Body Movements and Laughter Recognition: Experiments in First Encounter Dialogues

Kristiina Jokinen
University of Helsinki, Finland
University of Tartu, Estonia
kristiina.jokinen@helsinki.fi

Trung Ngo Trong
University of Helsinki, Finland
trung.ngotrong@helsinki.fi

Graham Wilcock
University of Helsinki, Finland
graham.wilcock@helsinki.fi

ABSTRACT

This paper reports work on automatic analysis of laughter and human body movements in a video corpus of human-human dialogues. We use the Nordic First Encounters video corpus where participants meet each other for the first time. This corpus has manual annotations of participants' head, hand and body movements as well as laughter occurrences. We employ machine learning methods to analyse the corpus using two types of features: visual features that describe bounding boxes around participants' heads and bodies, automatically detecting body movements in the video, and audio speech features based on the participants' spoken contributions. We then correlate the speech and video features and apply neural network techniques to predict if a person is laughing or not given a sequence of video features. The hypothesis is that laughter occurrences and body movement are synchronized, or at least there is a significant relation between laughter activities and occurrences of body movements. Our results confirm the hypothesis of the synchrony of body movements with laughter, but we also emphasise the complexity of the problem and the need for further investigations on the feature sets and the algorithm used.

CCS Concepts

•Computing methodologies → Computer vision; Discourse, dialogue and pragmatics;

Keywords

laughter; gesturing; body movement; deep learning

1. INTRODUCTION

Human non-verbal behaviour, gesturing and body posture, is related to the person's internal activation and spoken activity, and is important in enabling smooth communication [13]. Speakers complement their utterances and control and coordinate interaction by gesturing, and by observing gesturing behaviour it is possible to make inferences about the partner's activity level, intentions, and emotions, and to predict the success of interaction. Interesting research on co-speech gesturing has been conducted in a neuro-cognitive framework [7], while numerous communication studies have focussed on modelling human engagement, entrainment, and mutual synchrony ([5, 15, 24], to mention a few).

Here we report work on automatic analysis of laughter and human body movements in a video corpus of human-human dialogues. The Nordic First Encounters video corpus [16] is a collection of dialogues where the participants make acquaintance with each other for the first time. We use the Estonian part of this corpus, which consists of 23 dialogues collected in the Multimodal Interaction (MINT) project [11]. This corpus has manual annotations of the participants' head, hand and body movements as well as of laughter occurrences.

Two questions addressed in this paper are (1) how well can simple video processing techniques be applied to conversational data, and (2) can the participants' movements be detected so as to enable studies on the co-occurrence and correlation between laughing and body movements in communicative behaviour. The first question aims to explore the robustness of the new technology in human behaviour studies. We use OpenCV video processing to automatically detect head and body movements, and then use deep learning techniques to predict if the person is laughing given a sequence of video features. A simple bounding box technique was used to detect each person's movements, but we are also experimenting with Optical Flow techniques which have been successfully applied to action recognition. The second question aims to exemplify the application of the technology to a current interaction task. The hypothesis is that there is significant correlation between movement occurrences and laugh activities, i.e. it is possible to predict whether a person is laughing or not given video features for a sequence of her/his movements.

The paper is structured as follows. Section 2 describes automatic annotation of body movements. Section 3 discusses research on laughter in dialogues. Section 4 presents our initial analysis and Section 5 describes improvements introduced by adding context information. Section 6 presents the results and finally, Section 7 draws conclusions and discusses future work.

2. BODY MOVEMENTS

We identify head and body coordinates using a variant of the algorithm described in [24]. This uses well-known techniques in video processing to extract the so called *bounding*

MA3HMI'16 , November 16 2016, Tokyo, Japan

ISBN 978-1-4503-4562-0/16/11...$15.00

DOI: http://dx.doi.org/10.1145/3011263.3011264

box around the participant's contour, and then heuristically devides it into three sub-contours that represent the participant's head, body and legs. As the participants are off-camera at the start of each video, we use the first frame to give background edges for background subtraction. For the remaining frames of the video we subtract the background edges to leave only the person edges. After reducing noise by morphological dilation and erosion we find the current frame's contours [18]. As the background edges have been subtracted, the two largest remaining contours are the two persons who have entered the scene.

