A 3-D Audio-Visual Corpus of Affective Communication

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, *Member, IEEE*, Thibaut Weise, and Luc Van Gool, *Member, IEEE*

Abstract—**Communication between humans deeply relies on the capability of expressing and recognizing feelings. For this reason, research on human-machine interaction needs to focus on the recognition and simulation of emotional states, prerequisite of which is the collection of affective corpora. Currently available datasets still represent a bottleneck for the difficulties arising during the acquisition and labeling of affective data. In this work, we present a new audio-visual corpus for possibly the two most important modalities used by humans to communicate their emotional states, namely speech and facial expression in the form of dense dynamic 3-D face geometries. We acquire high-quality data by working in a controlled environment and resort to video clips to induce affective states. The annotation of the speech signal includes: transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. We employ a real-time 3-D scanner to acquire dense dynamic facial geometries and track the faces throughout the sequences, achieving full spatial and temporal correspondences. The corpus is a valuable tool for applications like affective visual speech synthesis or view-independent facial expression recognition.**

Index Terms—**Audio-visual database, emotional speech, face tracking, visual speech modeling, 3-D face modeling.**

I. Introduction

WITH their increasing capabilities, computers are already part of our lives, yet they will not seamlessly blend into them until they will be able to recognize and simulate affective states, fundamental capabilities in human-human communications. Interacting with an artificial agent never will be perceived as natural unless the machine can guess the user's emotional state and react accordingly.

The field of affective computing has seen a boost in recent years [1]. Algorithms for both recognition and synthesis of emotional states are being developed; however, there is still a need for corpora of affective communication, needed for training and evaluating such systems. Acquiring these corpora is challenging, as human affective displays are multimodal, rare, and highly context- and culture-related. A first question is which modalities should be captured. The research community has converged towards the idea that affect-aware systems should use several modalities, in a way imitating humans. Emotional cues can be extracted from physiological measurements (e.g., [2]), but their invasiveness can influence the subject's state. Humans can easily guess someone's affective state only from cues such as facial expressions, voice modulation, and body pose, but speech and facial expression appear to encode most of the information used by humans to communicate emotions [3].

Another aspect to be taken into consideration when acquiring affective data is the desired degree of naturalness. A good trade-off between quality and naturalness needs to be found: corpora collected in controlled environments are by definition unnatural, but moving towards unconstrained settings increases the amount of noise. Many of the studies on affective computing concentrated so far on posed data, which is proven to differ from spontaneous behavior [4]–[6]. In cases where the accuracy of the data is crucial, e.g., for computer graphics purposes, induction methods represent a good compromise, and the literature is rich of examples where videos [7], still photographs [8], music [9], or manipulated games [10] have been used to elicit emotions. These methods are not a replacement of pure naturalism, but they are well established and have shown to evoke a range of authentic emotions in a laboratory environment [11], especially videos [12].

The evaluation and annotation of the recorded data requires some definition of emotion, which is still an open issue in itself. The majority of works on affective computing so far are limited to the six basic emotions, based on the cross-cultural studies on facial expressions of Ekman [13]. These few discrete categories actually stand for a family of emotions and are bounded to Ekman's stringent criteria on what emotions are [14]. An alternative are continuous representations where affective states are mapped onto a low-dimensional space, e.g., a 2-D space based on activation (strength) and evaluation (positive vs. negative) [15]. Collapsing the multidimensional space of possible emotional states onto a homogeneous low-dimensional space inevitably incurs in information loss, and different ways of performing the collapse will lead to different results. Such representations are also not intuitive and difficult to use for inexperienced users.

We present a novel multimodal corpus, aimed at the research fields of automatic synthesis and recognition of expressive visual speech. Together with speech, we acquire high-quality

Manuscript received December 14, 2009; revised March 24, 2010; accepted May 16, 2010. Date of current version September 15, 2010. This work was supported in part by the SNF fund Vision-Supported Speech-Based Human Machine Interaction (200021-130224). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Caifeng Shan.

G. Fanelli, J. Gall, and L. Van Gool are with the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland (e-mail: gfanelli@vision.ee.ethz.ch; gall@vision.ee.ethz.ch; vangool@vision.ee.ethz.ch).

H. Romsdorfer is with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria (e-mail: romsdorfer@ieee.org).

T. Weise is with the Laboratoire d'Informatique Graphique et Géométrique, EPFL Lausanne, Lausanne, Switzerland (e-mail: weiset@vision.ee.ethz.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2052239

dense dynamic 3-D facial geometries. The 3-D information is highly desirable in the mentioned research fields for its informative power, allowing to extract features more easily and reliably than for example 2-D video. Because of the necessary recording setup, we settle for elicited emotions and resort to video clips to induce affective states, as it was done, among others, in [16]. While the video clips provide a context in the spirit of film-based induction methods, the repetition of the emotional sentences serves in itself as an eliciting method [17]. We also introduce a consistency check by asking our speakers to evaluate the emotion in the video clip. Similarly to [18], we label the corpus using a list of affective adjectives to be weighted according to their perceived strength, allowing multiple labels for each sentence. Both the eliciting videos and the recorded data were evaluated by independent raters through online surveys.

The proposed corpus is valuable for applications like emotional visual speech modeling, but also for view-independent facial expression recognition, or audio-visual emotion recognition. The corpus will be made available for research purposes.

