Method for visualisation and analysis of hand and head movements in sign language video

Matti Karppa[1]*, Tommi Jantunen*[2]*, Markus Koskela*[1]*, Jorma Laaksonen*[1]*, and Ville Viitaniemi*[1]

[1]Department of Information and Computer Science,
Aalto University School of Science, Espoo, Finland
[2]Sign Language Centre, Department of Languages, University of Jyväskylä, Finland
{matti.karppa,markus.koskela,jorma.laaksonen,ville.viitaniemi}@aalto.fi
tommi.j.jantunen@jyu.fi

Abstract

This paper presents a method for the visualisation and motion analysis of hand and head movements in videos containing sign language and gestures. The method detects the parts of the person's bare skin on a video with an adaptive colour model, characterises the shapes of the hands and the head with a point distribution model, and tracks their motion separately by using the Kanade-Lucas-Tomasi algorithm and active shape models. The quantitative results are visualised in ELAN annotation software. The method is demonstrated in the paper in terms of its relevance to the annotation and analysis of sign language.
Index Terms: sign language, gesture, skin detection, point distribution models, active shape models, motion analysis

1. Introduction

This paper outlines the second development stage of our system for visualisation and semi-automatic analysis of motion in videos containing continuous signing and gestures produced with the hand and head articulators. The first version of the system [1] detected the different parts of the informant's bare skin on the video (the two hands and the head), tracked the combined motion of these parts of the body with the Kanade-Lucas-Tomasi algorithm [2], and represented the different characteristics of the motion of these parts with five statistical descriptors (the total amount of motion points, the total amount of horizontal and vertical motion, and the length of velocity and acceleration vectors). The functionality of the system was tested with Finnish Sign Language (FinSL) data. The linguistic analysis of the test results showed that signs involved more horizontal and vertical motion of the articulators than transitions between signs (i.e., signs were more controlled), that the total amount of motion was higher during transitions than during signs (i.e., transitions were more holistic), and that the combined speed of the three articulators was slower during signs than during transitions. The results agreed with the findings made in sign language phonetics and linguistics concerning the nature of signs and transitions [3, 4, 5] and we argued this agreement to be a positive indicator of the validity of the method.

The second version of the system has several new features and improvements over the first version. The major one of these is the capability of tracking the motion of different skin-coloured articulators separately, a feature that was specifically called for in the evaluation of the first system [1]. The other major improvements include the utilisation of active shape models [6] in motion tracking, and the option to import and visualise the quantitative results in ELAN annotation software [7], developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands[1]. The new system is presented in detail in this paper with reference to FinSL.

The main motivation behind the present work is that sign language and gesture researchers still lack a feasible way of visualising the motion of the hands and the head in their primary research material, video. In spoken language research, the option to visualise the speech signal is present in most analysis programs (e.g. *Praat*), and the visualisation unquestionably helps the researcher to observe systematicity and distinctions in the speech data in a way that is not possible by merely listening to a person producing language. For this same reason we consider it mandatory to develop corresponding technology also for the purpose of sign language and gesture research.

A consequential further motivation for the present work comes from the demands created by the new emerging field of corpus-based sign language research. Building a representative sign language corpus requires one to collect a large amount of sign language video material which is then subjected to detailed annotation, for example, in ELAN. Annotation, in turn, relies heavily on information on how the articulators move: for example, the identification and annotation of signs from a video is normally done by observing the visible changes in the direction of the movement of the signer's active hand [8, 9]. By making the relative changes in the movements of the articulators more visible in ELAN, our system directly contributes to and supports the annotation process and the creation of representative sign language corpora.

Finally, the system enables the researcher to conduct a type of motion analysis of signing and gesturing that has traditionally been feasible only if the data has been produced in predetermined laboratory settings using complex motion capture equipment and software. The downside of motion capture data has always been that it requires the informants to wear data gloves or other types of markers on their body [5, 10], which in many cases may lead to an unnatural production of signs and gestures. With our computer vision-based method, laboratory settings may be entirely avoided in the motion analysis of signs and gestures: ideally, the system allows researchers to work with videos containing natural (e.g. discourse) data and to base their claims more on the natural use of sign language and gestures.