For each person, we divide the person contour vertically into three regions for head, body and legs. This exploits the fact that in these videos the persons are always standing. As the top region of the contour contains the head, we can find a very precise subcontour of the head within that region. The middle region contains the upper body, and we can find a horizontally accurate subcontour for the body, arms and hands within that region. The lower region contains the legs, but this subcontour is relatively unreliable (see below). We draw labelled bounding boxes around the head, body and legs contours, and add a time stamp, as shown in Figure 1.

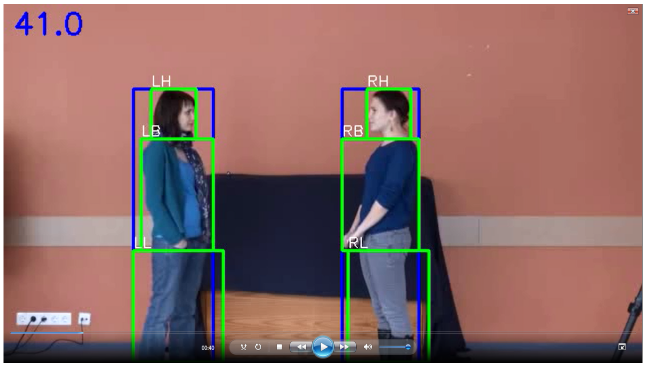

Figure 1: Video frame with bounding boxes for the head, body, and legs for the left and right participant.

In Figure 1, blue bounding boxes enclose the full persons, and green bounding boxes enclose the heads, bodies and legs. The green boxes are labelled LH (left person head), LB (left person body), LL (left person legs) and similarly RH, RB, RL for the right person head, body and legs.

Figure 2 shows a video frame with a set of bounding boxes for the head, body and legs during the participants' laugh. It shows how the participants lean towards each other and their head is also pushed foward. It also reveals that the vertical division of the body and legs regions is only approximate: the hand of the person on the right crosses the boundary between the body region and the legs region. The current algorithm does not include specific hand-tracking which is proposed in [17].

Of the three separate bounding boxes (the head, body and legs) for each participant, we use only the head and body bounding boxes for detecting laugh segments. As mentioned, the bounding box for legs is noisy and unstable, and if excluded, the main patterns in the signal remain unchanged. Moreover, feet and leg movements seem a rather distinct activity from laughing. The unstable legs bounding box also influences the the full person contour, so the bounding box for the whole body is also omitted from our studies.

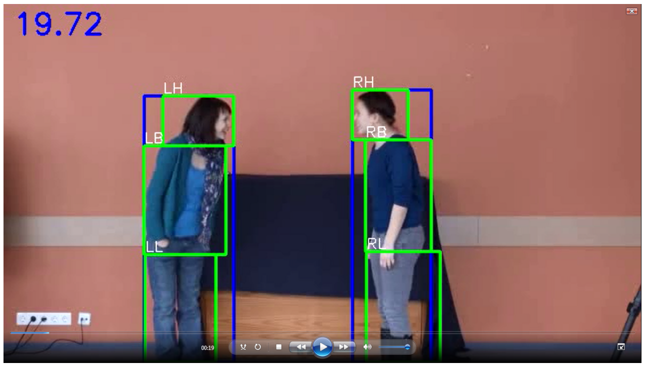

Figure 2: Video frame with the bounding boxes during laughter.

3. LAUGHTER

Laughter has been much studied in speech research with the focus on its acoustic properties, in particular to categorize various forms of laughter for emotion recognition [1, 21, 22]. In recent studies on social signals and their correlation with the interactional context it is shown that the timing of laughter is correlated with the interaction flow and it conveys information about the underlying discourse structure [4, 3]. Higher amounts of laughter occur in topic transition moments than in topic continuation moments and when the temporal distance from the topic boundary increases, laughter becomes more likely to occur. [9] studied laughter and engagement and noted that a significant change in the amount of laughter occurs at fifteen seconds around the topic changes.

