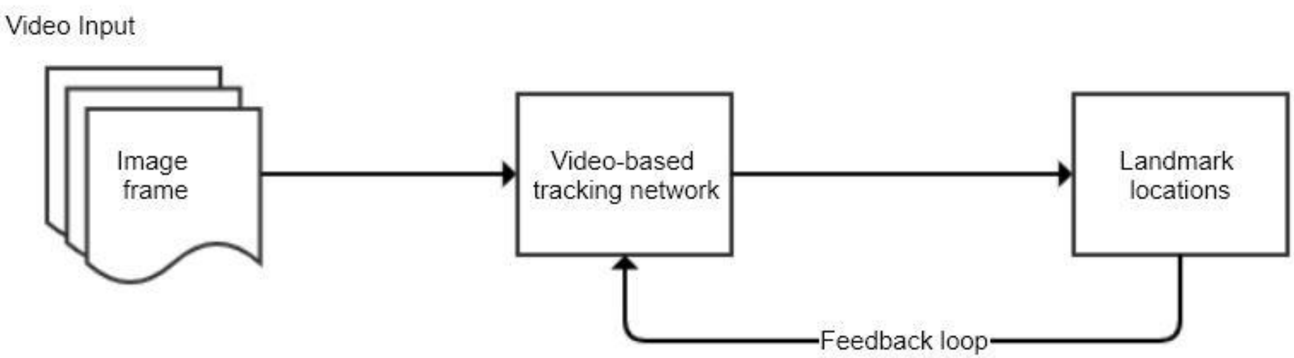


Image source: https://towardsdatascience.com/pytorch-basics-how-to-train-your-neural-net-intro-to-rnn-cb6ebc594677

Method

- Video – Sequence of frames with temporal connection
- Sequence data? (Use RNN)
- Give frames in temporal order, detect landmarks, feedback along with next frame.

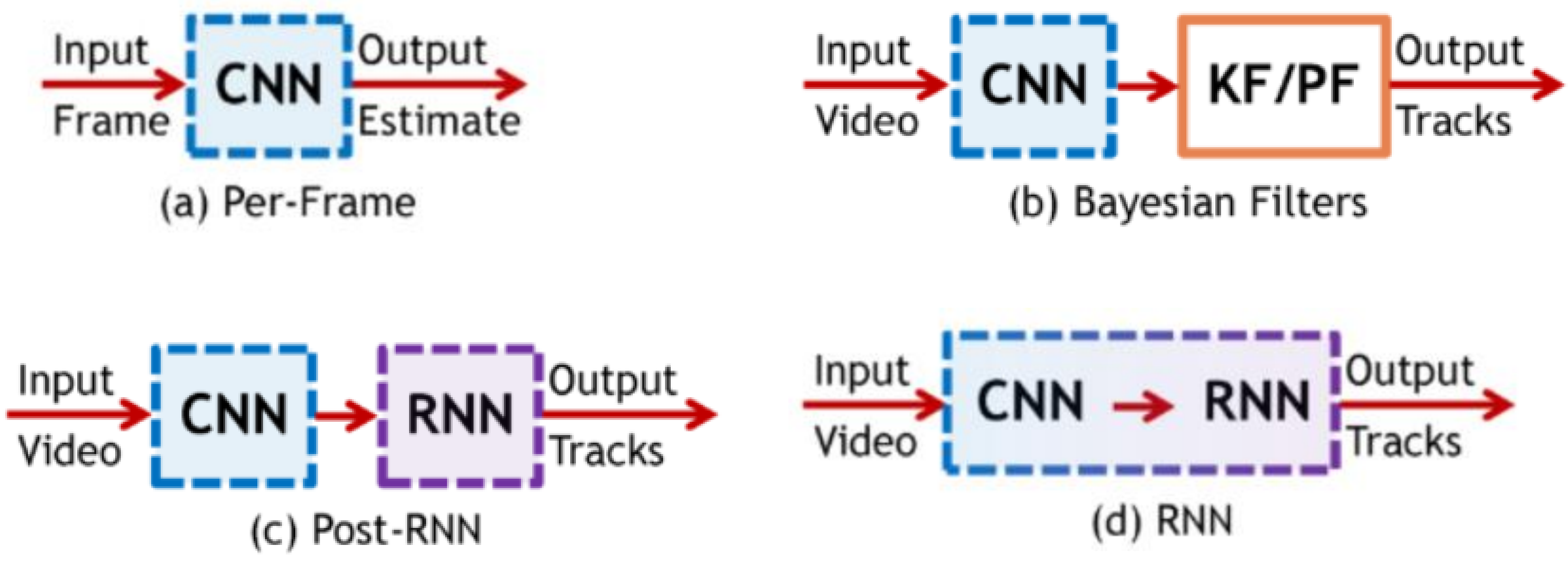


Why Deep Learning?

