Recognition of Human Body Movements for Studying Engagement in Conversational Video Files

Martin Vels
Institute of Computer Science
University of Tartu
Estonia
martin.vels@ut.ee

Kristiina Jokinen
Institute of Computer Science
University of Tartu
Estonia
kristiina.jokinen@ut.ee

Abstract

This paper investigates object recognition techniques to automatically detect human behavior in video conversations. The ViBe background subtraction algorithm, together with standard image processing techniques is applied to conversational videos where two people meet for the first time, and the results show the usefulness of the technique in human communication analysis. By detecting the conversational participants and analyzing their conversational styles through the detected body movements, we can visualize, and draw conclusions concerning the participants' engagement in the communicative activity. The paper discusses these novel observations that show the synchrony and engagement in the participants' behavior.

1 Introduction

The main questions in interaction studies are how people engage themselves in the interaction, how does their focus of attention work (where it is directed to and how it changes over time), and how their body movement, gestures, and other multimodal communication signals help in interaction management (Goodwin, 1981; Kendon, 2004). Humans use several communicative systems ("modalities") and several physical "carriers" of the messages of these systems (vocal sounds and visible hand movements) (Chellappa et al., 1997). The various modalities are not used independently of each other in communication, however, but usually one modality depends on the other modalities, and the communicated information is interpreted in a holistic way, using the signals from all modality channels simultaneously (Jokinen and Wilcock, 2012). Thus, in order to understand human multimodal communication it is necessary to identify those signals that convey communicative meaning, and to understand their intended meanings in the given context. In this respect visual signals and especially human body movements play a central role, besides verbal communication, facial expressions, eye-gazing, breathing, etc. in order to express meanings, emotions, and attitudes, and to coordinate the interaction in general (see e.g. Argyle, 1975)

Concerning the recognition of body movements, relevant questions deal with how humans segment and assign meanings to objects and scenes in their visual field, and what are the appropriate techniques that would enable automatic segmentation and interpretation of visually presented information. Such work contributes to a better understanding of human visual information processing, and also of the requirements for developing intelligent systems and smooth user interactions with such systems. It also opens up new possibilities for building various applications that deal with personal digital devices. It is evident that interaction strategies are important regardless whether the communication takes place face-to-face, via audio-only, or via both video and audio. Advancement of technology in digital devices, like cameras, smartphones, and wearable computers like Google Glass, has made more holistic communication capabilities possible both for the users and the systems, and thus human movement detection is crucial to facilitate natural and flexible interaction.

However, detecting humans (Santhanam et al., 2012) and their gestures (Mitra and Acharya, 2012) from videos is a non-trivial task algorithmically. The quality of the detection depends on the quality of

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

the video, the illumination and background of the scene, the position of the human compared to the video camera (i.e., facing the camera, standing in profile), what persons in the scene are wearing (color, patterns), occlusion (e.g., hand is behind the body, thus not visible for other party), etc. On the other hand, manual annotation of video data is a common exercise which, however, requires a lot of resources. Work on automatic behaviour annotation using video data concerns e.g.gesture and gesture expressivity (Caridakis et al. 2006; Oikonomopoulos 2006), hand, head and body movements (Allwood et al. 2007), as well as synchrony and body posture analysis using depth cameras (Michelet et al. 2012; Baur et al. 2013). However, this work differs from our approach since we do not only seek to provide and compare annotations with respect to human gesturing, but to study how well the image processing techniques can be applied and modified in order to recognize human body movement in natural conversational interactions.

In this paper we use the video collection of Estonian First Encounter Dialogues from the MINT (Multimodal INTeraction) project (Jokinen and Tenjes, 2012). We have created a technical solution that visually identifies human body movement on video files and tags them with descriptive and quantitative information, thus reducing the work needed for annotating videos manually. We use the ViBe background subtraction algorithm, together with standard image processing techniques to automatically detect human body in the natural conversation dialogue data and we create a diagram representation of the whole video that expresses changes in the detected human body location in the given scene. We also compare the performance of the algorithm to the manually annotated data so as to explore the applicability of the algorithm in human communication studies such as gesture and posture movements, synchrony, and engagement.