Laughter occurrences have been classified into different types, such as mirthful and embarrassed, while the main division is usually between free laughter (voiced rhythmic laughing) or speech-laugh (laughing is simultaneous with speaking). [10] found significant overlap with speech and laugh in the corpus of student and student-teacher conversations, and that although there was no difference in the length of free laughter compared with that of speech-laugh, the duration of embarrassed laughs was significantly longer than that of all other types of laughter (mirthful, breathy, polite). [19] used a four-way classification, with the most common distinction between the spontaneous mirthful laugh and polite laugh, which together account for 80% of laughs.

Few studies, however, concern the correlation of laughter and other multimodal communication signals. In this paper we are interested in the correlation of body movements with laughter, hypothesising that laughter occurrences and body movement are synchronized, i.e. there is a significant co-occurrence relation between laughter and occurrences of body movements.

4. INITIAL FRAME ANALYSIS

Preprocessing of the data requires alignment of laughter annotations and video frames. Video frames are provided at fixed points in time (frequency 20 frames per second), but laughter segments are annotated for particular time durations, e.g. a laughter from time stamp 9.02 to 9.93 means the participant laughed for 0.91 seconds. In order to align laughter annotations with the frames, we copied the laughter labels onto all video frames from the start of the laughter till the end of the laughter, e.g. in the above example from 9.02 till 9.93. As a result, each video frame is labelled either *0* for not-laugh or *1* for laughter.

We selected three videos for the experiments, and for the construction of training and test sets, we collected all the frames from the videos (the length of the videos varies from 5 minutes to 6 minutes 40 seconds, so each video has about 7000 frames given the sampling frequency of 20), then aligned the frames with the laugh/non-laugh marking, and finally collected all the frames together in one big dataset. This dataset contains about 21,000 samples, and after shuffling, about 2/3 is used for training and 1/3 for testing.

Our initial experiments used straightforward frame-based models to correlate data points with laughter values. We first tried models using raw bounding boxes as features (autocorrelation), but this did not work. We then characterised the movement by focussing on the differences between the bounding boxes in two separate frames. Following gesture studies (e.g.[17]), we calculated speed and acceleration for the movement. The timestamp differences **t-9** and **t-25** indicate the best heuristic values found via experimental trials to select the consecutive frames when calculating the values from a simpler system to a more complicated one. The following feature sets were tried:

- speed of change: difference between bounding boxes at time **t-1** and **t**
- acceleration of change: difference of the difference between bounding boxes at time **t-1** and **t**
- cumulative differences, and cumulative differences of differences from time **t-25** to **t**
- cumulative differences, and cumulative differences of differences from time **t-9** to **t**

However, these features did not work well except for the last feature set, which seemed to work for some videos. Figures 3 and 4 show the analysis of head and body movement features for two of the videos. The red lines are the frames which are labelled as laugh activities. The features in the figures are as follows:

- **d1** is the cumulative differences. This is the speed with which the bounding box is changing over time.
- **d2** is the differences of the differences. This is the acceleration of the speed.
- **Person_d1** and **Person_d2** are combinations of moving signals of Head and Body related to **d1** and **d2**.

The movement peaks represent high-activity regions, and mostly appear at the beginning when the two people shake hands. The activity decreases towards the end of the interaction. A few larger movement patterns can be observed in the middle of the videos, indicating that standing participants are not completely still during the interaction but move their body and hands. However, these movements need not be related to laughter, but may indicate pointing or waving gestures. We conclude that although these plots do not clearly visualise individual laughing events, they do show the overall joint patterns of the participants' movement activity.