II. Related Work

One way to categorize databases for training and evaluating affection-aware systems is based on whether the recorded emotions are naturalistic, artificially induced, or posed. A comprehensive overview of the existing audio-visual corpora can be obtained from [1], [19], and [20]. In the following, we list some of the available datasets, with a specific focus on affective communication.

The HUMAINE Network of Excellence has been an important step forward in the field of affective computing, producing a collection of databases [20] containing a large number of audio-visual recordings divided into naturalistic and elicited.

Among the naturalistic databases, the Vera am Mittag dataset [21] consists of recordings from a German TV talk show, containing spontaneous emotional speech coming from authentic discussions. Most of the data was labeled by a large number of human evaluators using a continuous scale for three emotion primitives: valence, activation, and dominance. The Belfast naturalistic database [22] contains TV recordings and interviews judged relatively emotional, annotated using the FEELTRACE [11] system. The EmoTV corpus [23] contains interactions extracted from French TV interviews, both outdoor and indoor, with a wide range of body postures.

Among the elicited datasets, the Sensitive Artificial Listener (SAL) database [20] contains audio-visual recordings of humans conversing with a computer. The SAL interface is designed to let the user work through a range of emotional states. The SmartKom database [24], comprises recordings of people interacting with a machine asking them to solve specific tasks provoking different affective states. In the Activity Data and Spaghetti Data sets [20], volunteers were recorded while, respectively, engaging in outdoor activities and feeling inside boxes containing various objects (e.g., spaghetti or buzzers going off when touched). The subjects recorded the emotions they felt during the activities.

The eWiz database [18] contains 322 sentences pronounced by the same speaker with varying prosodic attitudes suggested by reading a text specifying the context. In [25], the EmoTaboo protocol is introduced, consisting in letting pairs of people (one being a confederate) play the game “Taboo” while their faces, upper bodies, and voices are recorded.

Going towards acted corpora, the GEMEP corpus [26] comprises recordings of the voices, faces, and full bodies of professional stage actors while uttering meaningless sentences, following the method of Banse and Scherer [27]. The set of displayed emotions is an extension of the six basic ones, and the actors were guided by reading introductory scenarios for each emotion. In [28], students were filmed while pronouncing a set of sentences, each representing one of eleven affective states, an extension of the six basic emotions.

Annotating video recordings is difficult and time consuming. For example, the popular Facial Action Coding System (FACS) labeling [29] takes a trained expert about 2 h for 1 min of video footage. An alternative are marker-based motion capture systems, used to obtain 3-D information; an example is the IEMOCAP database [30], where actors were recorded in dyadic sessions with markers on the face, head, and hands while performing affective communication scenarios. Motion-capture techniques were also employed to record actors engaged in affective speech for corpora aimed at visual speech modeling for synthesis purposes, as in [31] and [32]. Despite the accuracy and robustness of such methods, placing markers on someone’s face is error prone and might influence the subject’s emotional state like other invasive physiological measurements.

When dense 3-D face geometry data is desired, most of the available datasets only target face recognition and therefore contain only still scans of neutral faces. The only exception is [33], where students were scanned while changing their facial expression from neutral to one of the six basic emotions, without speaking.

To our knowledge, there are no available datasets combining audio and dense 3-D facial deformations of affective communication.

III. Data Acquisition

In order to simultaneously record audio-visual speech data, we employed the real-time 3-D scanner described in [34] and a studio condenser microphone. To keep the noise level as low as possible, we acquired the data in an anechoic room, with walls covered by sound wave-absorbing materials. Fig. 1 shows the setup, with a speaker being scanned while watching an eliciting video on the screen.

A. Corpus Definition

Our database consists of 40 short English sentences extracted from feature films. The clips were selected by the authors, trying to cover a wide range of emotions and ensuring that the speech was clear, without music or other voices in the background. The movie clips do not just contain the sentence to be pronounced, but are longer (about 30 s on average) and are supposed to build the emotional state in the viewer. Our volunteers satisfied the sole requirement of being native English speakers: a total of 14 subjects, eight females and six males, aged between 21 and 53 (average 33.5). Each sentence was recorded twice: with and

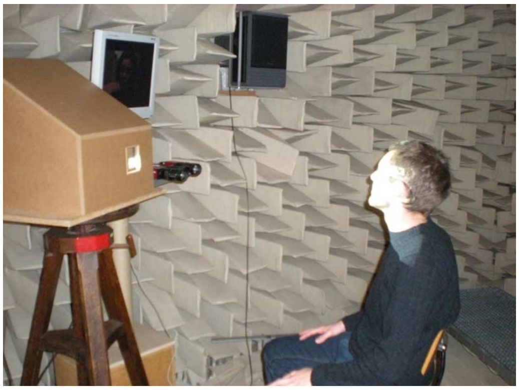


Fig. 1. Recording setup: one speaker sits in front of the 3-D scanner in the anechoic room while watching one of the eliciting videos clips.

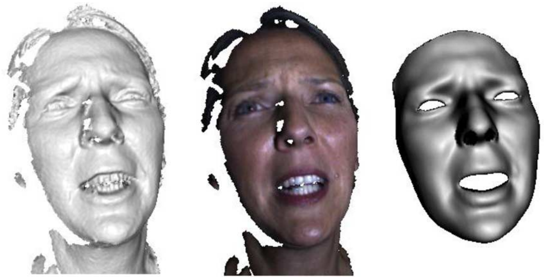


Fig. 2. From left to right, the image shows the 3-D reconstruction of a person's face, the corresponding texture mapped on it, and the personalized face template deformed to fit the specific frame.

without emotion. In total, we recorded 1109 sequences, 4.67 s long on average.