- State-of-the-art in image processing tasks
- Operate directly on data
- Learn more generic features directly from data
- Computational efficiency
- Domain knowledge is never obsolete

Using RNNs

- **RNNs and Bayesian filters are operationally similar**



(a) Per-Frame

(b) Bayesian Filters

(c) Post-RNN

(d) RNN

Image source: Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

Using RNNs

(d) RNN

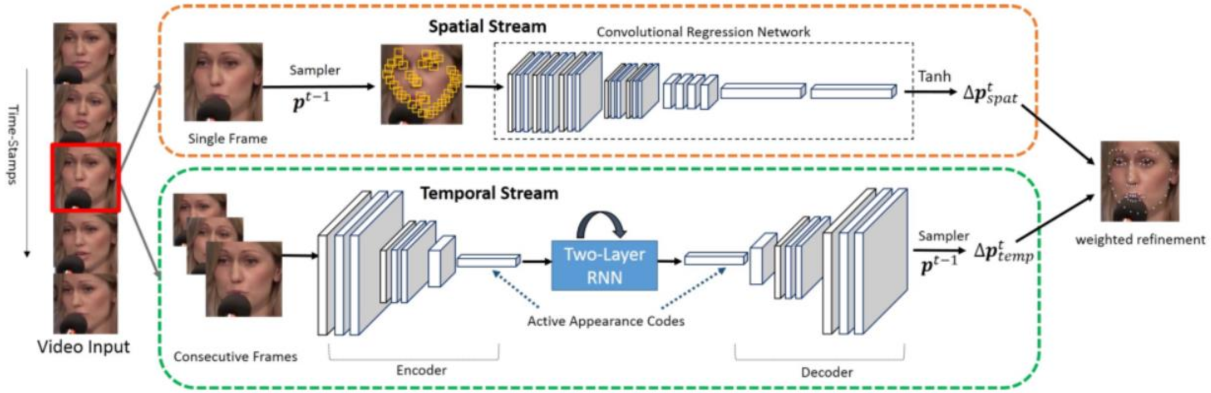
- FC-RNN is used to retain generalization of pretrained CNN
- Trained end-to-end
- SynHead dataset for head pose estimation
- 300-VW dataset for facial landmark localization
- L2 loss function
- Evaluation - Area Under the Curve (AUC), Failure Rate (FR %)

Image source: Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

Two-stream network

- Exploit both appearance information from still frames(spatial) and temporal information across frames(temporal)
- Spatial stream – Image pixels (still) -> landmark locations
- Temporal stream – Compress video as active appearance codes(whole face changes across frames)

Two-stream network



- **Spatial stream transforms local facial patches to shape residuals used to refine current face shape from previous.**
- **Temporal stream – Encoder-decoder network with 2-layer RNN. Capture facial dynamics in temporal dimension**
- **Final prediction is a weighted fusion of spatial and temporal streams shape updates**

Image source: Hao Liu, Jiwen Lu, Jianjiang Feng, Jie Zhou: Two-Stream Transformer Networks for Video-based Face Alignment(2017).