[1] http://www.lat-mpi.eu/tools/elan/

2. The method

The analysis of hand and head movements in this work is based on bottom-up visual analysis. The process consists of a number of distinct steps: first the face of the person on the video is detected, then all skin regions are located and represented as blobs, the body parts are segmented and modelled using point distribution models and active shape models, and finally the motion of the body parts is tracked and represented using statistical descriptors. Figure 1 illustrates these processing stages.

The method is presented and demonstrated with FinSL video data. The analysed recording had been made prior to our current project and the possibility of automated analysis had not been taken into account at the time of recording. The shooting of the video is close to frontal and the quality of the material is PAL DV, 720×576 pixels and 25 frames/s at 3.6 Mbytes/s. The total duration of the recording is approximately one minute. The signing represents a semi-rehearsed monologue and describes the attitudes of younger deaf generation towards traditional deaf clubs.

2.1. Face detection

The analysis begins with face detection (Figure 1b), for which we use the Viola-Jones cascade face detector [11] available in OpenCV[2]. Due to the recording setup, the face direction is mostly frontal and the location of the single face in the video is restricted to the upper centre part of the view. In order to further avoid false positives, the face is expected to be found at approximately the same location in any two consecutive frames. Presently, the centroid of the face is allowed to move at most 25 pixels between two frames, and the area enclosed within the face rectangle is allowed to change up to 50 percent. If a face cannot be detected in a sequence of frames, its location is interpolated between the preceding and following successful detections.

2.2. Skin detection and region extraction

After the face detection has been performed, skin-coloured regions are located by using a detector based on multiple multivariate Gaussian distributions in the HSV colour space Figure 1c). The face is assumed to be mostly skin-coloured, so the pixels used to train the detector are chosen from an elliptical area within the face detection results from the first frame. Colour vectors are then clustered using K-means, and out of the K clusters, $R < K$ most populous ones are selected to be used with the model. Inverse covariance matrices Σ^{-1}, square roots of determinants of covariance matrices $|\Sigma|^{1/2}$ and mean vectors μ are computed for each of the R distributions. The actual detections are performed on a per-pixel basis. The probability density value of the skin model is computed for each pixel in each of the R distributions, and if the value exceeds a predetermined threshold θ in any of the R distributions, the pixel is classified as a skin pixel. In our current setup, typical parameter values have been $K = 9$, $R = 4$ and $\theta = 1.86 \times 10^{-7}$.

Postprocessing steps include applying morphological opening to reduce pixel noise caused by randomly misclassified pixels in the background, and to smoothen the borders of the skin-coloured regions. In addition, contour finding is employed to locate and fill holes in skin regions. Such holes tend to occur especially in the face because of eyes, lips and eyebrows.

Once the rudimentary skin detection has been performed, and a corresponding binary mask has been obtained, interconnected skin pixel regions (blobs) are extracted from the mask (Figure 1d). A basic sequential algorithm is used to find 4-connected pixel regions, after which a size threshold is applied, and too small regions are removed. This step reduces the number of misclassified skin areas, as most of those areas are very small.

Finally, the interconnected pixel regions are given crude identities as body parts. Currently, these identities are the head, the left hand, the right hand, or any combination thereof. The estimated identities have been colour-coded in Figure 1d: green for the head, and red and blue for the right and left hands, respectively. Criteria for deciding the identity include region centroid displacement from the detected location of the face, the number of distinct regions, and centroid locations with respect to other regions. Ideally, three distinct regions are found: one for the face, and two other blobs below it, sufficiently far apart. However, when two or more body parts are occluding one another, the regions will appear merged, and in such cases, the region is marked ambiguous, and a combined identity is assigned to it.