The rest of the paper is structured as follows. Section 2 introduces the MINT dataset in more detail. Section 3 describes the algorithms used for human body detection. Section 4 presents the results of the experiments and discusses them with respect to manually annotated data. Finally, Section 5 provides discussion and interpretation of the results, draws conclusions and describes plans for future work.

2 MINT Dataset

The MINT (Multimodal INTeraction) dataset contains 23 videos of the Estonian First Encounters Dialogues. Each of these videos is approximately 5 minutes long and contains two people having a conversation. The videos were recorded by three SonyHDR-XR550V cameras and three external Sony ECM-HW2 wireless microphones. The full 1920x1080 HD quality was used for the recordings. The Sony Vegas Pro 11 software was used to cut, edit and merge, and sync raw video clips to create video files that combined all three cameras, and to export videos to SD format that could be read by further analysis of the files. The resolution of the compressed avi files is 640x360 pixels, with 25 fps. These dialogues were filmed with three cameras from different angles. The first video contains the interaction shown from the center (see Figure 1), the second video shows half frontal view of the person on the left (see Figure 2), and the third video contains half frontal view of the person on the right (see Figure 3).

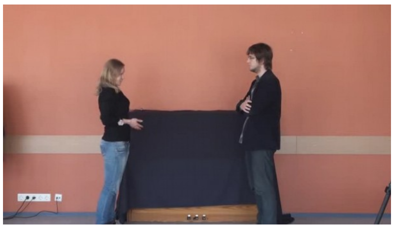

Figure 1: Center view of the scene.

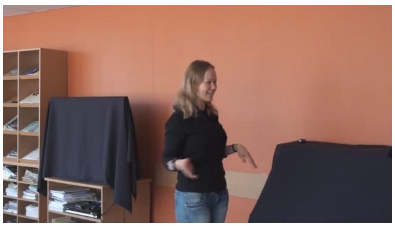

Figure 2: Left view of the scene.

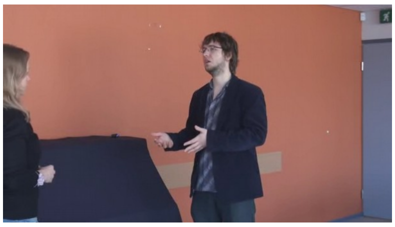

Figure 3: Right view of the scene.

There were 23 different people participating in these scenes, 11 female and 12 male. Each person participated in 2 videos, with different conversation partners in both cases.

3 Person Detection

We analyzed all the center-view videos using a background subtraction technique to detect persons in these videos. After the detection of persons we generated a diagram for each conversation which shows how persons moved back and forth (horizontal movement in the scene) during the conversation. This gave us a quick overview of the activity of the persons in the video. Unfortunately the other possible views (left and right) were not suitable for detecting horizontal human movements properly as the people in these scenes were filmed from a certain angle and not facing camera directly.

3.1 Background subtraction

To detect persons in video we used the ViBe (visual background extractor) algorithm (Barrich and Droogenbroeck, 2011) as background subtraction technique to remove all the image content except for the moving object, which in our case are the moving persons. Background subtraction is a widely used technique for image segmentation and there are various ways how this can be done. We used the ViBe algorithm because it is one of the fastest and most accurate background subtraction algorithms currently available. It was capable of processing videos with resolution of 640x360 pixels and 25 fps in real time.

The basic idea behind background subtraction is trivial. We have the static background frame and the current video frame. We compare these two frames pixel by pixel, remove all the pixels that are the same (background) on both frames, and retain the ones that are different (foreground).

Usually the clean background frame needs to be modeled. This means that we have to build a model of the background, containing more than only one frame and compare our current frame to the whole model to achieve the segmentation of the moving objects. There are several ways to build the background model. Often the background model is built using several sequential frames of the video, which means that the initialization of the model can take a long time (several seconds) and also keeping the model takes a lot of memory. If the frame rate of the video is high (30+ fps), then using this technique means that a lot of memory and computational power are needed. E.g., one gray-scale frame with size 640x480 pixels requires 300kB. Each second of this video is 9MB. Also, even if we build our model from such a video, we still need to analyze about 5-10 seconds of the video (150-300 frames) until we get some idea of moving parts in the video to build a decent model for our background.