B. Recording Protocol

Each speaker sat alone in the anechoic chamber, in front of the scanner and the microphone, while the authors could give instructions and control the recordings from a separate room.

For the first part of the corpus, the speaker was asked to read the sentences from text displayed on a computer screen, trying to keep a neutral tone. In a second stage, the speaker watched the eliciting video and was asked to rate its emotional content by means of a paper questionnaire, as explained in Section V-A. The videos could be seen more than once if requested. In order to capture the emotional version of each sentence, the speaker was finally asked to repeat the sentence using the emotional tone perceived from the video.

IV. Data Processing

In order to minimize labeling efforts, we processed the data using state-of-the-art methods for both the auditory and visual modality.

The real-time 3-D scanner [34] is employed to capture detailed 3-D geometry and texture of the performances of each speaker at 25 fps, as shown by the first two images in Fig. 2. The 3-D reconstruction presents a mean squared root error of about 0.5 mm. We achieve full spatial and temporal correspondences of the 3-D data thanks to the two-step procedure introduced in [35]: First, a generic mesh is warped to fit the reconstructed 3-D model of each speaker's face in neutral pose, i.e., with the mouth closed and all facial muscles relaxed. Second, the template is automatically tracked throughout all sequences of the speaker, using both geometric and texture constraints to drive the optimization. This step is not only vital for facial expression analysis and synthesis, but it also allows the normalization of the rigid head movements.

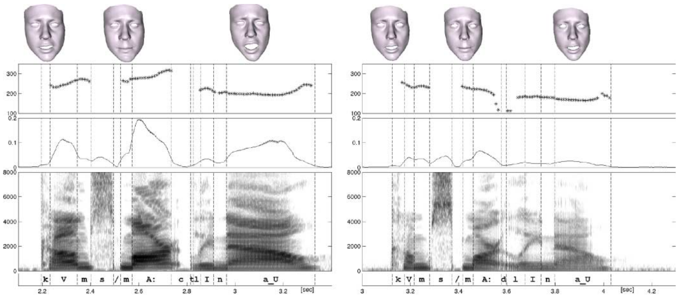


Fig. 3. Comparison of (left) emotional and (right) neutral versions of the same phrase, pronounced by one speaker. Spectrogram, signal intensity contour, fundamental frequency contour, and sample 3-D faces are shown from bottom to top. The emotional utterance shows higher overall signal intensity and a clear rising of the fundamental frequency contour at the second syllable, in contrast to the low falling one of the neutral utterance. Also, syllable nucleus durations are longer in the emotional state. Differences in facial deformations can also be noted, especially in the last part of the phrase.

Different affective states are manifested in speech by changes in the prosody; see [36] for an overview. Speech prosody can be described at the perceptual level in terms of pitch, sentence melody, speech rhythm, and loudness, features which correlate with physically measurable quantities like fundamental frequency (F_0), segment duration, and signal intensity. The annotation process necessary for obtaining the physical prosodic parameters of the utterances includes a number of steps: First, the sentence's text is transcribed into the phonological representation of the utterance; then accurate phone segmentation, fundamental frequency (F_0) extraction, and signal intensity estimation are achieved by analyzing the speech data. For all the above steps, we applied fully automatic procedures provided by SYNVO Ltd. [37].

Some of the extracted audio-visual features are shown in Fig. 3: For the same utterance, facial geometry, spectrogram, signal intensity, and fundamental frequency contour are compared between (left) emotional and the (right) neutral version.

V. Evaluation

In order to assess the quality of the corpus, we resorted to human observers for evaluating both the eliciting movie clips (Section V-A), and the acquired data in the form of videos containing renderings of the 3-D template tracking coupled with the original audio signal (Section V-B). In Section V-C, a preliminary analysis of the data is presented.

A. Eliciting Videos Evaluation

The speakers themselves were asked to rate the induction videos, just after having watched them and before pronouncing the emotional version of the sentences. A paper form was filled out, giving grades between 0 and 5 to a set of 11 suggested emotional labels ("Negative", "Anger", "Sadness", "Stress", "Contempt", "Fear", "Surprise", "Excitement", "Confidence", "Happiness", and "Positive"), where 0 means "I don't know", 1 corresponds to "Not at all", and 5 to "Very". An additional field

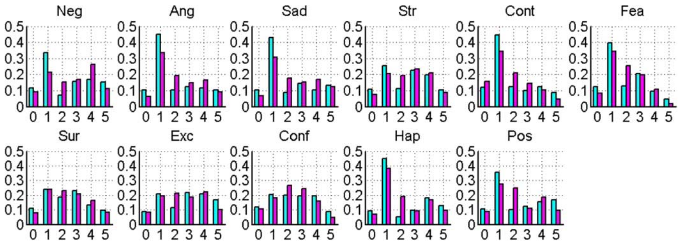
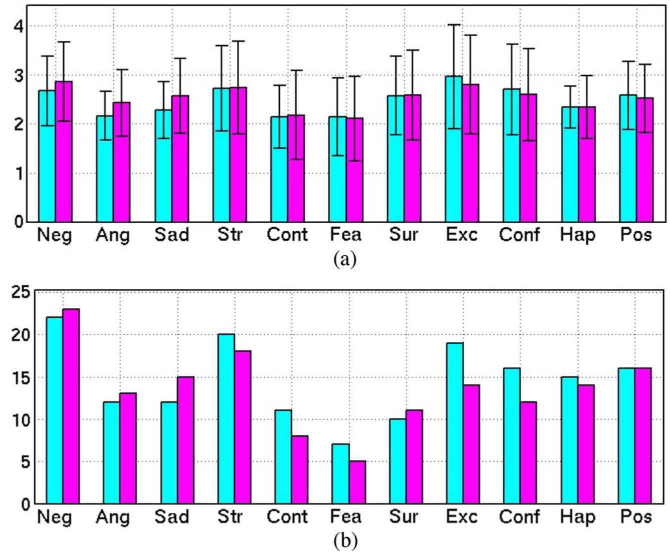


