3D Vision Technology for Capturing Multimodal Corpora: Chances and Challenges

Gabriele Fanelli[1], Juergen Gall[1], Harald Romsdorfer[2], Thibaut Weise[3], Luc Van Gool[1]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]Signal Processing & Speech Communication Laboratory, Graz University of Technology, Austria
[3]Laboratoire d'Informatique Graphique et Géométrique, EPFL Lausanne, Switzerland
{gfanelli, gall, vangool}@vision.ee.ethz.ch, romsdorfer@tugraz.at, thibaut.weise@epfl.ch

Abstract

Data annotation is the most labor-intensive part for the acquisition of a multimodal corpus. 3D vision technology can ease the annotation process, especially when continuous surface deformations need to be extracted accurately and consistently over time. In this paper, we give an example use of such technology, namely the acquisition of an audio-visual corpus comprising detailed dynamic face geometry, transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. By means of the example, we will discuss the advantages and challenges of integrating non-invasive 3D vision capture techniques into a setup for recording multimodal data.

1. Introduction

Multimodal corpora are an important resource for studying and analyzing the principles of human communication. Besides speech, the visual modality encodes probably the most important cues for humans to perceive communicative behavior like hand gesture, facial expression, or body posture. The recent efforts to collect audio-visual corpora are reflected in the literature, see (Zeng et al., 2009; Douglas-Cowie et al., 2007; Cowie et al., 2005), for instance.

The most labor-intensive part for acquiring a multimodal corpus is the annotation of the data, in particular for the visual modality. Although there are coding schemes like the Facial Action Coding System (FACS) (Ekman and Friesen, 1978) or annotation tools like ANVIL (Kipp, 2008) supporting the manual labeling of video sequences, the commonly used 2D video recordings inherently lead to a loss of information by projecting the 3D content onto the 2D image plane and discarding the depth information. Fortunately, 3D displays and capture technologies have emerged as commercial products. One example are invasive methods that place markers on the human body or face to capture the movement. However, the attachment of markers is not only time-consuming and often uncomfortable for the subject to wear, it can also significantly change the pattern of motion (Fisher et al., 2003). In general, invasive methods cannot be used for the acquisition of authentic data. Non-invasive vision systems provide a valuable alternative solution. Similar to the human vision system, they capture the scene with two or more 2D sensors to obtain the depth information. Current systems estimate the 3D surface deformation of the human body (Gall et al., 2009) or the face (Weise et al., 2009a) with minimal manual effort, as shown in Figure 1. In contrast to discrete annotation schemes like FACS which divides the state space into 32 facial muscles activation units, the continuously captured 3D data streams do not suffer from quantification artifacts as they capture the strength of the deformation accurately and consistently over time.

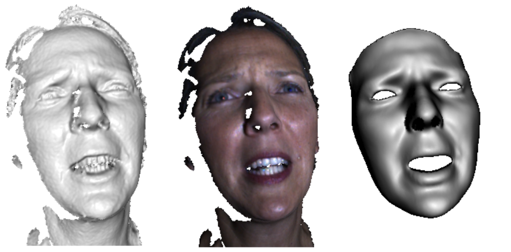

Figure 1: From left to right, the image shows the 3D reconstruction of a person's face, the corresponding texture mapped on it, and the spatio-temporal consistent representation of the face surface.

In this paper, we give an example for the use of 3D vision technology for the acquisition of an audio-visual corpus that comprises detailed dynamic face geometry, transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. By means of the example, we will discuss the advantages and challenges of integrating non-invasive 3D vision capture techniques into a setup for recording multimodal data. The corpus will be released for research purposes at the end of the year.

2. 3D Face Capture System

Our goal is to capture the face geometry of a speaker over time and represent the data such that it serves as annotation for the visual modality. The first step requires the reconstruction of the depth map of the subject's face for each frame, as shown in Figure 1. To this end, we employ the 3D scanner described in (Weise et al., 2007). The system combines stereo and active illumination based on phase-shift for robust and accurate 3D scene reconstruction. Stereo overcomes the traditional phase discontinuity problem and motion compensation is applied in order to remove artifacts in the reconstruction. The system consists of two high-speed grayscale cameras, a color camera, and a DLP projector without the 4-segment color wheel (RGBW), so that it projects three independent monochrome images at 120 Hz (sent as the R, G, and B channel). The two monochrome cameras are synchronized and record the three images. The color camera is also synchronized, but uses a longer exposure to integrate over all three projected images and thus capture the texture. In our current setting, the system is very stable at 25 Hz although the recording could be performed at up to 40 Hz. The acquisition rate is a limitation of many vision systems since higher frame rates require short exposure times and thus very bright light. Movements which are faster than the acquisition rate cannot be captured, nevertheless, the system accurately captures texture and the depth map of the face, as shown by the first two images in Figure 1.