2.3. Feature point tracking

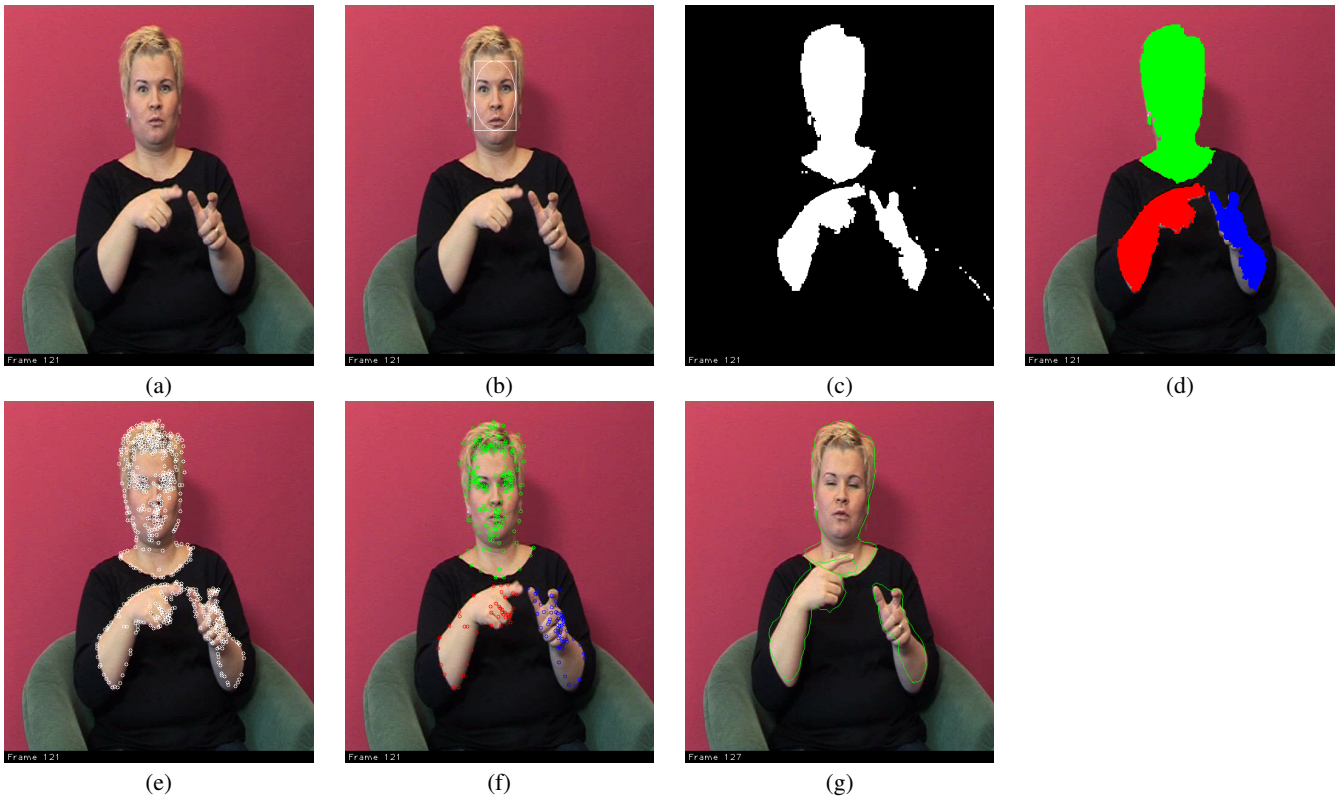
After skin colour detection, we track the skin areas with local motion in the video stream by using the Kanade-Lucas-Tomasi (KLT) algorithm [2], which is based on first detecting distinctive corner pixel neighbourhoods and then minimising the sum of squared intensity differences in corresponding small image windows between successive video frames (Figure 1e). If the appearance of the pixel neighbourhood changes too much, for example, due to occlusion or complex 3-D motion, we consider the motion point to be lost. To replace any lost motion points and to track any new areas of motion, we detect and initiate new distinctive pixel neighbourhoods in each video frame. The tracked points are either matched to some of the previously identified body parts, or ignored if no match is found. In Figure 1f, the assigned body part identities have been indicated with different colours.

2.4. Point distribution model

A separate point distribution model (PDM) with $M = 60$ landmark points is constructed for describing each of the three modelled body parts. PDMs are a commonly used method for statistical shape description. There the shape is described in terms of how the positions of landmark points identified from an object typically vary relative to each other. PDMs are usually learned from a collection of training shapes by first identifying the landmark points, aligning the landmarks of different training examples, and performing principal component analysis (PCA) of the landmark coordinates.

