Detecting Body, Head, and Speech in Engagement

Martin Vels and Kristiina Jokinen

Institute of Computer Science
University of Tartu
J. Liivi 2, 50409 Tartu, Estonia
{martin.vels,kristiina.jokinen}@ut.ee

Abstract. This paper describes ongoing work related to the analysis of videos, and to recognising of body, head and legs of conversing partners. The spoken utterance transcripts are used to mark the speech and divided into different types: propositional, backgrounding, and laugh segments. We report promising results which can be linked to the conversational participants experience and engagement in the interaction.

Keywords: keywords should be written here ...

1 Introduction

During a conversation, the listeners are performing body movements, at all times. These movements demonstrate the speakers intentions, interests, and feelings.

Based on this affirmation, our hypothesis is that many of the gestures performed by the speaker can occur when the latter tends to emphasise, to highlight the parts of the communicative situation that it judges relevant, those that it wants to draw the attention of its audience to.

The short paper is structured as follows. We introduce our data in Section 2, and discuss its cleaning and method in Section 3. We present results in Section 4, and discuss them with future prospects in Section 5.

2 Interaction and gesticulation

Conversation must be seen not as alternating monologues but as a social system. Any utterance is produced in some sort of social situation, it is produced under the guidance of some pragmatic aim, it plays a role in the interactional setting, it has a content that is being conveyed, etc. Dialogue makes significant social or interpersonal demands as well as semantic and syntactic ones. The interlocutors must, without any formal structure or rules, manage to organize their conversation, co-ordinate their contributions, and calibrate their meanings as they go along. The intrinsic problem in dialogue thus conceived is that, although both partners must remain involved, only one person can talk at once. Whenever a speaker has the floor, there exists the possibility that the conversation could veer off into monologue. One solution to the problem is for the speaker to involve the listener regularly. The speaker can do this by inserting phrases such as "You know?" or "As you just said", or even "What do you think?" However, the frequent use of such verbal by-play would constantly interrupt the verbal narrative, so nonverbal means of seeking or maintaining listener involvement is well suited to this function. It is proposed that interactive gestures, for all of their many specific forms and meanings, constitute a class with the common function of including the listener and thereby counteracting the beginning of a drift toward monologue that is necessarily created every time one person has the floor. Such gestures, especially when delivered simultaneously with the verbal narrative, can efficiently exert countervailing force in the direction of dialogue

Gesture may represent some aspects of the content. However, gestures are not simply symbols, entities for carrying meaning about something else, but physical actions with their own distinct properties for example, they occur at specific moments in time and at particular points in space (Goodwin 1986). A gestural sign is formed by the cognitive system that is also used in the movement of the body in the physical environment. Even the gestures that do not have an obvious enactment component, such as abstract deictic gestures, are also formed by spatio-motoric thinking. The production of abstract deictic gestures, which point to a seemingly empty location in front of the speaker or move as if to track a moving object, could be related to the ability to orient our body parts (e.g., gaze and the hand) toward a target in the physical environment, and to the ability to track the target when it moves (Kita 2000: 17-18).

Gesticulation is often an important component of the utterance unit produced, in the sense that the utterance unit cannot be fully comprehended unless its gestural component is taken into consideration. In many instances it can be shown that the gesticulatory component has a complementary relationship to what is encoded in words, so that the full significance of the utterance can only be grasped if both words and gesture are taken into account (Kendon 1986). The word gesture serves as a label for that domain of visible action that participants routinely distinguish and treat as governed by openly acknowledged communicative intent.

Dialogue in conversation is collaborative. That is, dialogue requires social processes, such as co-ordination and calibration, in addition to the individual processes of language production and comprehension. Any utterance is produced in some sort of social situation; it is produced under the guidance of some pragmatic aim; it plays a role in the interactional setting; it has a content that is being conveyed, etc.

Some aspects of the content may be represented by a gesture. Gestures depicting a path of movement, a mode of action, relations in space between objects or entities are what McNeill (1992) has called iconic gestures. The content that is represented need not be descriptions of actual or possible actions, events, spatial relationships, but may be as if entities, actions, spatial relationships that serve as metaphors for concepts at any level of abstraction (cf. McNeill 1992; Calbris 1990; Kendon 1993). An iconic gesture is typically placed at the onset of or just prior to the speech unit to which it relates (Kendon 1983). It means, that the gesture foreshadows that unit. It aids listeners in the operation of understanding by enabling foresight. Iconic gestures project upcoming components of talk (Streeck 1988).

