Comparing Computer Vision Analysis of Signed Language Video with Motion Capture Recordings

Matti Karppa*, **Tommi Jantunen**†, **Ville Viitaniemi***, **Jorma Laaksonen***,
Birgitta Burger‡, **Danny De Weerdt**†

*Department of Information and Computer Science,
Aalto University School of Science, Espoo, Finland,
firstname.lastname@aalto.fi

†Sign Language Centre, Department of Languages,
University of Jyväskylä, Finland,
tommi.j.jantunen@jyu.fi, danny.deweerdt@jyu.fi

‡Department of Music, Finnish Centre of Excellence in Interdisciplinary Music Research,
University of Jyväskylä, Finland
birgitta.burger@jyu.fi

Abstract

We consider a non-intrusive computer-vision method for measuring the motion of a person performing natural signing in video recordings. The quality and usefulness of the method is compared to a traditional marker-based motion capture set-up. The accuracy of descriptors extracted from video footage is assessed qualitatively in the context of sign language analysis by examining if the shape of the curves produced by the different means resemble one another in sequences where the shape could be a source of valuable linguistic information. Then, quantitative comparison is performed first by correlating the computer-vision-based descriptors with the variables gathered with the motion capture equipment. Finally, multivariate linear and non-linar regression methods are applied for predicting the motion capture variables based on combinations of computer vision descriptors. The results show that even the simple computer vision method evaluated in this paper can produce promisingly good results for assisting researchers working on sign language analysis.

Keywords: Sign language, Motion capture, Computer vision, Multivariate regression analysis

1. Introduction

When analysing sign language videos, linguists routinely segment the stream of signing into signs and inter-sign transitions (for a discussion, see (Jantunen, 2013)). The segmentation has been traditionally done by observing, from the video, the visible changes in the direction of the movement of the signer's active hand, corresponding to the moments when the speed of the hand is at its slowest. Recently, many researchers have started to enhance the segmentation process with quantitative measurement concerning the hand movement (e.g. (Duarte and Gibet, 2010; Jantunen, 2013)). For this task, the most accurate method has always been considered to be motion capturing.

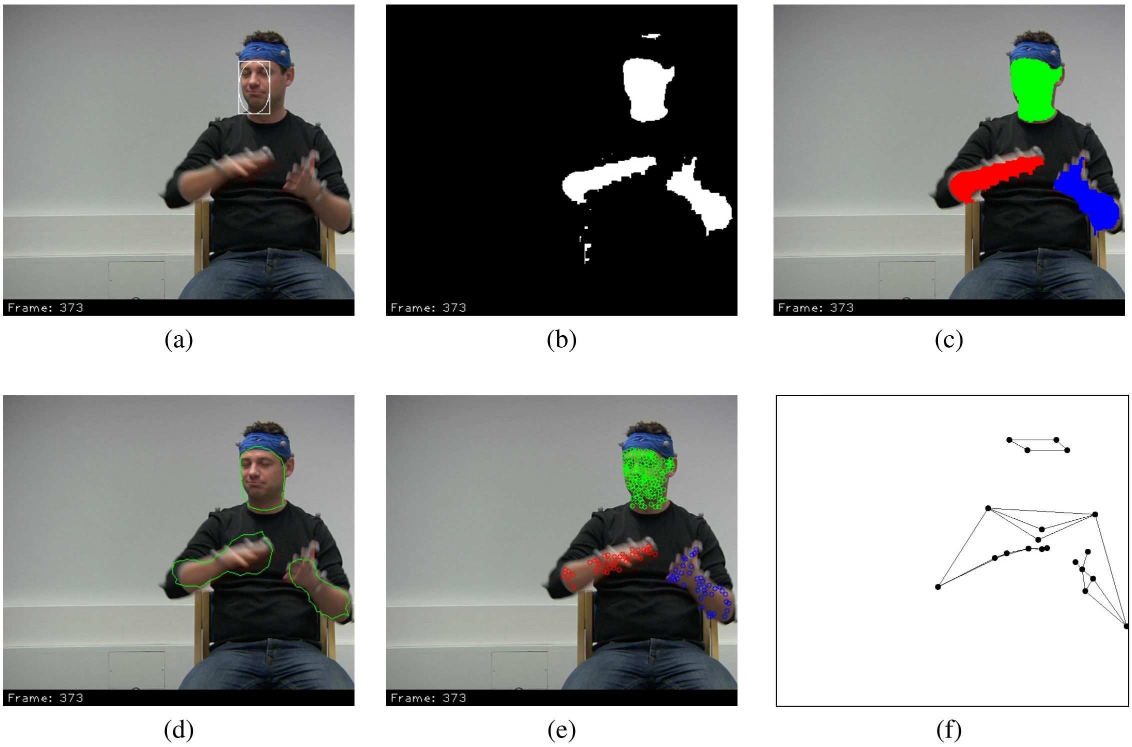
However, because motion capture cannot be used for prerecorded material and is always tied to laboratory settings, we have in our previous work introduced a computer-vision-based method that enables researchers to track and measure the motion of the hand and other articulators on the basis of the video only (Jantunen et al., 2010; Karppa et al., 2011). In this paper, we evaluate the accuracy of this method by comparing its results to the speed measurements obtained through motion capture. The comparison is based on one 52-second-long recording of continuous signing in Finnish Sign Language, collected with the motion capture equipment. During recording of that data, the movements of the signer were also recorded with a digital video camera directly facing him, and our computer-vision-based motion analysis has been applied to this video. After calculating a number of features describing the motion of the articulators in the video material, these values were qualitatively compared with their motion capture counterparts. Finally, a quantitative analysis was performed by calculating correlations between the motion capture measurements and the video-based motion values and their multivariate regression combinations. The results show an encouragingly good agreement between the motion capture and video-based data.