The depth map of a face is a cloud of 3D points that cannot be directly used for analyzing changes in the face geometry as this requires full spatial and temporal correspondences of the 3D data. For example, a point of the left eyebrow should be also part of the left eyebrow for all captured faces, i.e., 3D points should maintain their semantic meaning among different scans. To achieve this goal, we use the two-step procedure introduced in (Weise et al., 2009a): First, a generic template mesh is warped to the reconstructed, expressionless, 3D model of the speaker. Second, the resulting personalized template is automatically tracked throughout all recorded sequences of the same speaker. The template ensures spatial consistency over time and between different subjects.

Preprocessing In order to build a person-specific face template, each speaker is asked to turn the head with a neutral expression and as rigidly as possible in front of the real-time 3D scanner. The sequence of 3D scans is registered and integrated into one 3D model using the online modeling algorithm proposed in (Weise et al., 2009b). Small deformations arising during head motion violate the rigidity assumption, but in practice do not pose problems for the rigid reconstruction. Instead of using the reconstructed 3D model directly, a generic face template is warped to fit it. Besides enabling a hole-free reconstruction and a consistent parameterization, using the same generic template has the additional benefit of providing full spatial correspondence between different speakers.

Warping the generic template to the reconstructed 3D model is achieved by means of non-rigid registration, where for each mesh vertex $\mathbf{v}_i$ of the generic template a deformation vector $\mathbf{d}_i$ is determined in addition to a global rigid alignment. This is formulated as an optimization problem, consisting of a smoothness term minimizing bending of the underlying deformation (Botsch and Sorkine, 2008), and a set of data constraints minimizing the distance between the warped template and the reconstructed model. As the vertex correspondences between generic template and reconstructed model are unknown, closest point correspondences are used as approximation similarly to standard rigid iterative closest point registration. A set of manually labeled correspondences are used for the initial global alignment and to initialize the warping procedure. The landmarks are mostly concentrated around the eyes and mouth, but a few correspondences are selected on the chin and forehead to match the global shape. The manual labeling needs to be done only once per speaker and takes at most a couple of minutes. The resulting personalized template accurately captures the facial 3D geometry of the corresponding person.

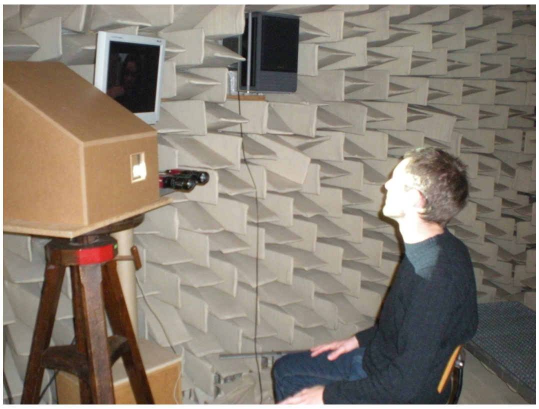

Figure 2: Recording setup: one speaker sits in front of the 3D scanner in the anechoic room while watching a video for instructions on the screen.

The diffuse texture map of the personalized template is automatically extracted from scans where the subject moves the head rigidly, by averaging the input textures. The face is primarily illuminated by the 3D scanner, and we can therefore compensate for lighting variations using the calibrated position of the projection. Surface parts that are likely to be specular are automatically removed. The reconstructed texture map is typically oversmoothed, but sufficient for the tracking stage.

3D Tracking The personalized face template is used to track the facial deformations of each performance. For this purpose, non-rigid registration is employed, in a similar manner as during the template creation. In this case, the distances between the template vertices and the 3D scans are minimized. To ensure temporal continuity, optical flow constraints are also included in the optimization, as the motion of each vertex from frame to frame should coincide with the optical flow constraints. During speaking, the mouth region deforms particularly quickly, and non-rigid registration may drift and ultimately fail. This can be compensated for by employing additional face-specific constraints such as explicitly tracking the chin and mouth regions, making the whole process more accurate and robust to fast deformations. Figure 1(right) shows a personalized model adapted to a specific frame of a sequence.