There have been various competing classifications of gestures in the literature, though the terminology has often been somewhat misleading. Kendon (1986, 1995), Scheflen (1973), Bavelas, et al (1992), and many thers clearly shows how body movements and the flow of speech are intimately linked within an individuals communication system and between interactants. While some behaviors may seem less integrated than others, verbal and nonverbal behaviors are unquestionably part and parcel of the same overall system of communication. Just like verbal communication, each body expression or vocal sign conveys a meaningful message, which can be received and processed by other people.

McNeill (1979) has found a close fit between the occurrence of a gesture and the occurrence of a speech unit expressing whole concepts or relationships between concepts. He reports that the peak of the gesture (that is to say, the most accented part of the movement which Kendon calls the stroke) coincides with what was identified as the conceptual focal point of the speech unit. McNeill has suggested that each new unit of gesture, at least if it is of the sort that can be considered representational of content, appears with each new unit of meaning. Each such gesture manifests, he suggests, a representation of each new unit of meaning the utterance presents (Kendon 1986b: 35). In his later works McNeill (1999) extends these ideas. He has put forward some positions about relations of gestures and speech. (1) Speech and gesture comprise a single system of meaning representation. Gesture does not derive from speech, or speech from gesture. Both derive from a deeper idea unit source that they represent co-expressively. (2) Imagery is part of utterance meaning. This does not mean that utterances automatically refer to imagery but imagery grounds categorial content. Dialectic implies that categorial content equally affects imagery, as the form of imagery changes in different linguistic systems. (3) Content motivates form in gesture. (4) The speech-gesture system shows that dynamic imagistic representations arise during speaking. These representations are part of the speakers online thinking for speaking (McNeill 1999). A smallest unit that retains the essential properties of a whole, in our case the whole of an image and a linguistically-codified meaning category, such as McNeill and his associates see in the speech-gesture window. They use the gestures semantic content and its synchrony with spoken linguistic segments to infer the speakers thought units.

3 Data

3.1 MINT

The videos used in this research are from the MINT (Multimodal INTeraction) project that deals with Artificial Intelligence and multi-modal agents [9]. One particular field where intelligent agents need a lot of development is the study of emotion and sentiment, not only in gestures, but language as well - for example in speech synthesis where it soon becomes important for an agent to learn different tones for communicating more effectively (Vainik 2014: 335).

The MINT dataset contains 23 videos of the Estonian First Encounters Dialogues where the speakers are unfamiliar with each other and they are expected to make acquaintance with their partner. They are expected to describe their likings to the partner but not to start emotional arguments on due to social politeness rules. Original Full HD (1920x1080px) videos were resized to 640x360px and 25 frames per second before the data processing described in following sections.

3.2 Head and body movement detection

Our goal was to segment out heads and bodies of the persons from the MINT conversational videos. As the first frame in the video did not contain any persons we used that as the background image. We converted it from RGB to grayscale and applied Canny edge detector [4] on it to receive a black and white image containing only edges present in that frame. Next, we iterated over all the frames of the video to segment out human head and body position coordinates. Our segmentation was done in following way: we started by converting original video frame into grayscale and applying Canny edge detector on the frame just as we did previously with the background frame. Next, we subtracted the background frame from the current frame. The resulting image contained only person edges. Next, we applied morphological closing operation [6] which is combination of morphological dilation [6] and erosion [6] operations applied one after another. We used closing operation to reduce possible noise and to get non-continuous contours. Next, we applied contour finding algorithm [15] and used the two contours that had largest areas. These two contours were the detected persons present in the frame. Once the human body locations were detected we cropped out the topmost region of the contour as the head and used contour finding algorithm once again only on that cropped out region. This way we were able to retrieve a very precise location and size of the head position in the frame. Next, we used the coordinates of the head to determine where the body started vertically in the whole body surrounding contour. We divided that contour into two parts vetically, top part contained torso and the bottom part contained legs. We applied contour finding algorithm once again on these frame regions to retrieve precise coordinates of the body and leg locations in the frame. A sample of the head and body segmentation steps of a single frame can bee seen on Fig. 1.

3.3 Noise removal

Once we got the coordinates of the human head and body positions and sizes in each of the video frames we applied noise removal techniques to that data as the segmentation of the video frames is not perfect and always contains some noise in it. During the noise reduction, all the data points where the head or body position coordinates in the sequential frames were differing more than certain threshold were replaced with median value of that particular coordinate series.