For the body part PDMs, the training shapes are chosen automatically from the video by examining the interconnected pixel regions extracted from the binary shapes after the skin region extraction. Blobs whose identities have been decided to be non-ambiguous are used for training the model. The bottommost point in the region is assumed to be either the base of the visible neck or the elbow and is therefore deemed relatively stable, thus the first landmark is placed there. The remaining landmarks are placed at equal intervals around the boundary of the region. The training shapes are aligned with each other using the algorithm presented in [12].

[2] http://opencv.willowgarage.com/wiki/



(a) (b) (c) (d)

(e) (f) (g)

Figure 1: *Processing stages in the video analysis: (a) frame of input video, (b) face detection, (c) detection of skin-coloured regions, (d) skin blob detection, (e) corner point detection, (f) feature point tracking, (g) fitting of active shape models.*

2.5. Active shape model

The point distribution models are used as a basis for active shape models (ASM) that track the body part poses and shapes between consecutive frames of video. Figure 1g illustrates the body part tracking with active shape models. Note that in this video frame, the signer's right hand is occluding her neck. Despite the occlusion, the ASMs are able to maintain the distinction between the body parts.

The ASM tracking employs the iterative algorithm presented in [6] for updating the pose and shape parameters of the ASM. The set of pose parameters consists of scale s, orientation ϕ and translation (x_c, y_c). The body part shape in turn is represented in the previously trained PDM parametrisation. In the algorithm, the parameters are first initialised to give an approximate fit. After this, the following iteration is repeated until convergence:

1. New target landmarks are selected.
2. Given the current estimate of PDM shape parameters, the pose parameters are updated so that the model would correspond to the selected landmarks.
3. Given the pose parameters, the PDM shape parameters are updated in turn.

Our implementation of the algorithm uses the interconnected pixel regions and history information when setting the initial shape and pose parameters; the shape is expected to change only a little between consecutive frames, so the shape parameters, resulting from fitting in the previous frame, are used as the initial shape. Initial model location parameters are set so that the bottommost landmark matches that of the skin region of the bodypart in question. The model is then rotated while preserving the location of the bottommost landmark point. Initial orientation is chosen so that the number of pixels at the intersection of the area enclosed within the boundaries of the shape and the skin region is maximised. If the region was deemed ambiguous, i.e. containing two or more disjoint skin regions, it is first split in half at its centroid. Splitting is done vertically when two hands are merged in the same region, and horizontally when the head is merged with one or two hands in the region.

When fitting, target landmark selection is performed using Sobel intensity gradients. Skin-coloured regions are assumed to have higher intensities than the background, so the desired points are expected to have strong gradients pointing towards the centroid of the region.

2.6. Quantitative measurements

Performing the Kanade-Lucas-Tomasi tracking and fitting the active shape models provide quantitative measurements of body part movements. Altogether, KLT and ASM measurements combined, we obtain a feature vector of 53 features per frame.

As to point feature tracking, the features are associated with the body parts obtained in the skin segmentation stage. Short-lived points are ignored; currently, to be included in computations, each point is required to have been tracked over three consecutive frames. Velocity and acceleration vectors are then computed for each point in each frame. For a tracked point i in frame f, denoted $\mathbf{x}_{i,f}$, these are defined as $\mathbf{v}_{i,f} = (\mathbf{x}_{i,f+1} - \mathbf{x}_{i,f-1})/2$ and $\mathbf{a}_{i,f} = \mathbf{x}_{i,f+1} - 2\mathbf{x}_{i,f} + \mathbf{x}_{i,f-1}$.

For each frame, five motion features are computed using these point features. These are the number of points tracked in the frame $D_1 = N$, the sum of horizontal velocity components $D_2 = \sum_{i=1}^{N} v_{x,i}$, the sum of vertical velocity components $D_3 = \sum_{i=1}^{N} v_{y,i}$, and the Euclidean norms of the sums of velocity and acceleration vectors $D_4 = ||\sum_{i=1}^{N} \mathbf{v}_i||$ and $D_5 = ||\sum_{i=1}^{N} \mathbf{a}_i||$.

