K-007

Communicating People Identification in Multimedia Streams - An Audiovisual Fusion Approach -

Tomasz M. Rutkowski*, Victor V. Kryssanov†, Koh Kakusho*, Michihiko Minoh*

1. Introduction

This paper discusses the problem of identifying the locations of communicating people in audiovisual recordings. The main goal of the presented study is thus to localize communicative/interactive events in multimedia (audio and video) information streams. Social characteristics of the communicators are estimated to analyze the communication situation as a whole. This allows us to combine various results obtained in social science and multimedia signal processing into one coherent framework. The ultimate goal of the presented study is to build a system, in which a computer could identify the communicators' position, and then track and evaluate the communication situation from a captured video. We employ the hybrid linear/transactional communication model that is linear in short time windows corresponding to the turn-taking and interlaced behavior. The turn-taking (role changing) between the communicators is a critical assumption in this model, and it allows us to identify communicating people in multimedia streams. The case of intentional communication is considered, which occurs when all the communicators are willing to interact. We introduce a measure of the communication efficiency as a combination (fusion) of mutual information estimates between two visual, two audio, and two pairs of audiovisual streams. Experiments are made, and the results obtained support our hypothesis about the possibility of the identification of communicating people based on the audiovisual fusion analysis.

2. The Approach

The hybrid linear/transacitonal communication model describes the communicators' audiovisual behavior and to their interactive behavior during a conversation [1]. The synchronization and interaction measures (efficiency-like) developed in our previous research on human communication can be used for the purposes of this study [6]. The applied communication model is linear in short time windows. The active (in short time windows) communicator - *the sender* - is supposed to generate more audiovisual flow with breaks, when the receiver responds. The passive (in short time windows) communicator - *the receiver* - is expected, on the other hand, to react "properly", not disturbing (overlapping with) the sender's communication activity. The turn-taking (role changing) between the sender and the receiver is a critical assumption in the hybrid model. The observation area is captured in our approach with two cameras and two stereophonic microphones. We propose a measure of the communication efficiency as a combination of four mutual information estimates between two visual (V_i), two audio (A_i), and two pairs of audiovisual streams (A_i, V_i). First, the two mutual information estimates are evaluated for selected video regions of interest (ROI), where communicators may be present, and also for speech occurrences, as follows:

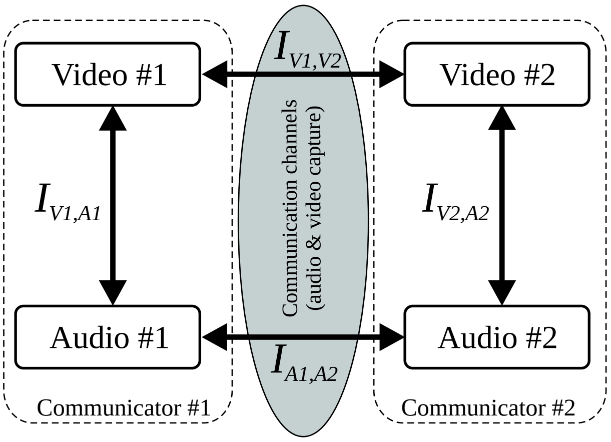

Figure 1: Communicating people identification. Mutual information estimates $I_{A1,V1}$ and $I_{A2,V2}$ between audio and visual streams of localized communicators account for the local synchronization, which permits us to identify communicating people. $I_{A1,A2}$ and $I_{V1,V2}$ are used to confirm the interactive activity.

$$I_{Ai,Vi} = \frac{1}{2} \log \frac{|R_{Ai}|\,|R_{Vi}|}{|R_{Ai,Vi}|}, \tag{1}$$

where $i \in \{1,2\}$, and R_{Ai}, R_{Vi}, and $R_{Ai,Vi}$ stand for the estimates of the corresponding covariance matrices. Next, the two mutual information estimates indicating simultaneous activity in same audio and video modes are calculated:

$$I_{\alpha1,\alpha2} = \frac{1}{2} \log \frac{|R_{\alpha1}|\,|R_{\alpha2}|}{|R_{\alpha1,\alpha2}|}, \tag{2}$$

where $[\alpha1, \alpha2] \in \{[V1, V2], [A1, A2]\}$, and $R_{\alpha1,\alpha2}$ are the empirical estimates of the corresponding covariance matrices related to different communicator activities. A conceptual graph illustrating this idea is shown in Figure 1. $I_{A1,V1}$ and $I_{A2,V2}$ evaluate the local synchronicity between the audio (speech) and video (mostly facial movements) flows from observed communicators. It is assumed that the sender has a higher synchronicity reflecting a higher activity. $I_{V1,V2}$ and $I_{A1,A2}$ are to detect possible crosstalks in same modalities (i.e. audio-audio and video-video) of the communicators. The latter pair is also used to detect possible overlapping in the activities, which would have a negative impact on the communication quality.

Since the recorded audio usually contains a lot of redundant information, a feature extraction procedure combined with compression should be performed. As speech is usually indispensable in communication, we borrow techniques from the speech recognition field [2].