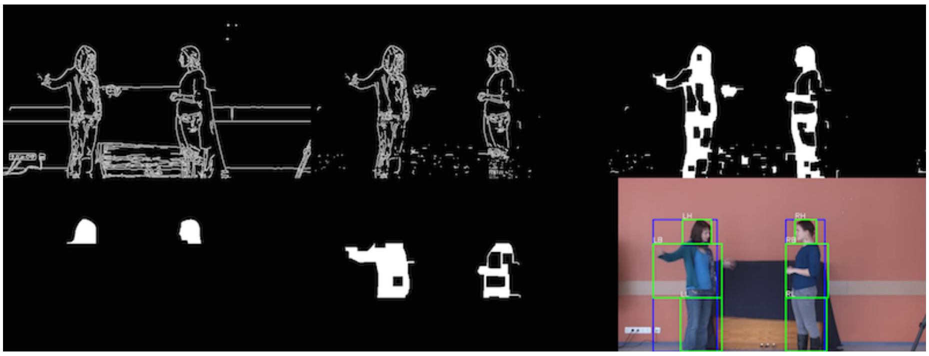

Fig. 1. Head and body segmentation steps: edge detection, background subtraction, closing, head contour, body contour, and final result with head, body and leg coordinate detected for both persons.

In addition, cases where the human body back coordinate value was greater than the human body front coordinate value were obviously incorrect and we used median coordinate values there instead as well.

After noise removal we ended up with 3 horizontal coordinates for each person per frame. These 3 coordinates were back and front coordinates of the human torso position and single coordinate for the head position. We only used the middle point value of the head coordinates as the human head doesn't change it's horizontal size. We named these coordinates in a following manner: LBB (left person body back), LBF (left person body front), LH (left person head), RBB (right person body back), RBF (right person body front), and RH (right person head). With these three coordinates for each person we were able to capture all the horizontal movements of the human head and body during the conversation. From this data it was possible to discover the moments when persons were performing some kind of hand gestures during the conversation. It appears that the segmentation and data cleaning was very successful as all the hand gestures, body and head movements could clearly be seen on the peaks of the body and head coordinates (see Fig. 2).

3.4 Speech data

In addition to automatic head and body position extraction from the conversational videos we also used manually annotated speech data. Timestamps were assigned to the articulated speech, laughs and non-articulated vocalisations (e.g, hmm, umm, ahem). Using this data we were able to draw a diagram of the whole conversation containing the speech (see Fig. 3). The whole 5-minute conversation is divided into 5 parts, one minute in each part. Each subplot is logically divided into two vertical sections, top section contains speech data for person standing on the right and bottom section contains speech data for person standing on the left during the conversation. In addition both of these sections are divided into 3

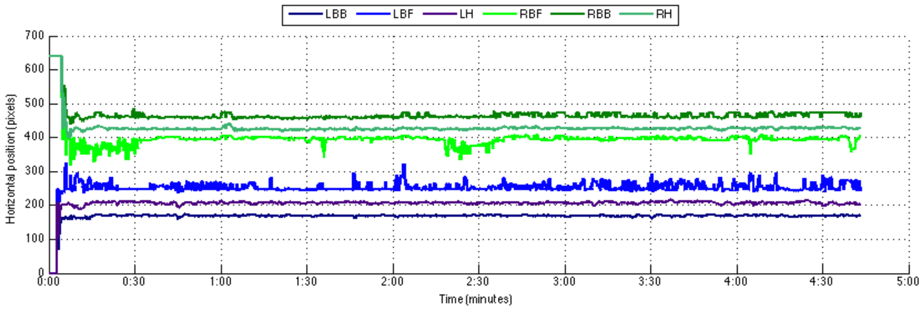

Fig. 2. Whole conversation with head and body movements of the person on the left in video scene shown on the bottom (LBF, LBB, LH) and the same info for the person standing on the right shown on the top part of the diagram (RBF, RBB, RH). A lot of hand movements can be seen from the LBF and RBF values.

sections, where the middle section contains actual speech bits, the sections closer to the middle of the diagram are showing the laughs and the sections outermost sections show the non-articulated vocalisations.