These five features are computed separately for each of the three body parts. As most points are rather short-lived, the present method of assigning body part labels to them may produce ambiguous results. Therefore, the values are computed twice: first by including a minimum set of relevant points, then by including all relevant ambiguous points. In addition, total values are computed, including all tracked points, disregarding identity information, leading to the total of 35 features.

As to the ASM, the current version of the program can track centroid locations (x_c, y_c) and orientation angles ϕ, measured as the angle between a chord from the bottommost landmark point to the farthest point in the model and the positive x-axis. Since each ASM has a body part identity, these measurements yield 18 features:

- *angular velocity of head or left/right hand* (3 in total)
- *horizontal/vertical components and the magnitude of the velocity vector of head or left/right hand* (9 in total)
- *area of the intersection of pixels enclosed by two ASMs* (3 in total)
- *angular direction of the part* (3 in total)

3. Demonstration of the method

The applicability of the method is demonstrated below from the perspective of its potential for visualising motion and relevance to the linguistic annotation and analysis of sign language. The results of the quantitative measurements are exported into a CSV file that, in turn, is imported to ELAN by using its *Linked Files* function. In ELAN, after the configuration of tracks (see [13]), the quantitative motion information is visualised as line graphs in trackpanels embedded in the Timeseries Viewer (see Figure 2). All motion information in the Timeseries Viewer (i.e., line graphs) is synchronised and time-alignable with the information shown in the Video Viewer (i.e., video) and in the Timeline Viewer (i.e., tiers and annotations).

Note that all annotations of signs shown in this demonstration have been made prior to importing the quantitative motion information to ELAN. Signs have been identified on the basis of video by following the manual method presented in [14]. Transitions in between signs have been identified at the same time with signs by using ELAN's *Create Annotations from Gaps* function.

The capability of the method to correctly track and visualise the changes in the movement of the signer's active hand, marking the beginnings and ends of signs (see [4, 5, 8, 9]), can be demonstrated with the FinSL sign MONEY in Figure 2. The sign MONEY occurs in the monologue as the first part of the co-compound meaning 'treasurer' (literally MONEY+CARE-TAKER 'the one who takes care of the money'). The sign is produced in a way that the flat handed active hand (palm orientation upwards) moves downwards and contacts the palm of the flat handed passive hand (palm orientation also upwards). In Figure 2, the downward movement of the active hand is represented by a falling slope of the blue line graph that describes the change in the amount of vertical motion of the active hand. As seen from the red line graph describing horizontal motion, the articulation of the sign MONEY involves also side-to-side movement of the active hand. In Figure 2, the blue-coloured vertical block shows which portion of the visualised motion information corresponds to the temporal domain of the sign MONEY.

Investigation of the signs CARE-TAKER and SECRETARY in Figure 2 reveals that the method tracks and correctly represents the changes in the amount of horizontal and vertical motion also in them. The movement of the sign CARE-TAKER is a single sweeping (up-to-down and side-to-side) arc-like movement. In the sign SECRETARY, the movement is repeated which explains the fall-rise-fall curve.

In general, the visualised motion information concerning the direction of the movement has undeniable value for the manual segmentation of continuous signing into signs. The changes in the direction of the movement of the articulators can be hard to notice by looking only at the video, but they are easily detected by observing the graphs representing the changes in the horizontal and vertical motion. The visualised information also aids in detecting the internal variation found in movements. For example, it is a well-known fact that head shakes used to negate FinSL sentences (see [15]) are phonetically reduced towards the end of the sentence. This is easily seen from the graphs describing the amount of horizontal and vertical head motion. Two examples of such graphs are given in Figure 3.

The FinSL sentence [YOUNG index INTEREST NO] in Figure 3 is a topic-comment structure [16] with an overall meaning 'The young are not interested (in taking care of the duties)'. The negative head-shake in which the head rotates from side to side begins towards the end of the index finger pointing that marks the topic [YOUNG index] 'the young' and lasts till the end of the following comment [INTEREST NO] 'are not interested'. As can be seen from the graph describing the amount of the horizontal motion of the head in Figure 3 (the upper track panel with red line), the amplitude of the head movement does not stay the same throughout the production of the sentence, but decreases towards the end.