2. Methods for analysing sign language material

2.1. Motion capture recordings

The motion capture data used in the experiment was recorded with an eight-camera optical motion capture system (ProReflex MCU120) at a frame rate of 120 Hz by tracking the three-dimensional positions of 20 small ball-shaped markers attached to the signer's upper torso, head, and each arm and hand as illustrated in Figure 1f. However, in the present study, only the data derived from the ulnar and radial wrist markers and the index finger marker of the active hand were used in the analysis.

The Matlab Motion Capture Toolbox (Toiviainen and Burger, 2010) was used for further processing of the data. After filling the gaps that had occurred during the recording, the active hand wrist centroid segment was calculated on the basis of the ulnar and radial wrist marker data. The final processing step included the calculation of the speed



(a) (b) (c)

(d) (e) (f)

Figure 1: (a)–(e) Processing stages in the video analysis: (a) face detection, (b) detection of skin-coloured regions, (c) skin blob detection, (d) fitting of ASM active shape models, (e) KLT interest point tracking. (f) The skeleton model used with the motion capture equipment.

(i.e., the magnitude or Euclidean norm of the velocity data) of the wrist segment. In the present study, the results of the computer-vision-based method were compared to this magnitude data, as well as to the speed of the right index finger marker.

2.2. Computer-vision-based analysis

Figure 1 illustrates the main processing stages of our computer-vision-based method that was used for extracting motion descriptors from the video in the present study. The signer's face was first detected (Figure 1a) by using the Viola-Jones cascade face detector (Viola and Jones, 2001). As the next step, skin-coloured regions were located by using a detector based on multiple multivariate Gaussian distributions in the HSV colour space (Figure 1b). From the skin detector output, interconnected skin pixel regions were extracted using elementary image processing operations. Heuristic rules were used for determining whether the regions corresponded to either of the hands or the face region of the signer, or their combinations (Figure 1c).

The computer-vision-based analysis extracts two complementary sets of motion descriptors from the video. Firstly, a separate point distribution model (PDM) (Cootes et al., 1992) was constructed for describing each of the three modelled body parts (both of the hands and the head). The point distribution models were used as a basis for active shape models (ASM) (Cootes et al., 1995) that track the body part poses and shapes between consecutive frames of video (Figure 1d), giving arise to a set of 18 descriptors for each frame.

As another motion analysis method, local motions in the detected skin regions were estimated by detecting distinctive corner points (Figure 1e) and tracking them with the Kanade-Lucas-Tomasi (KLT) algorithm (Shi and Tomasi, 1994) . For each frame of the video, the movements of the tracked points were summarised with 35 descriptors.