3.5 Head and body movements combined with speech

Finally we combined speech data together with head and body movement data into a single diagram (see Fig. 4) to see the correlations between these modalities. These diagrams contain six line-plots (3 per person) containing back and front locations of both conversation partners' upper body as well as the head locations of these persons as described in section 3.3. In addition, there are also three categories of speech data shown next to head and body position coordinates. Speech data is named in following manner: LS (left person speech), LL (left person laugh), LV (left person vocalisation), RS (right person speech), RL (right person laugh), and RV (right person vocalisation).

4 Results

From the video and speech analysis diagrams we were able to detect many interesting interaction patterns (see Fig. 5) between conversation partners and could also see immediately which person was laughing a lot (see Fig. 3), who was more dominating in the sense that he/she was speaking more, or who performed many body movements during the conversation. Also, we were able to see if a person was using his hands a lot during the speaking (see Fig. 7) or how frequently non-articulated vocalisations were used by a person. In addition, it was easy to find irregular gestures like person pointing somewhere in the back (see Fig. 6), waving her hands around (see Fig. 8), or bending her whole body forward (see Fig. 9).

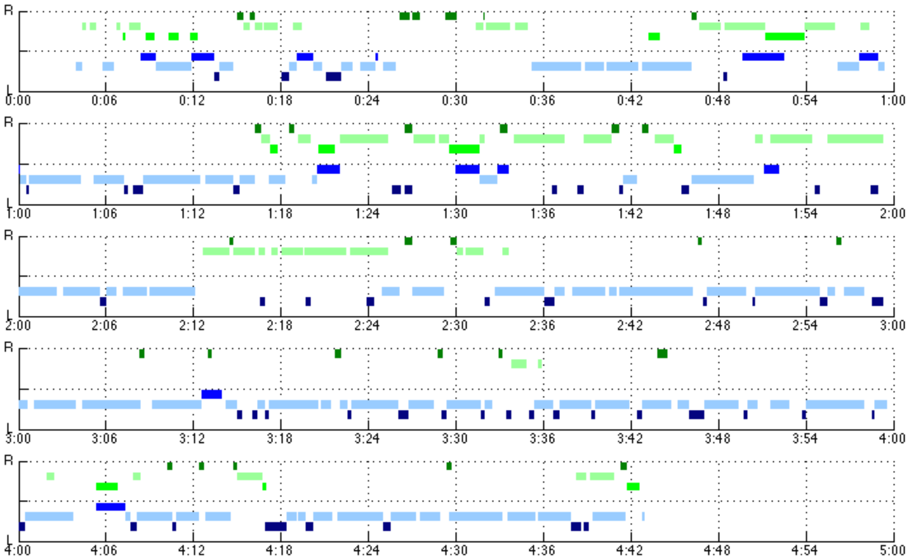

Fig. 3. A speech diagram of a 5-minute dialog separated into five one-minute pieces. Each one-minute piece contains top part for right conversation partner (using green colours) and bottom part for left conversation partner (using blue colours). Each conversation is divided into 3 parts: darkest colour represents non-articulated vocalisations, average tone is for the speech, and the brightest tone shows laughs.

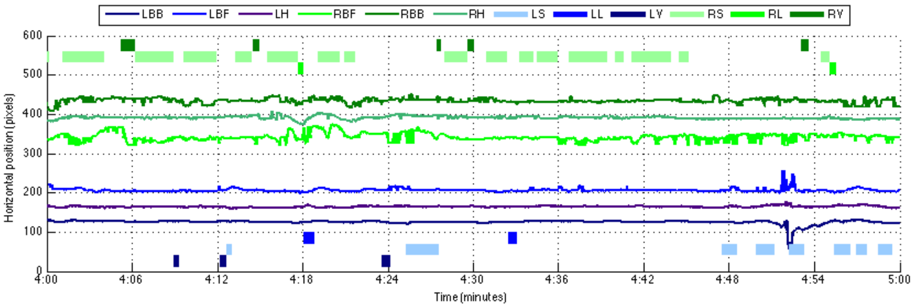

Fig. 4. An example of one minute long excerpt of a conversation where head and body movements of the conversation partners are combined with their speech data into a single diagram.

