Technology and Disability 20 (2008) 109–119
IOS Press

Toward computational understanding of sign language

Christian Vogler[a,*] and Siome Goldenstein[b]
[a]*Gallaudet Research Institute, Gallaudet University, Washington, DC, USA*
[b]*Instituto de Computação – UNICAMP, Caixa Postal 6176, Campinas, SP 13084-971, Brazil*

Abstract. In this paper, we describe some of the current issues in computational sign language processing. Despite the seeming similarities between computational spoken language and sign language processing, signed languages have intrinsic properties that pose some very difficult problems. These include a high level of simultaneous actions, the intersection between signs and gestures, and the complexity of modeling grammatical processes. Additional problems are posed by the difficulties that computers face in extracting reliable information on the hands and the face from video images. So far, no single research group or company has managed to tackle all the hard problems and produced a real working system for analysis and recognition.
We present a summary of our research into sign language recognition and how it interacts with sign language linguistics. We propose solutions to some of the aforementioned problems, and also discuss what problems are still unsolved. In addition, we summarize the current state of the art in our sign language recognition and facial expression analysis frameworks.

1. Introduction

Nowadays is an interesting time to be deaf. On the one hand, there has been much progress in the areas of audiology and assistive listening technologies. On the other hand, computer technologies, webcams, and the Internet have finally delivered on their promise, and video conferencing and video-based communication are becoming ubiquitous among deaf sign language users all over the world, especially so in North America and Western Europe. Video phones have almost completely displaced text telephones for the deaf, and video blogs ("vlogs") frequently outnumber the number of deaf-related text blogs on sites such as DeafRead.[1]

Despite these advances the deaf and hearing worlds are still very much separate, partly because of the language barriers – many deaf people do not have an intelligible voice, and some struggle with the written language, as well. Conversely, many hearing people do not have the time or inclination to learn sign languages. An automatic sign language recognition and translation system could go a long way toward tearing down these barriers.

Facilitating communication is, however, not the only application of sign language recognition: Sign language linguists would love to automate the tedious task of transcribing the fine details of sign language data for further analysis. Sign languages could become a component of human-computer interaction in the same way that speech recognition could. Gestures are an integral part of human communication, and as sign languages are more structured, many researchers hope to apply concepts from sign language recognition to gesture recognition. Sign language learners could conceivably use the computer for training and feedback on how the movements are performed correctly. Last, but not least, sign language recognition could also facilitate the archiving and retrieval of video-based communication of the deaf, which at present is ephemeral.

Although computational sign language processing is an exciting area and holds much untapped potential, the field is far behind the field of spoken language processing, and particularly speech recognition. The reasons are many, but can broadly be classified into problems

*Address for correspondence: C. Vogler, Gallaudet Research Institute, Gallaudet University, 800 Florida Avenue, NE, Washington, DC 20002, USA. E-mail: christian.vogler@gmail.com.
[1]http://www.deafread.com.

ISSN 1055-4181/08/$17.00

with representing the complexities of the language adequately in the computer (Section 2) and problems with interpreting human actions on video, especially facial expressions (Section 3).

In this paper we discuss some of the difficulties associated with sign language modeling and representation, and propose some solutions for American Sign Language (ASL), in addition to noting some unsolved problems. Furthermore, we present a system for analyzing the human face, which exhibits great potential for augmenting sign language recognition with hitherto grammatical information that manifests on the signers' faces. We conclude with a summary of experimental results from both ASL recognition and facial expression analysis that highlight the current state of the art of our recognition framework (Section 4).

2. Sign language modeling

Sign language recognition is much harder than it seems. On a first, superficial glance, the speech recognition field seems to have done all the hard work already, and all that seemingly remains to be done is transferring the state of the art to the field of sign language recognition. On a closer look, however, things are not so simple, because the internal structure of signed languages is markedly different from the structure of spoken languages. In the past, many sign language recognition systems have made the tacit assumption that sign language utterances are conceptually a string of gestures, similar to the way spoken words can be strung together. In the following we show several examples that show why this assumption is not tenable, and then discuss sign language modeling approaches.

2.1. *Sign language utterances are not a string of gestures*

Consider the English sentence "Ann blames Mary." A common way to represent it in sign language is via glosses, where each sign is represented by its closest English equivalent in all-capital letters. Thus, this particular example would be represented as ANN BLAME MARY.

Research into sign language recognition has, with very few exceptions, followed the same tradition. However, basic glosses drop many important pieces of information that are present in the signed sentence, among them grammatical markings for agreement. For the English sentence to be grammatical, the subject "Ann" must be followed by a verb in the third person singular, which is indicated by the "s" at the end of the verb "blames;" that is, the verb agrees with the subject. Analogously, in ASL the verb "BLAME" must be modified in order to agree with both the subject and the object, but this information is not indicated in the gloss above. Based on it, it is easy to form the mistaken impression that the sign language translation is simply the sign for ANN, followed by the citation form of the sign for BLAME,[2] followed by the sign for MARY.

