Multimodal Analysis of Upper-Body Gestures, Facial Expressions and Speech

Serkan Özkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez, Engin Erzin
MVGL, Koç University
Rumelifeneri Yolu, Sarıyer, Istanbul, Turkey
E-mail: serozkul@ku.edu.tr, ebozkurt@ku.edu.tr, sasta@ku.edu.tr, yyemez@ku.edu.tr, eerzin@ku.edu.tr

Abstract

We propose a multimodal framework for correlation analysis of upper body gestures, facial expressions and speech prosody patterns of a speaker in spontaneous and natural conversation. Spontaneous upper body, face and speech gestures exhibit a broad range of structural relationships and have not been previously analyzed together to the best of our knowledge. In this study we present a multimodal database of spontaneous conversation. Within this database to identify cross modal correlations, we first perform unsupervised temporal segmentation of each modality using hidden Markov model (HMM) structures. Then, we perform correlation analysis based on mutual information measure, which is experimentally computed over the joint histograms of recurrent temporal segments (patterns) of modalities.

Keywords:

1. Introduction

An ideal system for automatic analysis and recognition of human affective information should be multimodal, as the human sensory system is. The integration of multiple sources of information would enhance the power for achieving a reliable emotional recognition; hence building multimodal databases is considered a very important issue for affective computing research. While this need is clearly acknowledged within the research community, very few large multimodal databases are available. Most of the databases deal only with speech or facial expressions and even when considering few more complete multimodal databases available they mostly combine audio and visual (facial expression) information, very few add upper-body gesture information.

Moreover, naturalness of the emotional database is another issue. Acquiring realistic emotional data is a challenging task. In fact, many of the databases available ask the subjects to act or pose emotions in order to extract speech and facial related features. In the last years, this lack of naturalism has been severely criticized. Recent research is oriented towards inducing emotions of speakers (elicited databases) or collecting real-life data.

There have been a large number of studies on emotion and non-verbal communication of facial expressions and also on expressive body movements. Yet, these studies were mostly based on acted basic emotions. HUMAINE database is one of the most comprehensive multimodal databases, which was collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. Acted emotional state recordings of anger, despair, interest, pleasure, sadness irritation, joy and pride incorporated facial expressions, body movements and gestures and speech. The database is also multi-lingual including recordings in languages English, French and Hebrew [1]. In another study, Gunes and Piccardi analyze upper body gestures during six acted emotional behaviors [2].

Some of the most successful efforts to collect new emotional databases have been based on broadcasted television programs. Some of these examples are the Belfast natural database [3], the VAM database [4] and the EmoTV1 database [5]. Likewise, movie excerpts with expressive content have also been proposed for emotional corpora, especially for extreme emotions (e.g., SAFE corpus [6]).

2. System Architecture of Multimodal Database Collection

Building the data corpora is crucial part of the applied scientific studies. The data corpus should provide the means for understanding all the aspects of a given process. It should also direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a well-designed data corpus (i.e. capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results. In this paper we present in detail the process of building our multimodal data corpus in the Multimedia, Vision and Graphics Laboratory (MVGL) of Koç University.

2.1 Recording Equipment

At MVGL, an automated multi-camera motion capture system is available to collect and analyze multi-view video data, which is primarily used for human body motion modeling applications. The motion capture system that we developed is based on 3D tracking of the markers attached to the person's body in the scene by making use of the multi-stereo correspondence information from multiple cameras. Portable camera, Drift HD170, is used as head attached camera to record facial expressions. The Drift HD170 supports 1080p HD recording at 30 frames per second and weighs 138g.

2.2 Recording Settings

We use the four cameras of the MVGL 8-camera optical

motion capture system, which are placed on a truss structure on the ceiling, positioned three meters away to capture the upper body movements of the participant in all directions for 3D tracking. These cameras are placed like a curve of a semicircle to prevent any marker getting out of the field of vision. This setting also increases the multi-camera calibration accuracy.

Moreover, during the recordings participants wear a black motion capture suit that is covered with optical markers, which are clearly visible in all lighting conditions. Totally 8 markers are used in the upper body tracking process that are attached to 8 human upper body parts (head, chest, 2x upper arm, 2x elbow, 2x forearm). The participant stands in the middle of the room during the recordings. We give the participant some degree of freedom in moving, but this freedom is limited to moving in place.

Speech is recorded with a high quality Sony ECM-166BMP lapel microphone worn by the participant close to mouth, so that movement of the participant would not cause volume alterations in the recordings. Moreover, it is small enough that the microphone does not occlude body markers. Lapel microphone and upper body recording system are both connected to a high capacity server that manages the synchronization between audio and video information. Speech of the actor is recorded at 16K sampling rate and stored in wav formatted files.

