

Towards the Automated Visualization and Analysis of Signed Language Motion – Method and Linguistic Issues

Tommi Jantunen¹, Markus Koskela², Jorma Laaksonen³, Päivi Raino⁴

¹ Department of Languages, University of Jyväskylä, Finland

^{2,3} Department of Information and Computer Science, Aalto University
School of Science and Technology, Finland

⁴ Sign Language Unit, Finnish Association of the Deaf, Finland

tommi.j.jantunen@jyu.fi, markus.koskela@hut.fi, jorma.laaksonen@hut.fi,
paivi.raino@kl-deaf.fi

Abstract

This paper presents a semi-automatic method for the visualization and analysis of motion in signed language from a digital video. The method detects the different parts of the signer's bare skin on the video, tracks the motion of the different parts of the body and represents the frame-wise motion of the parts of the body with statistical descriptors. Results of an experiment are demonstrated and discussed. The main linguistic issue deals with the relationship between signs and transitional sequences. Evidence is provided for the claim that lexical-phonological and transitional sequences are qualitatively different in terms of their motion characteristics.

Index Terms: sign language, motion analysis, transitional movement

1. Introduction

In this paper, we present, first, a technological method that enables a sign language researcher to graphically represent and semi-automatically analyze signed language *motion* – i.e. the movements of the hands and other body parts, such as the head and the whole body – from digital video material containing natural signing. Second, we demonstrate and briefly discuss, from a linguistic point of view, the results of an experiment in which this method was used. The linguistic discussion focuses especially on the motion characteristics of (lexical-phonological) signs and intersign transitions (cf. spoken words and their transitions).

Several aspects of motion have been claimed to manifest *prosody* in signed language. For example, the more prominent production of the manual movement of signs has been associated with stress [1], and different nonmanual movements (e.g. eye blinks, head nods, and body leans) have been analysed as indicators of prosodic units in signed language (e.g. prosodic words and phrases) [2], [3]. It has also been argued that signed language motion in general is prosody [4]. In this paper, we do not directly address the question what counts as prosody or is prosodic in signed language. However, we suggest that the method we present here can contribute positively to the investigation of this question in the future.

The main motivation behind the present work is that, unlike research on spoken language, research on signed language still lacks a feasible way of *visualizing* data on the level maximally close to linguistic signal, that is, on the level of *phonetics*. By data visualization we refer to the automatized or semi-automatized process in the data handling phase in which a researcher collapses the linguistic material numerically and converts it into graphical diagrams (cf. the use of *Praat* and comparable software on spoken language

research; for some earlier attempts on signed language data visualization, see [1], [5], [6]). We consider the data visualization process to be a key methodological step in the phonetic analysis of any language: visualization allows the researcher to observe systematicity and deviations in the data (e.g. rhythmic patterns of the moving hands and body and standard deviations from these) in a way that is not possible by merely observing a person producing language (whether in person or on video with time-aligned annotations; cf. ELAN, <http://www.lat-mpi.eu/tools/elan/>).

A further motivation for the work is our conviction that the phonetic analysis of signed language should be based as much as possible on natural (e.g. discourse) data. Hitherto, the modelling and in-depth linguistic analysis of motion has been feasible only if the signed data has been produced in pre-determined laboratory settings using complex motion tracking equipment and software (e.g. [4], [5]). In principle, the method presented in this paper makes it possible to entirely avoid laboratory settings in the phonetic study of signed language. Already at this pilot stage of development our method allows one to make complex motion analysis of many types of digital video material, making it viable for future phonetic signed language research to cover more natural language use.

2. The method

The analysis of a signed language video in this work is based on computer vision analysis of a quantitative nature. The stages in the process are widely used and robust algorithms. The analysis consists of three steps: first the skin regions of the subject are detected, then the motion of the skin regions is tracked, and finally frame-wise motion is represented using statistical descriptors.

2.1. Skin area filter

We begin the analysis by applying a color-based skin color filter to limit the tracking of motion to the hands, arms and face of the subject (see Figure 1a). This increases the accuracy of the estimation of relevant motion, as, for example, any slight occasional movement of the clothes of the subject is eliminated from further analysis.