A more accurate rendition of this sign language utterance is given by [17, p. 64], shown below, and in Fig. 1.

$$\begin{array}{l} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \underline{\text{head tilt}_{\text{i}}} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \underline{\text{eye gaze}_{\text{j}}} \\ \text{ANN}_{\text{i}}[+\text{agr}_{\text{i}}]_{\text{AgrS}}\,\underline{[+\text{agr}_{\text{j}}]_{\text{AgrO}}\quad {}_{\text{i}}\text{BLAME}_{\text{j}}\ \text{MARY}_{\text{j}}} \end{array}$$

The $[+\text{agr}_{\text{i}}]_{\text{AgrS}}$ and $[+\text{agr}_{\text{j}}]_{\text{AgrO}}$ show that both subject and object agreement are present in this sentence. In particular, the beginning of the verb BLAME must be in the same location as the sign for ANN, and the end of this verb must be in the same location as the sign for MARY. The citation form of this verb, in contrast, begins at the chest and ends in the space in front of the signer.

This rendition also shows that the head tilts toward the subject and that the eyes gaze at the object while the verb and the object are signed. Although in this particular sentence these facial markings are optional [17], facial expressions and markings play an important role in signed languages and must consequently be addressed by a complete sign language recognition system. Even in this particular example, they could provide important disambiguation information to the recognition system.

The way in which the verb "blame" is modified is just one of many ways in which spoken and signed languages are inflected; that is, the words and signs are modified to express additional information, which include, but are not limited to, agreement, number, tense, and so on. Inflection causes problems for recognition systems, because they must be able to recognize all the different ways in which a word or sign can appear as part of a sentence. If we naïvely modeled each possible appearance separately, we would quickly run into scalability problems, because we would simply have to train the recognition system on too many examples

[2] Citation form means the way in which a sign is typically represented in a dictionary, similar to how English verbs are typically represented by infinitive forms.

(a) Anticipation: Begins spelling ANN before locating the person near the shoulder.

(b) Locating Ann near the shoulder while finger spelling the final "N".

(c) Subject agreement: Beginning of BLAME is located at same place as ANN.

(d) Agreement via nonmanual markings while BLAME is performed: Head tilt toward subject (ANN), eye gaze toward object (MARY).

(e) Object agreement: MARY is located at end of the BLAME movement. Note that the head tilt and eye gaze are still being held.

(f) Finger spelling final "Y" of MARY.

Fig. 1. Anticipation and agreement in the sentence “Ann blames Mary.” Source: NCSLGR [18].

of each of the possible appearances, in order for this approach to work [31,32].

Speech recognition, as well as sign language recognition systems, can in principle, avoid this problem altogether by breaking down the words or signs into smaller parts – their constituent phonemes. Words consist of a sequence of sounds, where assigns consist of a collection of handshapes, locations, orientations, movements, and non manual markings, such as facial expressions [27]. Following the traditions of both spoken and sign language linguistics, we call these smaller units “phonemes.” Because their numbers are limited in both spoken and signed languages, it is, in principle, possible for a recognition system to be trained on recognizing the individual phonemes, and then to recognize new words and signs simply by stringing them together accordingly [4,13,14,20,21,32].

2.2. Simultaneity in signed languages

Modeling signs in terms of their constituent phonemes does not, however, solve all scalability problems. The reason is that, unlike in spoken languages, phonemes cannot simply be assembled sequentially to form signs. Rather, several aspects of the sign’s configuration can change simultaneously, as shown in Fig. 2.

To get an idea of the magnitude of the problem that recognition systems face here, it is illuminating to consider the completely naïve approach first: what about simply tossing all possible combinations of phonemes together, without regard for linguistic constraints and interdependencies between phonemes? Then the recognition system would have to look at all possible combinations of handshapes, orientations, locations and movement types of both respective hands, respectively. This approach leads to a combinatorial explosion, as can easily be seen by multiplying all the numbers of possible respective phonemes together. If, for example, we assume that there are 40 distinct handshapes, 8 distinct hand orientations, 8 wrist orientations, and 20 body locations in ASL, a number that seems to be an underestimate, if at all, then the number of possible combinations, for both hands, would be (40 * 8 * 8 * 20)2, which is more than one billion.

Of course, it would be ludicrous to assume that ASL really exhibits that many combinations, as is evident from the many linguistic constraints, such as handshape constraints [2], and dependencies between handshape, hand orientation, and location (e.g., signs that touch a part of the signer’s body with the thumb or a specific finger have only a few orientations that are even physically possible). Coming up with a good computational

(a) Starting out with the "A"-handshape.

(b) The "A"-handshape makes contact with forehead while the left hand starts moving into position.

(c) The "A"-handshape changes to the "L"-handshape for both hands, while they simultaneously move to make contact.