3. Example: Audio-Visual Corpus

In order to acquire and annotate data for an audio-visual corpus, we have integrated the 3D face capture system into a setup for recording audio data. To obtain clean audio data, the setup is placed in an anechoic room with walls covered by sound wave-absorbing materials. For the audio recordings, we use a studio condenser microphone placed in front of the speaker. Since the cooling fans of the projector for the 3D face capture system become very noisy, we have built a wooden enclosure, open at the back to allow the heat out and equipped with a non-reflecting glass at the front to emit the light. The enclosure reflects most of the noise

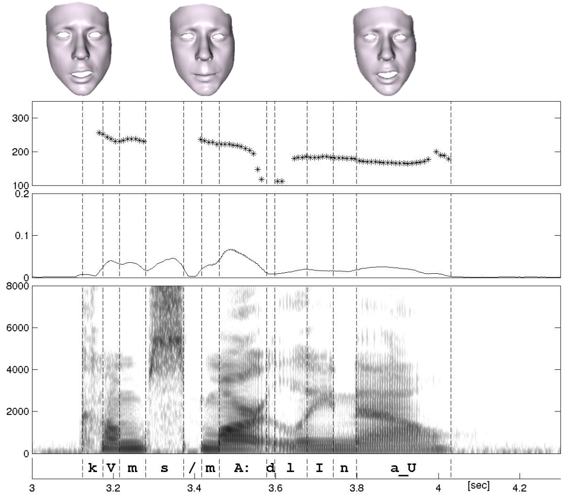

Figure 3: Phone sequence, spectrogram, signal intensity contour, and fundamental frequency contour of the speech signal, plus sample faces are shown from bottom to top.

to the back where it is absorbed by the walls. The signal-to-noise ratio (SNR) of the recorded audio stream is high enough to perform automatic speech segmentation, which is described in Section 3.2. Figure 2 shows the setup, with a volunteer being scanned while watching a video for instructions on the screen. The computers for processing the data are placed in a separate room, from where the system can be fully controlled.

3.1. Corpus

Our corpus currently comprises 40 short English sentences spoken by 14 native speakers (8 females and 6 males, aged between 21 and 53) who volunteered to have their voice and facial movements recorded. Each person was required to sit alone in the anechoic room and asked to pronounce each sentence. For synchronization purposes, the volunteers clapped their hands in front of the cameras before uttering the sentences. The clapping is automatically detected in the audio stream and used for cutting the 3D video stream. This introduces a maximum temporal discrepancy of 20 ms between the audio and video stream. A hardware synchronization could improve the accuracy. We have also recorded a corpus for affective speech, described more in detail in (Fanelli et al., 2010).

3.2. Audio Annotation

The auditory modality of the corpus is best annotated by means of the speech prosody and the sequence of phones. Speech prosody can be described at the perceptual level in terms of pitch, sentence melody, speech rhythm, and loudness. The physically measurable quantities of a speech signal are the following acoustic parameters: fundamental frequency (F_0), segment duration, and signal intensity. F_0 correlates with pitch and sentence melody, segment duration with speech rhythm, and signal intensity with loudness. Figure 3 shows an example of the annotated audio-visual corpus.

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 extraction, and signal intensity estimation are achieved by analyzing the speech data. For the extraction of these prosodic quantities, we applied fully automatic procedures provided by SYNVO Ltd. In the following, we give an overview of the extraction procedures for fundamental frequency, signal intensity, and segment duration.

vowels	i: I U u: e @ q 3 3: V O: A A: Q
diphthongs	@_U a_I a_U e_I E_@ I_@ O_I o_U U_@
consonants	p p_h b t t_h d k k_h g m n N r f v T D s z S Z x h j w l
affricates	t_S d_Z
pauses	c_u c_v /

Table 1: Segment types of English phones and speech pauses used for transcription of the speech data of the audio-visual corpus.

Transcription The phonological representation contains the sequence of phones for the sentences in the audio-visual corpus, the stress level of syllables, the position and strength of phrase boundaries, plus the indicators of phrase types. Initial phonological representations of the sentences are obtained by applying the transcription component of the SYNVO text-to-speech synthesis system to the text version of the corpora. See (Romsdorfer and Pfister, 2007) for a description of such a transcription component. These initial phonological representations contain the standard phonetic transcription of the sentences (Table 1).

The phonological information (phrase type, phrase boundary, and sentence accentuation) of these automatically generated representations is then adapted to the speech signals. Neural network-based algorithms are employed for automatic phrase type, phrase boundary, and syllable accent identification. Detailed information on this procedure can be found in (Romsdorfer, 2009).

