Human Behaviours-based Automatic Depression Analysis using Hand-crafted Statistics and Deep Learned Spectral Features

Siyang Song[1], Enrique Sanchez[2] , Linlin Shen[3], Michel Valstar[1]
1. School of Computer Science, The University of Nottingham, UK
2. Samsung AI Centre, Cambridge, UK
3. College of Computer Science and Software Engineering, Shenzhen University, China

Background

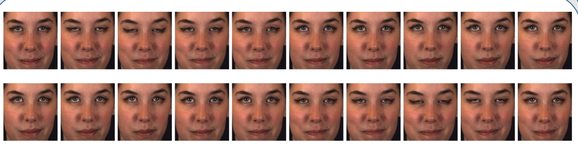
Background: 1. Temporal information is crucial to human facial behaviour understanding; 2. Facial actions are continuous and smooth process; 3. The same facial actions of different people are similar.

Research Gap: In some scenarios, only a still image is available, resulting in the performance of state-of-the-art methods for facial expression recognition or affect estimation degrades substantially.

Motivation: the goal is to propose an approach that can infer generic facial temporal information from a single face image.

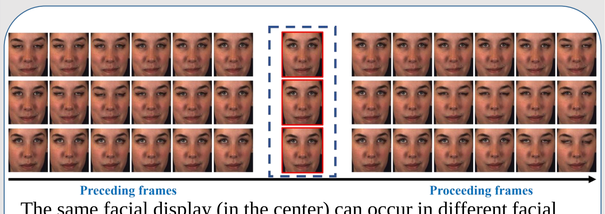
Symmetric and ambiguous of facial actions

Symmetric pattern



The lower facial action is the temporally reversed counterpart of the upper one, where both sequence can be a plausible facial action.

Ambiguous pattern



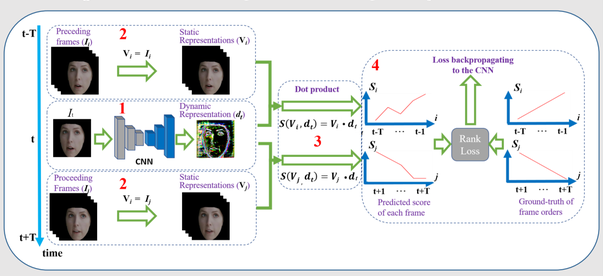
The same facial display (in the center) can occur in different facial actions.

While we could directly use dynamic image/optical flow/MHI as target representations for the proposed dynamic learning task, we observe that facial actions display a **symmetric** and **ambiguous** temporal pattern, that could lead to weak representations, as the ``predictive" task. In other words, having very different descriptors for similar inputs is known to make the learning process harder.

Methodology

Target

The goal is to generate a $d_t = f(I_t)$, where I_t is a still face image and d_t is the required dynamic representation that can summarize the dynamics around the I_t. In particular, we propose a DR that is targeted with ranking not only the preceding frames, but also the proceeding frames. In other words, the DR is chosen to be a kernel that can rank both past and future frames, based on their temporal positions relative to the given input face image. This way, the modeling of symmetric and ambiguous facial action patterns can be partially addressed.



Self-supervised training

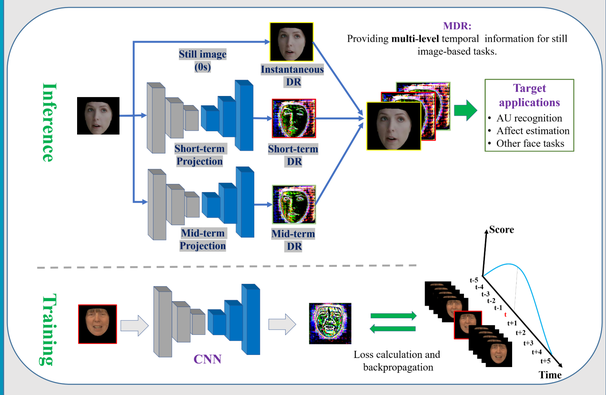
During training, we are given a set of sequences from which we can have access to the adjacent frames of a given image I_t. In (1), the given image I_t is forwarded to the network we aim to learn, that produces a DR d_t. We can measure the ranking capabilities of d_t by projecting it onto the preceding and proceeding frames (2). To rank the frames, we compute the difference between pair-wise scores, each computed as a dot product between the generated DR and the corresponding preceding or proceeding frame (3). These scores are used to compute a Rank Loss, which allow us to measure the extent of which the current d_t is correctly ranking the frames within the sequence. We can backpropagate the Rank Loss w.r.t. the parameters of the network that has produced the DR d_t (4). This way, the network not only learns to produce a correct representation d_t, but also contributes to define it.

