Coding Hand Movement Behavior and Gesture with NEUROGES Supported by Automatic Video Analysis

Oliver Schreer[1], Stefano Masneri[1], Harald Skomroch[2] , Hedda Lausberg[2]

1 Fraunhofer Heinrich Hertz Institut, Einsteinufer 37, 10587 Berlin, Germany

2 Deutsche Sporthochschule Köln, Am Sportpark Müngersdorf 6, 50933 Köln

Introduction

In many different research disciplines in humanities, very large multimodal corpora are being processed and analysed in order to solve quite a large variety of different research questions. Not only in the gesture community, but also in psycholinguistics and psychology, the way how human act with their hands and body is of relevance. To solve the different research questions, a detailed annotation of multimodal data (video and speech) is performed. This annotation is basically performed manually by human raters and it is usually a very time consuming process. In empirical analysis of hand movement behaviour and gesture, the coding system NEUROGES and the annotation tool ELAN are very useful tools in this research domain [1][2]. Although, the manual annotation process is supported very well, it is still a very time consuming process to get at least the basic assignment such as start and end of hand movements to perform the seemingly most simple step, i.e., to segment the ongoing flow of behaviour into natural units. We present an automatic video analysis tool that supports the researchers in this exhaustive exercise inspecting and analysing videos that substantially facilitates the segmentation and annotation processes. Many significant events in human body motion can be detected quite robust by modern video analysis tools and therefore speed up the annotation process by a serious amount. The NEUROGES coding system provides a seven steps algorithm to analyse hand movement behaviour including gestures. It consists of three modules progressing from segmenting and annotating behaviour according to Activation, Structure, and Focus (Module I) to analyzing the relation of the right and left hands concerning Contact and Formal Relation (Module II) to finally assessing the Function and Type of the movements (Module III). The developed video analysis tool is able to support the annotation in a fully automatic way and therefore can speed up the overall annotation process significantly.

Challenges and solutions for automatic video analysis

A useful video analysis tool must fulfil a number of different challenges in order to be ease of use by users in humanity research. Some of the key challenges are:

- The algorithms must be able to cope with different number of persons;
- varying backgrounds must be taken into account in terms of arbitrary colour, texture, and motion as well as moving cameras;
- the algorithms should be able to extract meaningful information without any prior knowledge of the scene;
- the algorithms must be able to cope with different video quality, different spatial and temporal resolution;
- common video formats must be supported.

The overall goal is to achieve processing of videos in a fully automatic way, i.e., without the need for further human assessment and to reach a maximum overlap with annotations by human raters. The presented tool is based on different video processing techniques to achieve the above mentioned goals. It is built on top of previous development in the context of the Avatech project [3][4]. The detection and tracking of hands is based on skin colour, which is a unique feature of humans. Together with motion information, the person's visible hands in the scene are detected and tracked. A face detection and tracking module provides necessary positional information and other information about the faces of the captured persons [5][6]. The skin-colour based hand tracking module provides a number of information for each frame such as:

- the position of the hand

- the speed of the hand movement in succeeding frames
- directional information of the hand movement in succeeding frames

This frame-based information is then post-processed to get information for longer temporal segments e.g.:

- the start and end of a hand movement
- the temporal sub-segments in which the hand moves in the same direction
- relational information between hands and between hands and head

Furthermore, the tool also provides a number of additional important information as follows:

- If hands are touching each other, they are assigned additionally as joined hands.
- In case a person is wearing a short-sleeves shirt (as shown in Figure 3), the tool automatically detects it and separates the arm from the hand region, to increase the accuracy of hand tracking.
- Quite often a person's hands are not moving in space, but only the fingers are moved. This is very important information, which can be gathered from video analysis as well. This kind of movement is called intrinsic motion and hands are annotated respectively.
- Furthermore, the rest positions of both hands are calculated and are adapted over time. Rest positions provide valuable information notably for nonverbal behavior and gesture researchers.

By using the result of the face detection module, a body part assignment is performed in order to relate the face of a person to the detected hands of the same person. In Figure 1, two examples are given where the different hands, the head and the estimated rest position are visualized. The differently coloured ellipses at the hand position assign the left and right hands.

