Dynamic versus Static Facial Expressions in the Presence of Speech

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

***Abstract*— Face analysis is an important area in affective computing. While studies have reported important progress in detecting emotions from still images, an open challenge is to determine emotions from videos, leveraging the dynamic nature in the externalization of emotions. A common approach in earlier studies is to individually process each frame of a video, aggregating the results obtained across frames. This study questions this approach, especially when the subjects are speaking. Speech articulation affects the face appearance, which may lead to misleading emotional perceptions when the isolated frames are taken out-of-context. The analysis in this study explores the similarities and differences in emotion perceptions between (1) videos of speaking segments (without audio), and (2) isolated frames from the same videos evaluated out-of-context. We consider the emotions happiness, sadness, anger and neutral state, and emotional attributes valence, arousal, and dominance using the MSP-IMPROV corpus. The results consistently reveal that the emotional perception of static representations of emotion in isolated frames is significantly different from the overall emotional perception of dynamic representation in videos in the presence of speech. The results reveal the intrinsic limitations of the common frame-by-frame analysis of videos, highlighting the importance of explicitly modeling temporal and lexical information in face emotion recognition from videos.**

I. Introduction

Emotions play a crucial part in our lives, influencing how we interact and communicate with others. Humans, as well as animals, can express and understand social signals associated with emotions with little effort, even if the externalization of emotion is subtle. These emotional skills are often missed or ignored in *human computer interaction* (HCI) affecting the capability of the systems. It is important to design algorithms that can mimic human emotion perception, which will radically transform the way we interact with existing systems. While we externalize emotional behaviors through multiple modalities, our effort in this study is on facial expression in videos in the presence of speech.

Important advancements have been made in *face emotion recognition* (FER) from static images with clear expressions [1]–[3]. An open challenge is to infer emotions from videos, especially when the subject is speaking. Several studies have considered speech as *noise*, selecting key frames with clear emotional displays [4], [5], or discarding facial features from the mouth area [6], [7]. A straightforward approach to process videos is to separately recognize emotional cues in each of its frames, aggregating the results at the segment level [8], [9]. We argue that this strategy is an oversimplified approach that has intrinsic limitations, ignoring contextual information, including lexical content, which is important in the analysis of facial expression. Speech articulation changes the appearance of the face creating a challenging interplay between acoustic and lexical information [10], [11], where the facial muscles are triggered to produce speech while externalizing emotional cues (Fig. 1). As a result, the emotional cues observed from isolated frames may lead to poor representations of the emotional content in a video.

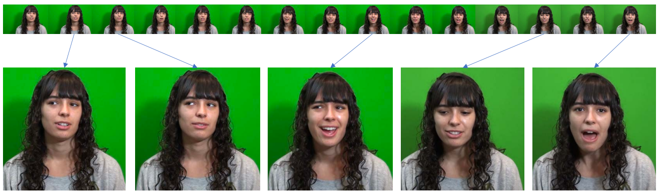


Fig. 1. Static representation of emotion in frames extracted from a video. Individual frames do not represent well the emotion of the video, especially in the presence of speech.

This study investigates the differences and similarities between (1) videos of speaking segments (without audio), and (2) isolated frames from the same videos evaluated out-of-context. Can frames sampled from the video and annotated out-of-order provide reliable information to correctly infer the emotional perception observed from the entire video segment? We hypothesize that the static representation of isolated frames is intrinsically limited, providing a poor representation of the emotional content perceived after watching the entire video. The analysis relies on a subset of the MSP-IMPROV corpus. We annotate the emotional content of frames from video segments of subjects engaged in dyadic interactions. The perceptual annotation of the frames is conducted out-of-order to remove contextual information (i.e., information gained by viewing the surrounding frames). We consider emotional categories (anger, happiness, sadness, neutral state), and emotional attributes (valence, arousal, and dominance). The annotations of the frames are compared with the annotations of videos without audio. The annotations are also compared with the results of a CNN-based face emotion classifiers trained to recognize categorical emotions from images trained on a separate dataset.

The results of this study support our hypothesis that the emotional perception of isolated frames provides a poor representation of the emotional perception of videos. We observed that the similarities and differences between static and dynamic facial expressions depend on the emotional category. While static images provide adequate representations for happy videos, the results consistently show important differences for angry videos. We analyze the temporal

This work was supported by NEC Foundation Of America

dynamic of the perceived emotion in images, showing that emotional displays fluctuate across time, where different frames extracted from the same video can be perceived with very different emotions. These differences can be explained, up to some extent, due to the presence of speech (Fig. 1). This result indicates that choosing an apex image to represent a video is also problematic. These results provide important insights for FER, challenging practices to process frame-by-frame the facial images in a video, and supporting practices that attempt to extract temporal information [12], [13] or compensate for lexical information [14]–[16]. The implications of this study are important. We need better algorithms that can reliably and dynamically disentangle emotional and lexical information conveyed in videos.