Two-stream network

- Tested on 300-VW and TF datasets
- Evaluation – Normalized RMSE and Cumulative Error Distribution plots
- Weighted fusion – β_1 and β_2 = 0.5 yields the best performance

LSTMs

- **LSTM is used to exploit spatial and temporal information**

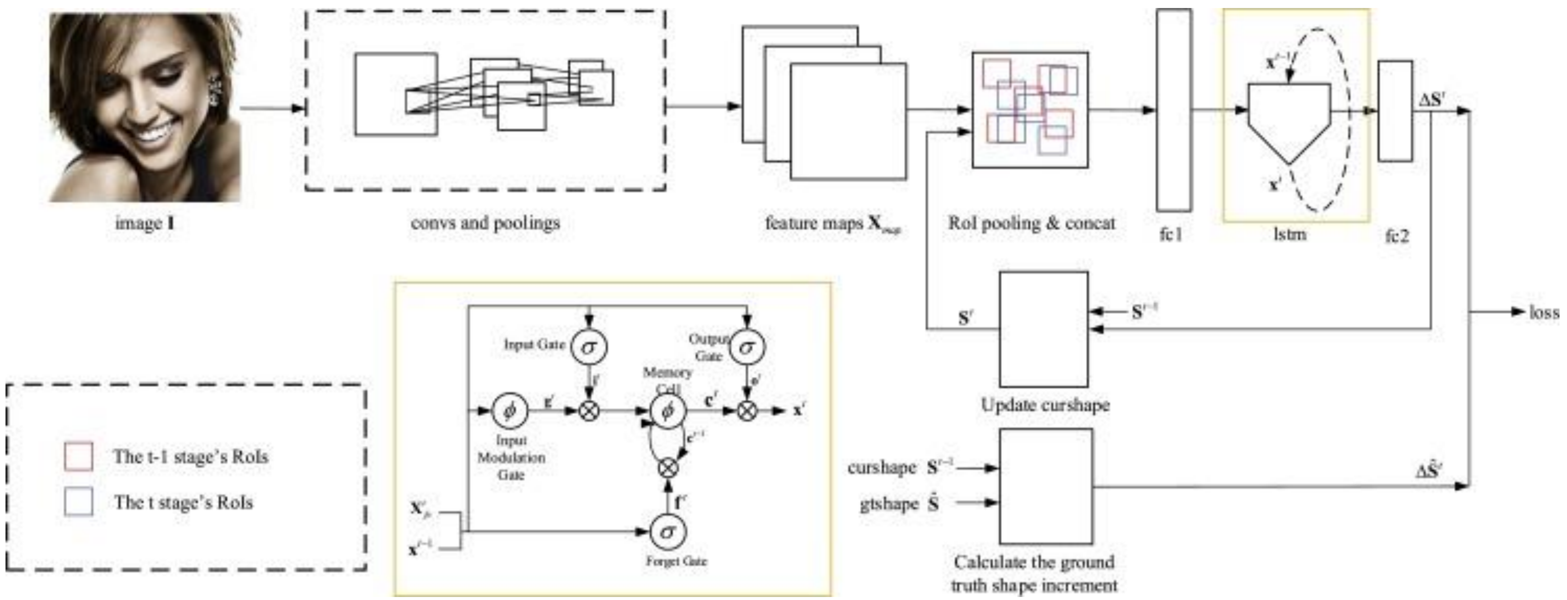
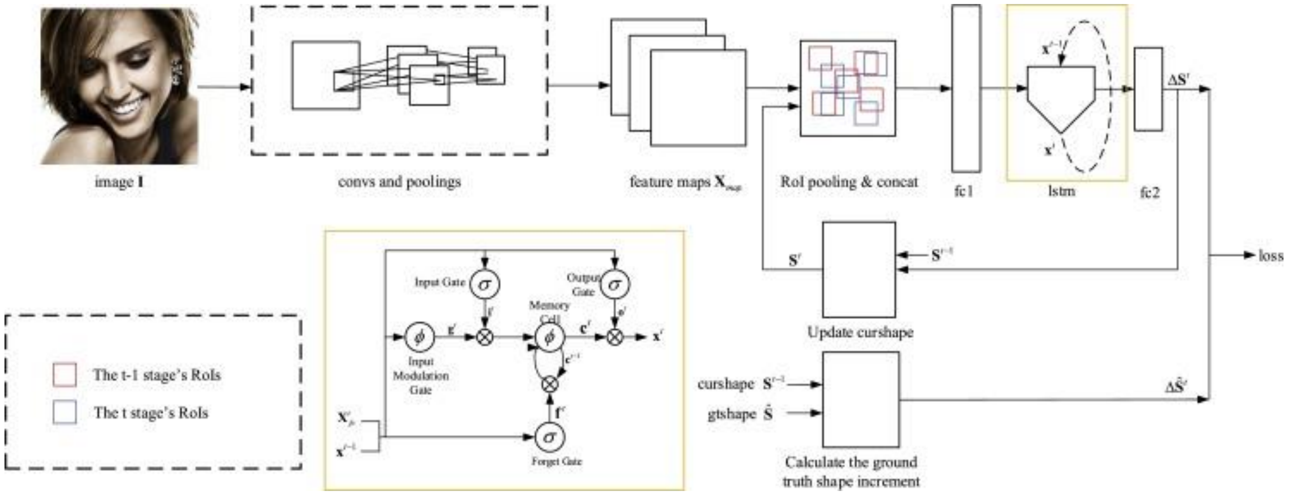