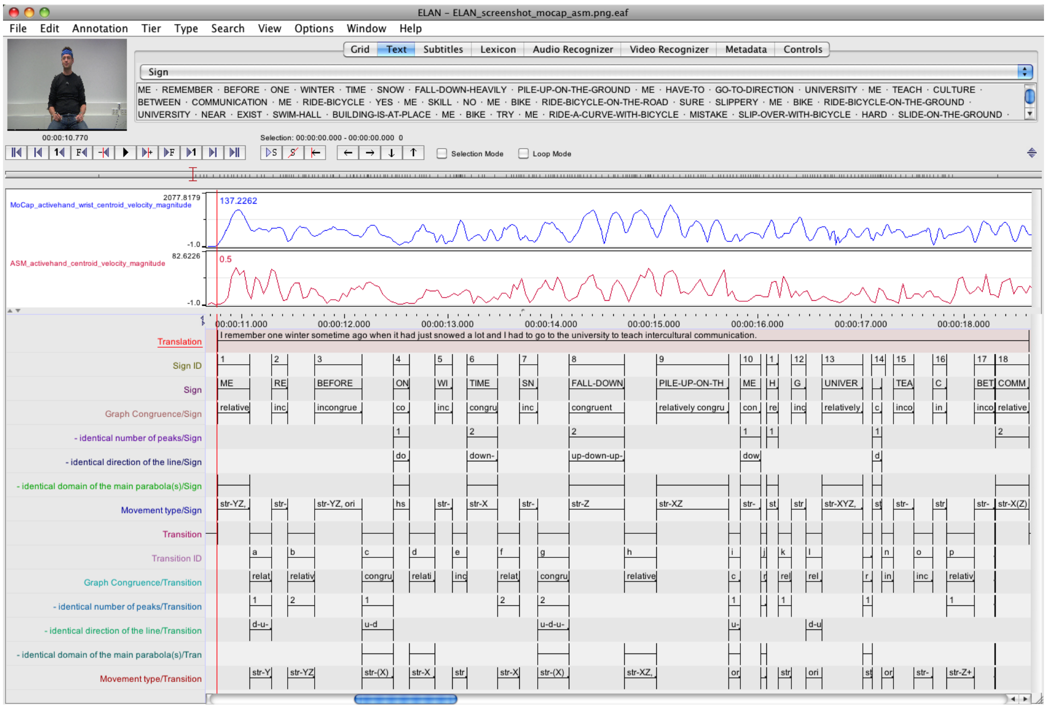
3. Comparison of video analysis results and motion capture recordings

3.1. Qualitative observations

The first step in comparing the velocity estimates produced by the two methods was qualitative inspection of the correspondence between the wrist velocity magnitude derived by motion capture and one of the statistics extracted by computer analysis: the active hand ASM velocity magnitude. For this purpose, velocity graphs describing both of them were imported into ELAN annotation software[1]. In ELAN, the graphs were time-aligned with the video of the signing and manually created annotation cells corresponding to signs and transitions.

The actual comparison was done by visually observing the degree of congruence of the line graphs during the first eight seconds of the signing, corresponding to the first full sentence of the story as seen in Figure 2. Three features were observed from the graphs for both signs (n=18) and transitions (n=17): the number of peaks, the direction of the line, and the domain of the main parabola(s) of the line. The two lines were treated *congruent* if all three features were identical, *relatively congruent* if one or two features

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



ELAN – ELAN_screenshot_mocap_asm.png.eaf

Sign

ME · REMEMBER · BEFORE · ONE · WINTER · TIME · SNOW · FALL-DOWN-HEAVILY · PILE-UP-ON-THE-GROUND · ME · HAVE-TO · GO-TO-DIRECTION · UNIVERSITY · ME · TEACH · CULTURE · BETWEEN · COMMUNICATION · ME · RIDE-BICYCLE · YES · ME · SKILL · NO · ME · BIKE · RIDE-BICYCLE-ON-THE-ROAD · SURE · SLIPPERY · ME · BIKE · RIDE-BICYCLE-ON-THE-GROUND · UNIVERSITY · NEAR · EXIST · SWIM-HALL · BUILDING-IS-AT-PLACE · ME · BIKE · TRY · ME · RIDE-A-CURVE-WITH-BICYCLE · MISTAKE · SLIP-OVER-WITH-BICYCLE · HARD · SLIDE-ON-THE-GROUND ·

Translation: I remember one winter sometime ago when it had just snowed a lot and I had to go to the university to teach intercultural communication.