Fig. 4. Histograms showing the evaluation of all movie clips as expressed by the (cyan-left) speakers and by the (magenta-right) users of the online survey. For each emotion, the bars indicate how many times (in percentage) that particular label was given the corresponding grade shown on the x-axis.



(a)

(b)

Fig. 5. Eliciting videos evaluated by the (cyan-left) speakers and by (magenta-right) users of the online survey; for each emotional label, (a) shows mean and standard deviation of the received grades, while (b) represents the number of sentences given an average grade >3 for that label.

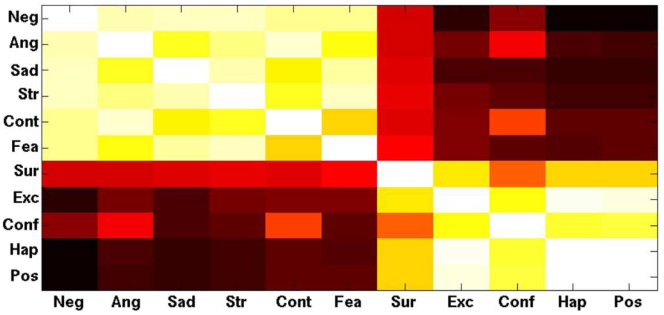


Fig. 6. Correlations between the affective adjectives, given the evaluations of the eliciting videos in the online survey. There is a high correlation (bright fields) among positive and negative emotions.

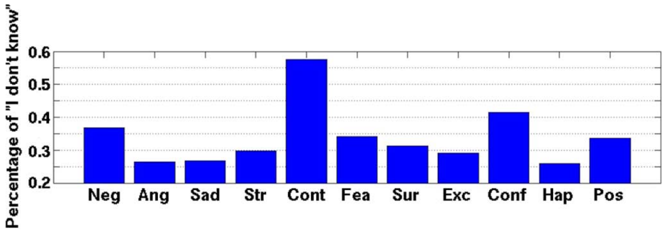


Fig. 7. For each label (on the x-axis), the number of times it was rated 0 (“I don’t know”) is shown in percentage of all the evaluations of the eliciting videos by the online survey. “Contempt” and “Confidence” were the labels of which people were least certain.

was provided, allowing the suggestion of new labels considered appropriate for the clip. The original list of labels was built starting from the six basic emotions and adding/removing labels by a preliminary screening of the eliciting videos (e.g., “Disgust” was never observed and thus removed). We do not claim that these labels represent the space of emotions in general, only that they are adequate for describing the selected video clips. The eliciting videos were also shown to a larger audience, by means of an online survey, presenting the same structure of the paper form given to the speakers. The order was randomized, allowing the user to quit the evaluation at any time. In total, 122 people took part in the survey (20.5% of which were native English speakers), labeling over 1000 video clips. The average inter-rater correlation[1] was 0.622 for the speakers and 0.646 for the online survey.

Figs. 4 and 5 compare the results of the two separate evaluations of the eliciting clips: the cyan bars on the left correspond to the answers given by the volunteers, while the magenta bars on the right to the results of the online survey. In Fig. 4, for each label, the histograms show how many times (in percentages of all movie clips) it was given the grade on the x-axis. The distributions of the grades are very similar, indicating that the laboratory environment had only a minor impact on the perception of affective states. In Fig. 5(a), for each emotional label, mean and standard deviation of its perceived strength are plotted over all sentences. Fig. 5(b) compares the number of sentences labeled as the corresponding emotion on the x-axis (i.e., with an average grade >3), giving an idea of the affective content of the eliciting videos. In general, we note a predominance of negative labels, and, for the online survey, a slightly higher standard deviation and a tendency to give higher grades to negative emotions. Fear and contempt were the least perceived affective states from our eliciting videos, while the most suggested additional labels were “Nervousness”, “Disappointment”, and “Frustration”.

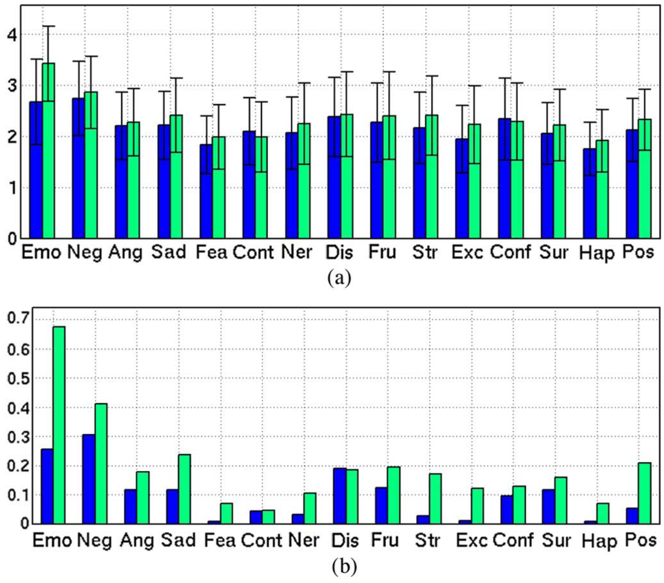
Some of the labels naturally depend on each others, as can be seen in Fig. 6, plotting the correlation between the evaluations of the online survey, where the brighter upper-left and lower-right corners indicate a high correlation among positive and negative states. Correlation can be noted between some of the basic emotions (e.g., “Sadness” and “Fear”, or “Surprise” and “Happiness”), indicating that a single label procedure based on the basic emotions would have been insufficient to describe the affective states present in our eliciting videos, and thus supporting our choice of an expanded label set.

