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/transacitonal 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:

\[
I_{A_i,V_i} = \frac{1}{2} \log \frac{|R_{A_i}| |R_{V_i}|}{|R_{A_i,V_i}|}, \tag{1}
\]

where \(i \in \{1, 2\}\), and \(R_{A_i}, R_{V_i}, \) and \(R_{A_i,V_i}\) 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:

\[
I_{\alpha_1,\alpha_2} = \frac{1}{2} \log \frac{|R_{\alpha_1}| |R_{\alpha_2}|}{|R_{\alpha_1,\alpha_2}|}, \tag{2}
\]

where \([\alpha_1,\alpha_2] \in \{[V_1,V_2], [A_1, A_2] \}\), and \(R_{\alpha_1,\alpha_2}\) 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. \(I_{A_1,V_1}\) and \(I_{A_2,V_2}\) 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. \(I_{V_1,V_2}\) and \(I_{A_1,A_2}\) 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].

![Figure 1: Communicating people identification. Mutual information estimates \(I_{A_1,V_1}\) and \(I_{A_2,V_2}\) between audio and visual streams of localized communicators account for the local synchronization, which permits us to identify communicating people. \(I_{A_1,A_2}\) and \(I_{V_1,V_2}\) are used to confirm the interactive activity.](image-url)
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 decorrelation 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 decorrelation 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 \( I_{A_1, V_1}, I_{A_2, V_2} \) is used to identify actively talking people, while the unimodal mutual information estimates \( I_{A_1, V_1}, I_{A_2, V_2} \) 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