II. Background

A. Related Work

We distinguish between experienced, expressed and perceived emotion. Experienced emotions are the true emotions that someone feels. Expressed emotions are the modulations and manipulations explicitly or implicitly externalized due to emotions. Perceived emotions are the emotions that others infer by observing the subject. We argue that affective computing solutions should be mostly focused on perceived emotions, where the goal is to replicate the human capability to interpret emotions from others. This paper analyzes the perception of emotion from static and dynamic facial representations.

The perception of emotions can be highly subjective. Emotional perceptual evaluations have shown that the inter-evaluator agreement tends to be low. Some of these differences are due to gender. Biele and Grabowska [17] found that the perception of emotion can differ from men and women while evaluating static and dynamic facial expression. Their results showed that women labeled emotions with higher intensities than men. The differences between static and dynamic facial representations can also be attributed to how we process facial expressions. Adolphs et al. [18] suggested that images and videos are processed, up to some extend, by different areas of the brain.

Studies have analyzed the information provided by images and videos of facial expressions using datasets where the subjects were not speaking. Cunningham et al. [19] showed that having a dynamic sequence of at least 100 ms, which retains the temporal order led to better results than considering static out-of-order images. Artificial videos created by shuffling the order of the frames achieved inferior performance, highlighting the importance of temporal information. Ambadar et al. [20] conducted perceptual evaluations under four conditions: an image representing the emotional of a video, a sequence of static images with correct temporal order, a sequence of static images with noise frames in between, and a sequence of the first and last frames. They found that the video representations unseparated by noisy frames provided the best accuracy in detecting emotions compared. The other four sets provided similar results. A contrasting result was presented by Gold et al. [21], where they argued that a single static image at the apex of an emotion can better represent a dynamic sequence. However, they considered short sentences where the subjects displayed posed expressions without speaking. While these studies have considered static and dynamic facial expressions, the datasets used in their analysis were controlled with pose expressions and without articulation. These conditions are not representative of expressive behaviors observed during naturalistic human interactions. Our study addresses this problem on less controlled conditions by considering the articulatory movements in the presence of speech.

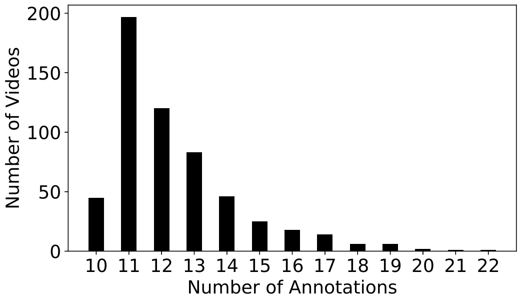


Fig. 2. Number of videos from the MSP-IMPROV corpus that have been annotated by a given number of annotators (video-only condition).

B. Database

We are relying on the MSP-IMPROV corpus [22] for this analysis, which is a multimodal emotional database. The corpus was recorded in six sessions of dyadic interactions between actors (i.e., 12 subjects). The participants improvised hypothetical scenarios that led one of them to say a target sentence in a given emotion. The set with the target sentences is referred to as the *target - improvised* set. The interactions were recorded using a high definition camera in front of each participant with a 1440 × 1080 resolution. The sessions were well illuminated with two professional LED light panels, using chroma-key green screens placed behind the participants.

The key feature of this corpus is that the *target - improvised* set of this corpus has been annotated with emotional labels under different conditions to study emotional perception [23]: (1) audiovisual presentations, (2) audio-only presentations, and (3) video-only presentations. Our analysis uses the video only presentation, since adding audio will introduce additional information that will prevent us from comparing the emotional perception inferred from videos and frames. There are 564 videos in total that have been annotated with a mean duration of 2.34sec.

The annotation of the emotional content in the videos was conducted with perceptual evaluations in *Amazon Mechanical Turk* (AMT), where multiple workers were recruited (see details in Mower-Provost et al. [23]). Each video was annotated with the categorical classes happiness, sadness, anger, neutral state and other. The videos were also annotated with attribute-based annotations with a 9-point Likert scale for valence (negative versus positive), arousal (calm versus active) and dominance (weak versus strong). The video only presentations have been annotated by at least 10 people.

