Facial Expression Recognition in the presence of Speech

Ali N. Salman
Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
ans180000@utdallas.edu

I. INTRODUCTION

Emotion recognition plays an important role in everyday human-to-human interaction. Facial expression recognition (FER) systems aim to accurately perceive the displayed emotion. Achieving human level perception of emotions can dramatically change human–computer interaction. While FER systems can accurately perceive posed expressions in images, it is still very challenging to classify emotions in videos, where speech articulations are introduced (see fig. 1). Objectives of the research: (1) study and compare human-performance in perceiving emotions in the static and dynamic representations, highlighting important differences. (2) Propose approaches to compensate for facial articulations introduced by speech and increase the accuracy of video-only FER systems.

II. DYNAMIC VERSUS STATIC FACIAL EXPRESSION IN THE PRESENCE OF SPEECH

Firstly, we conduct an evaluation of whether the emotional perception of several isolated frames in a video is a good representation of the emotional perception of the entire video. Our hypothesis is that emotions observed from isolated frames provide a poor representation of the emotions in a video. These differences can be explained, up to some extend, due to the presence of speech (Fig. 1). For this analysis, we are relying on the MSP-IMPROV corpus, a multimodal emotional database. The key features of this corpus is that a portion of it has been annotated under different conditions: (1) audiovisual presentation, (2) audio-only presentations, and (3) video only presentation [1]. Each video is annotated with happiness, anger, sadness, neutral or other. Also, a Likert scale of 1-7 was used for valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). For this analysis, we compare the video only presentation with annotations of isolated frames, extracted at 3 frames per second. For the isolated frames, each frame is annotated by 5 people, with the same 5 classes and Likert scale for valence, arousal, and dominance as the video only representation.

The study compares the emotional perception of isolated frames (static representation) and the emotional perception of video segments (dynamic representation). We create five sets in this analysis. The first two sets correspond to annotations provided by evaluators for the video-only condition. The third set consists of the annotations of the isolated frames, which are extracted from the same videos. The fourth set corresponds to the results of a facial expression recognition model created in our laboratory. The fifth set corresponds to randomly selected emotional classes (i.e., random choice). For categorical classes, each of these sets creates a five-dimensional distribution for happiness, sadness, anger, neutral state and other. We describes these sets in more details in this section.

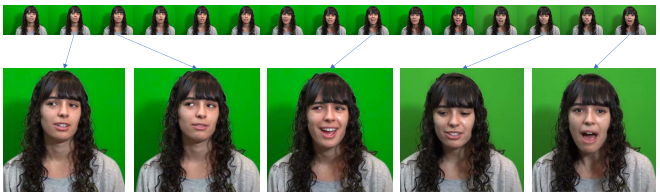

Fig. 1. Static representation of emotion in frames extracted from a video. Individual frames do not represent well the emotion of the video, especially in the presence of speech.

A. The GROUND Set

We randomly selected five annotations from each video. The annotations from the remaining videos are used to estimate the ground truth labels. Since all the videos are annotated by at least 10 independent evaluators, each video has from 5 to 17 annotations. We use this set to estimate the ground truth label after removing the aforementioned evaluations. For each video, we normalize by the number of evaluators to obtain a distribution.

B. The REFERENCE Set

The second set is used as reference. It corresponds to the annotations obtained from the five evaluations per video that were originally removed to estimate the ground truth label for the video-only condition (Sec. II-A). This set is used to compare the ground truth labels with labels provided to the same videos by independent annotators (e.g., inter-evaluator agreement). We also normalized the annotations to obtain a distribution.

C. The FRAME Set

This set corresponds to the annotations provided by raters to isolated images. We extract frames from the corresponding videos at a rate of three frames per seconds. In total, we have 4,723 frames, which are annotated with emotional labels with perceptual evaluations conducted on crowdsourcing using an identical approach used to annotate the video-only data. Since the frame-by-frame approach to process video often ignores the relationship between frames, we shuffle the presentation of the frames in the evaluation, removing temporal information. Each frame is annotated by five evaluators.