Conclusion

We empirically validated the capacity of the DRs to rank unseen frames in test time, as well as their contribution to the face-related tasks. We illustrated that the generated DRs can be used indistinctly for the tasks of Action Unit intensity and dimensional affect estimation, attaining state of the art results. In addition, we validated that a network trained with a rank loss function generalizes better to unseen images than a model trained using pre-defined representations

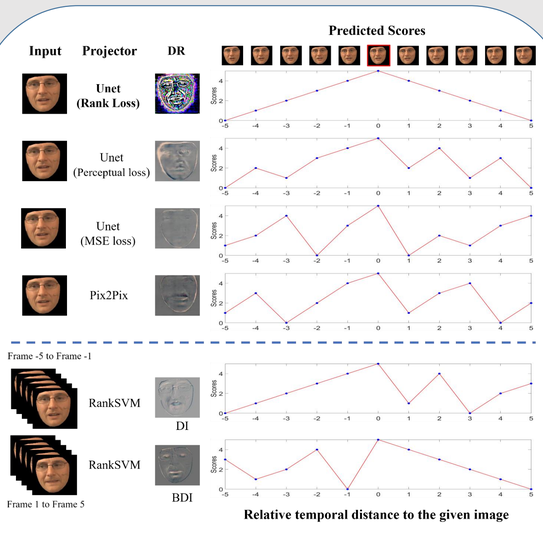
Applications

We propose a novel approach to the modeling of temporal face dynamics from still images (top). Our approach can be used to infer several time-length dynamics, and further combined to enhance face-related tasks such as AU intensity estimation and dimensional affect estimation. During training (bottom), a set of videos is used to learn the DRs, without explicitly generating a fixed set of target representations.



Specifically, we generate a multi-scale set of DRs, each capturing a different temporal scale by using a unique window length. This combination allows the proposed approach reaching state-of-the-art results in the tasks of AU intensity estimation and dimensional affect estimation. We first note that the generated d_t are 3-channel tensors, no matter the choice of T. Thus, we can train a different model for different values of T, and combine the outputs before applying them to further related tasks. Herein, we will explore the use of a **Single Dynamic Representation** (**SDR**), using just the generated DR, and the use of a **Multi-level Dynamic Representation** (**MDR**), which combines the output of networks trained using different time lengths T.

Results



Examples of ranking frames using DRs generated by different methods. The networks of Pix2Pix, Unet(P) and Unet(MSE) were trained using dynamic facial image [2] as the target. Meanwhile, the RankSVM uses at test time the adjacent frames to compute the kernel.

	AU	6	10	12	14	17	Avg.
ICC	CCNN-IT [44]	0.75	0.69	0.86	0.40	0.45	0.63
	2DC [38]	0.76	0.71	0.85	0.45	**0.53**	0.66
	VGP-AE [45]	0.75	0.66	**0.88**	0.47	0.49	0.65
	HG-HMR [37]	**0.79**	0.80	0.86	0.54	0.43	0.68
	Pix2Pix* [17]	0.59	0.62	0.68	0.29	0.31	0.50
	Unet(P)* [43]	0.55	0.65	0.65	0.30	0.26	0.48
	Unet(MSE)*	0.56	0.63	0.66	0.29	0.26	0.48
	SDR+HG-HMR	0.78	0.80	0.85	0.47	0.45	0.67
	MDR+HG-HMR	0.77	**0.83**	0.87	**0.62**	0.49	**0.72**
MSE	CCNN-IT [44]	1.23	1.69	0.98	2.72	1.17	1.57
	2DC [38]	0.75	1.02	0.66	1.44	0.88	0.95
	VGP-AE [45]	0.82	1.28	0.70	1.43	0.77	1.00
	HG-HMR* [37]	**0.77**	0.92	0.65	1.57	0.77	0.94
	Pix2Pix* [17]	1.22	1.31	0.85	1.90	0.92	1.24
	Unet(P)* [43]	1.53	1.08	1.07	1.62	0.95	1.25
	Unet(MSE)*	1.09	1.55	1.18	2.12	1.15	1.42
	SDR+HG-HMR	0.88	0.84	0.75	1.90	0.60	0.99
	MDR+HG-HMR	0.99	**0.79**	**0.64**	**1.34**	**0.48**	**0.85**

TABLE I: AU intensities estimation results on BP4D dataset. ∗ denotes results obtained by our own implementation

[1]. Bilen, Hakan, et al. "Dynamic image networks for action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
[2]. Song, Siyang, et al. "Dynamic Facial Models for Video-based Dimensional Affect Estimation." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.
[3]. Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.