TABLE I
PERFORMANCE OF OUR FER SYSTEM IMPLEMENTED WITH THE VGG-FACE NETWORK USING THE AFFECTNET CORPUS.

Emotion	Precision	Recall	F1-score
Happiness	0.87	0.90	0.89
Anger	0.76	0.70	0.73
Sadness	0.75	0.70	0.72
Neutral	0.64	0.70	0.67
Average	0.75	0.75	0.75

Figure 2 shows the number of videos annotated by a given number of evaluators, ordered from 10 to 22. A large number of annotators allows us to estimate meaningful distributions of the emotional content conveyed in these videos.

III. METHODOLOGY

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 describe these sets in more detail in this section.

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 a 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. III-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 second. 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 videos (Sec. II-B). 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.

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 [24]). 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. [25] 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.

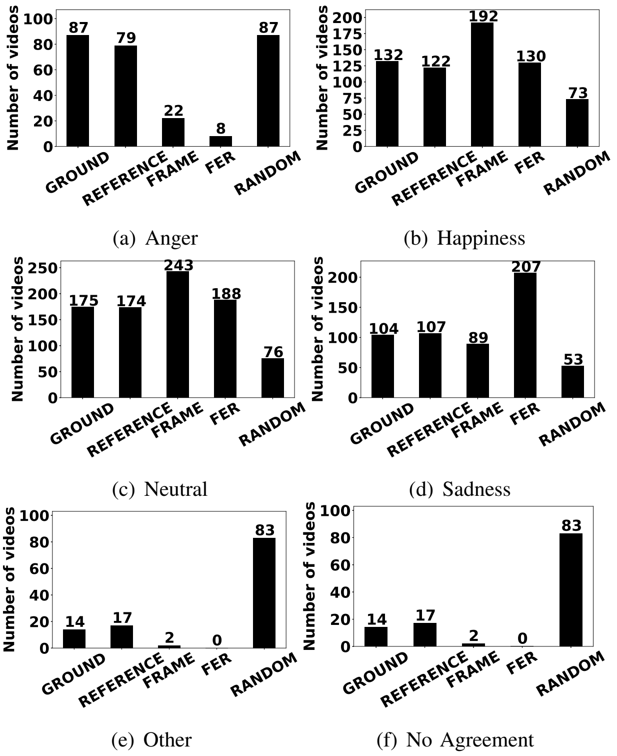
Table I provides the precision, recall and F1-score of our FER system. This model achieves an F1-score of 75% on the development set (our testing set) of the AffectNet corpus. As a reference, Mollahosseini et al. [24] achieved an F1-score of 57% on an eight-class problem. To use the model on images from the FRAME set, we extract the face from the image using the Dlib library [26]. The face images are aligned and resized to 224×224, using the resulting image as the input of our FER system. We transformed the activations of the output layers into a distribution. The final output is a distribution considering the results across all the frames. Since our FER model does not have the class other, we set this value always to zero.

E. The RANDOM Set

The fifth set corresponds to selecting an emotion for each frame at random. After normalization, we estimated a distribution for each video. This set is used as a reference for the metrics when we compare two uncorrelated emotional distributions.

IV. STATIC AND DYNAMIC REPRESENTATIONS ANALYSIS

This section analyzes the labels assigned to each of the five sets described in Section III. We compare the perception of emotion from facial expressions from frames and the entire videos, highlighting the importance of dynamic information.



(a) Anger (b) Happiness

(c) Neutral (d) Sadness

(e) Other (f) No Agreement

Fig. 3. Number of videos per emotion for each set. The emotion with the highest frequency is selected for each video. No agreement is reached when the highest two frequencies are equal.

The analysis includes categorical (Sec. IV-A) and attribute-based (Sec. IV-B) emotion representations. We also consider the temporal dynamics of the labels assigned to frames (Sec. IV-C). The study analyzes the phone dependency in the evaluation of emotions in the frames (Sec. IV-D). We also analyzed one video as an example (Sec. IV-E).

A. Analysis of Categorical Emotions

For categorical emotions, our first analysis compares the overall distribution of consensus labels assigned in each of the five sets. Figure 3 shows these distributions for four emotional categories, sentences without agreement and sentences labeled as other. First, we notice that the distributions of the labels for GROUND and REFERENCE are very similar across emotions. However, the distributions of the labels for GROUND and FRAME are very different. We notice that anger is underrepresented in the FRAME and FER sets, compared to annotations obtained from videos (GROUND and REFERENCE). This result shows that anger is an emotion that is hard to recognize in static images without temporal information. In contrast, sadness (FER) and happiness (FRAME) are over-represented. A potential explanation is that the recognition of certain emotions may rely on specific cues that are extended over time (e.g., smiling). Other emotions may require more complex temporal coordinated face movements that may be activated at different times (e.g., brow lowering, mouth tightening).