Fig. 2. Simultaneous events in ASL: The handshape changes at the same time as the respective hands move downward and upward to make contact. Source: NCSLGR.

model for what interdependencies are possible, however, is very difficult. Currently, the best try seems to be to collect statistics on a large corpus of utterances in a signed language, such as with the annotated ASL corpus from the National Center for Sign Language and Gesture Resources at Boston University [18], but how such statistics could be applied in practice is still an open problem.

In the absence of reliable statistics on the interdependencies of simultaneous events, another approach is to assume that the combinations of phonemes are stochastically independent [30,32,33]. As a consequence, the various contributions of the handshape, orientations, locations, and movements are independent of one another, and thus it is no longer necessary to look at all possible combinations, as long as each group of phonemes can be recognized reliably on its own. Figure 3 shows a schematic example of how the sign for BLAME from Fig. 1(c)–(e) is represented with this approach in multiple channels of information; details can be found in [32, 33]. Although this assumption causes the recognition system to lose information, for our pilot project with a 22-sign vocabulary it still led to higher recognition rates than the ones that could be obtained by ignoring the simultaneous combinations of phonemes – that is, it is better to assume that the handshape and the hand movements are independent from each other, rather than to look at only the hand movements and to ignore the handshape (see also the experiments in Section 4.1).

2.3. Discussion and limitations of current modeling approaches

Although breaking down signs into smaller parts holds great promise, there are still many open research problems and technical difficulties to overcome. We already discussed the topic of simultaneous events in signed languages, but there are also difficulties with the concept of phonological modeling of signed languages itself. Unlike for spoken languages, there is no consensus on what exactly constitutes a phoneme in ASL, and what is the basic structure of a sign, as evidenced by the myriad of different phonological models in the linguistics literature [5,16,22–24]. Each model has its own advantages and disadvantages [14,31], and as of yet there has been no comparison of the different models in the sign language recognition literature. In addition, linguistic analysis abstracts away from the physical characteristics of the data signal and what constitutes a phoneme from a linguistic point of view may not be an optimal representation of a machine learning point of view, as linguistic commonalities do not always manifest themselves in similar looking data signals. For this reason, some approaches use statistical clustering, instead, to identify and extract the common parts of different signs directly from the data [3,4,34]. The advantage of this approach is that it may lead to higher recognition rates, as the physical characteristics of the sign language utterances are reflected better in the recognition system. On the other hand, because there are no linguistic counterparts to the clusters, it becomes much harder to model rules in the language, such as the ones that dictate when the beginning and ending location of a sign are swapped [16].

Another unsolved problem is that in natural sign language utterances by native signers the locations of the signs can be greatly influenced by the context of the surrounding signs. Figure 4 illustrates this phenomenon: the person signs the sequence JOHN LIKE CHOCOLATE. Normally, the sign for LIKE is performed with a movement straight away from the chest, as shown in part (g) of the figure. However, in this particular

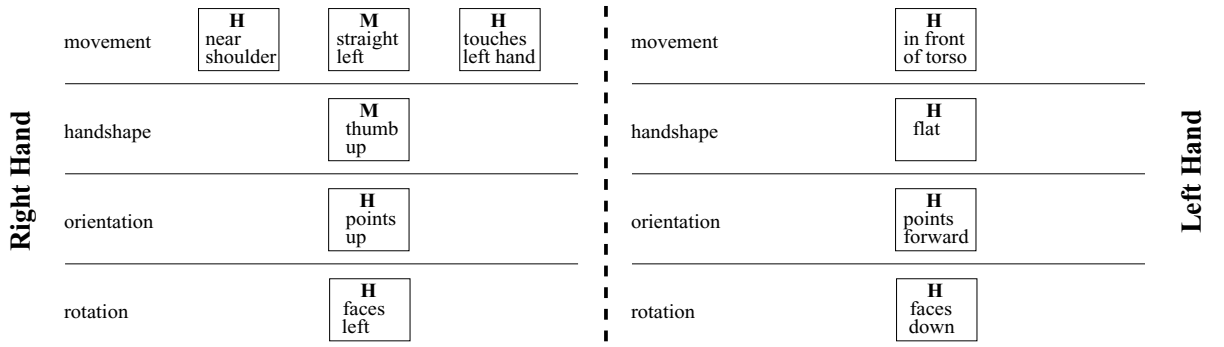

Fig. 3. Our representation of the sign for BLAME (cf. Fig. 1). Movements, handshapes, and orientations are represented independently of one an-other, each of which can consist of multiple events in sequence, such as the movement of the right hand. Boxes denoted with H represent static aspects of the sign's configuration, whereas boxes denoted with M represent a configuration change, such as the hand moving to a different location.

context with CHOCOLATE – the sign for which is performed on top of the left hand in front of the body – the movement belonging to LIKE disappears and is absorbed into a movement straight from the chest to the left hand.

