SPATIOTEMPORAL ATTENTION BASED DEEP NEURAL NETWORKS FOR EMOTION RECOGNITION

Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn
School of Electrical and Electronic Engineering, Yonsei University, Korea
{easy00, kso428, srkim89, khsohn}@yonsei.ac.kr

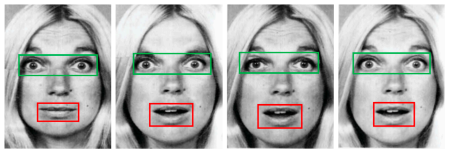
Introduction

Goal

- This paper describes recognizing dimensional emotion based *spatiotemporal attention method* via convolutional neural networks.

Motivations

- Cost aggregation step is required to regularize matching costs from neighboring pixels with an explicit kernel function.
- Most existing cost aggregation methods optimally design the edge-aware weights in a hand-crafted manner.
- Example, (4-type of surprise)



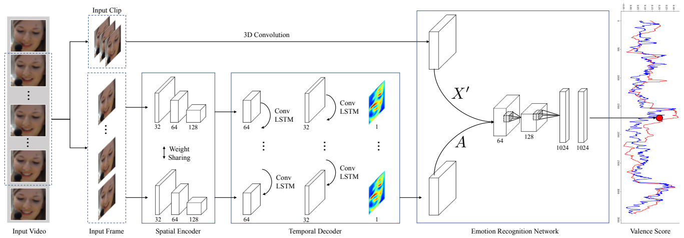
- Categorical emotion cannot cover full range of emotion
- mouth (red box) and eye (green box), which are emotional sailent parts within facial videos is essential for recognizing emotion robustly.

The key aspect of the proposed method

1) Extracting the features of each frame with spatial associations using 2D-CNNs
2) Estimating spatiotemporal attention of the video using Convolutional LSTM (ConvLSTM)
3) The dimensional emotions of each frame are estimated by leveraging 3D-CNNs to encode both appearance and motion information simultaneously

Proposed Method

Network Architecture



Spatiotemporal Attention Network

→ ***Spatial Encoder***

- To take spatial correlation into consideration, we propose the feature encoder of 2D-CNNs

→ ***Temporal Decoder***

- Utilizing ConvLSTM modules that encode the temporal correlation across inter-frames while preserving the spatial structure over sequences.

→ ***Spatiotemporal Attention Inference***

- Soft attention manner : attention is multiplied to 3D convolutional feature activations.

$$X'' = A \odot X$$

Emotion Recognition Network

- Estimate a dimensional emotion for the facial video by leveraging the spatiotemporal attention
- Employ 3D-CNNs to deal with temporal information, which simultaneously consider spatial and temporal correlations across the attention-boosted features X'' and directly regress the emotion.

Experimental Results

Quantative Results

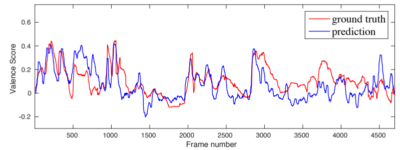
- **Component-wise analysis**

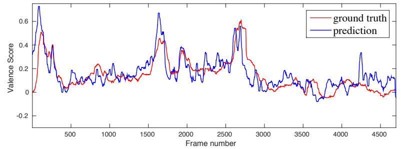
2D-CNN	3D-CNN	STA	RMSE	CC	CCC
✓			0.113	0.426	0.326
	✓		0.104	0.510	0.493
	✓	✓	0.102	0.572	0.546

- **Evaluation on RECOLA**

Method	RMSE	CC	CCC
Baseline [26]	0.117	0.358	0.273
CNN [1]	0.113	0.426	0.326
CNN + RNN (≈ 1 sec.) [1]	0.111	0.501	0.474
CNN + RNN (≈ 4 sec.) [1]	0.108	0.544	0.506
LGBP-TOP + LSTM [29]	0.114	0.430	0.354
LGBP-TOP + Bi-Dir. LSTM [15]	0.105	0.501	0.346
LGBP-TOP + LSTM + ϵ-loss [30]	0.121	0.488	0.463
CNN + LSTM + ϵ-loss [30]	0.116	0.561	0.538
3D-CNN + STA (≈ 4 sec.)	**0.102**	**0.572**	**0.546**

- **Estimated graph on RECOLA**

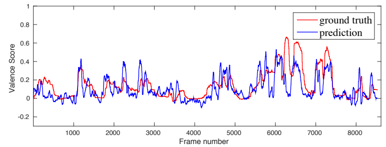


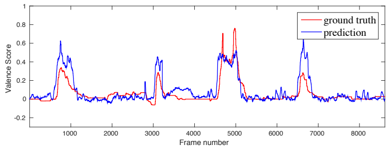


- **Evaluation on AV+EC`17**

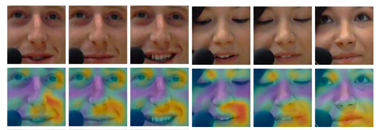
Method	RMSE	CC	CCC
Baseline [31]	-	-	0.400
CNN [1]	0.114	0.564	0.528
CNN + RNN (≈ 4 sec.) [1]	0.104	0.616	0.588
3D-CNN + STA (≈ 4 sec.)	**0.099**	**0.638**	**0.612**

- **Estimated graph on AV+EC`17**





Visualization of attention



Conclusion

- Propose dimensional emotion recognition framework that lever- ages the spatiotemporal attention of video frames.
- Consider only spatial appearance and temporal motion for the facial video sequence simultaneously using 3D-CNNs.