File Edit Annotation Tier Type Search View Options Window Help
Grid
Text
Subtitles
Lexicon
Audio Recognizer
Video Recognizer
Metadata
Controls
00:00:10.770
Selection: 00:00:00.000 - 00:00:00.000 0
Selection Mode
Loop Mode
2077.8179
MoCap_activehand_wrist_centroid_velocity_magnitude
137.2262
-1.0
82.6226
ASM_activehand_centroid_velocity_magnitude
0.5
-1.0
00:00:11.000
00:00:12.000
00:00:13.000
00:00:14.000
00:00:15.000
00:00:16.000
00:00:17.000
00:00:18.000
Translation
Sign ID
Sign
Graph Congruence/Sign
- identical number of peaks/Sign
- identical direction of the line/Sign
- identical domain of the main parabola(s)/Sign
Movement type/Sign
Transition
Transition ID
Graph Congruence/Transition
- identical number of peaks/Transition
- identical direction of the line/Transition
- identical domain of the main parabola(s)/Tran
Movement type/Transition

Figure 2: ELAN screenshot of the first eight seconds of the video used in the qualitative analysis. The curves show the wrist marker speed W$|v|$ and the ASM centroid speed ASM $|v|$.

were identical, and *incongruent* if none of the features were identical.

The main results of the qualitative analysis are given in Table 1. The results show that, of the total of 35 analysed sequences, two thirds fell into categories *congruent* and *relatively congruent*, and one third was classified as *incongruent*. Reflecting perhaps the qualitative difference between signs and transitions (Jantunen, 2013), the graphs associated with signs included more cases of pure congruence and incongruence whereas graphs associated with transitions showed mostly relative congruence. However, for both signs and transitions, incongruent cases were in minority, and the most congruent feature in the graphs was the one describing the overall shape of the main parabola(s).

A more detailed analysis of the data revealed that movements along the depth dimension were the primary cause of incongruence in the results; all the incongruent cases—both signs and transitions—included this type of movement whereas in the most congruent cases such movements were not noticeably present (see Figure 2). This was expected (Jantunen et al., 2010; Karppa et al., 2011) as the computer-vision-based method operates in the 2D space lacking the dimension of depth, inherently present in the 3D motion capture.

	signs	transitions
congruent	5	3
relatively congruent	5	11
incongruent	8	3

Table 1: Graph congruence for signs and transitions.

3.2. Quantitative analysis

In the quantitative part of the analysis, we first calculated the numerical correlation between the wrist marker speed W$|v|$ and the active hand active shape model velocity ASM $|v|$, the same quantities that were already studied in the qualitative analysis. Then, we extended the analysis to contain all the 53 video-based descriptors, including the horizontal and vertical interest point speed descriptors KLT $\sum v_x$ and KLT $\sum v_y$, respectively. From the motion capture measurements we additionally used the active hand index finger marker velocity components Fv_x and Fv_y.
We studied exhaustively the agreement between all pairs of the above-mentioned motion capture measurements and video-based descriptors alone, and also formed multivariate regressors from the descriptors for predicting the values of the motion capture measurements. The level of agreement was measured with energy-normalised correlation and used for assessing the usefulness of the descriptors. To achieve this, all motion capture measurements and predictions were z-normalised by subtracting their average values and normalising their variance to unity. As the multivariate regression methods require training samples, the data was divided into training and test parts and the correlations were measured only in the latter half.

3.3. Correlations between individual variables

Here, each computer-vision-based descriptor was correlated individually against every motion capture variable. A selection of most relevant correlation coefficients is presented in Table 2.
The highest correlation (0.507) with respect to the wrist marker speed was shown by the descriptor measuring the active hand ASM centroid velocity, which is also the de-

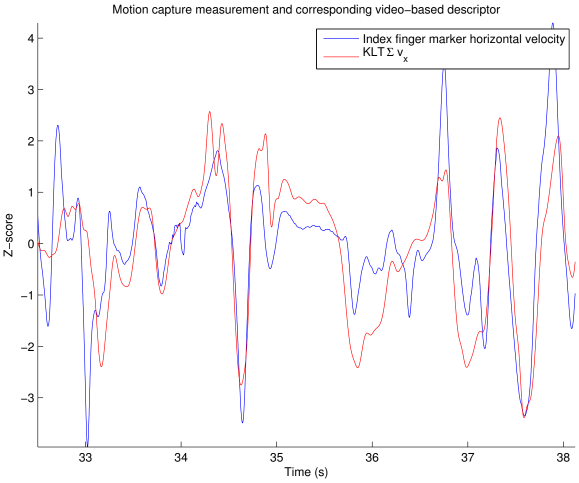


Figure 3: The index finger horizontal motion capture velocity component $\mathrm{F}v_x$ and the sum of horizontal KLT interest point velocity components KLT $\sum v_x$ plotted aligned.