The ViBe algorithm (Barrich and Droogenbroeck, 2011) that we were using for segmentation introduced a novel idea where the background model is initialized from a single frame using 8-neighborhood of each of the frame pixel and randomly choosing 20 instances of these neighbor-pixels to build a background model.

The next problem with the background model is to keep it up-to-date when time passes and the scene changes. One of the most widely used techniques is to remove the oldest samples of pixel values from the background model and adding new values from the newer scenes. This method, even if it seems the most natural, may still not be the best solution. Namely, the fact that some of the pixels in the background model are old does not automatically mean that these pixels are no longer correct background representatives. Thus the ViBe algorithm that we are using, had a different approach — it replaces the values in the background model randomly. This technique gave this algorithm a better response speed in case of changing scenes but at the same time not removing correct background pixels from the model just because of the age.

Segmentation itself was simple, comparing each pixel of the current frame to according pixel in our background model and when certain amount of model representatives were close enough, we considered the pixel to be background, otherwise, it was foreground. Euclidean distance was used to determine the closeness of the pixel value to according background model values. It is possible to work in RGB or gray-scale images, the only algorithmic difference is that in case of gray-scale, we only compare one value of the pixel (intensity in the range of 0..255), but in case of RGB-video, we need to compare all the three color-components of the pixel (red, green and blue) to determine the distance between two pixel values. As the results of the segmentation were similar in both gray-scale and RGB-images, we used the gray-scale version because of less computational complexity involved.

3.2 Erosion and Dilation

Finally, when the frame was segmented, we had a black-and-white image (see Figure 4), where zero means background and one means foreground. Now, as there are usually some kind of illumination changes in frames which result in noise in our segmented image, we used two image processing techniques to get rid of the noise. Namely, we used erosion (Gonzales and Woods, 2010) (see Figure 5), which helped us remove all the single pixels. Next, we employed a dilation operation (Gonzales and Woods, 2010) (see Figure 6) operation, which helped us make the interesting objects larger and remove some of the small gaps between close parts of the objects. Erosion and dilation are the morphological operations. Dilation adds pixels to the boundaries of objects in images and erosion removes pixels of object boundaries. The amount of pixels added or removed is determined by the structuring element used during these morphological operations. In our case we used 3x3 square structuring element.

3.3 Object detection

After the segmented image was eroded and dilated it was possible to use a contour detection algorithm (Suzuki and Abe, 1985) available in OpenCV library. This algorithm returns a set of contour areas, which are hierarchically arranged. We used the top-level hierarchy and found the top left and bottom right coordinates of the contours, which had certain size area. This way we were able to find the positions of the moving persons in the frame that were larger than a certain predetermined threshold (see Figure 7).

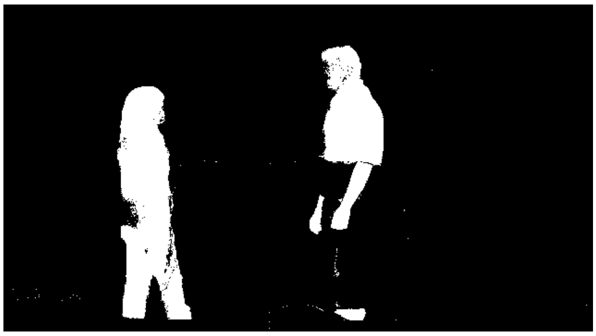

Figure 4: Segmented image.