We also compare the agreement between the labels using the F1-score metric. We consider the labels from the GROUND set as the ground truth. The consensus labels obtained using the other four sets are used to estimate the F1-score. A score close to 1 indicates a high agreement between the labels. Table II shows the results. We first compare GROUND and REFERENCE labels. Both of these labels are obtained by perceptual evaluations of videos without audio. Although the same data was provided to the annotators, the F1-score is only 0.77, on average. The emotion with the highest performance is happiness, which indicates that happiness might be a less ambiguous emotion when relying only on facial cues. We achieve an F1-score of 0.56 when we compare GROUND and FRAME (27% relative decrease from the F1-score between GROUND and REFERENCE). When we compare GROUND and FRAME, happiness is also the emotion with the highest F1-score (0.79). The annotators were able to identify the dominant emotion in the videos from isolated frames. In contrast, we observe that anger is the emotional category with the lowest F1-score (0.22) when we compare the labels between GROUND and FRAME. The main problem is the recall rate, where many images from angry videos are not perceived as anger. This result indicates that the perception of anger relies more on dynamic information and, therefore, it is very difficult to consistently perceive angry faces from isolated frames. This result agrees with the study of Cauldwell [27], which indicated that the perception of anger was significantly different when the conversations were evaluated in-context or out-of-order, without contextual information. The F1-score for the labels provided by our FER system (0.49) is lower than the ones obtained with perceptual evaluations of isolated frames (0.56). However, the trends across emotions are very similar indicating that the frame-by-frame approach without considering the relationship between consecutive frames cannot properly represent the dominant emotion perceived in the video. This result is particularly clear for anger.

TABLE II

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

TABLE III

AVERAGE EUCLIDEAN DISTANCE BETWEEN THE DISTRIBUTIONS PROVIDED BY THE FIVE SETS COMPARED IN THIS ANALYSIS. SMALLER VALUES INDICATE THAT THE DISTRIBUTIONS ARE CLOSER.

L2 norm	GROUND	REFERENCE	FRAME	FER	RANDOM
GROUND	0.00	0.34	0.42	0.55	0.68
REFERENCE	0.34	0.00	0.44	0.57	0.69
FRAME	0.42	0.44	0.00	0.47	0.52
FER	0.55	0.57	0.47	0.00	0.74
RANDOM	0.68	0.69	0.52	0.74	0.00

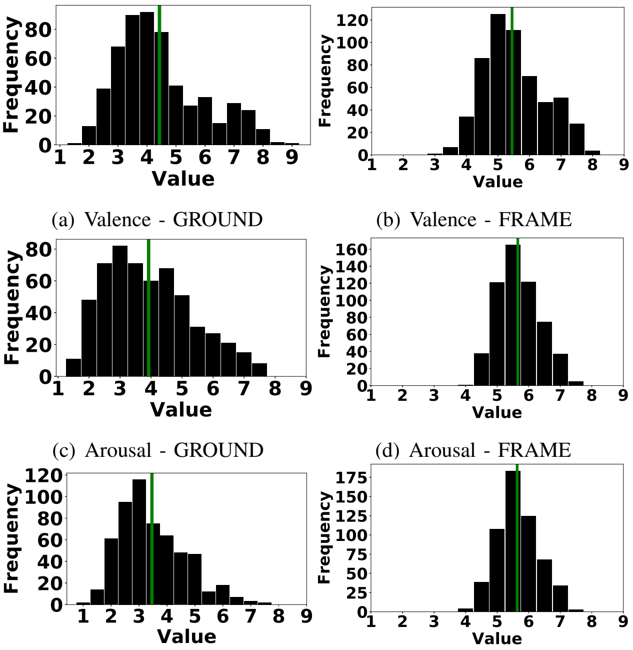
We also estimated the average *Euclidean distance* (ED) between the distributions collected from these five sets. Table III gives the ED across the five sets. When we compare the ground truth (GROUND) with the REFERENCE set, we observe the lowest ED (d =0.30), as expected since both sets correspond to annotations of the entire videos. This value is our reference. The ED increases when we compare the ground truth with the distributions of either the FRAMES with d =0.42 (23.5% relative increase) or the FER model with d =0.55 (61.8% relative increase). As a reference, the ED between the GROUND and RANDOM sets is d = 0.68. These results indicate that the emotional perception of isolated frames is not representative of the emotional perception after watching the entire video. These results support our hypothesis that isolated frames provide a weak representation of the emotions in a video. Modeling temporal and lexical information is important.