Figure 1. Results of hand and head tracking and related body part assignment.

Coding of hand movement behaviour

The positional, temporal, and relational information resulting from automatic video analysis are then transformed into corresponding value of the NEUROGES coding system. This coding system is divided in three modules as described above. Following the coding algorithm the continuous stream of hand movement behaviour is first segmented into *movement* units and *rest* units (Step 1). The *movement* units are then taken as the basis for the Structure category units (Step 2), the resulting Structure units are used as the basis for the Focus assessment (Step 3), etc. Consequently, the segmentation of behaviour becomes more and finer grained during the assessment process. As there is no pre-segmentation when starting with Module I, most segmentation is conducted in Module I, which imposes the biggest effort for manual annotation. (It is obvious that the finest grained annotation belongs to the first module, which results at the same time in the biggest effort for manual annotation.) Therefore, automatic video analysis efficiently contributes to an overall effort reduction by segmenting and further annotating the hand movement behaviour. In Figure 2, the values of the Module I categories Activation, Structure, and Focus are depicted. The red marked values are directly derived from the available results of the automatic video analysis tool. The resulting annotations are stored in an xml-file following the notion of the multi-media annotation tool ELAN and can be directly imported.

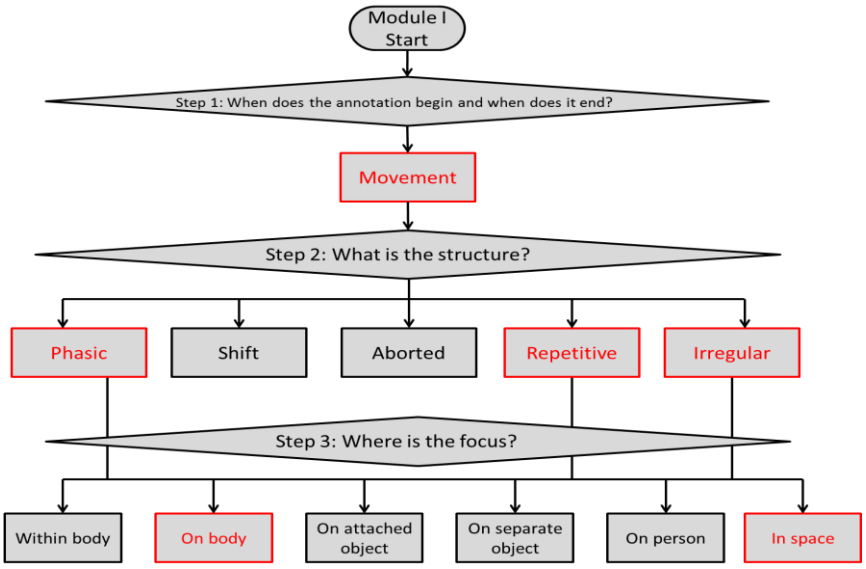

Figure 2: NEUROGES Module I with the categories Activation, Structure and Focus (according to [2])

It is important to note that the available video analysis results can be used for many different kinds of higher level annotations. For example, in the gesture community, the annotation of preparation, stroke, and retraction phases is important, which can be gathered from our results as well. Researchers are also interested about the position of the hands related to the body as defined in the McNeill gesture space [7]. As the position of the body is tracked continuously, the hands position can also be annotated relative to the body according to McNeill's definition. In Figure 3, an example image is given showing rectangles that relate to the gesture space. The inner rectangle represents the centre-centre part of the gesture space.