If a recognition system expects to identify the sign for LIKE by its typical movement, it will be unable to recognize it in this particular context, so the system must be aware of how context affects and changes the signs. Unfortunately, at this moment, we do not yet know enough about how this process works, so we do not have enough information to adapt a recognition system accordingly.

2.4. Bridges between gestures and signs

Computational models of sign languages must also be prepared to deal with a much more fundamental problem than a simple choice of phonological model. Although signed languages are now viewed as full languages in their own right, there has been increasing awareness that gestures and signs are blended in the same modality [15,25].

For example, there exist classifier signs [27], which are signs with a distinct handshape, such as one representing a person or a vehicle, which then trace out a path describing the movements or actions that the referent takes. The handshape is a linguistic element of sign languages, subject to the grammatical constraints thereof, whereas the path is free-form, and thus gestural in nature. Phonological modeling of such free-form paths is unlikely to succeed, because unlike with regular signs, there are very few constraints on what form the movements can take, so it may not be possible to build them up from a small number of basic units.

Hence, future research will have to come up with alternatives for recognizing and interpreting classifier signs. On the other hand, classifier recognition also constitutes the area where cross-pollination between sign language recognition and gesture recognition is the most likely to yield valuable new results.

3. Facial expressions

No sign language recognition system would be complete without a way to recognize and interpret facial expressions. In Fig. 1 in Section 2.1 we already noted an example where the face indicates grammatical agreement. In addition, facial expressions pro-vide, among others, the grammatical markings for negation, questions, and topics [17]; in other words, they constitute a large part of the grammar of signed languages. Some facial movements are easy to detect, such as head tilt and eye gaze, whereas others are more subtle, such as squinting, eyebrow raising, and cheek puffing.

The latter are particularly difficult for a computer to detect and track accurately on video, especially in the presence of occlusions – situations where the hands cover the face during a sign language utterance. In light of these difficulties, it is not surprising that, to date, there is very little published work on specifically recognizing facial expressions in sign languages. In the following we discuss our approach toward video based face tracking.

3.1. Face tracking

In order to interpret what happens in sign language, it is not enough to know about the static properties of

(a) JOHN (b) LIKE (c) CHOCOLATE

(d) JOHN (e) LIKE (f) CHOCOLATE

(g) Citation form of LIKE

Fig. 4. Context affects the locations and movements of signs in the sentence JOHN LIKE CHOCOLATE. The movement of LIKE (b, e), which normally is a movement straight away from the chest (g), is absorbed into the transition from the chest to the locus of CHOCOLATE (c, f). Source: NCSLGR.

the face. In a computational sign language application, the computer must be able to observe the dynamics of facial expressions, as mentioned above. Visual input is preferable, since cameras are cheap, and webcams facing the user are now a common accessory in laptops.

Extracting information from images lies in the realm of computer vision. As images change in a video sequence, tracking methods aim to recover the variation of abstract underlying values, also known as parameters, that describe in a succinct set of number show the object of interest is moving. Tracking is hard – the computer does not have a global understanding of the scene as we humans do, and usually only knows localized color information.

The most basic tracking techniques follow the movement of two dimensional positions over local patches the image, such as the corner of an eye, or a spot on the hand. However, these are not enough to understand the complex global movements of the head and the face. There are two general approaches to recovering more global information from localized image patches: 3D deformable models and 2D active shape models (ASMs) and their cousins, active appearance models (AAMs). Both use localized image information, but constrain the individual movements to fit a globally coherent behavior. They differ in how they obtain and enforce the constraints.

With 3D deformable models, we use a human's expertise to design a 3D model of a face, with a set of parameters that control the muscles' actions, and in turn facial expressions, at any moment. All the special effects in contemporary movies provide testimony that face models can be as realistic and convincing as necessary. We need to adapt the 3D model every time that we try to track a hitherto unknown person, but this needs to be done only once per person, not per video sequence.

The tracking procedure attempts to estimate the value of the muscles actuation at every moment. Although the mathematical procedure is quite elaborate, the intuition behind it is very simple. The idea is to model the connections between the 3D model and the corresponding image points via springs. As the image changes, the springs stretch and compress and reach an equilibrium point that holds the best possible muscle configuration. Because we have a large number of these springs, measurement errors from individual image points are averaged out. Although in Fig. 5 we see only four schematic associations between the model and the image, our tracking system usually follows between 60 and 120 of these correspondences, which are selected automatically from high-contrast regions in the images.