2.2. Motion points tracking

After skin color detection, we track the areas with local motion in the video stream by using an algorithm based first on detecting distinctive pixel neighborhoods and then minimizing the sum of squared intensity differences in small image windows between successive video frames [7]. This enables us to track the locations of these distinctive pixel neighborhoods,

or *motion points*, over an extended period of time (see Figure 1b). If the appearance of the pixel neighborhood 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 neighborhoods in each video frame.

2.3. Motion descriptors

By tracking the locations and identities of motion points, we obtain a frame-wise representation of the relevant motion in the signed language video. In this work, we calculate and use five statistical descriptors, denoted $D1-D5$, indicating different characteristics of overall motion in the video material.

First, we record the number of currently tracked motion points in frame f . For a motion point to be included, it must also be detected in the previous ($f-1$) and following ($f+1$) frames. We denote the number of tracked points in frame f as N_f and use this as a motion descriptor:

$$D1 = N_f \quad (1)$$

Each of the N_f motion points in the image has a corresponding *location vector* $\mathbf{d}_f(i) = (x_f(i), y_f(i))$, where $1 \leq i \leq N_f$, a *motion vector* $\mathbf{v}_f(i) = (vx_f(i), vy_f(i))$, and an *acceleration vector* $\mathbf{a}_f(i) = (ax_f(i), ay_f(i))$. The motion and acceleration vectors are computed as

$$\mathbf{v}_f(i) = \frac{1}{2}(\mathbf{d}_{f+1}(i) - \mathbf{d}_{f-1}(i)) \quad (2)$$

and

$$\mathbf{a}_f(i) = \mathbf{d}_{f+1}(i) - 2\mathbf{d}_f(i) + \mathbf{d}_{f-1}(i) \quad (3)$$

We also use the total amount of motion both horizontally and vertically as motion descriptors:

$$D2 = \sum_{i=1}^{N_f} vx_f(i) \quad (4)$$

and

$$D3 = \sum_{i=1}^{N_f} vy_f(i) \quad (5)$$

In addition, we record the vector sums of motion and acceleration in a frame, and use the lengths of these vectors as motion descriptors:

$$D4 = \left| \sum_{i=1}^{N_f} \mathbf{v}_f(i) \right| \quad (6)$$

and

$$D5 = \left| \sum_{i=1}^{N_f} \mathbf{a}_f(i) \right| \quad (7)$$

2.4. Current limitations

The proposed method for analysing signed language videos can make a rough representation of frame-wise motion of signed language. There are, however, some limitations to what

the approach can produce. In the current system, the skin color detector does not make any distinction between the hands and the face of the subject. The resulting motion description (see Figure 2) is therefore a mixture of motion caused by both manual and nonmanual articulators. To differentiate between the face and hands separate detectors should be used.

In the current setup we cannot adequately estimate the motion of articulators moving towards or away from the camera. This can be considered as a limitation, as the articulation of signed language contains such motion. However, we hypothesize that enough of the articulated motion can be captured for a successful analysis. The accuracy of the analysis can, too, later be increased, for example, by using multiple cameras with overlapping fields of view or a more sophisticated 3-D model for the articulated motion.

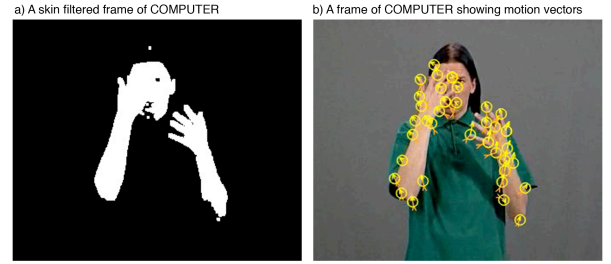


Figure 1: An illustration of the motion tracking algorithm. a) The skin area filter removes all non-skin pixels (shown here in black) and preserves the skin-colored pixels (white). b) The motion point algorithm tracks the locations of distinctive pixel neighborhoods, shown as yellow circles, in successive video frames. The motion vectors of the tracked points are shown as yellow arrows. Stationary motion points are not shown.

3. Demonstration of results

The method is demonstrated in this paper with a video clip from *Suvi*, the online dictionary of Finnish Sign Language (<http://suvi.viittomat.net>). The clip, example 3 in *Suvi*'s article 1038, contains a short story with five signs glossed concatenatedly: BOY, INDEX, COMPUTER, HOBBY, PLAY-JOYSTICK ('The boy is really interested in playing computer games.').