TABLE I

USING THE GROUND LABELS, THE TABLE LISTS THE F1-SCORE ACHIEVED WITH THE CONSENSUS LABEL DERIVED FROM THE OTHER FOUR SETS. EACH VIDEO IS REPRESENTED BY THE AVERAGE OF ITS AGGREGATED ANNOTATIONS.

Label	Set	Precision	Recall	F1-score
Happiness	REFERENCE	0.91	0.84	0.87
	FRAME	0.67	0.97	0.79
	FER	0.78	0.77	0.78
	RANDOM	0.29	0.16	0.16
Anger	REFERENCE	0.73	0.67	0.70
	FRAME	0.55	0.14	0.22
	FER	0.50	0.05	0.08
	RANDOM	0.16	0.16	0.16
Sadness	REFERENCE	0.77	0.79	0.78
	FRAME	0.66	0.57	0.61
	FER	0.40	0.79	0.53
	RANDOM	0.21	0.11	0.14
Neutral	REFERENCE	0.72	0.72	0.72
	FRAME	0.54	0.77	0.63
	FER	0.55	0.59	0.57
	RANDOM	0.29	0.16	0.20
Average	REFERENCE	0.78	0.76	0.77
	FRAME	0.61	0.61	0.56
	FER	0.56	0.55	0.49
	RANDOM	0.24	0.15	0.17

We add all the evaluations assigned to one frame and normalize their value to obtain the emotional distribution of the frame. Then, we add all the evaluations assigned to one video. We obtained the distribution of a video after normalizing by the total number of frames.

D. The FER Set

In the analysis, we also want to compare the emotional content obtained by processing the isolated frames using an automatic FER system. For this purpose, we trained a FER system to recognize the emotional classes happiness, sadness, anger and neutral state from static images.

The classifier is trained with images from a separate dataset (AffectNet corpus [2]). The corpus contains images of faces in the wild, which have been annotated with categorical classes and emotional attributes (arousal, valence and dominance). We use 20% of the training set as a validation set, using the development set suggested for this corpus to test our classifier. The architecture of the classifier relies on the VGG-Face model proposed by Parkhi et al. [3] for face recognition. We used the weights of the *Convolutional Neural Network* (CNN) in the VGG-Face model as the initial weights of our model to predict the emotions. We added three fully connected layers with 512, 512, and 256 nodes, respectively. Then, we add a softmax output layer. During training, only the fully connected layers were trained, freezing the parameters of the VGG-Face model. Finally, we under-sample the training data to achieve a uniform distribution across emotional classes.

We first analyze the categorical emotions. We aggregate all the annotations over the whole video and consider the highest frequency as the label. Using the GROUND set as our ground truth we calculate the F1-score (see table I). Overall the average F1-score while comparing GROUND and REFERENCE sets is 0.77. The F1-score for FRAME and FER systems are 0.57 and 0.49, respectively. There is a decrease of 27% for human annotations and 36% for the FER system compared to video-only annotation. When looking at the average of each individual emotion, we notice that the FER system is always behind the human annotation (FRAME). Additionally, the comparisons between static and dynamic differ drastically between each emotional class. For example, the accuracy between REFERENCE, FRAME, and FER for happiness is relatively close, compared to other emotions. Additionally, the accuracy of the anger emotion is poor for FRAME and FER achieving only 0.22 and 0.08 F1-scores, respectively, compared to 0.70 for REFERENCE. This comparison shows that some emotions, such as happiness, can provide enough cues in the static representation to be correctly classified. It also shows that some emotions, such as anger, rely greatly on the temporal cues perceived in the dynamic representation.

TABLE II

EUCLIDIAN DISTANCE BETWEEN LABELS IN THE VALENCE, AROUSAL AND DOMINANCE SPACE. THE COMPARISON INCLUDES THE GROUND, REFERENCE, FRAME AND RANDOM SETS. THE LABELS IN EACH SET ARE AGGREGATED AT THE VIDEO LEVEL.