In the phonetic analysis of sign languages, the velocity and acceleration information of different articulators that we are able to calculate from the video is especially useful. Figure 4 shows an example of this type of information. The graph in the figure describes the speed (i.e. the magnitude of the velocity vector, or the Euclidean norm) of the active hand calculated on the basis of active shape model data.

The graph in Figure 4 shows how both signs and transitions occurring in continuous signing tend to contain a roughly parabolic-shaped speed curves. Visual observation has revealed that corresponding curves characterise signs and transitions also when the measurements are based on traditional motion capture data [5, 10].

4. Conclusion

In this paper, we have presented an improved version of our computer vision based method that allows researchers to semi-automatically visualise and analyse the motion information in videos containing continuous signing and gestures. The method detects the parts of the person's bare skin (the hands and the head, i.e. the articulators) on the video with an adaptive colour model, characterises the shapes of the articulators with a point distribution model, and tracks the motion of the articulators by using Kanade-Lucas-Tomasi algorithm and active shape models. The quantitative results are visualised using the ELAN an-

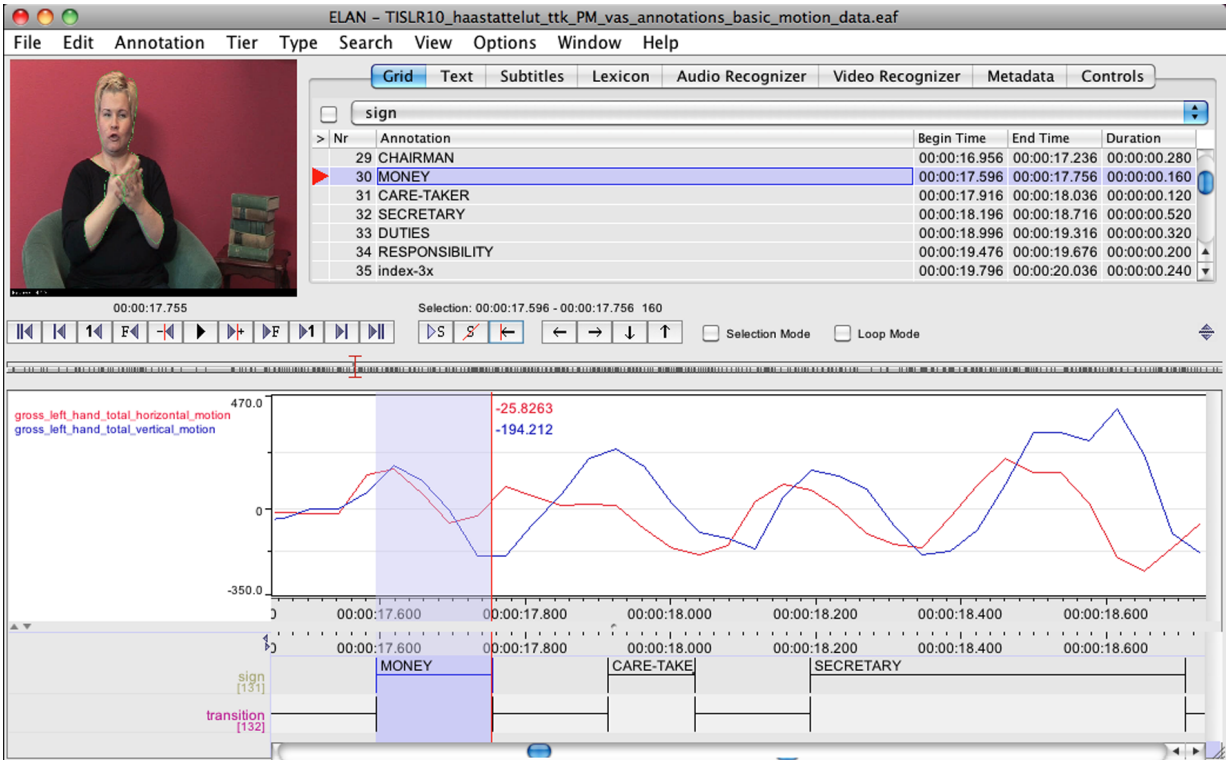


Figure 2: *ELAN screenshot showing a selection of different viewers. The trackpanel in the Timeseries Viewer contains linegraphs visualising motion information about the horizontal (red line) and vertical (blue line) motion of the signer's active hand. In the figure, the active hand is labelled as* left hand *because it is located on the left side of the video frame.*

notation software. The main advantage of the visualisation is that it allows the researcher to observe the fine distinctions in the video data more accurately than just by looking at a person producing sign language or gestures.

Although the method is still in its developmental phase, it can already be used as a tool in the annotation and analysis of sign language and gestures. The demonstration of the method with the FinSL data showed that the method efficiently captures the changes in the direction of the movements of the articulators and, consequently, it can be used to enhance the identification and annotation of signs from sign language video. Furthermore, the