The motion information of the story is visualized in five different diagrams in Figure 2. The blue curve (y value) in diagrams represents a) the amount of horizontal (i.e. descriptor $D2$) and b) the amount of vertical motion ($D3$) of the articulators, c) the number of tracked motion points ($D1$), d) motion vector length ($D4$), and e) the combined acceleration value of the articulators ($D5$), respectively.

The lexical-phonological parts of each sign, represented by vertical red bars, were manually determined and time-aligned into the diagrams. The signs were identified by observing changes in the movement of the hands and arms; the moments of time when there was significantly less or relatively no movement of the hands and arms (cf. the parts of the video signal where the frames were not blurred) were considered to be the beginning and end points of lexical-phonological movements and signs, given that these parts also corresponded to the semantic boundaries indicated by the intuitions of native signers. Sign internal units such as syllables are not marked in Figure 2.

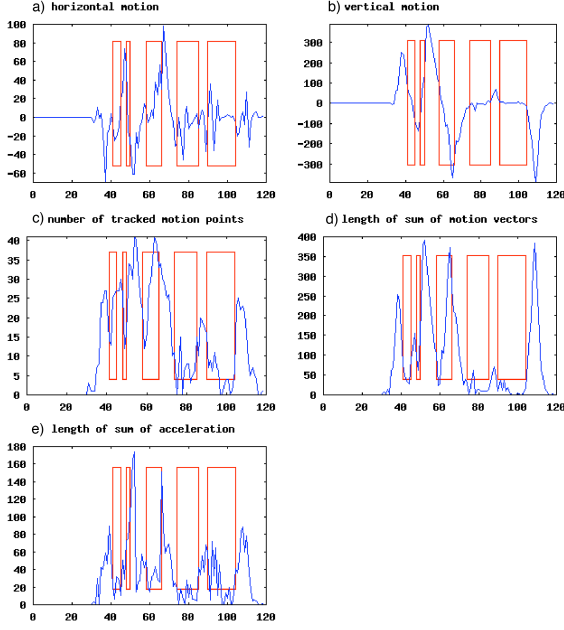


Figure 2: Visualized motion information from Suvi's example 1038/3. The blue curve (y value) in each diagram represents a) the amount of right-left (positive-negative, respectively) and b) up-down (positive-negative, respectively) motion of the tracked articulators (the hands, arms and face), c) the number of tracked motion points, d) motion vector length, and e) the combined acceleration value of the articulators, respectively. All motion descriptors are drawn per the first 128 frames of the signed sequence. Vertical red bars indicate the five lexical-phonological signs.

3.1. Horizontal and vertical motion

In general, changes in the horizontal and vertical position indicated as D2 and D3 in our analysis mapped well to the boundaries of manually identified signs. However, the boundaries were not unambiguous in all cases. For instance, concerning the first two signs, BOY and INDEX, there is no clear-cut juncture in the horizontal and vertical motion between them. Also the beginning and the end of the third sign, the two-handed compound COMPUTER (literally KNOWLEDGE+MACHINE), display an uneven alignment of the manually identified sign boundary and the tracked motion boundary. In the first example, the moment where the hand has produced the sign BOY and continues its uninterrupted leftward and downward motion to the following sign INDEX is of special linguistic interest. The uninterrupted movement from BOY to INDEX makes it reasonable to argue that the two signs form one prosodic word, where the first sign is the head and the second a clitic suffixed to it [8]–[10].

The syllable count of each sign is visualized best in Figure 2a. The signed syllable can be defined as a sequence of sign stream that corresponds to one sequential phonological movement within a sign [4], [11]. In Figure 2a, such a unit corresponds roughly to one major upward or downward directed sequence of the motion curve inside a sign bar. Following this criterion, the signs represented by the first three bars classify as monosyllables, the compound COMPUTER being a borderline case [12]; the last two signs count as disyllables.

Concerning the last two signs, the amount of horizontal

and vertical motion in their final syllable is smaller than the amount of motion in their first syllable; the amount of motion tends to reduce towards the end of multisyllabic signs. The method also captures the phrase-final lengthening phenomenon [4], [13] in the very last syllable of the last sign.

Overall, the horizontal and vertical motion in signs exhibits more variation than that in transitional sequences between the signs, i.e. in intersign transitions. This agrees with the widely held claim that only sign-internal lexical-phonological movements, but not sign-external transitional movements, are modifiable [14], [15].