Image source: Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

LSTMs



- **Input – Image and Initial face shape**
- **Output – Predicted shape increment for the initial face shape**
- **Input -> several conv + max pooling -> ROI pooling for initial face shape -> concat -> FC layer -> LSTM -> predicted shape increment**
- **Update initial shape according to predicted shape increment**

Image source: Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

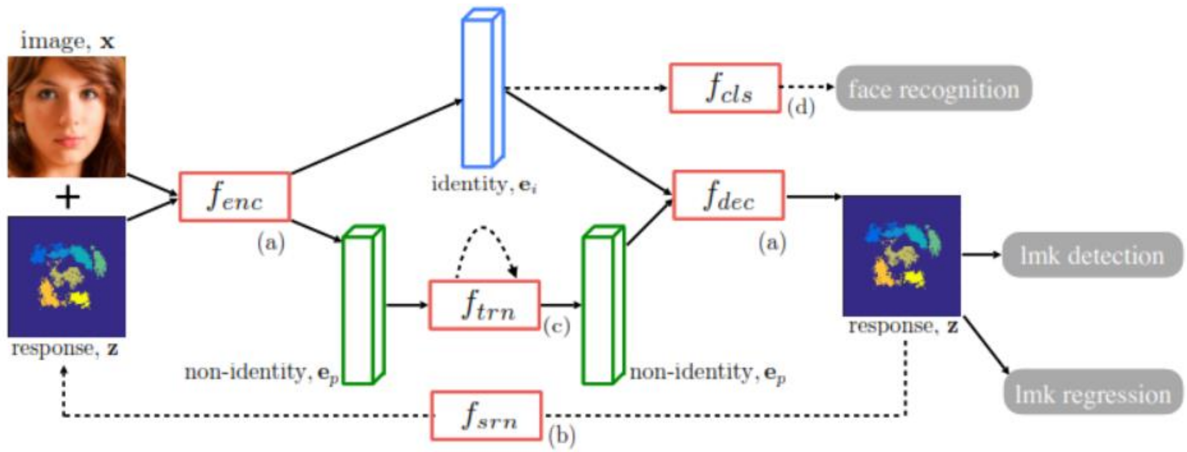
LSTMs

- **Landmark detection method**
- **Trained on COFW, LFPW, Helen, AFW**
- **Evaluation – Point-to-point RMSE**
- **Runtime – 18ms**

Encoder-Decoder Network

- Encoder – Image pixels -> Low dimensional feature space
- Decoder – Features in low dimensional space -> facial landmark heatmaps
- Feedback loop between the output(facial points) and the input

