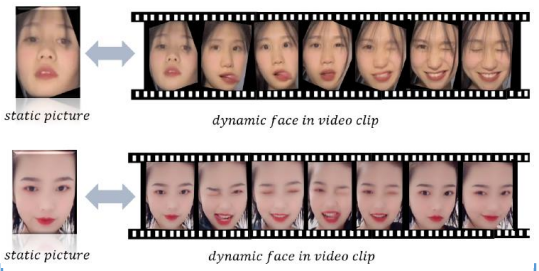
Two-Stream Temporal Convolutional Network for Dynamic Facial Attractiveness Prediction

Nina Weng[1], Jiahao Wang[2], Annan Li[2*], Yunhong Wang[2]
Shenyuan Honors College of Beihang University, Beijing, China[1]
School of Computer Science and Engineering, Beihang University, Beijing, China[2]
wengnn@buaa.edu.cn

Introduction

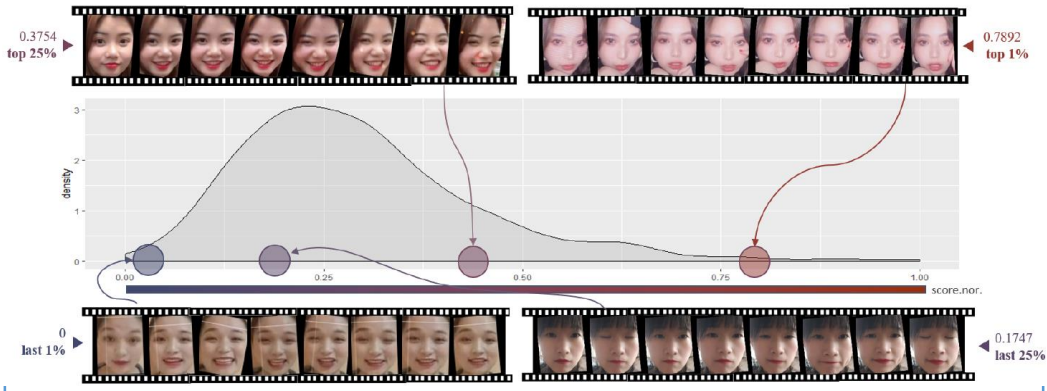
In the field of facial attractiveness prediction, while deep models using static pictures have shown promising results, little attention is paid to dynamic facial information. Meanwhile, the increasing popularity of short video apps creates an enormous demand for facial attractiveness prediction from short video clips. In this work, we target on the dynamic facial attractiveness prediction problem. A large-scale video-based facial attractiveness prediction dataset (VFAP) with more than one thousand clips from TikTok is collected. We propose a two-stream temporal convolutional network (2S-TCN) to capture dynamic attractiveness features from both facial appearance and landmarks. We employ attentive feature enhancement along with specially designed modality and temporal fusion strategies to better explore the temporal dynamics. Extensive experiments on the proposed VFAP dataset demonstrate the superiority of 2S-TCN.

Motivation



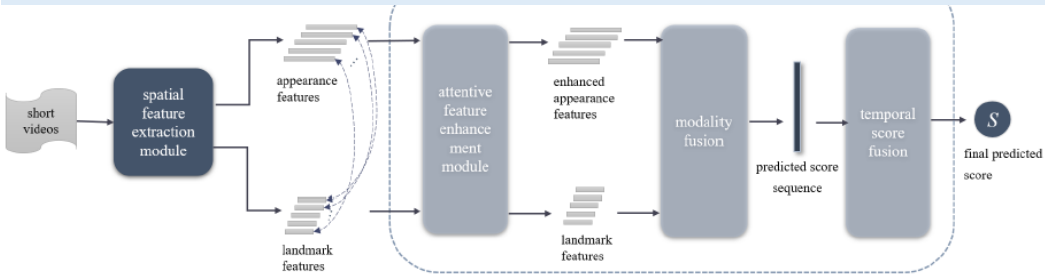
The continuous face sequence or video is naturally more competent in expressing facial attractiveness compare with static pictures. On one hand, psychological and neuroscience studies have shown that temporal cues play an important role in the perception of human face. Dynamic face sequences are better at conveying emotions and expressions. On the other hand, the increasing popularity of short video apps creates an enormous demand for FAP from short video clips.

Video-based Facial Attractiveness Prediction Dataset



All videos in VFAP are collected from TikTok. After carefully manual selection, we build the VFAP dataset with 1,430 short videos. The dataset is divided into four subsets according to the different TikTok channels. The average frame number of each video is about 314, and the whole dataset contains over 449k facial frames. Our attractiveness score is generated according to the interactive behaviors in social media, including pressing the like button, leaving a comment and forwarding the video to others. The ranking of facial attractiveness can be observed in the figure above, where four examples sampled from the top 1%, 25% and last 1% are presented.

Two-Stream Temporal Convolutional Network



As shown above, we first extract frame-level facial appearance and landmark features in the spatial feature extraction module. The proposed 2S-TCN model then performs temporal modeling on these spatial features. Within 2S-TCN, the appearance features are augmented by the attentive feature enhancement module. Modality and temporal score fusions are successively performed to get the predicted attractiveness score.

Experimental Results

Method	S0	S1	S2	S3	All
AlexNet	0.08920	0.01559	0.03243	0.13180	0.01388
Resnet-18	0.11078	0.00787	0.03016	0.12721	-0.00240
ResNeXt-50	0.12934	0.06878	0.02485	0.12898	-0.00388
2S-TCN	**0.38621**	**0.26273**	**0.32138**	**0.38699**	**0.18965**

In order to validate the significance of temporal information in this task, we compare our proposed dynamic framework with three state-of-the-art static FAP methods (i.e. AlexNet, Resnet-18, ResNeXt-50), which are selected from the top performers of the SCUT-FBP5500. It can be observed from the results that the static methods do not predict well on the dynamic facial data. In comparison, our 2S-TCN model has much higher performance in both single-subset and all-subset experiments, demonstrating the importance of temporal modeling. We also conduct extensive ablation studies in order to validate the design of our 2S-TCN.