In addition, a remote-controlled head-mounted HD camera captures participant's facial expressions. The camera is mounted on a helmet worn by the participant and it can capture whole facial gestures. Since mount system with cam is light-weighted, the participant can easily move and rotate his/her head in any direction comfortably. The head camera records audio as well, but it is only used for synchronization purposes with the upper body recording system.

Figure 1: Four camera views with head camera on the subject and a sample view from head mounted camera

2.3 Recording Environment

The recordings are performed in the MVGL multi-camera studio with good lighting conditions. A good lighting condition means that there is enough diffuse light to leave no shadows on the participants face. All cameras are focused on the participant and the head-mounted camera is just adjusted in order to get a clear view of face. Sample views from multi-camera and head-mounted system are given in Figure 1. The background and the floor of the room are covered with black color.

2.4 MVGL-MUB: Multimodal Upper-Body Database

The Multimodal Upper-Body (MVGL-MUB) Corpus that we have collected currently consists of 42 recordings from five pre-defined scenarios in Turkish with 7 participants from Turkey. A summary of the multimodal database is given in Table 1.

Scenario	Conversation (Monologue/Dialogue)	Head-Cam (Y/N)	Record Count
1	M	Y	16
2	M	Both	9
3	M	Both	7
4	D	N	6
5	D	N	4

Table 1: Summary of multimodal database, Average record length is 3-5 minutes per recording.

We have used different types of scenarios in our records to extract characteristic features from gestures. Each scenario is designed for natural and transparent interaction of participants within the recordings, so that gesture profile can be used in characterization and synthesis of gestures in synchrony with speech data. A summary of the scenario descriptions are given as following:

Actor Id	Gender (Male/Female)	Age	Background
1	M	20-25	M.Sc. Student
2	M	20-25	M.Sc. Student
3	F	20-25	M.Sc. Student
4	M	20-25	M.Sc. Student
5	M	20-25	M.Sc. Student
6	F	25-30	Ph.D. Student
7	M	40-45	Professor

Table 2: Attributes of the subjects acted in the scenarios

Scenario 1: Storytelling a memory
The first scenario consists of an exciting event that the subject/participant faced with. Each subject tells an incident about his/her experience avoiding pretended gestures in spontaneous manner.
Scenario 2: Storytelling a documentary
The second scenario is storytelling a documentary film. Each subject watches the same documentary film and talks about it to another person in front of the cameras.
Scenario 3: Storytelling a fairy tale
In the third scenario, subject is asked to read/watch a fairy tale from a text/video, and tells the story as he/she remembers.
Scenario 4: Conversation with an agent
In the fourth scenario, subject is given a text from phone call between an angry client and customer support service employee who is asked to act part of the client and forced the subject to make overpowered gestures.
Scenario 5: Watch & Tell
As for the fifth case, subject is expected to watch and tell thoughts about various previously prepared videos and images.

3. Multimodal Analysis

Our multimodal analysis system consists of feature extraction, unimodal segmentation and correlation analysis parts. First, we extract features for each modality. Then, we use HMM classifiers to segment patterns in an unsupervised fashion. Finally, we employ mutual information for the correlation analysis of these modalities.

3.1 Feature Extraction

3.1.1 Upper Torso

The motion capture process is a marker-based approach where a set of distinguishable color markers is attached to the joints of the participant. As a result of optical marker tracking process, 3D world coordinates for each frame and each marker are calculated by using the camera calibration parameters. After getting the 3D marker movement data, smoothing operations are applied to marker data to eliminate noises and it is converted to joint angles in Motion Builder.

For a given frame in the video sequence, a set of N images can be obtained from the N cameras. We define the upper body gesture feature vector, f_k for frame k to include the Euler angles associated with the 3D joint angles with their first difference:

$$f_k = [\theta_k, \phi_k, \psi_k],$$

where θ_k, ϕ_k and ψ_k are the Euler angles in the x, y and z axes, respectively representing posture of the speaker at frame k. The resulting feature vector is 3 dimensional for each chosen joint of the speaker. We consider joint angles of 5 body parts (left arm, left forearm, right arm, right forearm and neck), eventually 15 dimensional feature vector F_k^u is used to represent upper body gesture for each frame:

$$F_k^u = [f_k^1, \dots, f_k^5]^T,$$

where f_k^j is feature vector for j^{th} body joint in frame k. After extracting body gesture features, we have temporally segmented all records into short action chunks (3–6 seconds long) and clustered these chunks into $N_g = 6$ groups by using unsupervised clustering. Then we observed each labeled action and tried to predict generic attributes of the clusters. These action groups are described below;

Label:	Description:
a	Raise hand: raise any hand from low to high altitude
b	Instant hand moves: make sudden and major hand moves
c	Two hand contact: unites both hands on front
d	Stand by: actor does no action
e	Hands down: Moves hands from high to low altitude
f	Circular hand action: make circular hand motion

Table 3: Gesture labels and their explanation

Each action chunk is stored in our gesture dataset in the following format:

$$S_i^g = (b_i^g, e_i^g, n_i),$$

where b_i^g and e_i^g are begin and end time of the gesture label n_i respectively.