L2 norm	Dimension	GROUND	REFERENCE	FRAME	RANDOM
GROUND	Valence	0.00	0.56	1.17	1.72
	Dominance	0.00	0.77	2.26	2.22
	Arousal	0.00	0.74	1.83	1.97
REFERENCE	Valence	0.56	0.00	1.20	1.74
	Dominance	0.77	0.00	2.33	2.29
	Arousal	0.74	0.00	1.88	2.00
FRAME	Valence	1.17	1.20	0.00	1.12
	Dominance	2.26	2.33	0.00	1.01
	Arousal	1.83	1.88	0.00	0.97
RANDOM	Valence	1.72	1.74	1.12	0.00
	Dominance	2.22	2.29	1.01	0.00
	Arousal	1.97	2.00	0.97	0.00

TABLE III

PERFORMANCE OF THE STATIC FER SYSTEM IN THE FEATURE EXTRACTOR MODEL. THE APPROACH IS IMPLEMENTED WITH THE VGG16 NETWORK USING THE AFFECTNET CORPUS.

Emotion	Precision [%]	Recall [%]	F1-score [%]
Happiness	89.8	91.0	90.5
Anger	76.7	71.2	73.9
Sadness	75.8	71.6	73.7
Neutral	63.7	70.1	67.0
Average	76.5	76.2	76.3

Another method for comparing FER is the use of VAD space (i.e., valence, arousal, and dominance). In this analysis, we compare all the sets, with the exception of the labels of the FER set, since the FER system was built to recognize categorical emotions. The results are shown in Table II. Once again, GROUND and REFERENCE are the sets with the smallest distances. The labels from the FRAME set are closer to the labels of the RANDOM set than to labels of the GROUND labels. This result holds for each emotional attribute in the VAD space. This result further supports our hypothesis that dynamic information is crucial for the perception of emotions. Evaluation of isolated frames leads to different emotional judgments.

III. Blind Lexical Facial Expression Recognition in the Presence of Speech

While models for FER from static images achieve high accuracy for posed expression. Emotion classification when the subject is speaking (i.e., speech articulation is introduced) is still an ongoing challenge. Models have been developed that need transcriptions and phoneme alignment. Our goal is to find methods that achieve higher accuracy on video only data (i.e., no transcriptions/phoneme). To achieve this goal we rely on the MSP-IMPROV [4], which contains 15 target sentences spoken in happiness, anger, sadness, and neutral emotions, from 15 different actors.

In our first attempt we created paired data for the training set (e.g., same lexical content with different emotions). We phonetically align the videos conveying the same lexical content, but with different emotions. Since we are interested in an emotional-to-neutral facial transformation, one of the video is emotional (happiness, anger or sadness), and the other is neutral, where the goal is to learn the mapping between images. With the paired data, we train a CycleGAN, as well as, Pix2Pix network [5], which is a model that uses GANs to find image-to-image transformations using paired data. Our data consists of paired images extracted from one actor speaking the same sentence while conveying two different emotions (happiness and neutral). Figure 2 shows that the Pix2Pix model was indeed able to transform an emotional image (Fig. 2.A) into an image closer to neutral without removing speech articulation information (Fig. 2.B). Notice that the reference image (Fig. 2.C) is used here only as a reference, since the model has not seen this image during training. After the transformation, we compared the pixels between the input and the output of the system (Fig. 2.D) and between the input and the reference image (Fig. 2.E). We notice that Figure 2.D has bright areas around the mouth indicating that the orofacial area was affected by the transformation. Unfortunately, both models struggle to find reasonable transformations when trained on two or more subjects. When introducing multiple subjects, the model is required to differentiate between individual differences, in addition to lexical and emotional differences.