B. Analysis of Attribute-based emotional Representation

The analysis also evaluates the difference in emotional perception from videos and isolated frames using attribute-based representations. The use of emotional attributes provides an alternative representation to describe emotion, complementing the information observed from categorical emotions. For example, it provides information to quantify differences in the emotional content of images or videos labeled with the same emotional class (i.e., within-class variability).

Our first analysis with emotional attributes is to explore the global distributions for valence, arousal, and dominance in the GROUND and FRAME sets (labels from the REFERENCE set are not included in this analysis as they have similar distribution as GROUND). Figure 4 shows the results, which reveal a shift in the perception of emotional attributes in the FRAME set. The isolated images are perceived as more active (arousal), more positive (valence) and more dominant (dominance) than the videos. Anger was the emotion with major differences in the perceptual evaluation of isolated frames and videos (Sec. IV-A). Anger is usually associated with low valence, which explains the shift in valence (i.e., images perceived more positive than videos). Interestingly, even emotional attributes for isolated images present clear shifts from the corresponding distribution for videos.

Another metric that we use to compare the differences in emotion perception in emotional attributes is the Euclidean distance between the labels in the VAD space (i.e., valence, arousal, and dominance). This analysis compares all the sets, except the FER set, since the FER system was built to recognize categorical emotions. The results are shown in Table IV. 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.



(a) Valence - GROUND (b) Valence - FRAME

(c) Arousal - GROUND (d) Arousal - FRAME

(e) Dominance - GROUND (f) Dominance - FRAME

Fig. 4. Distribution for valence, arousal, and dominance using the GROUND and FRAME sets. For FRAME, the value for a video is the frame annotation average. The vertical green line represents the mean value.

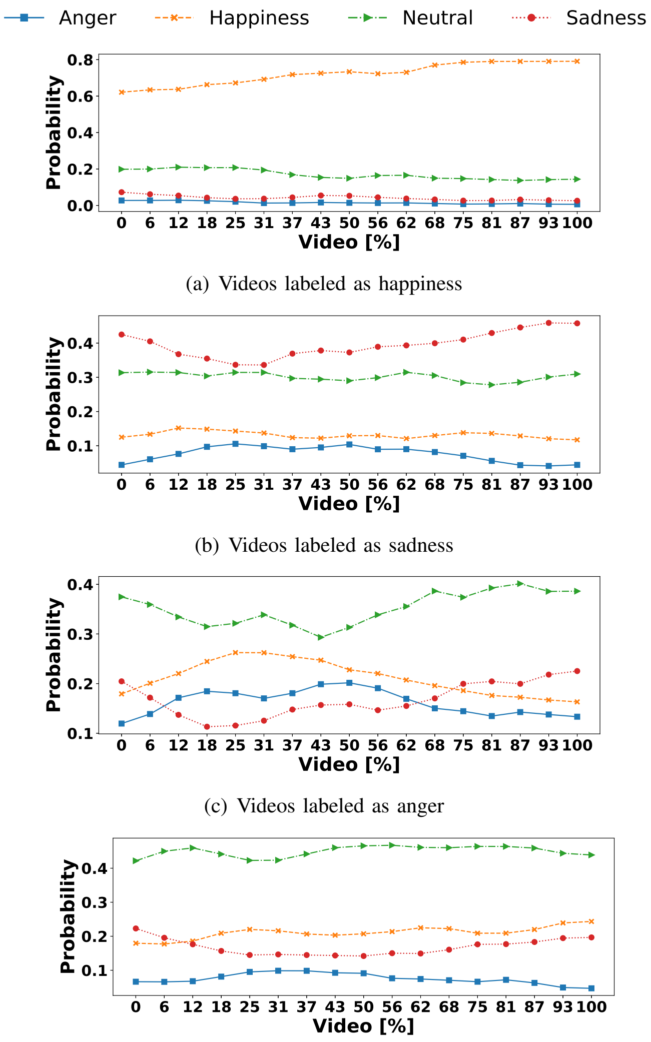
TABLE IV

EUCLIDEAN 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

C. Temporal Analysis

This part of the analysis compares the temporal evolution of the emotion. For this purpose, we consider the emotional evaluations for the isolated images, analyzing the average



(a) Videos labeled as happiness

(b) Videos labeled as sadness

(c) Videos labeled as anger

(d) Videos labeled as neutral

Fig. 5. The temporal average distribution in the FRAME set for videos classified as (a) happiness, (b) sadness, (c) anger, and (d) neutral state.

trends (GROUND and REFERENCE sets do not have any temporal information). Since the videos do not have the same durations, we align the videos by interpolating and extrapolating the scores provided to each of their frames. The videos are stretched or compressed such that they have the same length. After the alignment, we average the temporal emotional curves obtained from the FRAME set.