We have observed and validated gesture labels by using our supervision tool Video Observer. This tool displays the action clips along with their labels so we monitored performance of the unsupervised clustering process. According to our observation, we repeated clustering process with better parameters which gives more meaningful clusters

3.1.2 Face

We use Active Appearance models (AAM) for facial feature tracking. AAM is composed of shape and texture components. In order to build an AAM, we first extract video frames demonstrating distinct facial expressions and then label these images with N points defining the main features. These N points describe the shape of a face on a frame, whereas texture component can be described as pattern of intensities or colors across an image patch. Given a set of labeled images, all shapes are aligned into a common co-ordinate frame. Then, principal component analysis (PCA) is applied. Similarly, to build the texture model PCA is applied to grey level information of shape-normalized images. Finally, shape and texture parameters are concatenated. Then, to eliminate any existing correlations again PCA is applied.

We use the AAM-API for the implementation and totally track 88 facial points, including eyes, eye-brows, nose,

mouth and chin. Then, we calculate center points for each eye by taking average of eye corner coordinates. The distance of eye-brow points to these centers are used as features. Additionally, we also calculate inner mouth point coordinates differences and concatenate to eye-brow feature set. Finally, we calculate the first order derivative of the feature vector and obtain 48 dimensional facial expression feature vector.

3.1.3 Speech Prosody

Voice characteristics at the prosodic level, including intonation, rhythm and intensity patterns, carry important clues for emotional states. In this study, pitch frequency, first derivative of pitch frequency is considered in speech labeling procedure. Pitch variations, during word pronunciation, indicate emphasis on the phoneme and based on these variations each pitch alteration is tagged automatically with the most suitable label m from $N_p = 6$ choices below:

Label:	Description:
H*	Peak Accent: a high pitch tone target on an accented syllable relatively to speaker's pitch range
L*	Low Accent: a low pitch tone target on an accented syllable relatively to speaker's pitch range
L+H*	Scooped Accent: a low tone followed by a relatively sharp rise on an accented syllable
!H*	Down Step High Tone: a clear step down onto an accented syllable from a high pitch tone
DEA	De-accented pitch
SIL	Silence: non speech gap with no intensity value

Table 3: Prosody labels and their explanation

The pitch alterations during each syllable or word pronunciations, is tagged with one of 6 labels above so a tag sequence with time intervals is obtained. We define speech prosody feature as S_i^p for i^{th} interval-tag pair.

$$S_i^p = \left(b_i^p, e_i^p, m_i\right),$$

where b_i^p and e_i^p are begin and end time of the prosody label m_i respectively.

Speech tagging process is done both in manual and automated fashion thus supervised and unsupervised prosody datasets are formed. For supervised dataset, we manually listened and labeled speech by using Praat software which is a program for acoustic analysis. Automated speech tagging is managed by our program Automated Speech Togger which is based on AuToBI Java toolkit [7].

3.2 Correlation Analysis

HMM classifiers are used to define re-occurring patterns on the unimodal streams of upper body, face and speech prosody. Then we employ mutual information for the correlation analysis of the modalities. The unimodal patterns are considered as discrete random variables, and the mutual information between two modalities is computed over the discrete event counts.

3.2.1 Mutual Information

A possible procedure to correlate the structure of the upper body, facial expressions and prosody streams is to match the corresponding segmentations – for example by counting the co-occurrences. Since speech and body gestures or facial expressions may have temporal correlation, we also consider time lag between modalities.

In discrete case, mutual information to correlate sequence of independent realizations of random variables *A* and *B* is defined as:

$$[Eq.1] \qquad I(A,B) = \sum_a \sum_b P(A,B)\, log \frac{P(A,B)}{P(A)P(B)},$$

where *P(A)* and *P(B)* are marginal densities and *P(A, B)* is joint density. Mutual information is always non-negative, and zero if and only if the variables are statistically independent. The mutual information takes into account the whole dependence structure of the variables, not only the covariance.

In our case, random variable *A* represents our gesture class and *B* represents prosody class. Thus mutual information can be computed using a co-existence matrix of prosody and gesture modalities.