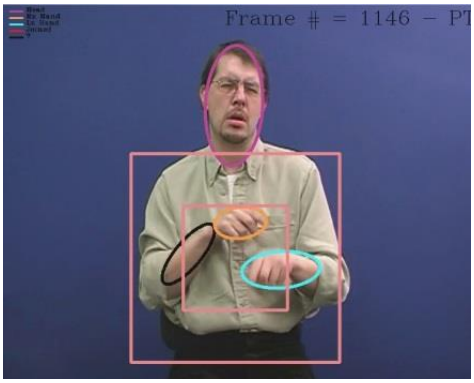

Figure 1; Visualization of gesture space

Evaluation

The evaluation has been performed by comparing the automatically created annotations with the ones created by a human rater. The ground truth information has been provided for two videos, for which the human rater has marked the start and end time of each *movement* unit (Activation category coding), each of which was then attributed a Structure value (Structure category coding). Currently, we restricted the automatic analysis and annotation to *phasic*, *repetitive, and irregular* units, since these are by far the most common occurring values in hand movement behaviour as analysed with the Structure category. For each frame in the two videos, it was checked, if there was an ongoing movement or not and then the values of precision, recall and F-measure (see [8]) have been calculated. Furthermore, the same procedure has been applied to evaluate the correctness of the assignment of the Structure value of hand movement. The results are summarized in Table 1. The algorithm achieves a very high accuracy in the detection of movement, while it is slightly less accurate when it comes to distinguish the type of hand movement performed.

The same behaviour arises when computing the temporal agreement for the Activation and Structure category: in the first case the average agreement between human rated annotations and automatically annotated ones is 0.77, while the modified Cohen's Kappa (computed with the EasyDiag tool [9]) has lower values, that is $\kappa = 0.11$ and κ-max = 0.56. We expect a notable improvement in these values once more ground truth annotations will be available.

	Precision	Recall	F1-Measure
Movement	91.2%	64.1%	75.3%
Phasic	48.5%	19.9%	28.3%
Repetitive	47.4%	40.2%	43.6%
Irregular	54.0%	64.0%	58.6%

Table 1: Performances on the detection and classification of structure of hand movement

Conclusion

A tool for automatic video analysis and annotation has been presented. The system described allows the researchers to save time by automatically detecting body parts and recognizing hand movement. The tool can be used in different research areas (i.e., gestural behaviour analysis, sign language analysis, interaction analysis including therapist patient interaction) and is capable to deal with a large variety of scenarios such as multiple persons, moving camera, short-sleeves tracking of hands and non-uniform background scenarios. The result of the analysis consists in a series of annotations representing the movements of the hands over time and the spatial relationship between the hands and the body. The higher level semantic annotations provided are designed to follow the NEUROGES coding system. The tool can run from within ELAN and the annotations it creates can also be exported as XML files for further analysis. First evaluations of the accuracy of the automatic annotations are promising, but further improvement is still required. Future work aims to improve the classification of hand movements and gestures, to make the system more robust (i.e., improve tracking in case of illumination changes or slow camera movements) and to add new types of annotations such as values from Modules II and III from the NEUROGES coding manual.

References

1. Lausberg, H., and Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, Instruments, & Computers, 41(3), pp.841-849. doi:10.3758/BRM.41.3.591.
2. Lausberg, H. (2013). NEUROGES - A coding system for the empirical analysis of hand movement behaviour as a reflection of cognitive, emotional, and interactive processes. In: Müller C, Cienki A, Fricke E, Ladewig SH, McNeill D, Tessendorf S (Eds.). Body – Language – Communication: An International Handbook on Multimodality in Human Interaction. Volume 1. Handbooks of Linguistics and Comunication Science: Berlin, New York: De Gruyter Mouton, pp 1022 - 1037.
3. Schreer, O. and Schneider, D. (2012). Supporting linguistic research using generic automatic audio/video analysis. Language Documentation & Conservation Special Publication No. 6 (2012): Potentials of Language Documentation: Methods, Analyses, and Utilization, ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 39–46.
4. Lenkiewicz, P. et al. (2012). AVATecH - automated annotation through audio and video analysis. In N. Calzolari (Ed.), Proceedings of the Eighth Int. Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp. 209-214. European Language Resources Association, May 23-25, 2012.
5. Kueblbeck, C. and Ernst, A. (2006). Face detection and tracking in video sequences using the modified census transformation, Journal on Image and Vision Computing, vol. 24, issue 6, pp. 564-572, 2006, ISSN 0262-8856.
6. http://www.iis.fraunhofer.de/de/bf/bsy/fue/isyst.html
7. McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought, Chicago: The University of Chicago Press, 1992.
8. Van Rijsbergen, C. J. (1979). Information Retrieval (2nd ed.). Butterworth.
9. Holle and Rein, in preparation.