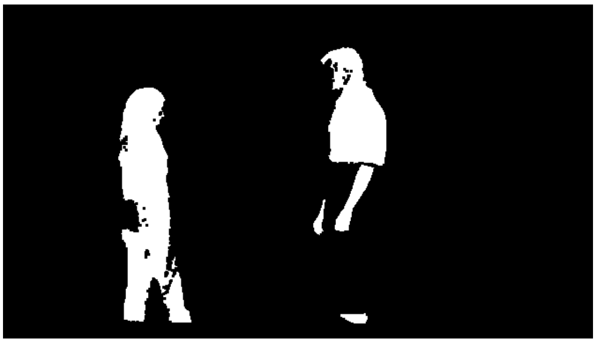

Figure 5: Eroded image.

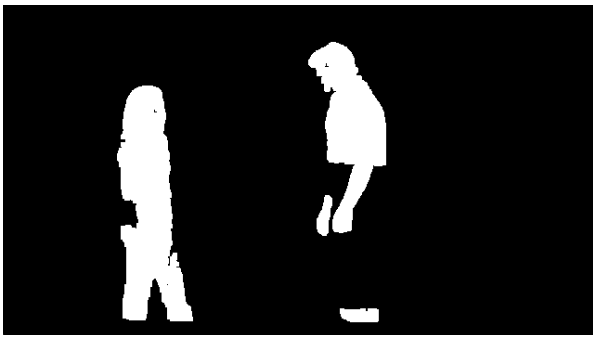

Figure 6: Dilated image.

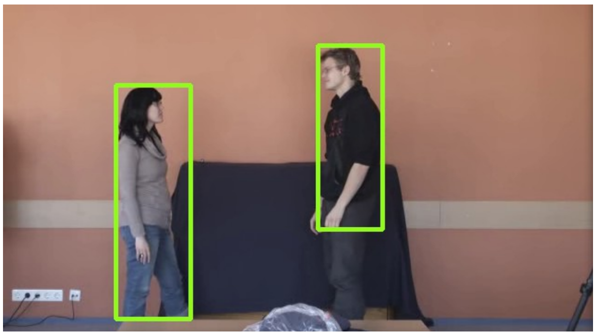

Figure 7: Image with moving persons detected.

4 Engagement in video files

4.1 Hand and body movement

We used left and right coordinates of the surrounding boxes around the persons (see Figure 7) found by the algorithm to draw a diagram that summarized the whole video (see Figure 8). We abbreviate the front and back coordinates of the surrounding box around the person on the left as LFC (left front coordinate) and LBC (left back coordinate). Similarly, the front and back coordinates of the surrounding box around the person on the right is abbreviated as RFC (right front coordinate) and RBC (right back coordinate). We also use the term "person front" when referring to the coordinates that correspond to the person's front and "person back" when referring to the coordinates that correspond to the person's back.