Figure 5 shows the results for categorical emotions. We group the videos using the emotional labels in the GROUND set, showing the mean curves obtained by the FRAME set for those videos. The figure shows that the emotions are not uniformly conveyed across time, where some regions are more emotionally salient than others. The perception of each emotional class fluctuates across the sentence. Since the frames are annotated out-of-order, contextual information is not considered in this evaluation. For videos labeled as happy (Fig. 5(a)), we observe that over 60% of the evaluators perceived the images as happy, confirming our finding that evaluators can reliable recognize happiness from isolated frames. The curve for happiness increases almost to 80% at the end of the video. This result agrees with studies suggesting that emotional cues for happiness tend to be emphasized at the end of the sentence [28]. For videos associated with sadness (Fig. 5(b)) and neutral state (Fig. 5(d)), the prominent emotions observed across time from the isolated images are the prominent emotions observed on the video. However, the proportion of evaluators who selected the *right* emotion is only around 40%. This pattern is not observed for anger (Fig. 5(c)), where the percentage of evaluators providing angry labels to the isolated images is less than 25%. This result confirms our previous findings about the challenges in detecting anger in isolated images without contextual information.

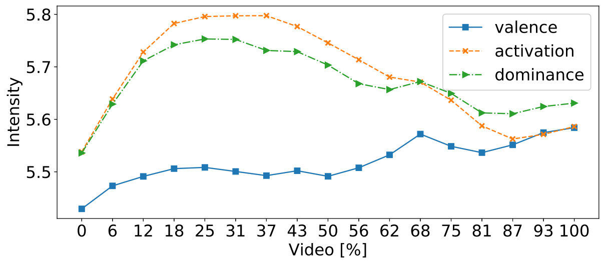


Fig. 6. Temporal average distribution of valence, arousal and dominance. The length of the extracted frames were normalized before averaging.

Figure 6 shows the results for attribute-based emotions. These curves are estimated across all the videos. This figure also shows that emotional information is not uniformly conveyed during a video. Arousal and dominance tend to increase in the first half of the videos (i.e., more active and dominant), reducing their value at the end of the video. A different trend is observed for valence, where the videos are perceived more positively at the end of the sentence. Understanding the intrinsic fluctuations of emotional behaviors is crucial to design robust FER systems.

D. Viseme Analysis

The key hypothesis in this study is that articulatory movements affect the perception of emotion. We expect that the perception of emotion from isolated images during certain phones will generate larger deviations from the global emotional perception on the video. We conduct an analysis at the phone level to evaluate this hypothesis.

We use the Montreal forced aligner toolkit [29], creating phone alignment. This information is used to assign each image in the FRAME set to a given phonetic category. Since the number of images per phone class is limited, we conduct this analysis at the viseme level, aggregating phonetic units that share similar visual appearance. We use the mappings between phones and visemes suggested by Lucey et al. [30]. We compare the Euclidean distance between the annotations of each frame and the annotations in the GROUND set (i.e., emotional labels assigned after watching the video).

Table V captures the average ED distances between GROUND and FRAME labels for each viseme. We notice that /sp/ (silence) has the second-lowest ED distance (0.63). This result is expected since silence contains minimal face articulation (i.e., reduced interplay between lexical and emo-

TABLE V

Viseme level analysis in the perception of emotion from static images. The table includes the coverage of the data for each viseme unit, the percentage of time that the predicted emotion matches the prominent emotion in the video, and the ED between the labels in FRAME (for those visemes) and GROUND sets.

Viseme	Coverage	Primary emotion	L2 Distance
ah	8.0%	39.3%	0.5849
sp	23.7%	44.1%	0.6371
er	1.9%	41.3%	0.6438
iy	9.1%	44.5%	0.6528
t	16.3%	46.0%	0.6719
ch	3.3%	43.3%	0.6740
ey	5.7%	41.4%	0.6787
x	4.9%	41.0%	0.6797
w	4.1%	36.0%	0.6929
k	14.0%	46.6%	0.7022
aa	1.6%	36.1%	0.7295
f	1.9%	37.3%	0.7593
uh	1.0%	38.2%	0.7598
p	4.3%	36.9%	0.7612