Fig. 7 tries to judge the suggested affective states: For each label, the bar represents how many times (in percentage of all evaluations) it was given the value 0 (“I don’t know”). “Contempt” and “Confidence” were given zeros most often, possibly being the states of which the observers were least certain.

B. Corpus Evaluation

In order to assess the quality of the acquired data, videos were created containing renderings of the tracked 3-D faces and the

[1] The Pearson product-moment correlation coefficient is used throughout the paper: $\rho_{xy} = cov(X,Y)/\sigma_x\sigma_y$.



(a)

(b)

Fig. 8. Evaluations of the (blue-left, read from text) neutral and (green-right, pronounced after having watched the eliciting video) emotional sentences of the corpus. For each label, mean and standard deviation of the received grades are plotted in (a), while (b) shows the percentage of sentences given an average grade >3 for that label. The plots show that the emotional part of the corpus was indeed evaluated as such by the anonymous observers.

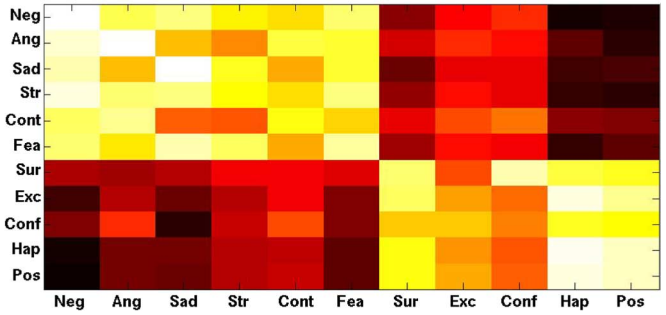


Fig. 9. Correlations between the evaluations of the (y-axis) eliciting videos and of the (x-axis) videos containing the renderings of the emotional sentences of the corpus. The correlation (bright fields) among positive and negative emotions is visible.

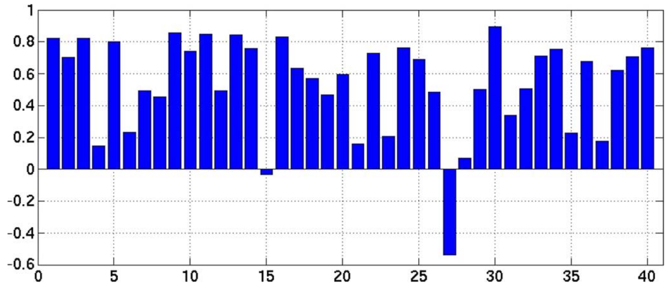


Fig. 10. For each corpus sentence, the correlation is shown between the average evaluation of the corresponding eliciting video and the average evaluation of the sentence pronounced by the speakers after watching the video. High values correspond to agreement in the evaluations.

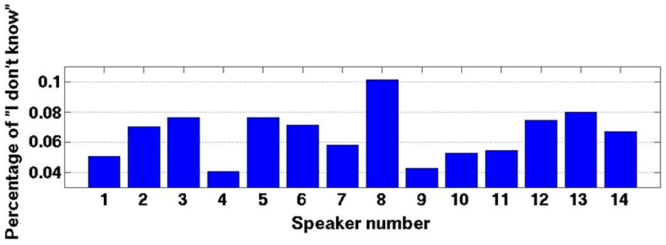


Fig. 11. Number of "I don't know" (in percentage) received by all emotional sentences pronounced by the speaker specified on the x-axis.

original audio signals. A new survey was designed, where the suggested emotional label set was enriched by the three states most commonly suggested during the evaluations of the eliciting videos ("Nervousness", "Disappointment", and "Frustration"), and by the additional label "Emotional". The anonymous users of the survey were presented with the sentences in a randomized order. The number of sequences to be judged being very large, it has not yet been possible to achieve a sensible number of evaluations for all; in the following analysis, only sentences which were rated at least three times have been considered.

The plots in Fig. 8 compare how the users of the survey (over 800 people) perceived the two parts of the corpus, i.e., the sentences (blue-left) read from text and (green-right) pronounced after watching the eliciting video. In (a), average grade and standard deviation over all sentences are given for each label, while in (b), the percentage of sentences which were given an average grade greater than 3 for the label on x-axis is shown. There is evidence of a general increase in the grades given to the emotional labels for the sentences pronounced after watching the eliciting videos, showing the effectiveness of the induction method; however, the result is unclear for labels like "Contempt" and "Confidence", supporting the intuition of Fig. 7.