3D deformable models can track any movement and deformation of the face, as long as the design was

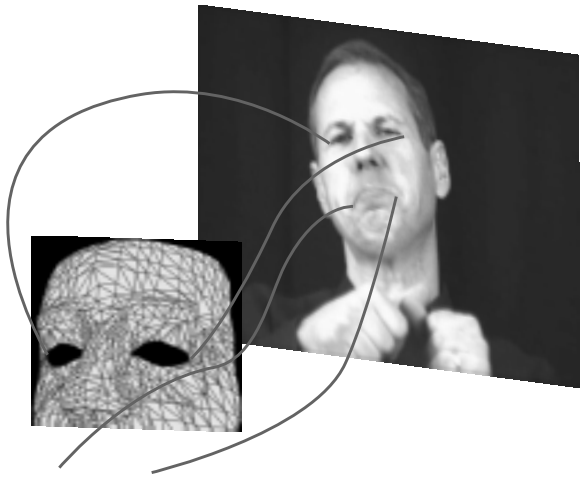

Fig. 5. The correspondences between 3D deformable model and image. The intuition of the mathematical procedure is that these connections are springs that pull the controlled parameters after the image movements.

general enough. Usually, we need between 12 and 24 parameters for a reasonable description of the face. Details on how to construct 3D deformable models and how to track them can be found in [9–12]. Their main disadvantage is that they have difficulties recovering when they lose track – for example, due to occlusions of the face by the hands –, but they fare well on hitherto unseen facial expressions.

ASMs and AAMs [7,8], in contrast, use a very different approach to constrain the motion of the individual image points. Instead of having a human to design 3D deformations, we collect hundreds, or even thousands, of representative images – each one in a different head and face configuration. These images are annotated by humans with the2D points that form the shape for each configuration (as shown in Fig. 6). With this information, and in the case of AAMs also with the texture on the face, we build a statistical model of the various shapes and configurations of the face. This distribution is then used to constrain and lock the motion of the points together.

Unlike 3D deformable models, they can recover from situations where they lost track, because the computer has "learned" what the face looks like in images. In addition, they also tend to capture the finer details of facial expressions, such as subtle eyebrow movements, better than 3D deformable models, and have been used successfully in assistive technology [6]. On the other hand, they fail when they encounter facial expressions that were not in the set of representative images, for which the statistical model was built.

In summary, the two approaches are complementary. In recent work they have been successfully combined, and hold great promise for taking the recognition of facial expressions in signed languages to the next level [29]; see also the experiments in Section 4.2.

3.2. Discussion and limitations of facial expression tracking

Despite the progress that the field has made in the past ten years, face tracking, like most other computer vision applications, is still a difficult problem with many failure cases. Nevertheless, we are now getting to a point where it is becoming usable for analyzing facial expressions, not only for sign language, but also for other applications, such as emotion recognition.

Yet, a potentially thorny problem that we will have to contend with in the near future is how to distinguish between facial expression that have a semantic meaning in a signed language, and the ones that are induced by emotions. For instance, if a person is upset, he or she will typically sign more forcefully, and the facial expressions also become more forceful, to the point that they look so different that the can confuse a recognition system, even though their basic meaning has not changed. How to separate out the effect that emotions have is still an unsolved problem.

4. Experiments

In the following we discuss the experiments that we ran on both our sign language recognition system and our face tracking system. These experiments reflect but one facet of the current state of the art. For a summary of other sign language recognition systems, this survey paper provides a good starting point [19].

4.1. Sign language recognition experiments

To illustrate the points that we made on phonological modeling in Sections 2.1 and 2.2, we now show the experimental results from a pilot study on a prototype of a sign language recognition system with a 22-sign vocabulary.

All signs were represented in terms of their constituent phonemes, with independent handshapes, locations, and movements, as described in [32]. For each phoneme, the recognition system stores a corresponding hidden Markov model (a type of statistical model well suited to recognition of speech and sign

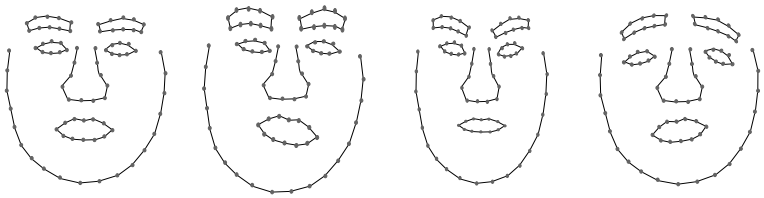

Fig. 6. Example of different shape configurations used to train an active shape model. The positions of the points have to be marked manually by a human for each corresponding example image.

Fig. 7. Three tracking examples from native signers with the tracked 3D model overlaid. The self-occlusions of the face by the hands have no effect on the tracking.

language) [20], the concatenation of which constructs signs and sign language sentences. Because accurate retrieval of handshapes from video is still a difficult research problem [1,26], we used a magnetic motion capture system in conjunction with a data glove to collect 499 ASL sentences overall. These ranged in length from 2–12 signs, which were constructed from the 22-sign vocabulary. We used 400 of these sentences to train the hidden Markov models, and we used the remaining 99 ones to evaluate recognition accuracy. Our evaluation criteria are sentence accuracy, which measures in how many sentences all signs were recognized flawlessly, as well as word accuracy, which measures how many signs were recognized correctly overall over all sentences.