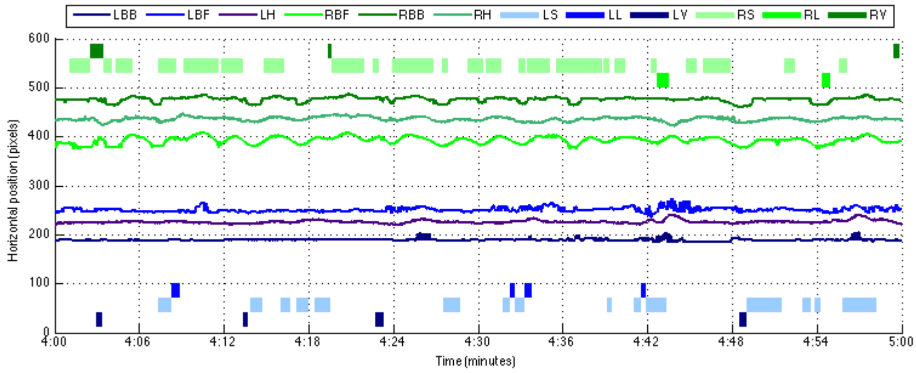

Fig. 5. Person standing on the right (green lines on the top) is rocking with the whole body back and forth rhytmically during the conversation.

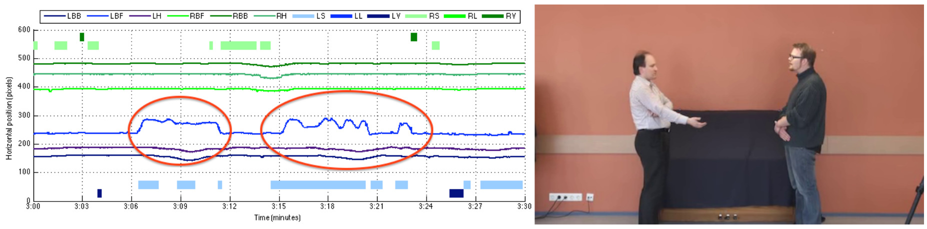

Fig. 6. Many spikes in the LBR coordinate means that this person is doing many repetitive hand gestures during speaking.

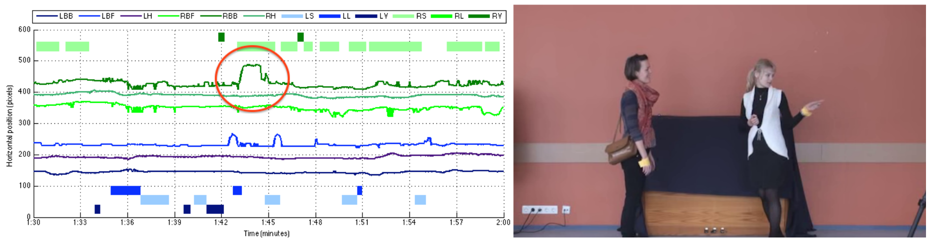

Fig. 7. Large spike in the RBB coordinate without the RH or RBF changes means that the person gestures somewhere behind her.

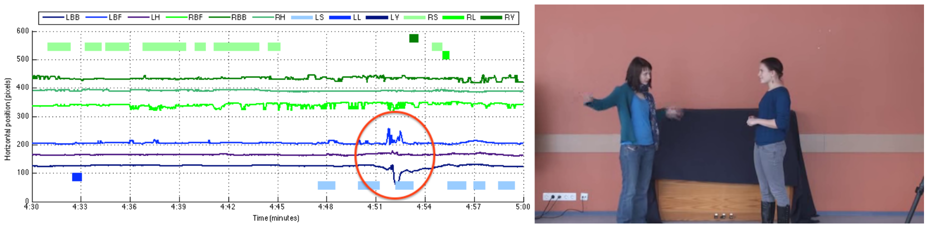

Fig. 8. Spikes in both LBR and LBF coordinates with unchanged LH coordinate mean that the person is waving her hands around.

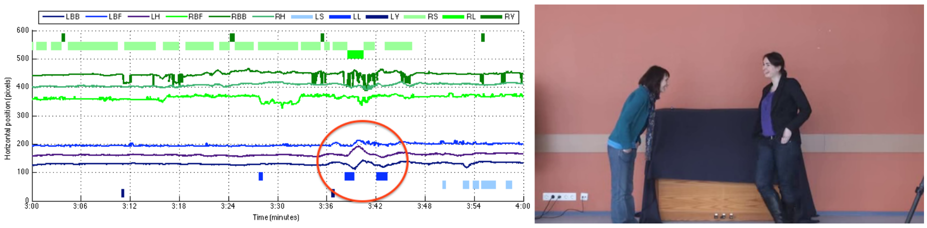

Fig. 9. Spikes in the LH and opposite spike in the LBB coordinates mean that the person is bending forward during the conversation.