Our second attempt aims to remove speaker dependencies using image-to-image transformations. To achieve this we consider an intermediate representation that is less dense and more generalizable than raw pixels. We use a 3D mesh, transforming emotional 3D mesh into neutral ones, instead of image-to-image transformations. This allows us to represent a frame as 512 3D points (1,683 points) a dramatic decrease from 128x128x1 (16,384 points) images. Figure 3 shows the proposed model which consists of three parts. The first block is the *feature extractor model*, which produces facial features from the original images using multiple layers of *convolutional neural network* (CNN). The second block is the *style extractor model*, which is the core contribution. This block compensates for the lexical variability, overcoming the noise introduced by speech articulation. The third block is the *fusion model*, which concatenates the feature representations from the first two models to predict emotions.

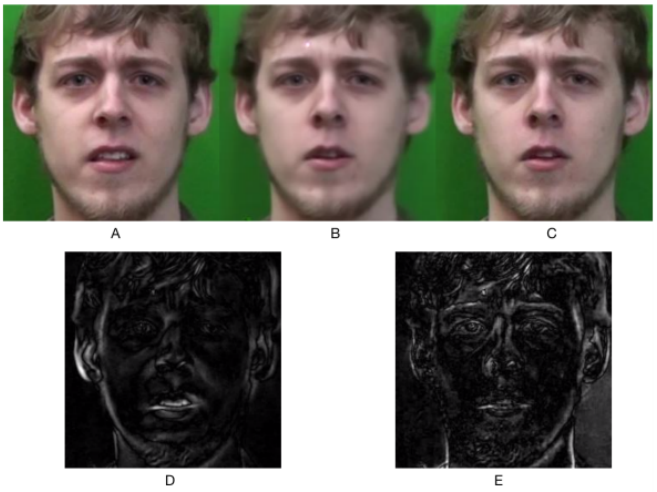

Fig. 2. Shows an example of the approach. (A) input image from the testing set. (B) output of the pix2pix model, (C) is the reference image that was paired with (A). Notice that the model has never seen this image. (D) Pixel difference between (A) and (B). (E) pixel difference between (B) and (C).

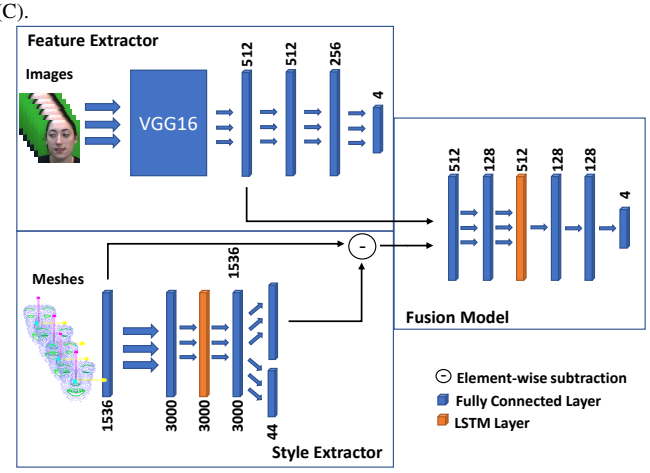

Fig. 3. Diagram of the proposed FER system for videos. The key contribution in this study is the style extractor, which aims to compensate for lexical information.

A. Feature Extraction Model

The purpose of this model is to extract a discriminative facial feature representation directly from images (Fig. 3). This vector will be combined with the feature representation produced by the style extractor model (Sec. III-B). The feature extraction model is static, processing each frame without considering temporal information (the style extraction model and the fusion model will incorporate temporal information). We use the VGG16 architecture [6] for our model. After the VGG16 network, we add three dense layers and a softmax layer with four neurons to predict the emotion in the images (happiness, anger, sadness, neutral state). The loss function for this model is the categorical cross-entropy loss. The feature representation that is passed to the fusion model is the first dense layer after the VGG16 max pooling layer (512 nodes, Fig. 3).

B. Style Extractor Model

The style extractor model creates a facial transformation from emotional (i.e., happiness, anger, sadness) to neutral

state for each frame. Then, the transformed frames are used as neutral reference to temporally and spatially contrast the original emotional frames. The resulting vector is expected to convey emotional information after compensating for lexical information. We use the Zface toolkit [7] to extract the 3D facial mesh and train the style extractor model.