3.2.2 Bimodal Co-Occurrence

To analyze structural relation between modals, we have counted temporal co-occurrence of each prosody label with corresponding gesture label(s) considering time lag τ. For each prosody entry S_i^g in the record, gesture entries are counted in the time interval $(b_i^g - \tau\ ,\ e_i^g + \tau)$. Hence, we formed ($\boldsymbol{N_p \times N_g}$) co-occurrence matrices $\boldsymbol{M(x,y)}$ for unsupervised and supervised gesture-prosody datasets.

To be able to use mutual information on co-occurrence matrices $\boldsymbol{M(x,y)}$, we normalized matrices with their sum. Thus each cell in the normalized matrix represents the joint probability *P(A, B)* of how both modals occur together*:*

$$[Eq.2] \qquad P(a,b) = \frac{M(a,b)}{\sum_{x\in B}\sum_{y\in A} M(x,y)}\ ,$$

After normalizing the co-occurrence matrix with the matrix sum, sum of all cells becomes 1. Thus normalized matrix becomes probability distribution function of multimodal coexistence. Sum of a prosody label row provides us the marginal density *P(B)* of the prosody label. In the same way, by summing a gesture label column, we obtain marginal density *P(A)* of the gesture label. On this basis, mutual information values are computed on each dataset for different time lag configurations.

4. Experimental Evaluations

We have calculated mutual information for gesture and prosody modals as stated in section 3.2.1. To test the performance of the datasets, we used unsupervised/supervised prosody dataset along with unsupervised gesture dataset. Besides different time lag parameters are used in the experiments to represent structural delay between body gesture and speech. Dataset performance results can be seen below:

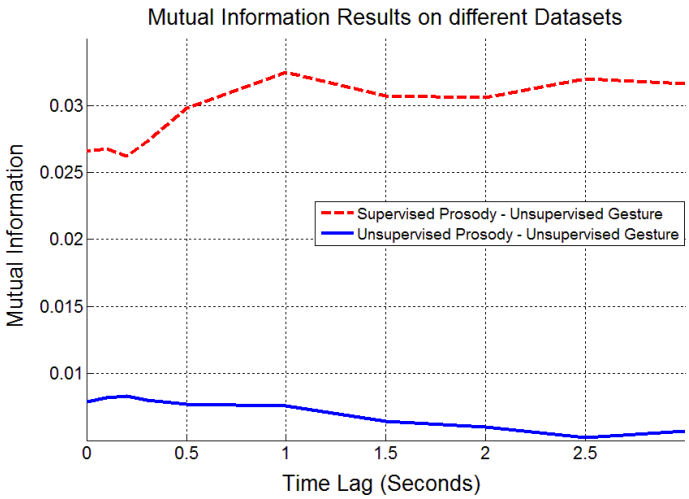

Figure 2: Mutual Information results for different datasets and time lag parameters

For our unsupervised prosody dataset, optimal mutual information value (0.008) is obtained at 0.3 second lag. Mutual information decreases as we move to high time lag values.

Prosody labels are automatically assigned based on pitch alteration. Also word structure or semantic is not considered in automatic speech tagging process. Thus wider and error-safe word intervals are tagged by our program. Since time-intervals for prosody labels are already long, increase in time lag parameter causes decrease in mutual information.

Alternatively, supervised prosody dataset promises far better mutual information with respect to unsupervised version. Optimal mutual information value (0.0325) is achieved at 1 second time lag and a promising candidate (0.0320) at 2.5 seconds. Since prosody labels are assigned manually in this dataset, higher mutual information and more realistic lag model is obtained. As in the real life, humans may start making correlated body gestures just before the speech. Regardless of speech semantics, our experiments show that there is a lagged correlation between pitch alterations of speech and body gestures.

5. Acknowledgements

This work was supported by Turk Telekom under Grant Number 11315-02.

6. References

[1] Humaine project portal: http://emotion-research.net/ (2011)

[2] Gunes, H., Piccardi, M.: Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction. p. 102. Beijing, China 2005

[3] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases (2003)

[4] Grimm, M., Kroschel, K., Mower, E., Narayanan, S.: Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49, 787–800, 2007

[5] Abrilian, S., Devillers, L., Buisine, S., Martin, J.: Emotv1: Annotation of real- life emotions for the specification of multimodal affective interfaces. In: In Proc. 11th International Conference on Human-Computer Interaction (HCI 2005). pp. 195–200. Las Vegas, Nevada, USA (2005)

[6] Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T.: The safe corpus: illustrating extreme emotions in dynamic situations. In: In Proc. First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation). pp. 76–79. Genoa, Italy (2006)

[7] Andrew Rosenberg, AuToBI project portal: http://eniac.cs.qc.cuny.edu/andrew/autobi (2012)