	W$\lvert v\rvert$	Fv_x	Fv_y
ASM $\lvert v\rvert$	**0.507**	-0.142	0.222
KLT $\sum v_x$	-0.175	**0.704**	0.227
KLT $\sum v_y$	-0.167	0.221	**0.673**

Table 2: Correlations between motion capture features (columns) and computer-vision-based descriptors (rows).

scriptor used in the qualitative analysis above and shown in Figure 2. This descriptor was closely followed by a interest point statistic measuring the average speed of tracked KLT interest points with a correlation of 0.475.

The index finger's horizontal and vertical velocity components had a strong match with the corresponding KLT interest point velocity component sums, indicating that those descriptors may be reasonably useful as such. This was indeed expected since the fingers contain more area suited for interest point detection than the wrist area. The horizontal velocity component descriptors are shown aligned for a portion of the data in Figure 3 and as a scatter plot in Figure 4.

Out of the eight most strongly correlating descriptors, seven were based on tracked points and only one on ASMs. It is interesting that some of the strongly correlating variables measure motion of the non-active hand, which thus seems to correlate with the motion of the active hand. Another interesting observation is that the number of tracked points has a strong negative correlation with the target velocity. This is explained by the fact that the tracker more often loses track of fast moving points. The magnitudes of correlation show that the descriptors may be reasonably useful even individually.

3.4. Multivariate regression

Regression was first performed linearly using different subsets of explaining variables. The best results were obtained using both interest point tracking and ASM descriptors of the active hand motion together. For the wrist marker speed W$\lvert v\rvert$, the predicted signal had correlation of 0.700, which is clearly stronger than that of any single descriptor. Sets of point-tracking-based descriptors outperform the sets of ASM descriptors, but combining both kinds of descriptors would appear to give the best results. On the contrary, for the index finger velocity components, Fv_x and Fv_y, the results did not improve very much compared to the univariate case; correlations for the regressed test set were 0.705 and 0.702, respectively. Neither shows a significant improvement, suggesting that the descriptors may be as useful as they can be on their own. Predictions made in this manner for the right hand index finger marker horizontal velocity can be seen in Figure 5.

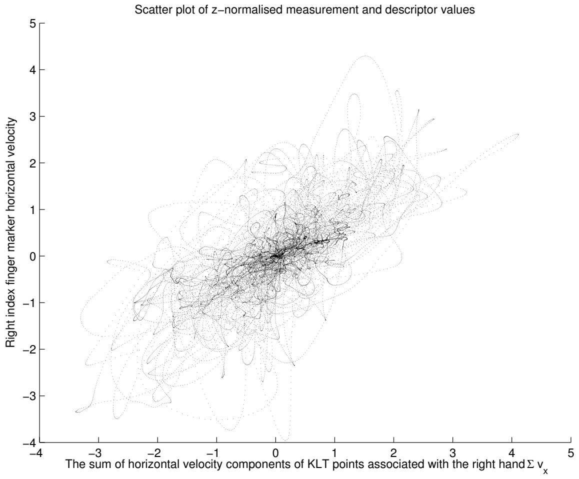


Figure 4: A scatter plot of the horizontal motion capture velocity of the active index finger Fv_x as the function of the corresponding KLT $\sum v_x$ descriptor.

Some of the descriptors turned out to be irrelevant and noise-like for the linear prediction task, and using all of the variables led to rather modest correlation of 0.439 for the

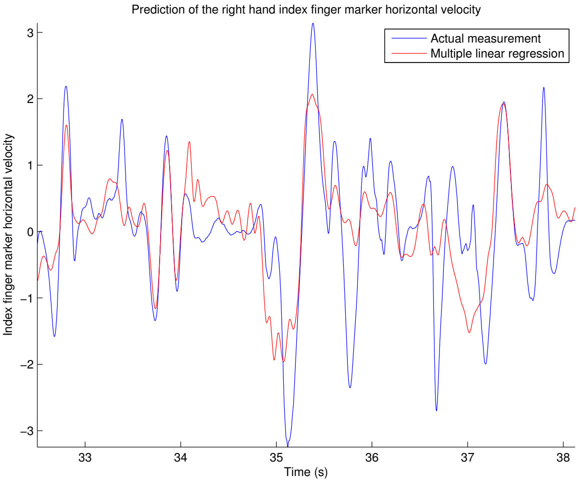


Figure 5: A part of the test data (blue) of the horizontal velocity component Fv_x of the index finger marker with the multiple linear regression plotted on it (red).