[*] Academic Center for Computing and Media Studies, Kyoto University, Japan

[†] College of Information Science & Engineering, Ritsumeikan University, Japan

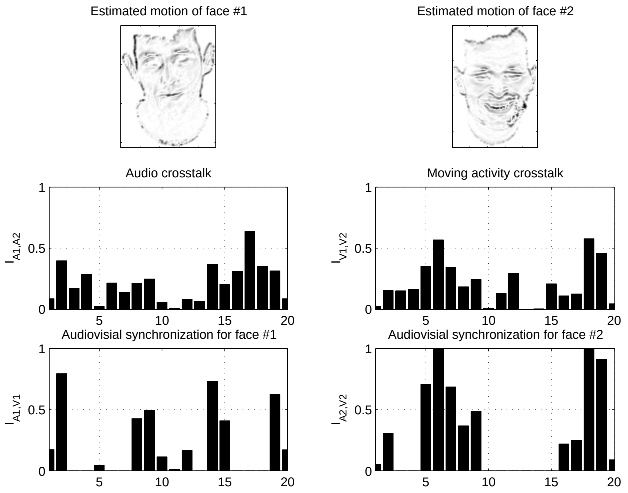

Figure 2: Application of the identification algorithm to to a real conversation. The two top plots show the motion features extracted from two facial areas of segmented videos. The audiovisual features responsible for local speech-related face motion of each person are plotted as $I_{A1,V1}$ and $I_{A2,V2}$. Plots $I_{A1,A2}$ and $I_{V1,V2}$ show the cross audio-audio and video-video synchronizations.

Specifically, we use the mel-frequency cepstrum coefficients (MFCC), since they represent important characteristics of the speech [2]. In the case of video, we have to obtain features compatible with the audio representations, which would carry information about the communication-related motion. To detect communicating humans in video, we combine search for faces and search for moving contours techniques. The visual flow of a captured scene is obtained as follows. For two consecutive movie frames, the temporal gradient is expressed as a smoothed difference between images convoluted with a two-dimensional Gaussian filter with an adjusted standard deviation [6]. The absolute value is taken to remove the gradient directional information, and to improve the movement capture. For the face tracking, two features are used: the skin color tracking (unfortunately, very sensitive to illumination variations) [7], and the moving face pattern using the non-negative matrix factorization method proposed in [4]. Features extracted in this way are localized and correspond to the intuitively perceived parts of communicators (i.e. face, eyes, nose, mouth contours, etc.). In the facial area, we can extract motion features that together with the audio information permit us to classify a person as talking, listening, or responding. To extract only the most significant information from motion frames and to compress it, we perform a two-dimensional digital cosine transformation (DCT), and only the first 24 DCT coefficients are used. When we consider the above feature set as independent samples with a multivariate probability distribution $p(A,V)$, an appropriate measure of audiovisual synchronicity or asynchronicity is the mutual information $I_{A,V}$ between random variables A and V. Since the probability distributions of A and V are unknown, an assumption about continuous distributions is made [3]. The feature vectors MFCC and DCT are, in this case, considered as samples with locally Gaussian multivariate distribution $p(A,V)$. The mutual information between visual and audio streams can then be estimated, using formulas (1)-(2). For practical reasons, we simplify the mutual information estimation procedure. A blind decorralation algorithm that was designed to minimize the shared mutual information among input signals is used [5]. The algorithm gradually decorrelates the output signals, leaving the possible crosstalk information trace inside the *demixing* matrix. The weights in the demixing matrix reflect the correlation between the audiovisual feature sets in different time windows for the chosen modalities. The higher the correlation between the audiovisual features for every member, the better the synchronization. The efficient communication process requires communicators should not talk at the same time and, in this case, the crosstalk coefficients are low. The adoption of the deccoralation algorithm permits us to find and track possible crosstalks between the communication channels.

3. Results and Conclusions

The employed hybrid communication model together with the four mutual information estimates allowed us to identify communicating people in recorded video streams, as shown in Figure 2. The audiovisual synchronicity evaluated in mutual information estimates $I_{Ai,Vi}$ is used to identify actively talking people, while the unimodal mutual information estimates $I_{A2,A2}$ and $I_{V1,V2}$ identify the interacting communicators, based on the interlaced behavior principle.

Thus the proposed audiovisual fusion approach proved reliable for localizing communicative events in multimedia streams, and it can be applied in various multimedia systems.

References

[1] R.B. Adler and G. Rodman. *Undestanding Human Communication.* Oxford University Press, 2003.

[2] C. Becchetti and L.P. Ricotti. *Speech Recognition.* John Wiley & Sons, Inc., Great Britain, 1999.

[3] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[4] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, October 21 1999.

[5] T.M. Rutkowski, A. Cichocki, and A.K. Barros. Speech enhancement using adaptive filters and independent component analysis approach. In *Proceedings of International Conference on Artificial Intelligence in Science and Technology, AISAT2000*, pages 191–196, Hobart, Tasmania, December 17–20 2000.

[6] T.M. Rutkowski, S. Seki, Y. Yamakata, K. Kakusho, and M. Minoh. Toward the human communication efficiency monitoring from captured audio and video media in real environments. In *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, pages 1093–1100, Oxford, UK, September 3–5 2003. Springer Verlag.

[7] L. Sigan, S. Sclaroff, and V. Athitsos. Estimtion and prediction of evolving color distributions for skin segmentation under varying illumination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, volume 2, pages 152–159. IEEE, 2000.