Fig. 9 compares the sentences uttered after watching the eliciting videos and the videos themselves by plotting the correlation between the subset of labels shared by the two surveys. Correlation is still noticeable among positive and negative emotions, but not as much as in Fig. 6, e.g., for "Confidence". Fig. 10 shows the correlation between the evaluations of an eliciting video and the evaluations of the corresponding sentence as pronounced by the speakers after watching the video. Most of the 40 utterances show high correlation (1 means full agreement), but some specific sentences show lower agreement, notably number 27, where apparently the emotional state perceived from the video was not similar to the one conveyed by the speakers' performance. This is not surprising since the eliciting videos are longer than the sentences in the corpus and thus can more easily build the emotional states in the viewer; also the absence of eyes, facial texture, and rest of the body makes the renderings of the tracked faces less effective in conveying the emotions.

Fig. 11 shows the number of times (in percentage) "I don't know" was chosen when evaluating the emotional sentences pronounced by the speaker specified on the x-axis. Speaker number 8 was given zeros about 10% of the time, appearing to be the least effective in conveying the affective states.

C. Data Analysis

In order to perform some preliminary studies on the acquired data and demonstrate possible uses of it, we proceeded by selecting as neutral the utterances with an average grade smaller than 3 for the label "Emotional", and the ones with "Emotional" mean grade greater than 3 as the remaining affective states. The plots in Fig. 12 show the relations of the affective adjectives and

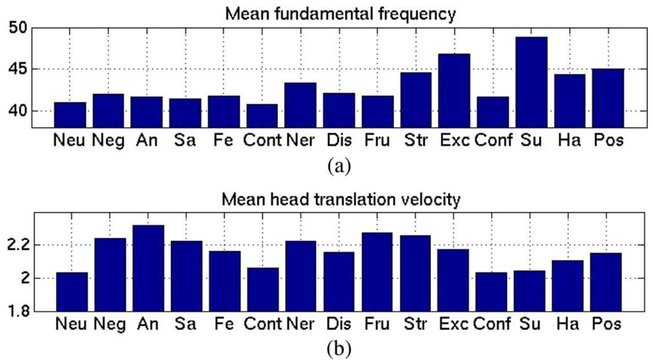


Fig. 12. (a) Fundamental frequency averaged over each proposed affective adjectives on the x-axis, i.e., over the corpus sentences which were given a mean grade >3 for that particular label. (b) Mean head translation velocity computed from sequences labeled as each of the adjectives on the x-axis.

simple audio and video features, averaged over all sequences labeled according to the above rule. In particular, Fig. 12(a) refers to the fundamental frequency, suggesting that positive emotions manifest themselves in higher values of (F_0). Fig. 12(b) shows the mean first derivatives computed over the magnitude of the rigid translations of the heads, i.e., mean velocity; it can be noted how emotional sentences present on average higher velocities, especially affective states like "Anger" and "Frustration". These plots indicate that a single feature is not enough to recognize the affective state but that already several low-level audio-visual cues can give some evidence for the affective state.

Fig. 13 demonstrates that some correlation exists between auditory and visual channel of our corpus. The plots show the correlation (over all acquired frames) between F_0, first 12 mel frequency cepstral coefficients $(m0 - m11)$, and mean Gaussian curvature calculated over the cheeks, mouth, and eyebrows regions (c_c, c_m, and c_b), for the sentences labeled as (a) "Sad", and (b) "Happy". As expected, there is strong correlation (bright areas) within features extracted from the same modality (especially for some of the audio features); however, correlation is also present between features extracted from different modalities. Note that the strength of the correlation between audio and visual features differs for the two labels.

Having at our disposal accurate phoneme segmentation and spatio-temporal correspondences among all facial scans, we can arrange the 3-D face scans into groups corresponding to particular phonemes, and thus build a statistical model of the phonemes' visual appearance (visemes). Fig. 14 shows the result of applying principal component analysis (PCA) to the scans corresponding to the phoneme "I", as uttered by the same subject. The three rows show the three main modes of variation observed in the data, with the average in the middle and the faces generated by setting the corresponding weights to -3 std. on the left, and, respectively, $+3$ std. on the right. It appears from the example that most of the variation spanned by the first modes corresponds to changes in the expressiveness of the speech (coarticulation effects are reduced to a minimum by selecting the central frame for each phone segment). This simple example shows the power of our facial representation, which paves the way for automatic visual speech synthesis and recognition.

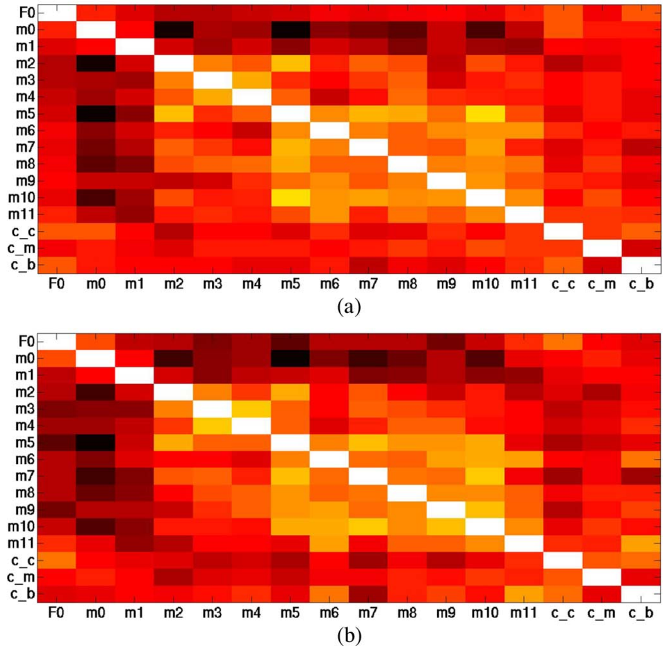


Fig. 13. Correlation between some audio features extracted from the speech signal (F_0, $m0 - m11$) and geometric features extracted from specific facial regions (c_c, c_m, and c_b) for (a) sad sentences and (b) happy sentences. Some correlation can be noted not only within the two modalities, but also between them.