3.2. Number of tracked motion points

In general, the number of tracked motion points, i.e. the motion descriptor D1 in Figure 2c, indicates the amount of movement features in the sign stream. In the example, major changes in the number of motion points occurred at the boundaries of signs and intersign transitions. Interestingly, the sequences with the maximal amount of motion points were either intersign transitions or short sequences of sign stream centering around sign-transition boundaries. We take this as evidence of the qualitative difference between the lexical-phonological movements and transitional movements mentioned above: lexical-phonological movements are more restricted and controlled while transitional movements are more free, containing (potentially) more moving body parts.

The beginning of the compound COMPUTER is accompanied by a small amount of motion points whereas the end of the compound contains a rather large number of them. This agrees with the traditional claim on American Sign Language that the beginnings of compounds are relatively light or unstressed and that the ends of compounds are heavy or stressed [16], [17]. However, we do not argue here for a direct correspondence between the number of motion points and linguistic stress: the number of motion points indicates only the relative amount of movement in the sign stream, the notion of linguistic stress being a more complex variable (e.g. [1], [18]).

With respect to the main levels of the number of tracked motion points, the five signs in the analyzed sequence can be seen as falling into two main groups. The main division occurs between the third sign COMPUTER and the fourth sign HOBBY. The change in the overall amount of movement features suggests that there is also a main constituent boundary in between these signs. The syntactic and semantic analysis supports this observation: the first three signs are best analyzed as forming a chain of two frame setting topic constituents (BOY+INDEX and COMPUTER) whereas the last two verbals (HOBBY and PLAY-JOYSTICK) form together a predicating constituent, i.e. the comment [19], [20].

3.3. Motion vector length

The curve in Figure 2d displaying motion vector length (D4) implies velocity. In the data, peak values were associated with both the boundaries of signs and intersign transitions, and with intersign transitions, not with the lexical-phonological signs *per se*. The distribution of peak values further supports the claim that lexical-phonological movements and transitional movements are qualitatively different: during lexical-phonological movements motion is slower than during transitional movements.

3.4. Acceleration

The most complex parameter calculated by our method is motion descriptor D5 in Figure 2e displaying the acceleration

of the movement of the tracked articulators. Acceleration issues should be of major importance in the study of signed language phonetics since acceleration peaks have been claimed to be the most perceivable parts of the signal (e.g. [5]) and they arguably play, consequently, a significant role also in the prosodic (e.g. rhythmic) patterning of signing (e.g. [21]).

In the data, acceleration peaks associated constantly either with the intersign transitions or with the boundaries of signs and intersigns transitions. The analysis of sign-internal movements revealed that acceleration peaks associated also with the sign-internal transitions between syllables and with their borders. Moreover, acceleration values within signs were always relatively lower than the values within intersign transitions. The distribution of acceleration peaks and levels agrees once again with the claim that phonological-lexical movements are qualitatively different from transitional movements.

In general, we consider that the acceleration data provide a very interesting basis for future research into sign language phonetics and phonology. Given that the current operationalization of perceptivity maxim as an acceleration peak is valid, and that the most perceivable moments associate with transitions, then the acceleration data make it reasonable to assume that, at least in terms of perception, transitions have a more important role in signed language than has usually been indicated in the literature. Obviously, more research on transitional movements and their role in signed language is needed.

4. General discussion and conclusion

In this paper, we have presented a technological method that allows a signed language researcher to visualize and analyze semi-automatically signed language motion from a digital video containing natural signing. Although the method is still in its early developmental phase, we have shown that it can already be used as a tool in the analysis of signed language phonetics. In the future the method will be developed further and tested with varying materials.

The main linguistic discussion has focused on the relationship between lexical-phonological movements and transitional movements. The results obtained through the method have agreed with the traditional claim that there is a qualitative difference between these two types of movements. We take this agreement to be both an indicator of the validity of the method and as providing further support for the claim we have discussed. In general, our data, especially on acceleration, suggests that the role of transitional sequences in signed language should be further investigated (cf. [1], [5]).

5. Acknowledgements

This work was supported by the Academy of Finland in the Finnish Centre of Excellence in Adaptive Informatics Research project. The work was also partly financed by the Academy of Finland under grant 134433. The authors wish to thank Eeva Yli-Luukko (Research Institute for the Languages of Finland), Eija Aho (University of Helsinki), Richard Ogden (University of York), and Onno Crasborn (Radboud University Nijmegen), as well as the anonymous reviewers of *Speech Prosody* 2010, for their valuable comments in preparing both the present work and the paper.