The input of the model is the 3D mesh of the emotional face. We pass the input (3D meshes) to a shared dense layer followed by a *long short-term memory* (LSTM) layer. The LSTM layer is added to model temporal information. The LSTM output is then passed to another shared dense layer. The output of the model is (1) the transformed neutral 3D mesh, which is a dense layer with 1,536 neurons, and (2) a softmax layer of 44 neurons, representing the one-hot encoding of the phone assigned to each mesh. The softmax layer facilitates learning phone-dependent features without requiring phonetic alignment during testing. The loss function is a linear combination of the mean squared error of the meshes, and categorical cross-entropy of the phones. Both losses are equally weighed. We use the difference between the emotional mesh (input) and predicted neutral mesh (output) as the style feature vector, which is then passed to the fusion model.

C. Fusion Model

The fusion model concatenates the feature representations provided by the feature extractor and style extractor models along the time axis. The goal of the fusion model is to predict an emotion for the entire sequence of frames. Each concatenated frame is passed into dense layers. Then, we use a LSTM layer to extract temporal information. The LSTM layer returns the last output of the sequence. This single feature vector is then passed to two dense layers followed by a softmax with four neurons, representing the emotional classes.

To assess the effectiveness of our proposed approach. We evaluate our architecture with and without the style extractor. Notice that both approaches include temporal information, since the fusion model is implemented with LSTM units. Table IV shows the F1-score of this model with and without the style extractor. The results clearly demonstrate the effectiveness of using the style extractor model, which leads to higher F1-score on all the emotions. On average, adding the style extractor improves the F1-score achieved by the FER model from 67% to 74%, which corresponds to a 7% (absolute) gain. The improvement is directly attributed to the additional features provided to the fusion model by the style extractor model. It is also worth noting that while the gap in performance on the training set is much closer (within 1-2%), the model with the style features achieves a higher accuracy on the validation and testing sets. Therefore, the features from the style extractor model also contribute to improve the generalization of the FER models.

TABLE IV

PERFORMANCE OF THE PROPOSED FER SYSTEM FOR VIDEOS ON THE TEST SET OF THE CREMA-D CORPUS. MODEL A INCLUDES THE STYLE EXTRACTOR. MODEL B DOES NOT INCLUDE THE STYLE EXTRACTOR.

Emotion	Precision		Recall		F1-score	
Model	A [%]	B [%]	A [%]	B [%]	A [%]	B [%]
Happiness	87.8	81.1	83.0	83.5	85.3	82.3
Anger	89.2	51.0	50.9	65.0	64.8	57.1
Sadness	78.6	83.0	60.5	52.3	68.4	64.1
Neutral	68.8	65.0	89.9	65.0	78.0	65.0
Average	81.1	70.0	71.0	66.4	74.1	67.1

IV. CONCLUSIONS AND FUTURE WORKS

We have been working to better understand the role that speech articulations play in changing the perception of emotion in dynamic and static visual-only data.

We have observed important differences when evaluating isolated images from videos, especially for certain emotions. While using a static FER system to recognize happiness might result in satisfactory performance, our analysis shows that using static a FER system to recognize anger will result in poor classification accuracy, even if the system has human-level performance. Overall, the labels from static representation fail to accurately predict the labels for dynamic representation. Additionally, we proposed a novel method to extract style without requiring phonetic alignment during inference. We found that the features from the style extractor model improve not only the FER performance, but also aid in the generalization of the model.

For future research, we would like to further analyze how each emotion behaves using dynamic and static representations. We would also study how different phones/viseme affect the perceived emotions. The results can help us further develop a reliable FER system. Additionally, we wish to explore different methods to separate the style (i.e., emotion) and the contents (i.e., speech). Separating the two can help in many tasks, not just FER. For example, knowing the style (i.e., emotion) and content (i.e., speech) can aid visual lip reading systems and voice activity detection systems.

REFERENCES

[1] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Procedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. [Online]. Available: https://doi.org/10.5244/c.29.41

[4] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016.

[6] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4845–4849.

[7] L. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–8.