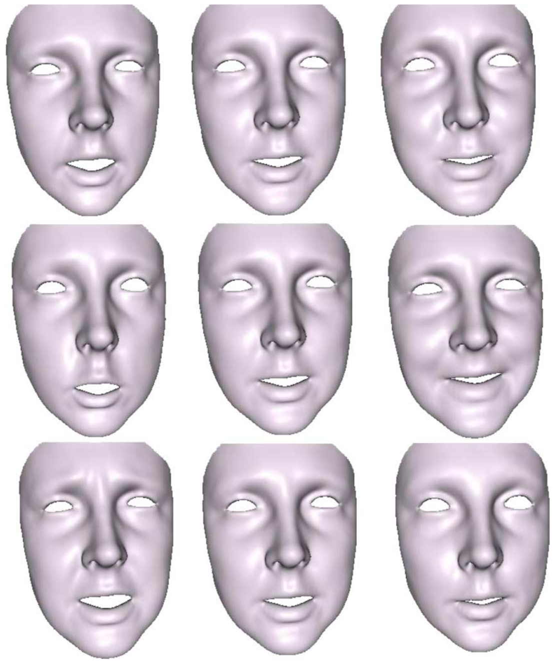


Fig. 14. First three modes of the PCA model of the phoneme "I". The middle column shows the average face, while the left and right columns represent the result of setting the mode's weight to -3 std. and $+3$ std., respectively.

VI. Discussion

In this work, we have presented a novel audio-visual corpus of affective speech and corresponding dense dynamic 3-D face geometries. The setup was designed for the acquisition of high quality data, targeting applications like visual speech modeling for synthesis and recognition purposes. The recordings of naturalistic emotions being unfeasible in the required studio environment, we resorted to eliciting videos to induce the affective states in the speakers. Our corpus stands out from all currently available datasets, which are either completely posed, limited to dynamic facial expressions, or lacking 3-D information.

The corpus comprises 1109 sentences uttered by 14 native English speakers, in the form of audio plus dense dynamic face depth data. For the speech signal, a phonological representation of the utterances, phone segmentation, fundamental frequency, and signal intensity are provided. The depth signal is converted into a sequence of 3-D meshes, providing full spatial and temporal correspondences across all sequences and speakers, a vital requirement for generating advanced statistical models targeting animation or recognition applications.

Although the evaluation shows that similar affective states are perceived by human observers when watching the eliciting videos and the processed data from the corpus, the used induction method is not a replacement of naturalism. This is the price to pay for high-quality data. Another limitation is the fact that the 3-D visual modality does not include eyes, eyelids, inner mouth, and other body parts beside the face. The raw 3-D data being part of the corpus, better templates could be used to track the faces and fill some of the above gaps.

The described corpus can be used to model coarticulation for the synthesis of emotional visual speech. Other applications include audio-visual emotion recognition, emotion-independent lip reading, or view-independent facial expression recognition. Our experiments indicate that the corpus might also be useful for studying the correlations between audio and facial features in the context of emotional speech. Because we intend to provide both the raw and the processed data, the corpus could also be used as a benchmark dataset, e.g., a comparison of different 3-D face trackers could consist in measuring the loss of information between the original and the processed data in terms of the perceived emotional content.

References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[2] R. W. Picard, "Toward computers that recognize and respond to user emotion," *IBM Syst. J.*, vol. 39, no. 3–4, pp. 705–719, 2000.

[3] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 9, pp. 52–55, 1968.

[4] C. M. Whissell*, The Dictionary of Affect in Language*, R. Plutchik and H. Kellerman, Eds. Reading, MA: Addison-Wesley, 1972.

[5] S. Frigo, "The relationship between acted and naturalistic emotional corpora," in *Proc. Corpora for Research on Emotion and Affect (LREC)*, 2006.

[6] M. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. ACM Int. Conf. Multimodal Interfaces (ICMI'07)*, New York, Nov. 2007, pp. 38–45.

[7] J. Gross and R. Levenson, "Emotion elicitation using films," *Cognit. Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[8] M. Bradley, B. Cuthbert, and P. Lang, "Picture media and emotion: Effects of a sustained affective context," *Psychophysiology*, vol. 33, no. 6, pp. 662–670, 1996.

[9] D. Clark, "On the induction of depressed mood in the laboratory: Evaluation and comparison of the velten and musical procedures," *Adv. Behav. Res. Therapy*, vol. 5, no. 1, pp. 27–49, 1983.

[10] K. Scherer, T. Johnstone, and T. Bänziger, "Automatic verification of emotionally stressed speakers: The problem of individual differences," in *Proc. Int. Workshop Speech and Computer*, 1998.

[11] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1–2, pp. 5–32, 2003.

[12] R. Westrmann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *Eur. J. Social Psychol.*, vol. 26, no. 4, pp. 557–580, 1996.

[13] P. Ekman, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[15] R. Craggs and M. M. Wood, "A two dimensional annotation scheme for emotion in dialogue," in *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[16] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," in *Proc. AFGR*, 2004.