The experimental results in Table 1 show that a respectable recognition accuracy can be achieved when the signs are broken down into smaller parts, which is also confirmed by [4]. In addition, they show that independently adding the handshape of the right hand, as well as adding the location and movement information from the left hand, significantly increases the recognition accuracy over using just one channel of information, even though the independence assumption is likely not valid in practice (cf. Section 2.2). The results also show that using three channels of information does not improve the recognition accuracy over using only two channels. One possible explanation is that the data set was not varied enough to allow the three-channel experiment to exhibit any advantages. Another possible explanation is that we have pushed the independence assumption to its limits, and that in order to improve recognition accuracy further, we have to collect statistical information on the interdependencies between phonemes.

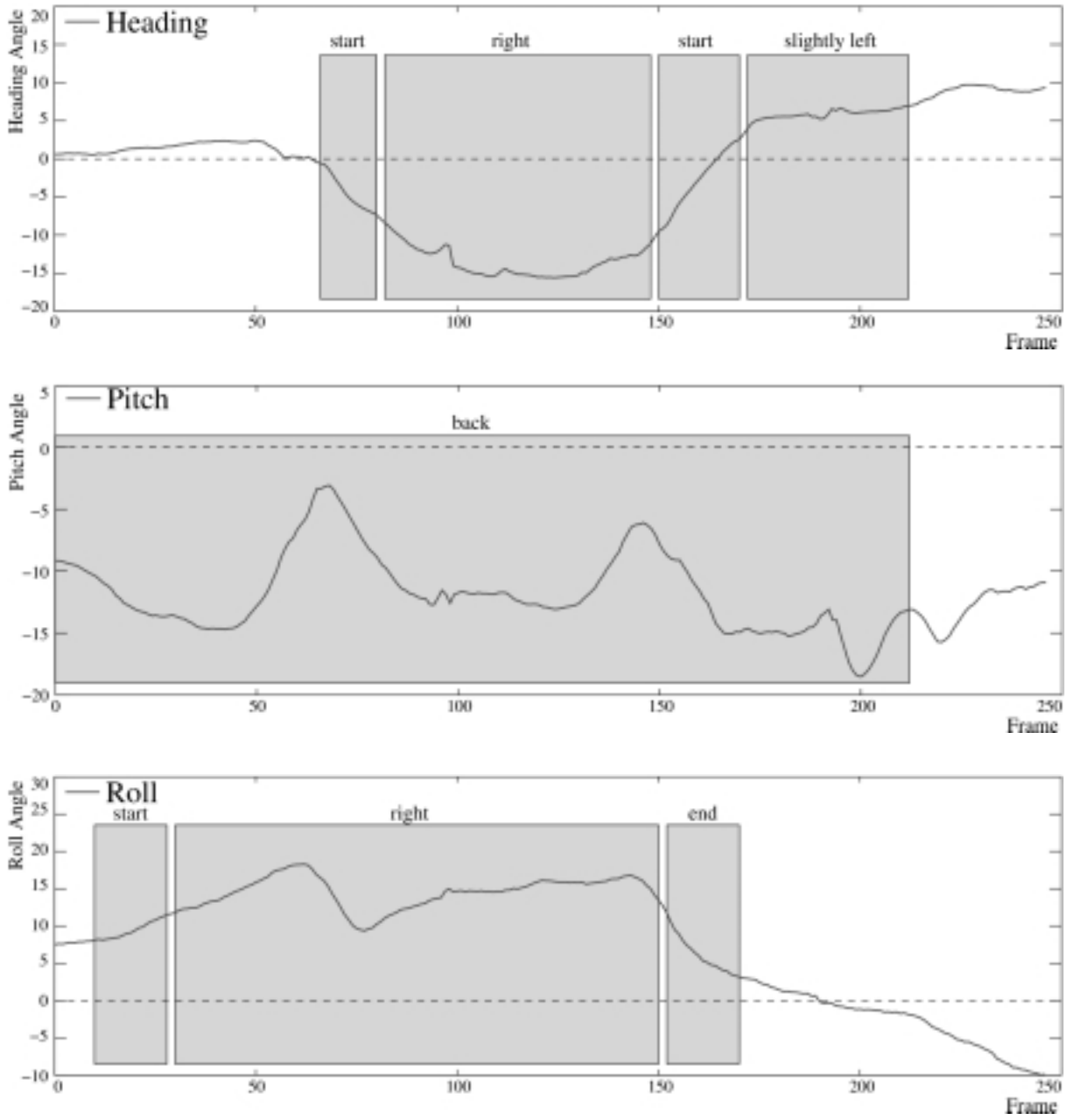

Fig. 8. Plot of the head movement angles compared to the expert human an-notations by the NCSLGR (labeled boxes) for the first sentence of the "Close Call" sequence in row 2 of Fig. 7. "Start" and "end" correspond to anticipatory head movements. In the top figure, "right" and "slightly left" indicate the respective head turns; in the middle figure, "back" indicates a head tilt backward; and in the bottom figure, "right" indicates a head tilt to the right.