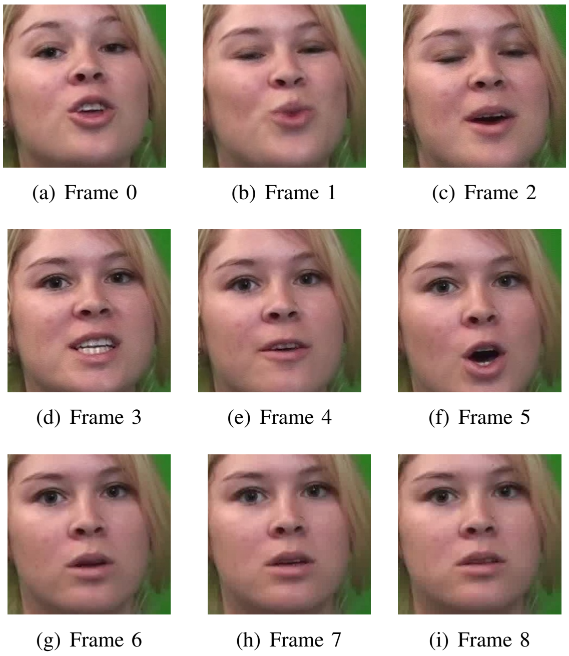
tional information). We also notice that the viseme with the highest ED distance is /p/. This result is expected since /p/ is a bilabial sound – a sound created by pressing and releasing the two lips. Because of this motion, the static emotional facial features for these images are incorrectly classified.

E. Case Study

To illustrate the mismatch between the perception of a video and the perception of isolated images of the video, we analyze one particular video in the corpus (video *MSP-IMPROV-S14A-F02-T-FM01*). Figure 7 shows the nine frames extracted from this video, which were annotated with emotional labels (FRAME set) and processed by our FER system (FER set). Figure 8(a) shows the results of the distributions of categorical emotions at the video level for the GROUND, REFERENCE, FRAME, and FER sets. The figure shows that the video was perceived as anger as the primary emotion by all the annotators (GROUND and REFERENCE sets). In contrast, the distribution for the annotations for the isolated images in this video shows a very flat distribution, where happiness and neutral are the most popular selections. The FER model recognizes most of the frames of the video as neutral. Figure 8(b) shows the temporal evolution of emotions perceived in the images, providing a distribution per frame. While this video was predominately perceived with the emotion anger, the proportion of labels for angry in the isolated images assigned by the raters is 40% or lower (with the exception of frame 3). The perception of anger in the video is not dominant in the facial expressions observed in the frames. Figure 8(c) shows the results for the normalized activations provided by the FER system for each frame. The probability of anger predicted by the model is almost zero for all the frames. This example shows that the emotional perception of a sentence is not necessarily the same as the perception of isolated frames.

V. Conclusion

This study considered the similarities and differences between emotional labels derived from facial expressions obtained after evaluating videos and isolated frames extracted from these videos. The key motivation in this analysis was to assess whether static representations from images are good approximations of dynamic representations inferred after watching the entire video when speech is present. The analysis in this study demonstrated important differences between emotional perceptions derived from videos and images, especially for angry videos.


(a) Frame 0 (b) Frame 1 (c) Frame 2

(d) Frame 3 (e) Frame 4 (f) Frame 5

(g) Frame 6 (h) Frame 7 (i) Frame 8

Fig. 7. Frames extracted from the video analyzed in Section IV-E. The analysis of these frames is presented in Figure 8.

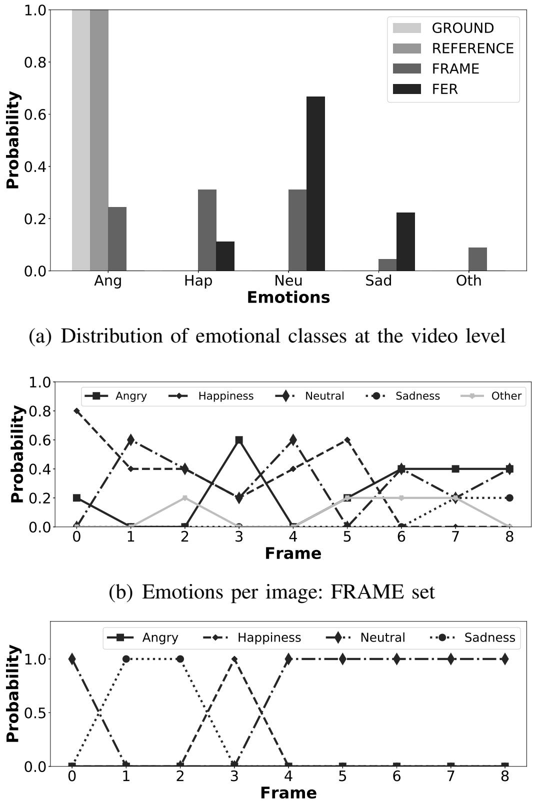
The results of this study have important implications for FER in videos, especially when the subjects are speaking. The common approach of analyzing frame-by-frames the images in a video without considering contextual information is problematic, having intrinsic limitations. Even if we can train an image-based FER system that perfectly replicates the emotional perception of human annotators, we may not be able to reliably predict the emotion in a video. Even selecting key frames in a video is problematic as the emotional content fluctuates over time creating a challenging interplay between lexical and emotional information that is reflected in the appearance of the face. Future research directions in this area should consider temporal models that capture contextual information, disentangling lexical and emotional information from facial expressions.