6. References

[1] R. B. Wilbur, "A experimental investigation of stressed sign production," *International Journal of Sign Language*, vol 1, no. 1, pp. 41–59, 1990.

[2] M. Nespor and W. Sandler, "Prosody in Israeli Sign Language," *Language and Speech*, vol. 42, no. 2–3, pp. 143–176, 1999.

[3] R. B. Wilbur, "Phonological and prosodic layering of nonmanuals in American Sign Language," In *The signs of language revisited: an anthology to honor Ursula Bellugi and Edward Klima*, K. Emmorey and H. Lane, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 215–244.

[4] D. Brentari, *A Prosodic Model of Sign Language Phonology*. Cambridge, MA: A Bradford Book, 1998.

[5] S. Wilcox, *The Phonetics of Fingerspelling*. Amsterdam: John Benjamins, 1992.

[6] P. Boyes Braem, "Rhythmic temporal patterns in the signing of deaf early and late learners of Swiss German Sign Language," *Language and Speech*, vol. 42, no. 2–3, pp. 177–208, 1999.

[7] C. Tomasi and T. Kanade, "Detection and tracking of point features," *Carnegie-Mellon University Tech. Rep. CMU-CS-91-132*, Apr. 1991.

[8] W. Sandler, "Cliticization and prosodic words in a sign language," In *Studies on the Phonological Word*, T. Hall and U. Kleinhenz, Eds. Amsterdam: John Benjamins, 1999, pp. 223–255.

[9] U. Zeshan, "Towards a notion of 'word' in sign languages," In *Word. A cross-linguistic typology*, R. M. W. Dixon and A. Y. Aikhenvald, Eds. Cambridge: Cambridge University Press, 2002, pp. 153–179.

[10] E. van der Kooij and O. Crasborn, "Syllables and the word-prosodic system in Sign Language of the Netherlands," *Lingua*, vol. 118, pp. 1307–1327, 2008.

[11] T. Jantunen and R. Takkinen, "Syllable structure in sign language phonology," In *Sign Languages: A Cambridge Language Survey*, D. Brentari, Ed. Cambridge: Cambridge University Press, to be published.

[12] T. Jantunen, "Tavu suomalaisessa viittomakielessä [The syllable in Finnish Sign Language]," *Puhe ja kieli*, vol. 27, no. 3, pp. 109–126, 2007.

[13] D. M. Perlmutter, "Sonority and syllable structure in American Sign Language," *Linguistic Inquiry*, vol. 23, pp. 407–442, 1992.

[14] T. Rissanen, *Viittomakielen perusrakenne*, University of Helsinki, Publications of the Department of General Linguistics, no. 12, 1985.

[15] D. M. Perlmutter, "On the segmental representation of transitional and bidirectional movements in ASL phonology," In *Theoretical issues in sign language research. Volume 1: linguistics*, S. D. Fischer and P. Siple, Eds. Chicago: The University of Chicago Press, 1990, pp. 67–80.

[16] E. Klima and U. Bellugi, *The Signs of Language*, Cambridge, MA: Harvard University Press, 1979, pp. 198–224.

[17] S. K. Liddell and R. E. Johnson, "American Sign Language compound formation processes, lexicalization, and phonological remnants," *Natural Language and Linguistic Theory*, vol. 4, pp. 445–513, 1986.

[18] G. Coulter, "Emphatic stress in ASL," In *Theoretical issues in sign language research. Volume 1: linguistics*, S. D. Fischer and P. Siple, Eds. Chicago: The University of Chicago Press, 1990, pp. 109–125.

[19] 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.

[20] T. Jantunen, "Tavu ja lause: tutkimuksia kahden sekventiaalisen perusyksikön olemuksesta suomalaisessa viittomakielessä [Syllable and sentence: studies on the nature of two sequential basic units in Finnish Sign Language]," Ph.D. dissertation, Department of Languages, University of Jyväskylä, Finland, 2008.

[21] G. D. Allen, R. B. Wilbur and B. B. Schick, "Aspects of rhythm in ASL," *Sign Language Studies*, vol. 72, pp. 297–320, 1991.