4.2. Face tracking experiments

As we said in Section 3, acquiring motion of the hands and the face from video is a hard problem, and existing approaches are still far from being robust. Nevertheless, recent progress is encouraging; for example, until recently, occlusions caused by the hands covering parts of the face threw off the entire system, but as the figures below show, this is not the case anymore [28].

In Fig. 7, we illustrate a few representative results of face tracking on American Sign Language sequences by three different native ASL signers [18], one female

Table 1
Comparison of recognizing one versus multiple independent channels of information. There were 99 sentences with 312 signs overall, drawn from a 22-sign vocabulary

Type of experiment	Sentence accuracy	Word accuracy
1-channel baseline: location + movement channel right hand	80.81%	93.27%
2 channels: location + movement channel both hands	84.85%	94.55%
2 channels: location + movement channel right hand, handshape right hand	88.89%	96.15%
3 channels: movement channels both hands, handshape right hand	87.88%	95.51%

and two male. These sequences are typical of what a recognition system has to expect, with real applications possibly having to manage even lower-quality images, especially from webcams. Aside from image quality, hair moving in front of the forehead, fast hand motions, and unfavorable lighting conditions, with some clearly saturated spots in the images, are also potential problem areas.

The results of the tracking process are quantitative measurements of how the head and the various parts of the face move. These measurements define the variation of the parameters over time, and help us understand the head motion and change in facial expressions. This process is illustrated in Fig. 8, where we compare the parameter variations with expert human annotations provided by the NCSLGR for the respective video sequences.

5. Conclusions and outlook

There remains much to be done before we have a sign language recognition system that can capture the full range of constructs that occur in natural conversations among native deaf signers. The high degree of simultaneity in signed languages, as well as the lack of a comprehensive computational model for signed languages, are serious stumbling blocks. Breaking signs down into smaller units, either computationally or linguistically is a promising avenue of research, as is devising ways to decouple the simultaneous aspects from one another, with a view toward ultimately developing a model for how they are interrelated.

The combination of 3D deformable models and 2D-learning active shape models gives us a new powerful tool for tracking and analyzing the facial expressions in signed languages. Thus, in the near future, it will become possible for recognition systems to tag utterances with additional grammatical and semantic information, which in turn will lead to improved recognition rates. Nevertheless, the fact remains that tracking humans on video is hard, and tracking the human hands is even harder. Future work will have to concentrate on both improving the tracking of the face and hands, andon recognizing facial expressions reliably, especially in the presence of emotions that change and distort how the expressions manifest.

Acknowledgments

The research in this paper was supported by NSF CNS-0427267, research scientist funds by the Gallaudet Research Institute, CNPq PQ-301278/2004-0, and FAPESP 07/50040-8. Carol Neidle provided helpful advice and discussion on the NC-SLGR annotations vis-a-vis the tracking results. Norma Bowers Tourangeau, Lana Cook, Ben Bahan, and Mike Schlang were the subjects in the video sequences that we discussed in this paper.

References

[1] V. Athitsos and S. Sclaroff, *Estimating 3d hand pose from a cluttered image*, In IEEE International Conference on Computer Vision (ICCV), 2003, 432–439.

[2] R. Battison, *Lexical borrowing in American Sign Language*, Linstok Press, Silver Spring, MD, 1978. Reprinted as Analyzing Signs, in: C. Lucas and C. Valli, *Linguistics of American Sign Language*, 1995, 19–58.

[3] B. Bauer and K.-F. Kraiss, Towards an automatic sign language recognition system using subunits, in: *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and T. Sowa, eds, International Gesture Workshop, volume 2298 of Lecture Notes in Artificial Intelligence, Springer, 2001, pp. 64–75.

[4] B. Bauer and K.-F. Kraiss, Video-based sign recognition using self-organizing subunits, in: International Conference on Pattern Recognition, 2002.

[5] D. Brentari, *A Prosodic Model of Sign Language Phonology, Language, Speech, and Communication*, MIT Press, Cambridge, MA, 1998.

[6] U. Canzler and K.-F. Kraiss, Person-adaptive facial feature analysis for an advanced wheelchair user-interface, in: *Conference on Mechatronics & Robotics*, (vol. 3), P. Drews, ed., Sascha Eysoldt Verlag, 2004, pp. 871–876.

[7] T.F. Cootes, G.J. Edwards and C.J. Taylor, Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*) **23**(6) (2001), 681–685.

[8] T.F. Cootes and C.J. Taylor, Active shape models – their training and application, *Computer Vision and Image Understanding* (*CVIU*) **61**(1) (1995), 38–59.

[9] Douglas DeCarlo and Dimitris Metaxas. Adjusting shape parameters using model-based optical flow residuals, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*) **24**(6) (June 2002), 814–823.