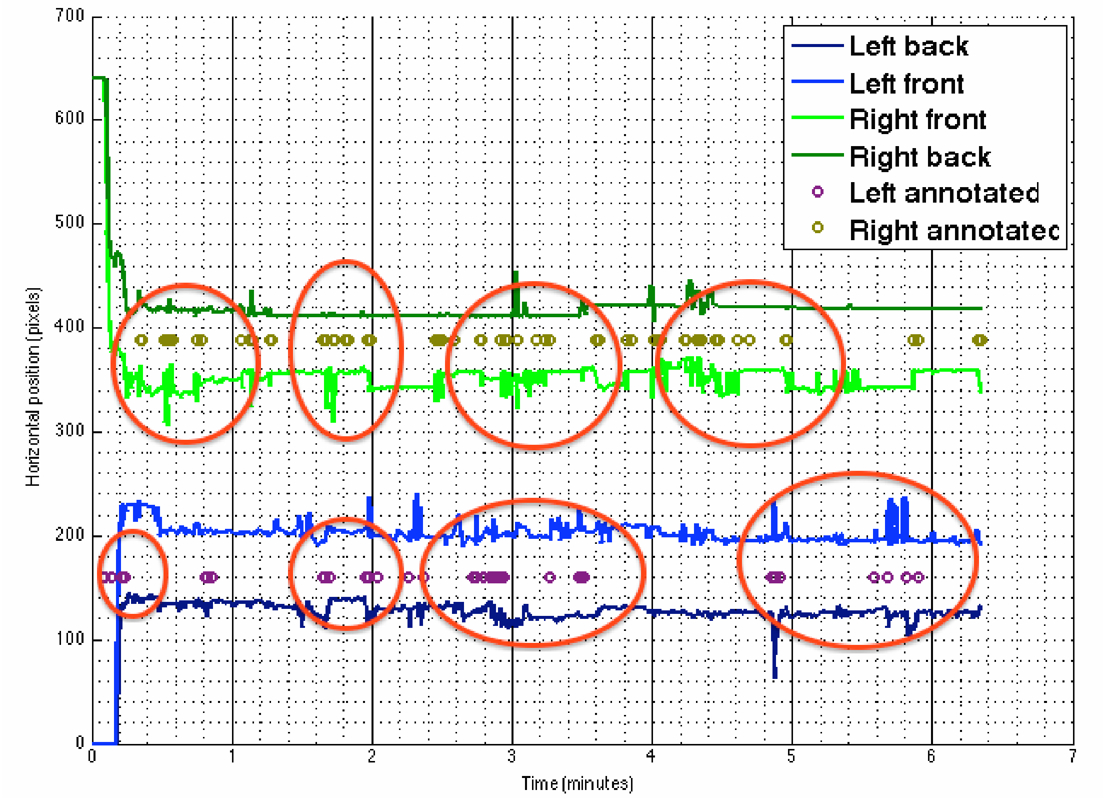

Figure 8: Person movement diagram of the whole video.

The diagram (see Figure 8) shows the movements of the left and the right person in time. Both conversation partners are described with two lines, first showing the back and next the front coordinate of the person. In the middle of these two lines, the hand movements of that particular individual are shown in small circles. Hand movement time was manually marked using annotation software—ANVIL (Kipp, 2001). By comparing the diagrams it can be seen that most of the hand movements can be detected using the front coordinates of the surrounding boxes of the moving persons. However, we also note that there are many cases where movements seem to be detected by the algorithm but are not annotated as communicatively important in the data. This is due to the fact that the size of the surrounding box does not change only due to hand gesturing but also due to person leaning over or moving leg etc. Sudden movements of the head or the whole body of the person are also detected in the changes of the coordinates. The LFC and RFC changes do not automatically indicate hand movements but need to be interpreted in the context of back coordinates.

It must also be emphasized that the comparison of the algorithm with manual hand gesture annotation is meant to visualize the functioning of the algorithm rather than to evaluate its performance against manually annotated data. The large number of "false positive" detections does not indicate the algorithm's oversensitivity to hand gestures, but rather, that the algorithm detects movement in general, and needs to be further tuned in order to detect hand gestures.

Comparing the manually annotated gesture tags with the automatically detected movements of the persons, we notice, however that in many cases it is possible to detect hand gestures from LFC and RFC. Especially if a person stands still, the back coordinate (BC) is steady, too, while the front coordinate (FC) changes rapidly because of the hand movement (see Figure 9 left back and left front). On the other hand, if the person moves frequently, then the FC does not reliably indicate the hand gestures (see Figure 9 right back and right front). As mentioned, the FC does not move only because of the hand gestures but also if the person moves her leg or head or bends forward. Thus, if we only look at the FC changes, the technique is ambiguous between the hand gesture, leg or head movement, and body bending.

We can also detect the handshake of the conversation partners in the beginning of the videos: in Figure 8 the persons do not shake hands, but in Figure 9 they do, as can be seen from the touching of the curves during the starting seconds of the video.