Fundamental Frequency Extraction F_0 values of the natural speech data of the prosody corpus are computed every 10 ms using a pitch detection algorithm based on combined information taken from the cepstrogram, the spectrogram, and the autocorrelation function of the speech signal, cf. (Romsdorfer, 2009). Signal sections judged as unvoiced by the algorithm are assigned no F_0 values.

Signal Intensity Extraction Signal intensity values of the natural speech data are computed every 1 ms. The root mean square value of the signal amplitude calculated over a window of 30 ms length is used.

Segment Duration Extraction An accurate extraction of phone and speech pause durations requires an exact segmentation of the natural speech data of the audio-visual corpus into adjacent, non-overlapping speech or pause segments, and a correct assignment of labels to these segments indicating their type.

Since the phonological representation contains the standard phonetic transcription of an utterance, it is convenient to use such transcription for automatic segmentation and labeling. However, a close phonetic transcription, indicating pronunciation variants made by the speaker, results in a much better segmentation and labeling.

Segment Types Segment types correspond to the phone types determined in the transcription. Plosives are additionally segmented into their hold and burst parts. While the burst part of a plosive is denoted by the same symbol used for the plosive phone type, a "c" denotes the hold part, also called closure or preplosive pause. Speech pauses corresponding to phrase boundaries are labeled with the symbol "/". For a plosive following a speech pause, no preplosive pause is segmented. Table 1 lists all segment types used for the transcription of natural speech data.

4. Chances and Challenges

State-of-the-art 3D vision technology opens new opportunities for acquiring and annotating the visual modality of multimodal corpora with minimal manual effort. We have shown by means of an example that 3D capture devices can be integrated into a setup for multimodal data acquisition. The problem of inference between the devices can often be solved sufficiently. In our example for instance, the microphone was placed outside of the field of view of the cameras and the visual capture devices were modified for noise reduction. A hardware synchronization between the capture devices for the various modalities is desirable although it is currently not implemented in our example where only the vision components are synchronized.

In contrast to marker-based systems, advanced multi-camera vision systems are non-invasive and consequently better suitable for capturing authentic data. Compared to 2D labeling schemes, they automate most of the annotation process, significantly reducing the costs for corpora acquisition. Due to the current trend of developing 3D hardware technology like cameras and displays for consumer applications, it's very likely that the prices for ready-to-use systems will drop over the next few years. The probably most appealing property of 3D vision technology is the ability to capture continuous surface deformations accurately and consistently over time. This provides a richer source of information than discrete annotation schemes.

However, there are also several limitations that need to be pointed out. The accuracy of the vision components usually degenerates at bad lighting conditions. In particular, arbitrary outdoor environments are very challenging for most systems. The speed of motion that can be captured is limited to the acquisition frame rate, which is typically in the range of 25 Hz and 40 Hz. The accuracy of non-invasive methods does not match yet the accuracy of marker-based systems. For instance, the 3D reconstruction shown in Figure 1 is sometimes noisy for the eyelids and the teeth which can result in errors around the eyes and for the estimated mouth shape. Resolving these problems is a challenging task for the future.

5. Acknowledgements

This work was partly funded by the SNF fund "Vision-supported speech-based human machine interaction".

6. References

M. Botsch and O. Sorkine. 2008. On linear variational surface deformation methods. *IEEE Trans. on Visualization and Computer Graphics*, 14:213–230.

R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388.

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. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Int. Conf. on Affective Computing and Intelligent Interaction*.

P. Ekman and W. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. 2010. Acquisition of a 3d audio-visual corpus of affective speech. Technical Report 270, ETH Zurich, January.

D. Fisher, M. Williams, and T. Andriacchi. 2003. The therapeutic potential for changing patterns of locomotion: An application to the acl deficient knee. In *ASME Bioengineering Conference*.

J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

M. Kipp. 2008. Spatiotemporal coding in anvil. In *Int. Conf. on Language Resources and Evaluation*.

H. Romsdorfer and B. Pfister. 2007. Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Communication*, 49(9):697–724.

H. Romsdorfer. 2009. *Polyglot Text-to-Speech Synthesis. Text Analysis and Prosody Control*. Ph.D. thesis, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January.

T. Weise, B. Leibe, and L. Van Gool. 2007. Fast 3d scanning with automatic motion compensation. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

T. Weise, H. Li, L. Van Gool, and M. Pauly. 2009a. Face/off: Live facial puppetry. In *Symposium on Computer Animation*.

T. Weise, T. Wismer, B. Leibe, and L. Van Gool. 2009b. In-hand scanning with online loop closure. In *IEEE Int. Workshop on 3-D Digital Imaging and Modeling*.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58.