[10] S. Goldenstein, C. Vogler and D. Metaxas, Statistical Cue Integration in DAG Deformable Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*) **25**(7) (2003), 801–813.

[11] S. Goldenstein, C. Vogler and D. Metaxas, 3D facial tracking from corrupted movie sequences, *In IEEE Conference in Computer Vision and Pattern Recognition* (*CVPR*) (2004).

[12] Siome Goldenstein, Christian Vogler, and Luiz Velho. Adaptive deformable models for graphics and vision, *Computer Graphics Forum* (*CGF*) **2** (December 2005), 729–741.

[13] C.-H. Lee, F.K. Soong and K.K. Paliwal, eds, *Automatic Speech and Speaker Recognition, Advanced Topics*, Kluwer Academic Publishers, Boston, MA, 1996.

[14] R.-H. Liang and M. Ouhyoung, A real-time continuous gesture recognition system for sign language, in: Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, 558–565.

[15] S. Liddell, Indicating verbs and pronouns: Pointing away from agreement, in: *The Signs of Language Revisited*, K. Emmorey and H. Lane, eds, Lawrence Erlbaum, 2000, pp. 303–320.

[16] S.K. Liddell and R.E. Johnson, American Sign Language: The phonological base, *Sign Language Studies* **64** (1989), 195–277.

[17] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan and R.G. Lee, *The Syntax of American Sign Language, Language, Speech, and Communication*, MIT Press, Cambridge, Massachusetts, 2000.

[18] C. Neidle and S. Sclaroff, Data collected at the National Center for Sign Language and Gesture Resources, Boston University, under the supervision of C. Neidle and S. Sclaroff. Available online at http://www.bu.edu/asllrp/ncslgr.html, 2002.

[19] A. Ong and S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*) **27**(6) (2005), 873–891.

[20] L.R. Rabiner, A tutorial on Hidden Markov Models and selected applications inspeech recognition, *Proceedings of the IEEE* **77**(2) (1989), 257–286.

[21] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., Englewoods Cliffs, NJ, 1993.

[22] W. Sandler, *Phonological Representation of the Sign: Linearity and Nonlinearity in American Sign Language*, Number 32 in Publications in Language Sciences. Foris Publications, Dordrecht, 1989.

[23] W. Sandler, Representing Handshapes, in: *International Review of Sign Linguistics*, W.H. Edmondson and R. Wilbur, eds, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1996, (Vol. 1)(5), pp. 115–158.

[24] W.C. Stokoe, Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. Studies in Linguistics: Occasional Papers 8. Linstok Press, Silver Spring, MD, 1960. Revised 1978.

[25] S. Taub and D. Galvan, Patterns of conceptual encoding in ASL motion descriptions, *Sign Language Studies* **2**(1) (2001), 175–200.

[26] G. Tsechpenakis, D. Metaxas and C. Neidle, Learning-based coupling of discreteand continuous trackers, *Computer Vision and Image Understanding* (*CVIU*) **104**(2–3) (2006), 140–156.

[27] C. Valli, C. Lucas and K. Mulrooney, *Linguistics of American Sign Language: An Introduction*, Gallaudet University Press, Washington DC, 2007.

[28] C. Vogler, S. Goldenstein, J. Stolfi, V. Pavlovic and D. Metaxas, Outlier rejection in high-dimensional deformable models, *Image and Vision Computing* **25**(3) (2007), 274–284.

[29] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein and D. Metaxas, The best of both worlds: Combining 3D deformable models with Active Shape Models, *In IEEE International Conference on Computer Vision* (*ICCV*) (2007).

[30] C. Vogler and D. Metaxas, Parallel hidden Markov models for American Sign Language recognition, *In IEEE International Conference on Computer Vision* (*ICCV*) Kerkyra, Greece, 1999, 116–122.

[31] C. Vogler and D. Metaxas, Toward scalability in ASL recognition: Breaking down signs into phonemes, in: *Gesture-Based Communication in Human-Computer Interaction*, A. Braffort, R. Gherbi, S. Gibet, J. Richardson and D. Teil, eds, volume 1739 of Lecture Notes in Artificial Intelligence, Springer, 1999, pp. 211–224.

[32] C. Vogler and D. Metaxas, A framework for recognizing the simultaneous aspects of American Sign Language, *Computer Vision and Image Understanding* (*CVIU*) **81**(81) (2001), 358–384.

[33] C. Vogler and D. Metaxas. Handshapes and movements: Multiple-channel ASL recognition, in: *Proceedings of the Gesture Workshop*, G. Volpe et al., ed., volume 2915 of Lecture Notes in Artificial Intelligence, Springer, 2004, pp. 247–258.

[34] C. Wang, W. Gao and J. Ma, A real-time large vocabulary recognition system for Chinese Sign Language, in: *Lecture Notes in Artificial Intelligence*, I. Wachsmuth and T. Sowa, eds, volume 2298, Springer, 2002, pp. 86–95.