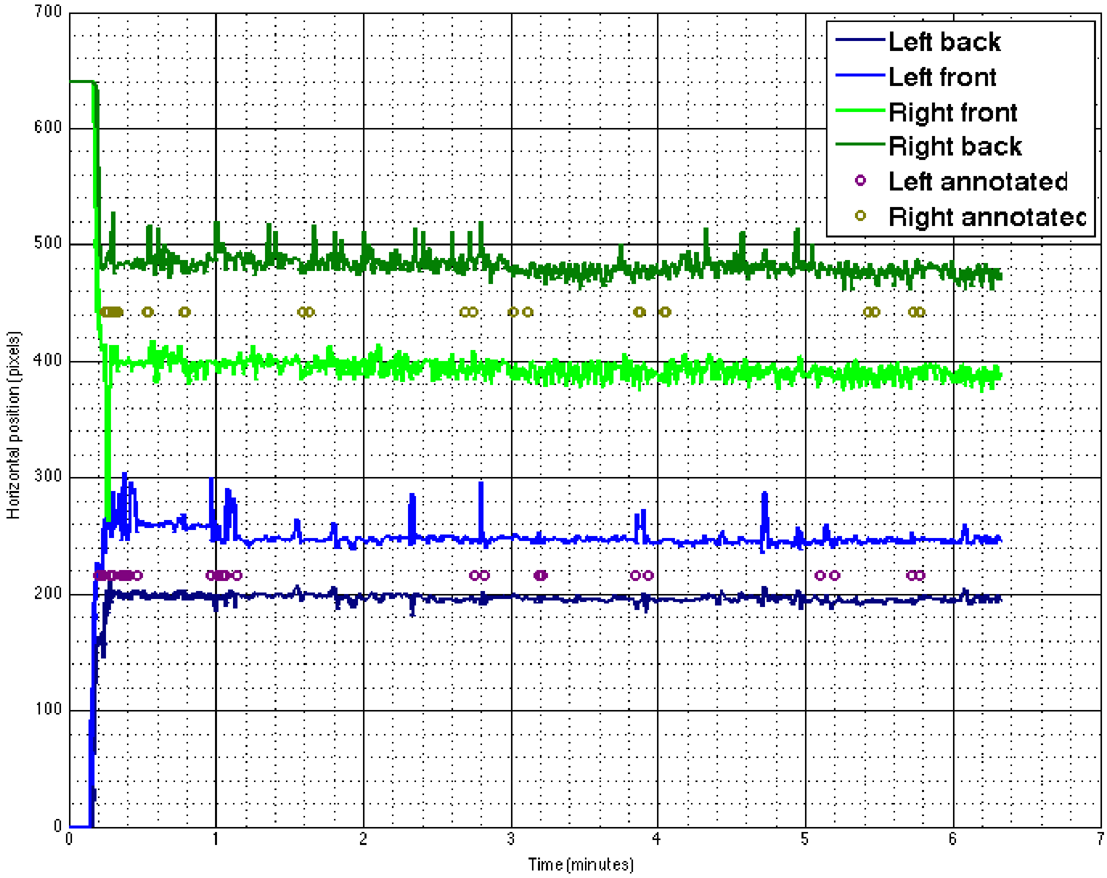

Figure 9: Rhythmic body movement of the person on the right.

4.2 Differences in individual body movements

The method is meant to study the human behavior as a whole, and another interesting behaviour that our diagrams can clearly visualize, is the differences between the conversational style of the interlocutors. We can see that if the partner is mostly standing still or performing many small movements during the conversation (left person in Figure 9). The rhythmic body movement, as can be easily seen in Figure 9 (right front), is an individual property that distinguishes people in the conversation. If the person is using her hands a lot during the conversation we can detect this by comparing the back and front coordinate movements (left person in Figure 10).

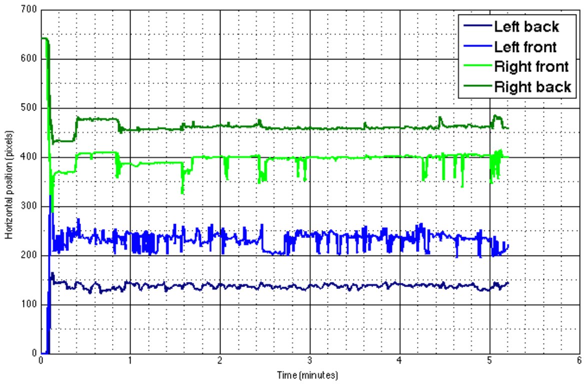

Figure 10: Frequent hand movement of the person on the left. (Front-coordinate changes while back-coordinate keeps still).