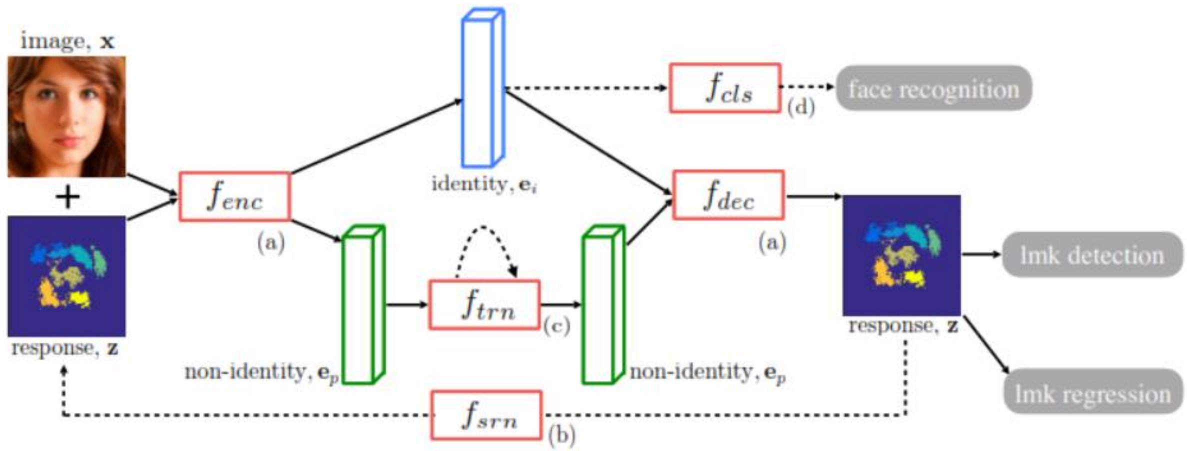
Encoder-Decoder Network



- Encoder -> Feature disentangling on low dimensional representation -> Decoder
- Disentangle temporal-variant and temporal-invariant factors
- Temporal-invariant: Person identity
- Temporal-variant: Pose, expression, illumination

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2017).

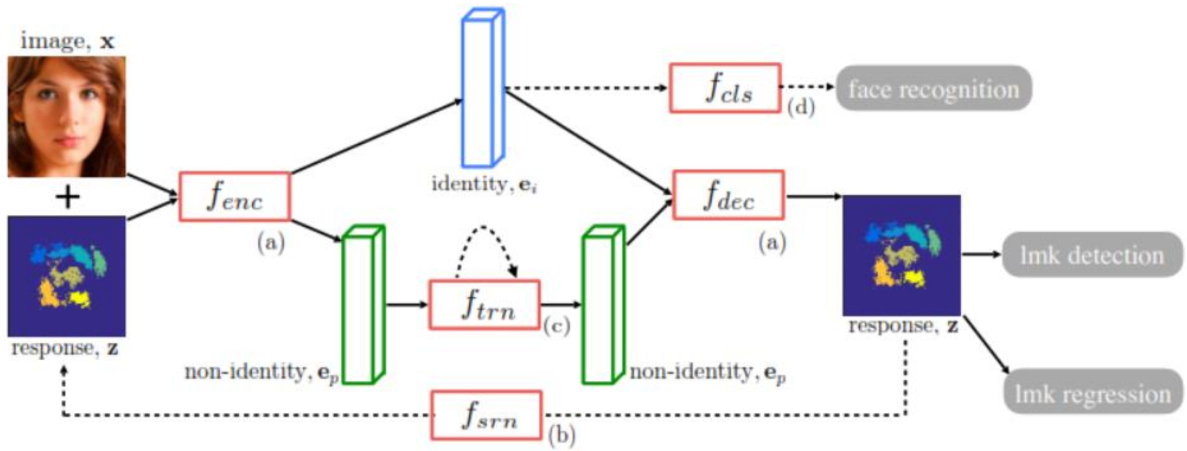
Encoder-Decoder Network



- **Spatial recurrent learning: Coarse-to-fine landmark search**
- **Feedback loop: Previous prediction + image**
- **Landmark detection: Detect major landmarks**
- **Landmark regression: Refine predicted locations from previous detection step**

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2017).

Encoder-Decoder Network



- **Temporal recurrent learning: Model non-identity factors (temporal-variant) using LSTM**

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2017).

Encoder-Decoder Network

- Evaluated on – AFLW, LFW, Helen, LFPW, TF, FM, 300-VW
- Evaluation – Inter-ocular distance normalized RMSE

Reinforcement Learning

- **Bounding box generation and facial landmark detection are heavily dependent**
- **Two agents – bounding box generation and landmark detection**