[17] E. Velten, "A laboratory task for induction of mood states," *Behav. Res. Therapy*, vol. 6, pp. 473–482, 1968.

[18] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in French: Data, model and evaluation," *Speech Commun.*, vol. 33, no. 4, pp. 357–371, 2001.

[19] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," *Neural Netw.*, vol. 18, no. 4, pp. 371–388, 2005.

[20] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Proc. 2nd Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, Lisbon, Portugal, 2007, pp. 488–500.

[21] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag German audio-visual emotional speech database," in *2008 IEEE Int. Conf. Multimedia and Expo.*, Apr. 23–26, 2008, pp. 865–868.

[22] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, no. 1–2, pp. 33–60, 2003.

[23] J. C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, "Manual annotation and automatic image processing of multimodal emotional behaviors," *Pers. Ubiq. Comput.*, vol. 13, no. 1, pp. 69–76, 2009.

[24] U. Türk, The Technical Processing in Smartkom Data Collection: A Case Study, LMU Munich, Jul. 2001, Tech. Rep.

[25] A. Zara, V. Maffiolo, J.-C. Martin, and L. Devillers, "Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics," in *Proc. ACII*, 2007, pp. 464–475.

[26] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," in *Proc. ACII*, 2007, pp. 476–487.

[27] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Soc. Psychol.*, vol. 70, pp. 614–636, 1996.

[28] L. S.-H. Chen, "Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction," Ph.D. dissertation, Univ. Illinois at Urbana-Champaign, Champaign, IL, 2000.

[29] P. Ekman and W. Friesen*, Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists, 1978.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, pp. 335–359, 2008.

[31] Y. Cao, W. C. Tien, P. Faloutsos, and F. H. Pighin, "Expressive speech-driven facial animation," *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302, 2005.

[32] K. Wampler, D. Sasaki, L. Zhang, and Z. Popovic, "Dynamic, expressive speech animation from a single mesh," in *Proc. Symp. Computer Animation*, 2007, pp. 53–62.

[33] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. FG*, 2008.

[34] T. Weise, B. Leibe, and L. V. Gool, "Fast 3D scanning with automatic motion compensation," in *Proc. IEEE CVPR*, 2007.

[35] T. Weise, H. Li, L. V. Gool, and M. Pauly, "Face/off: Live facial puppetry," in *Proc. SCA*, 2009.

[36] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective Information Processing*. New York: Springer, 2008.

[37] H. Romsdorfer, "Polyglot text-to-speech synthesis. Text analysis and prosody control," Ph.D. dissertation, Comput. Eng. Netw. Lab., ETH Zurich (TIK-Schriftenreihe Nr. 101), Zurich, Switzerland, Jan. 2009, No. 18210.

Gabriele Fanelli received the Laurea (M.Sc.) degree in computer engineering from the University of Rome Sapienza, Rome, Italy, in 2006, after developing his master's thesis at the Image Coding Group, Linköping University, Linköping, Sweden. He is currently pursuing the Ph.D. degree at the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland.

Since 2007, he has been a Research Assistant and at the Computer Vision Laboratory, ETH Zurich. His research interests include facial features detection and tracking, audio-visual speech recognition, and visual speech synthesis.

Juergen Gall received the B.Sc. degree from the University of Wales Swansea, Swansea, U.K., in 2004, the M.Sc. degree in mathematics from the University of Mannheim, Mannheim, Germany, in 2005, and the Ph.D. degree in computer science from the Saarland University and the Max-Planck-Institut für Informatik, Saarbrücken, Germany, in 2009.

He worked for the Machine Learning and Perception group at Microsoft Research Cambridge, Cambridge, U.K., in 2008. Since 2009, he has been a Postdoctoral Researcher at the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland. His research interests include interacting particle systems, markerless human motion capture, object detection, and action recognition.

Harald Romsdorfer (M'09) received the M.Sc. degree in electrical engineering from Vienna University of Technology, Vienna, Austria, in 2001 and the Ph.D. degree in electrical engineering from ETH Zurich, Zurich, Switzerland, in 2009.

During his time at ETH Zurich, he developed the worldwide first polyglot text-to-speech synthesis system polySVOX. Since 2009, he has been a Senior Research Associate at Graz University of Technology, Graz, Austria, where he is pursuing research in the area of microphone array-based speech processing. In 2007, he founded a company specialized on electronic circuit design in Gmunden, Austria, where he was CEO from 2007 to 2009. Since 2009, he is CEO and founder of SYNVO in Zurich, a company developing a new generation of text-to-speech solutions.

Thibaut Weise received the M.Eng. degree in computing from Imperial College London, London, U.K., in 2005 and the Ph.D. degree in electrical engineering from ETH Zurich, Zurich, Switzerland, in 2009.

In 2010, he joined the Computer Graphics and Geometry Laboratory at EPFL, Lausanne, Switzerland, as a Postdoctoral Researcher. His main research interests include 3-D scanning, modeling, and animation, with a strong focus on faces.

Luc Van Gool (M'95) received the electro-mechanical engineering degree from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1981.

Currently, he is a Professor at the Katholieke Universiteit Leuven and the ETH in Zurich, Switzerland. He leads computer vision research at both places, where he also teaches computer vision. He has authored over 200 papers in this field. He has been a program committee member of several major computer vision conferences. His main interests include 3-D reconstruction and modeling, object recognition, and tracking and gesture analysis. He is a co-founder of five spin-off companies.

Prof. Van Gool received several Best Paper awards.