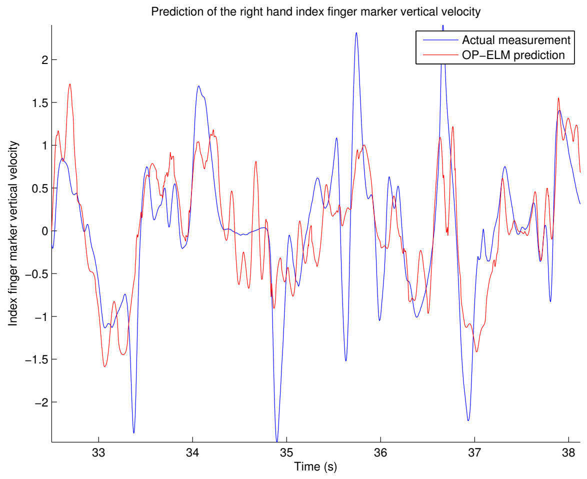


Figure 6: A part of the test data (blue) of the vertical velocity component $\mathrm{F}v_y$ of the index finger marker with the OP-ELM prediction plotted on it (red).

wrist velocity, which is worse than with the best individual regressor variables. The non-active hand descriptors do not contribute to the prediction quality when also the active hand descriptors are used, even though they do correlate somewhat strongly with the target variable individually.

Next, non-linear multivariate regression with the OP-ELM (Miche et al., 2010) method was used for predicting the motion capture variables. For the three motion capture variables, $\mathrm{W}|v|$, $\mathrm{F}v_x$ and $\mathrm{F}v_y$, the corresponding correlations when regression was performed with OP-ELM were 0.707, 0.704, and 0.706, respectively. In terms of correlation, the results were thus very similar to those with linear regression, which might indicate that linear methods are powerful enough for extracting all the information present in the computed motion descriptors. Predictions for the finger marker vertical velocity can be seen in Figure 6.

4. Conclusions

The qualitative and quantitative analysis of the data demonstrates that the presented computer-vision-based motion analysis produces promisingly accurate results. The descriptors based on interest point tracking and an active shape model contain complementary information that can be usefully combined to improve the quality of the motion analysis and the prediction of the articulator speeds.

The results show that our computer-vision-based motion tracking method is already an effective supportive tool in the annotation and analysis of sign language. The method tracks the motion of articulators at an accuracy encouragingly similar to that of traditional motion capture technology. The results of the qualitative analysis suggest that the main differences in the correspondences of measurements based on these two methods may eventually be fairly predictable, i.e. caused by dimensional differences.

Our plan is to develop the method further and test it with varied video material. This work will include, for example, a more detailed modelling of articulators through which we expect to obtain more precise information concerning, for example, hand-internal movements and facial gestures.

5. Acknowledgements

This work has been funded by the following grants of the Academy of Finland: 140245, Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL); 251170, Finnish Centre of Excellence in Computational Inference Research (COIN); 134433, Signs, Syllables, and Sentences (3BatS); 118616, Finnish Centre of Excellence in Interdisciplinary Music Research.

6. References

Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. 1992. Training models of shape from sets of examples. In *Proceedings of the British Machine Vision Conference*.

Timothy F. Cootes, David H. Cooper, Christopher J. Taylor, and Jim Graham. 1995. Active Shape Models - Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January.

Kyle Duarte and Sylvie Gibet. 2010. Corpus design for signing avatars. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, and A. Schembri, editors, *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 73–75. ELRA.

Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Päivi Rainò. 2010. Towards 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.

Tommi Jantunen. 2013. Signs and transitions: Do they differ phonetically and does it matter? *Sign Language Studies*, 13(2). forthcoming.

Matti Karppa, Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Ville Viitaniemi. 2011. Method for visualisation and analysis of hand and head movements in sign language video. In C. Kirchhof, Z. Malisz, and P. Wagner, editors, *Proceedings of the 2nd Gesture and Speech in Interaction conference (GESPIN 2011)*, Bielefeld, Germany. Available online as `http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Jantunen.pdf`.

Yoan Miche, Antti Sorjamaa, Patrick Bas, Olli Simula, Christian Jutten, and Amaury Lendasse. 2010. OP-ELM: Optimally-pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 21(1):158–162, January.

Jianbo Shi and Carlo Tomasi. 1994. Good features to track. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '94)*, pages 593–600, June.

Petri Toiviainen and Birgitta Burger, 2010. *MoCap Toolbox Manual*. University of Jyväskylä, Finland. Version 10.9.2010.

Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages I:511–518.