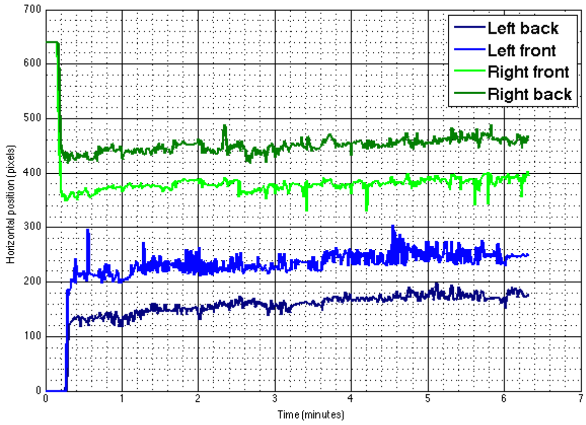

Figure 11: Movement trends. Both persons are moving to the right of the scene during the conversation.

4.3 Synchrony and irregularities

Finally, it is also possible to detect larger trends in position changes of the interlocutors during the conversation, e.g. if one individual is moving away from the other and the second one is following. A nice example of this kind of synchrony can be seen in Figure 11: the participants slowly move to the right of the scene which can be seen from the rising curves. As the participant on the left moves forward, the right participant moves further to the right trying to keep the same distance. The participants thus intuitively adapt their position so that the speaking distance remains constant. The participants' awareness to maintain a comfortable speaking distance intuitively can be used as an indirect measure of synchrony between the participants and of their adaption to the conversational situation. Such subtle movement may not be obvious by looking at the videos only, but our technique makes it concretely visible.

Another possible use of the algorithm is to identify large irregularities in positions of the conversation partners. For instance, if a participant performs a large step away from the regular position, this can be seen as a clear change in the magnitude of the movement among the normal moving of the participants (see Figure 12 and Figure 13).

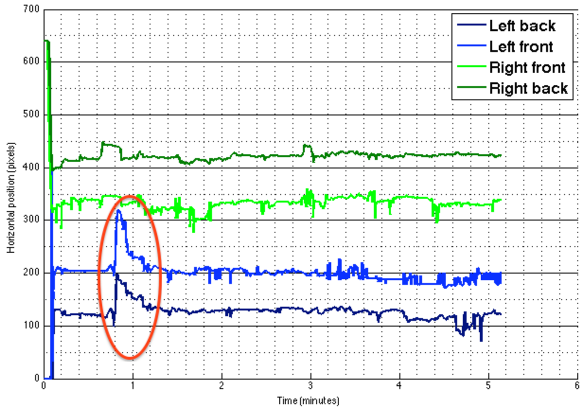

Figure 12: The person on the left performs a sudden movement towards the other person.

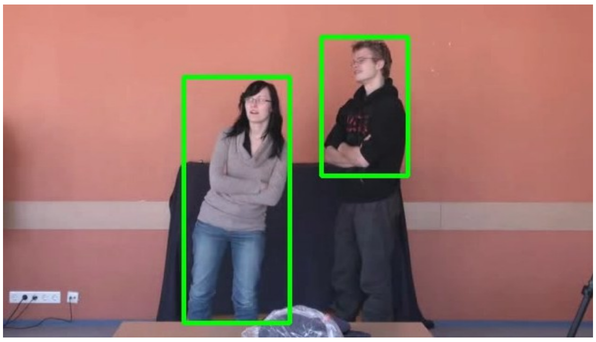

Figure 13: The person on the left is performing an unordinary movement.

5 Conclusion

In this paper we studied human behavior in communication and used image processing techniques to recognize body movement in video conversations. The algorithm focuses on the changes in human movements through the front and back coordinates of the box that surrounds the detected human body. We compared the movement visualization diagrams by the algorithm with the manually annotated gesture tags, and noticed that the algorithm can indicate hand gesturing, but it needs to be further developed to distinguish hand gestures from other movement types which also cause the coordinate of the surrounding box change. However, the technique provides an easy and helpful way to compare the participants' individual conversation styles. Moreover, as a novel contribution, the method can also show the participant's movement trends and irregularities in their behavior, which can effectively be used to study synchrony and adaption between the participants.