demonstration showed that the method is very capable of tracking and visualising the internal variation of head movements, and that it can track and visualise also more complex aspects of the motion of the articulators, such as the speed (i.e., velocity magnitude) of the active hand. Traditionally, such motion tracking has required the use of complex motion capture equipment and laboratory settings that in many cases result in unnatural production of data; for example, data gloves or retroreflective markers attached to the hands are known to block the proper articulation of many contacting and two-handed signs.

In the future the method will be developed further and tested with varying video materials. This work will include a more detailed modeling of articulators through which we expect to obtain more precise information concerning, for example, hand-internal movements and facial gestures. Such information is needed in our planned attempts to develop a more automatised mechanism for the annotation of larger sign language corpora. The future work will also include a calibration of the method with traditional motion capture data. We expect this work to further improve the veracity of the results of the method so that the method could be eventually launched as a tool to serve the needs of the community.

5. Acknowledgements

This work was financed in part by the Academy of Finland under grants 134433 and 140245 and as the Center of Excellence in Adaptive Informatics Research project.

6. References

[1] T. Jantunen, M. Koskela, J. Laaksonen, and P. Rainò, "Towards the automated visualization and analysis of signed language motion: Method and linguistic issues". In *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010, pp. 1–4.

[2] J. Shi, and C. Tomasi, "Good features to track". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '94)*, Jun 1994, pp. 593–600.

[3] S. Wilcox, *The Phonetics of Fingerspelling*. John Benjamins, 1992.

[4] R. E. Johnson, and S. K. Liddell, "A segmental framework for representing signs phonetically". *Sign Language Studies*, vol. 11, no. 3, pp. 408–463, 2011.

[5] T. Jantunen, "Signs and transitions: Do they differ phonetically and does it matter?". *Sign Language Studies*, vol. 13, no. 2, 2013, (forthcoming).

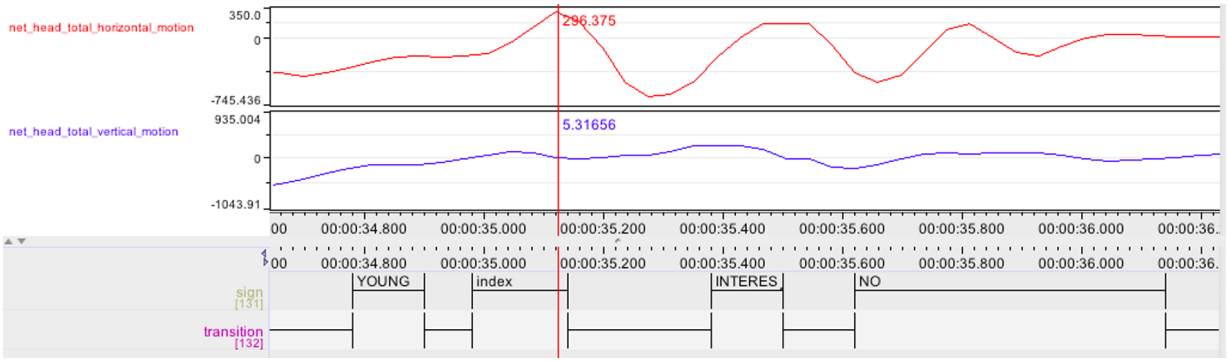


Figure 3: *A screenshot from ELAN showing line graphs for horizontal (top, red) and vertical (bottom, blue) motion of the head.*