References

[1] H. Soyel and H. Demirel, “Facial expression recognition using 3d facial feature distances,” in *Image Analysis and Recognition*, M. Kamel and A. Campilho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 831–838.

[2] X.-P. Huynh, T.-D. Tran, and Y.-G. Kim, “Convolutional neural network models for facial expression recognition using bu-3dfe database,” in *Information Science and Applications (ICISA) 2016*, K. J. Kim and N. Joukov, Eds. Singapore: Springer Singapore, 2016, pp. 441–450.

[3] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” *CoRR*, vol. abs/1511.04110, 2015. [Online]. Available: http://arxiv.org/abs/1511.04110



(a) Distribution of emotional classes at the video level

(b) Emotions per image: FRAME set

(c) Emotions per image: FER set

Fig. 8. Comparison of the distribution for one of the videos of the MSP-IMPROV (Fig. 7 shows the extracted frames). (a) aggregated distributions of emotions assigned to the video, (b) distributions assigned to the isolated images, and (c) normalized activation per frame by the FER model.

[4] S. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 534–537.

[5] C. Shan and R. Braspenning, "Recognizing facial expressions automatically from video," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. Augusto, Eds. Boston, MA: Springer, October 2010, pp. 479–509.

[6] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[7] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, September 2003.

[8] S. Kahou and et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550. [Online]. Available: http://doi.acm.org/10.1145/2522848.2531745

[9] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160–187, July 2003.

[10] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[11] ——, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 43–47.

[12] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015. [Online]. Available: https://doi.org/10.1109/iccv.2015.341

[13] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognition*, vol. 75, pp. 33 – 40, 2018, distance Metric Learning for Pattern Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317300948

[14] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *ACM International Conference on Multimedia (MM 2014)*, Orlando, FL, USA, November 2014, pp. 27–36.

[15] S. Mariooryad and C. Busso, "Facial expression recognition in the presence of speech using blind lexical compensation," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 346–359, October-December 2016.

[16] ——, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013, pp. 1–6.

[17] C. Biele and A. Grabowska, "Sex differences in perception of emotion intensity in dynamic and static facial expressions," *Experimental Brain Research*, vol. 171, no. 1, pp. 1–6, Jan. 2006. [Online]. Available: https://doi.org/10.1007/s00221-005-0254-0

[18] R. Adolphs, D. Tranel, and A. R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain and Cognition*, vol. 52, no. 1, pp. 61–69, Jun. 2003. [Online]. Available: https://doi.org/10.1016/s0278-2626(03)00009-5

[19] D. W. Cunningham and C. Wallraven, "Dynamic information for the recognition of conversational expressions," *Journal of Vision*, vol. 9, no. 13, pp. 7–7, Dec. 2009. [Online]. Available: https://doi.org/10.1167/9.13.7

[20] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, May 2005.

[21] J. M. Gold, J. D. Barker, S. Barr, J. L. Bittner, W. D. Bromfield, N. Chu, R. A. Goode, D. Lee, M. Simmons, and A. Srinath, "The efficiency of dynamic and static facial expression recognition," *Journal of Vision*, vol. 13, no. 5, pp. 23–23, 04 2013. [Online]. Available: https://doi.org/10.1167/13.5.23

[22] 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.

[23] 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.

[24] 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.

[25] 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

[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[27] R. Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 127–131.

[28] H. Wang, A. Li, and Q. Fang, "F0 contour of prosodic word in happy speech of Mandarin," in *Affective Computing and Intelligent Interaction (ACII 2005)*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. Picard, Eds. Beijing, China: Springer Berlin Heidelberg, October 2005, vol. 3784, pp. 433–440.

[29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 498–502.

[30] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Australian International Conference on Speech Science & Technology (SST 2004)*, Sydney, NSW, Australia, December 2004, pp. 265–270.