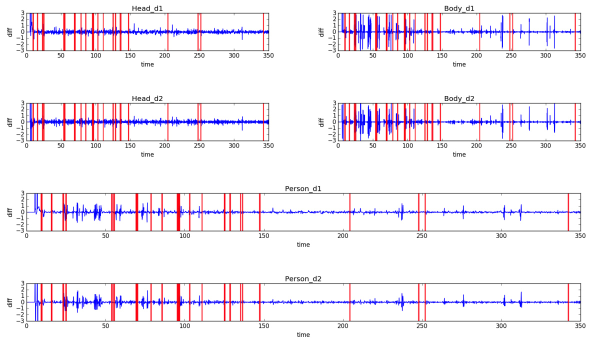

Figure 3: Head and body movements in Video 1, with laughter frames marked with red lines.

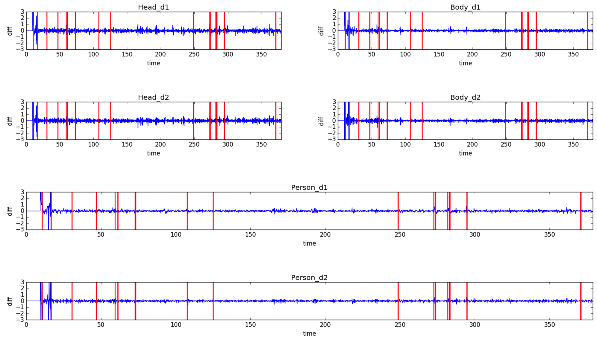

Figure 4: Head and body movements in Video 2, with laughter frames marked with red lines.

5. ADDING CONTEXT

It is understandable that the simple frame-based approach does not work for recognizing laughter and gesture correlation: individual frames are only short stretches of laughter, and obviously larger contextual information is needed. From the dialogue point of view, a good indicative feature of the context would be the topic of the conversation since more laughter occurs at topic changes [2]. On the other hand, [6] distinguished different phases in laughing which are connected with breathing and phonetic properties of laughter. Furthermore, it is also clear that if there is correlation between laughing and body movements, the context of these movements in which the laughing co-occurs is important. [13] analyses hand gestures as consisting of preparatory, stroke, and post-stroke (retraction) phases, so it is reasonable to assume that laughter-related movements would follow this kind of structure as well, although, of course, the actual timing of gesturing and laughing is complicated, and would require more systematic studies.

The present approach is based on a bottom-up video signal analysis, so we use the frames as the context to capture the pre- and post-laugh events. The relevant new information includes the context where a laughter signal can be observed for a couple of frames before the actual laughter, or where there occurs post-laugh signals after the laughing. Hence, we group frames into a sequence of frames which we call a *super-frame*, and do the preprocessing as before (d1 + d2) for all the super-frames. After some experimentation, the group size of 10 was found as an optimal size.

Figure 5 shows the results of taking the context information into account. The circles highlight the signal fluctuation when one of the participants laughed.

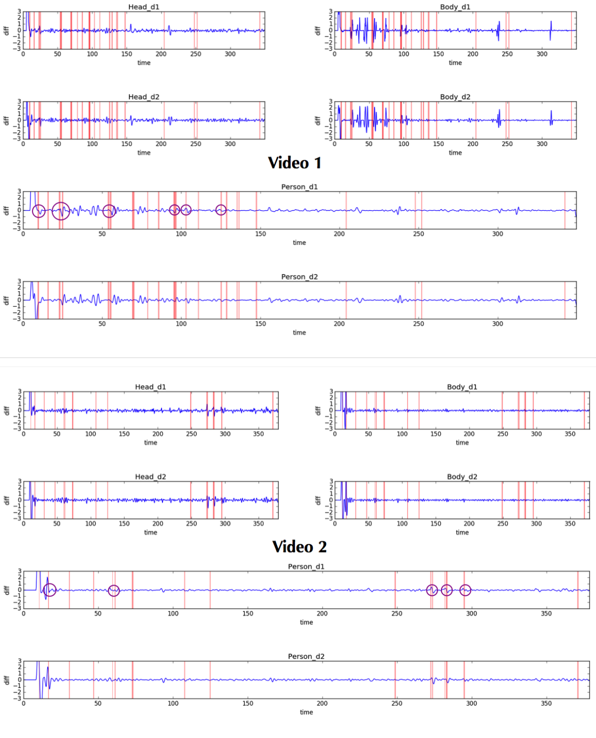

Figure 5: Head and body movements in Videos 1 and 2 when the contextual information is added.