5 Gestures and Speech

Kendon has showed that in continuous discourse, speakers group tone units into higher order groupings and so we can speak of a hierarchy of such units, and gesture phrases may be similarly organized. For example, over a series of tone units linked intonationally or by an absence of pauses into a coherent higher order grouping, the co-occurring gesture phrases are also linked (Kendon 1998). There remains a controversy about the way in which gesture as an activity is related to speech. Some investigators appear to consider it simply as a kind of spill-over effect from the effort of speaking, others see it as somehow helping the speaker to speak, yet others see it as determined by the linguistic choices a speaker makes as he constructs an utterance. An opposing view is that gesture is a separate and distinct mode of expression with its own properties, which can be brought into a cooperative relationship with spoken utterance, the two modes of expression being used in a complementary way (see Kendon 1998).

6 Discussion and Future Work

This paper started to explore.

Mancini et al. (2007) analyze human body movements in order make the virtual character to respond to the users expressive behavior appropriately.

The results show that the methods can be applied with fairly good classification results, and even though most sentences are neutral, the speakers are mostly positive when showing sentiment.

References

1. Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. In Martin, et al. (Eds.). Multimodal Corpora for Modelling Human Multimodal Behaviour. Language Resources and Evaluation, 41 (3–4), pp. 273–287 (2007)
2. Battersby, S.: Moving Together: the organization of Non-verbal cues during multiparty conversation. PhD Thesis, Queen Mary, University of London (2011)
3. Campbell, N., Scherer, S.: Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity. Proceedings of Interspeech. Makuhari, Japan (2010)
4. Canny, J.: A Computational Approach to Edge Detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6), pp. 679-698 (1986)
5. George Caridakis, Amaryllis Raouzaiou, Elisabetta Bevacqua, Maurizio Mancini, Kostas Karpouzis, Lori Malatesta and Catherine Pelachaud: Virtual Agent Multimodal Mimicry of Humans. Language Resources and Evaluation, 41 (3–4), pp. 367-388 (2007)
6. Gonzales, Rafael C. and Woods, Richard E.: Digital Image Processing (3rd edition). Pearson Education, Inc., pp. 652-661 (2010)
7. Jokinen, K.: Gestures in Alignment and Conversation Activity. Proceedings of the PACLING Conference. Sapporo,Japan, pp. 141-146 (2009)
8. Jokinen, K.: Constructive Dialogue Modelling:Rational Agents and Speech Interfaces. Chichester: John Wiley (2009)
9. Jokinen, K. and Tenjes, S.: Investigating Engagement - Intercultural and Technological Aspects of the Collection, Analysis, and Use of Estonian Multiparty Conversational Video Data. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mar (Ed.). Proceedings of the Eight International Conference on Language Recourses and Evaluation (LREC12) (2764 - 2769). Istanbul, Turkey: European Language Resources Association (ELRA) (2012)
10. Kendon, A.: Some Functions of Gaze Direction in Social Interaction. Acta Psychologica 26, pp. 22-63 (1967)
11. Kendon, A.: Gesture: Visible action as utterance.New York: Cambridge University Press. (2004)
12. Kendon, A.: Spatial organization in social encounters:the F-formation system, Conducting Interaction: Patterns of behavior in focused encounters. Studies in International Sociolinguistics, Cambridge University Press (1990)
13. Maurizio Mancini, Ginevra Castellano, Elisabetta Bevacqua and Christopher Peters.: Copying Behaviour of Expressive Motion. Lecture Notes in Computer Science, volume 4418. Computer Vision/ Computer Graphics Collaboration Techniques. Third International Conference, MIRAGE 2007. Proceedings. pp. 180-191. (2007)
14. Patrizia Paggio, Jens Allwood, Elisabeth Ahlsn, Kristiina Jokinen and Costanza Navarretta.: The NOMCO Multimodal Nordic Resource Goals and Characteristics. In Calzolari, N, Choukri, K. Maegaard, B, Mariani, J., Odijk, J, Piperidis, S, Rosner, M. and Tapias, D. (Eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) Valetta, Malta May 19-21. ELRA. (2010)
15. Suzuki, Satoshi and Abe, Keiichi: Topological Structural Analysis of Digitized Binary Images by Border Following. CVGIP 30 1, pp 32-46. (1985)