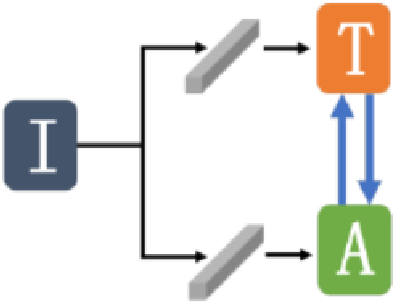
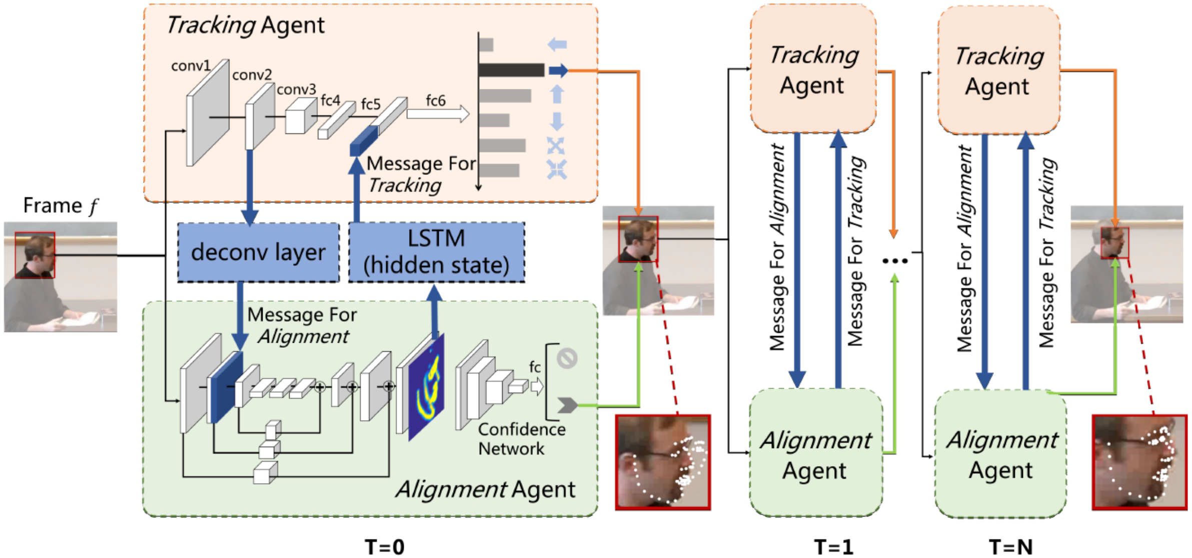


Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

Reinforcement Learning



- **Communication channels between agents (Deep Q-learning)**
- **Current state initialized to terminal state of previous frame**
- **Agents decide sequence of actions to adjust bounding box and regress landmarks simultaneously based on observed state and received messages**

Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

Reinforcement Learning

- **Go to next frame when landmarks are finalized**
- **State – Current image region extracted by bounding box**
- **Action – Tracking agent(Movement), Alignment agent(stop/continue iterations)**

- **Reward – Landmark detection accuracy improvements**

Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

Reinforcement Learning

- Evaluated on Category 3 of 300-VW
- Supervised learning stage -
  - Alignment agent trained on 300-W
  - Tracking agent trained on 300-VW
- Reinforcement learning stage –
  - Whole network trained on 300-VW
- Evaluation – Normalized RMSE and cumulative error distribution curves
- DADRL-3D – Trained with 3D data from 3D Menpo. More robust to large pose

Comparison

Approach	Evaluated on dataset	Evaluation metrics
Dynamic Facial Analysis using RNN	300-VW	AUC, FR
Two stream transformer network	TF, 300-VW	RMSE, CED plot
Face alignment recurrent network	COFW, LFPW, IBUG, 300-VW	RMSE
Recurrent encoder-decoder network	TF, 300-VW	RMSE
Dual agent deep reinforcement learning	300-VW	RMSE and CED plot

Comparison

Approach	300-VW		TF	
	RMSE(68 landmarks)	**RMSE(7 landmarks)**	**RMSE(68 landmarks)**	**RMSE(7 landmarks)**
Dynamic Facial Analysis using RNN	6.16			
Two stream transformer network	5.59			**2.13**
Face alignment recurrent network	5.9			
Recurrent encoder-decoder network	5.15	**5.29**	**2.77**	2.89
Dual agent deep reinforcement learning	**3.09**			

References

1. Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

2. Hao Liu, Jiwen Lu, Jianjiang Feng, Jie Zhou: Two-Stream Transformer Networks for Video-based Face Alignment(2017).

3. Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

4. Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2017).

5. Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

THANK YOU

QUESTIONS?