We continued with a statistical analysis using these features. Five different unsupervised clustering algorithms were applied, each of which using a different principle for clustering the dataset. Principal Component Analysis [12] searches for the dimensions that maximize the variances within the data points, preserving as much information as possible in the reduced dimensions. t-distributed Stochastic Neighbor Embedding [23] concentrates on similarities between data points and constructs a low-dimensional embedding that minimizes the Kullback-Leibler divergence to the joint probabilities of original distribution. Isomap Embedding [20] and Spectral Embedding are non-linear dimensionality reduction methods. The first one estimates the intrinsic geometry of data based on a rough estimate of each data point's neighbours, while Spectral embedding constructs a similarity matrix of data points and forms new dimensions from the eigenvalues of this matrix. Multidimensional scaling [14] is a popular technique for visualizing the similarity of individual cases. The algorithm tries to preserve the distances between objects in N-dimensional space as well as possible.

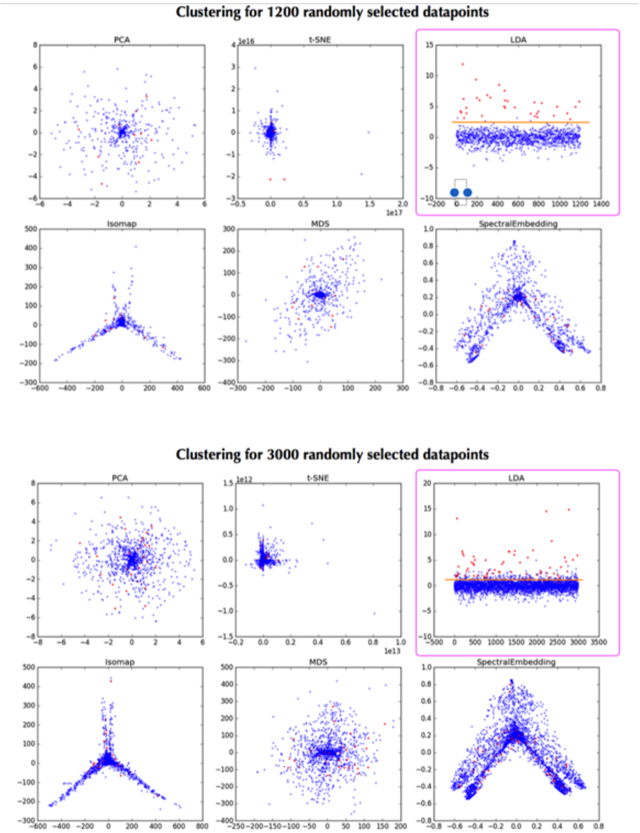

Figure 6: Visualizations of 6 clustering algorithms. From left to right upper row: Principal Component Analysis, t-distributed Stochastic Neighbor Embedding, Linear Discriminant Analysis. Lower row: Isomap Embedding, Multidimensional scaling, Spectral Embedding.

We also used one supervised classification algorithm, Linear Discriminant Analysis (LDA) [2]. LDA optimizes its subspace by maximizing the differences between classes, and by minimizing the differences within classes.

Figure 6 shows visualisations of the statistical relations of the data points with the laughter labels using the five different clustering algorithms and the supervised LDA. The red dots represent "laughing" data points, while the blue ones are "non-laughing" data points.

6. RESULTS

If the data points are mixed together, it is difficult to recognize specific groups (i.e. laugh vs. non-laugh). We can see from Figure 6 that none of the five clustering algorithms can distinguish the data points well, i.e. there are no obvious differences between features representing "laughing" samples and "non-laughing" samples.