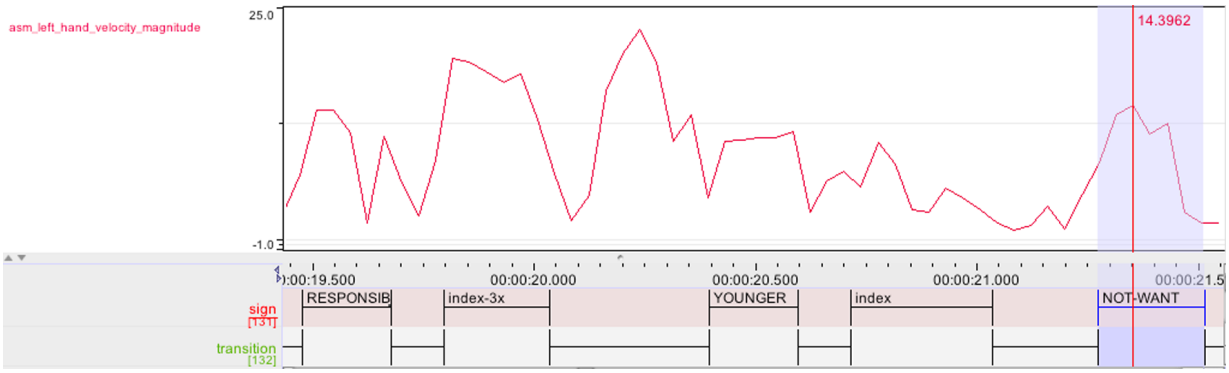


Figure 4: *A screenshot from ELAN showing a line graph for the variation in active hand speed (i.e. velocity magnitude) in continuous signing.*

[6] T. F. Cootes, D. H. Cooper, C. J. Taylor, and J. Graham, "Active Shape Models - Their training and application". *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan 1995.

[7] O. Crasborn, and H. Sloetjes, "Enhanced ELAN functionality for sign language corpora". In *Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages* at 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco, May–Jun 2008, pp. 39–43.

[8] O. Crasborn, and I. Zwitserlood, "Annotation of the video data in the 'Corpus NGT'". Department of Linguistics, and Centre for Language Studies, Radboud University Nijmegen, The Netherlands. Online publication http://hdl.handle.net/1839/00-0000-0000-000A-3F63-4, 2008.

[9] T. Johnston, "Guidelines for annotation of the video data in the Auslan corpus". Department of Linguistics, Macquarie University, Sydney, Australia. Online publication http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf, 2009.

[10] K. Duarte, and S. Gibet, "Corpus design for signing avatars". In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* at 7th Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta, May 2010, pp. 73–75.

[11] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 2001.

[12] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Training models of shape from sets of examples". In *Proceedings of the British Machine Vision Conference*, 1992.

[13] B. Hellwig, D. Van Uytvanck, M. Hulsbosch, A. Somasundaram, and M. Tacchetti, *ELAN—Linguistic Annotator, Version 4.1.0. Manual*. Online publication http://www.mpi.nl/corpus/manuals/manual-elan.pdf, 2011.

[14] T. Jantunen, "A comparison of two linguistic sign identification methods". In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* at 7th Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta, May 2010, pp. 129–132.

[15] L. Savolainen, "Interrogatives and negatives in Finnish Sign Language: an overview". In *Interrogative and negative constructions in sign languages*, U. Zeshan, Ed. Nijmegen, Ishara Press, 2006, pp. 284–302.

[16] T. Jantunen, "Fixed and free: Order of the verbal predicate and its core arguments in declarative transitive clauses in Finnish Sign Language". *SKY Journal of Linguistics*, vol. 21, pp. 83–123, 2008.