The results can be used to gain deeper understanding of human movements and body posture in natural interactions. The method can be used by annotators to find interesting gestures that may not be obvious from videos. They can also consolidate their annotations and check consistency of the annotations with respect to the automatically recognized movements.

The technique is based on empirically collected objective data and applying automatic signal analysis to the data, then comparing and combining this analysis with human perception and top-down analysis of annotated data. The work will thus also contribute to technical development of movement detection, and automatic scene analyses.

We will continue work on these lines to improve the algorithm on the conversational data, especially focusing on the specification concerning gesture recognition. We will also investigate further the synchrony of the participants as observed through their movements

References

Allwood, Jens, Cerrato, Loredana, Jokinen, Kristiina, Navarretta, Costanza & Paggio, Patrizia. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Martin, J.C. et al (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation, 41(3–4), 273–287, Springer.

Argyle, Michael. 1975. *Bodily Communication*. London: Routledege.

Barnich, Olivier and Van Droogenbroeck, Marc 2011. M., *ViBe: Universal Background Subtraction Algorithm for Video Sequences.* Image Processing, IEEE Transactions on, Vol. 20, pp 1709-1724.

Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner and Elisabeth André. 2013. *NovA: Automated Analysis Of Nonverbal Signals In Social Interactions*, HBU'13 Proceedings of the Third international conference on Human Behavior Understanding

Caridakis, G., Raouzaiou, A., Karpouzis, K., Kollias, S. 2006. *Synthesizing Gesture Expressivity Based on Real Sequences*, Proceedings of the LREC 2006 Conference

Chellappa, Rema, Chen, Tsuhan and Katsaggleos, Angelo 1997. *Audio-visual interaction in multimodal communication*, IEEE Signal Processing Mag., pp 37-38.

Gonzales, Rafael C. and Woods, Richard E. 2010. *Digital Image Processing (3rd edition).* Pearson Education, Inc.

Goodwin, Charles 1981. *Conversational Organization: Interaction between Speakers and Hearers.* Academic Press, New York.

Jokinen, Kristiina and Tenjes, Silvi, 2012. *Investigating Engagement - intercultural and technological aspects of the collection, analysis, and use of the Estonian Multiparty Conversational video data*. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey.

Jokinen, Kristiina and Wilcock, Graham, 2012. *Multimodal Signals and Holistic Interaction Structuring*. Procs of the 24th International Conference on Computational Linguistics (COLING). Mumbai, India.

Kendon, Adam. 2004. *Gesture: Visual Action as Utterance.* Cambridge University Press.

Kipp, Michael. 2001. Anvil - *A Generic Annotation Tool for Multimodal Dialogue*. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.

Michelet, Stephane, Karp , Koby, Delaherche , Emilie, Achard , Catherine, and Chetouani , Mohamed, 2012. *Automatic Imitation Assessment in Interaction*, HBU'12 Proceedings of the Third international conference on Human Behavior Understanding.

Mitra Sushmita and Acharya, Tinku 2007. *Gesture Recognition: A Survey.* Trans. Sys. Man Cyber Part C 37, 3, pp 311-324.

Oikonomopoulos , Antonios, Patras , Ioannis, and Pantic , Maja, *Spatiotemporal Salient Points for Visual Recognition of Human Actions*, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 36, No. 3, June 2006

Santhanam, T., Sumathi, C. P. and Gomathi, S. 2012. *A survey of techniques for human detection in static images*. In Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology (CCSEIT '12). ACM, New York, NY, USA, 328-336.

Suzuki, Satoshi. and Abe, Keiichi, 1985. *Topological Structural Analysis of Digitized Binary Images by Border Following.* CVGIP 30 1, pp 32-46.