By contrast, the Linear Discriminant Analysis found a separating line between laughs and non-laughs but only if the dataset was small (1200 data points); with larger data (3000 samples), the two classes are significantly mixed. In the first case the red dots (samples labelled as laughing) are also spread all over the horizontal axis, which indicates a heterogeneously distributed dataset and strong complexity of the cases. Unlike the other applied algorithms, LDA is a supervised algorithm which means the label of each sample is already known. The highlighted plots in Figure 6 refer to the LDA analysis.

We trained 14 different models on bounding boxes, but all provide poor performance. As Figure 6 shows, the best model uses LDA to transform the data to a more distinguishable space. Hence we created a pipeline: [transform data using LDA] -> [train classifier to discriminate 2 classes].

This algorithm performs decently on the training set, but LDA fails to capture the complexity of all the laughing samples (i.e it cannot transform the test data to a more separable space). This is reasonable because: 1) LDA is a linear algorithm, 2) there are significant differences within the samples that are labelled as laughing, 3) there exist ambiguities between laughing and non-laughing frames (indicated by all the unsupervised algorithms mixing them up).

7. CONCLUSIONS AND FUTURE WORK

In this paper we have studied correlation and co-occurrence of laughter and body movement in first encounter interactions, and tried to answer the questions of how well simple video processing techniques can be applied to conversational data, as well as whether the participants' movements can be detected in this way for further studies on communicative behaviour. We can conclude that laughing bears a relation to head and body movement but that the details of co-occurrence need more studies.

We used bounding boxes to provide a rough estimate of the participants' movement, and concluded that although they give helpful information showing some correlation with laughter activity, bounding boxes alone are not enough for a laughter detection algorithm alone.

When building models for predicting laughter activity, misclassification between normal activity and laughter may occur because: 1) there exist significant differences among laughter samples; for example laughter at the beginning of (first-encounter) conversations is a formal way of greeting each other, while it is also an expression of joy, or occurs quietly without any other actions; 2) there are ambiguities between laugh and non-laugh signals; for example people have a wide variety of head and body movements within any laughter activity.

To remedy the roughness of bounding boxes, we are also investigating the use of dense optical flow [8] for laughter recognition. Figure 8 shows the dense optical flow in the same frame with laughter shown in Figure 7. We hope to provide results from this approach in due course.

Future work concerns more detailed analysis of laughter occurrences, taking into account various types of laughter to see if body movements are related to particular types of laughter. We wish to compare results in larger cultural contexts and study differences in social signals, laughter and multimodal communication using deep learning with visual and speech features in intercultural communication.

We also plan to do further analysis on facial expressions, since smiling is closely related to laughing, and can be one of the pre- and post-laugh activities, besides having its own functioning in human communication. The bounding boxes can be very useful for extracting the pixels of heads and bodies. However, since there are large amounts of facial datasets available, we intend to leverage these data to enhance our model by having more powerful facial features.

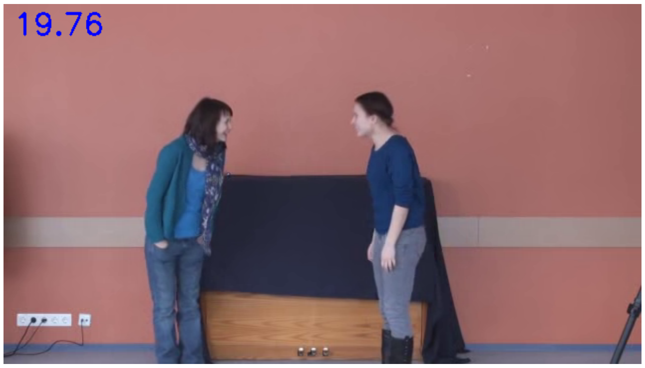

Figure 7: A video frame where laughter occurs.

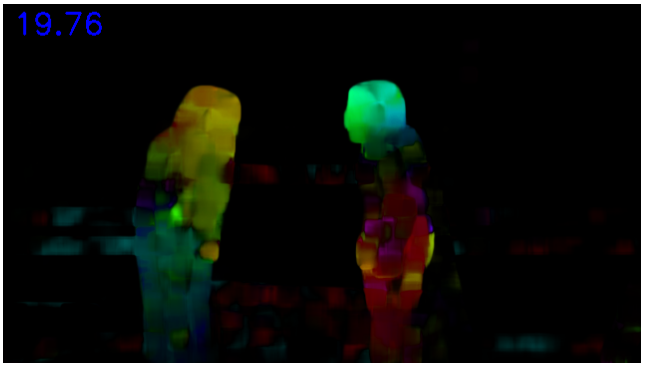

Figure 8: Dense optical flow at the same frame as Figure 7.

8. ACKNOWLEDGMENTS

The work is funded by the Academy of Finland through the project Fenno-Ugric Digital Citizens (grant n°270082). The first author would also like to thank support of the Estonian Science Foundation project (grant n°IUT 20-56).

9. REFERENCES

[1] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal of Acoustic Society of America*, 110:1581–1591, 2001.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003.

[3] F. Bonin. *Content and Context in Conversations: The Role of Social and Situational Signals in Conversation Structure.* PhD thesis, Trinity College Dublin, 2016.

[4] F. Bonin, N. Campbell, and C. Vogel. Time for laughter. *Knowledge-Based Systems*, 71:15–24, 2014.

[5] N. Campbell and S. Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *Proceedings of 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Makuhari, Japan, 2010.

[6] W. Chafe. *The Importance of Not Being Earnest: The feeling behind laughter and humor*. John Benjamins, Amsterdam, 1977.

[7] A. S. Dick, S. Goldin-Meadow, U. Hasson, J. I. Skipper, and S. L. Small. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, 30(11):3509–3526, 2009.

[8] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In J. Bigun and T. Gustavsson, editors, *Image Analysis (Lecture Notes in Computer Science 2749)*, pages 363–370. Springer, 2003.

[9] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Laughter and topic transition in multiparty conversation. In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2013)*, pages 304–308, Metz, 2013.

[10] K. Hiovain and K. Jokinen. Different types of laughter in North Sami conversational speech. In *Proceedings of the LREC Workshop Just Talking - Casual Talk among Humans and Machines*, Portoroz, 2016.

[11] K. Jokinen and S. Tenjes. Investigating engagement: Intercultural and technological aspects of the collection, analysis, and use of Estonian multiparty conversational video data. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, 2012.

[12] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[13] A. Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[14] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, Beverly Hills and London, 1978. Sage University Paper series on Quantitative Application in the Social Sciences, 07-011.

[15] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova. Acoustic-prosodic entrainment and social behavior. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montreal, Canada, 2012.

[16] C. Navarretta, E. Ahlsén, J. Allwood, K. Jokinen, and P. Paggio. Feedback in Nordic first-encounters: a comparative study. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, 2012.

[17] P. Saatmann and K. Jokinen. Experiments with hand-tracking algorithm in video conversations. In *Proceedings of 2nd European and 5th Nordic Symposium on Multimodal Communication*, Tartu, Estonia, 2014.

[18] S. Suzuki and K. Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30:32–46, 1985.

[19] H. Tanaka and N. Campbell. Acoustic features of four types of laughter in natural conversational speech. In *Proceedings of XVIIth ICPhS*, Hong Kong, 2011.

[20] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[21] J. Trouvain. Segmenting phonetic units in laughter. In *Proceedings of XVth ICPhS*, pages 2793–2796, Barcelona, 2003.

[22] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[23] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9:2579–2605, 2008.

[24] M. Vels and K. Jokinen. Recognition of human body movements for studying engagement in conversational video files. In *Proceedings of 2nd European and 5th Nordic Symposium on Multimodal Communication*, Tartu, Estonia, 2014.