Why don't you see what I mean?

Prospects and limitations of current automatic sign recognition research

Gineke ten Holt*, Petra Hendriks†, Tjeerd Andringa‡

*Artificial Intelligence, University of Groningen. Currently: Information and Communication Theory Group, Delft University of Technology

†Center for Language and Cognition Groningen, University of Groningen

‡Artificial Intelligence, University of Groningen

Abstract

This article presents an overview of current automatic sign recognition research. Based on a review of recent studies as well as on own research, several problem areas are identified that hamper the successful recognition of signed utterances by a computer. Some of these problems are shared with automatic speech recognition, while others seem to be particular to automatic sign recognition. These latter problems include context-dependency, determining the basic units of modeling, distinguishing signs from gestures, movement epenthesis and repetition within signs. As a possible solution to these problems, it is suggested that bottom-up processing should be supplemented with top-down processing.

1. Introduction

Imagine entering a post office. You want to send an important package through the express mail, so it is vital that the post office clerk understands exactly what you want. But you are deaf and a native signer, and the clerk does not speak or understand sign language – which is the case for most people around the world. This makes communicating your precise wishes troublesome. Although many deaf people are excellent at speaking and lip reading, this is still second-best to communicating in your native language. Write your words down? This has the same drawback, since writing is an encoding of a different (viz. spoken) language. The best solution would be to have your own, portable sign language interpreter. Just put your palmtop on the desk, aim the built-in digicam at yourself and the microphone at the desk clerk, and have your signs translated into words and the clerk's words translated into signs. Does this sound futuristic? It is somewhat, but it is not as unrealistic as you may think.

In the United Kingdom, a trial was held with TESSA, a signing virtual post office clerk, part of the ViSiCAST project (Cox, Lincoln, Tryggvason, Nakisa, Wells, Tutt, and Abbott, 2002[1]). Such virtual humans are called *avatars*. The human clerk could speak a phrase into the computer system, and a deaf person could then view TESSA signing that phrase on a screen. The words spoken by the clerk were automatically recognized, translated into British Sign Language and then signed by the avatar. At the time of the trial, the post office clerk could only choose from about 500 fixed sentences, the translations of which were fixed sign sequences. TESSA's sign language was created with the aid of video material of real signers. However, some preliminary steps were taken in the ViSiCAST project towards creating sign language synthetically. This

process, called *sign synthesis*, involves generating avatar signs based on descriptions of signs rather than on image data, and combining signs into sentences as needed instead of selecting prerecorded sentences. TESSA was generally received well by deaf people and post office staff. Currently, several similar projects are developing signing avatars (for example Paula[2], Vcom3d[3], the Dutch Gebarennet[4], and Chen, Gao, Fang, Yang, and Wang, 2003[5]). These projects all use synthesized sign language.

However, sign synthesis is only one half of the process of automatic sign language translation. The other half is automatic sign recognition (ASLR).[6] Sign recognition seems to lag behind sign synthesis when it comes to results, since it cannot yet produce the practical applications that are developed in sign synthesis. But if you want to make a portable sign language interpreter, you need sign recognition as well as sign synthesis. And this is not the only motivation for conducting sign recognition research. The results of sign recognition could also be used in sign language learning aids, human-computer interaction and virtual reality applications.

So there are many reasons to conduct research on automatic sign recognition. Why then are the results less impressive than in sign synthesis? Partly, this is because, in general, recognition is more difficult than synthesis. Earlier work in the related field of speech synthesis and automatic speech recognition (ASR) has shown that synthesizing understandable speech is easier than recognizing it (Huang, Acero, and Hon, 2001, pp. 4-6). The reason is that recognition has to deal with natural language with all its variation, incompleteness and background noise. For instance, a speech recognizer must be capable of understanding every possible variation in the pronunciation of the word *tomato*, from an Englishman's version to a Texan's. A synthesizer on the other hand only has to

produce one variation. There are five features of a sign that carry meaning: handshape, palm orientation, location, motion, and the non-manual component. In each of these features, variation is possible, so to handle natural sign language a recognizer must learn to understand every variant.

This article presents an overview of current ASLR research. Many of the first attempts at ASLR borrowed methods from automatic speech recognition, which has been studied since the 1950s. Sign recognition research only started around 1995 (Waldron and Kim, 1995). We will therefore first discuss the general problems of automatic recognition. These are problems that both ASR and ASLR have to deal with. Then we will evaluate current ASLR methods. In section 4, we shall discuss several problems that are specific for processing sign language. These are only recently becoming apparent and often have no counterpart in spoken language recognition. Finally, we will speculate about the future of ASLR and how research could possibly overcome the limitations that currently exist in this field.

2. General problems of automatic speech and sign recognition

As already mentioned in the previous section, speech and sign recognition are not easy to model computationally. There are certain basic difficulties which must be dealt with by natural language recognizers, whether they work with speech or with sign language. These difficulties are discussed below.

One problem that was already mentioned is **variation**: different people can pronounce or sign the same word quite differently. And even a single person may pronounce or sign a word differently in different circumstances (for instance when he or

she is agitated or happy).[7] These phenomena are called inter-speaker/signer and intra-speaker/signer variability, respectively.

Co-articulation resembles variation in that it is also a problem of sounds and signs taking on a different form. With co-articulation, adjacent sounds and signs change because they overlap each other, to the point where they are actually performed simultaneously. A related problem is the fact that **signs in context influence one another**: for example, start- and end location of a sign can shift under the influence of the end- and start locations of the previous or following sign, respectively. Another example is anticipation, where the non-dominant hand moves into position for a two-handed sign. It is difficult for a recognizer to identify such varying signals as variants of the same sound/sign, yet still discriminate well enough to distinguish between sounds/signs that differ only slightly.

Another problem pertains to **finding word boundaries in continuous speech/signing**. In a continuous stream, it is not easy to detect word boundaries, because words are not separated by silence. Anyone who has ever heard someone speaking in a completely unfamiliar language will have experienced this firsthand. So a recognizer cannot know for sure where the word boundaries in its data are. Recognizers that need to know where a word starts and ends in order to identify it (note that not all of them do, see section 3) will have difficulty dealing with continuous signed or spoken language. This problem raises the question of whether there is a one-to-one mapping between concepts and their physical form, since concepts (as represented by words) are not distinct entities in the physical data stream.

For automatic speech recognition, **background noise** has always been a problem. This is because sound is transparent: all background noise mixes with the speech sounds. With current methods it is often difficult to separate speech and background. Recognition is even more troublesome when two people are talking at the same time. A microphone will record the sounds from both sources as one mixed signal, which is difficult to process later on. For sign recognition, the background is less of a problem, since images are not transparent. Someone signing behind the primary signer's back does not hamper the recognition process. The secondary signer's signs are blocked by the primary signer and hence are not recorded by a camera aimed at the primary signer. But other forms of interference can be troublesome for ASLR. For instance, if the signer wears clothes in the same color as his or her hands, or if the visible background is cluttered with skin-colored objects or other people, the hands will be hard to find and track automatically. The reason is that, in the field of computer vision, finding an object in an image is usually done by finding a surface of the appropriate size and color. If clothing and hands are of almost the same color, or if there are skin-colored objects in the background, this becomes very tricky. Another form of interference is the occlusion of parts of the sign, for instance when one hand blocks the other from view.

Yet other problems have to do with modeling the language. There are different ways of handling automatic speech or sign recognition, but almost all of them rely on comparing sounds/sign data with stored examples. To do this, one has to model some unit at some representational level (word level, syllable level, sound level) in the language, and then match the observed data with these models. The first idea is usually to model entire words. This is because words are highly salient linguistic units for humans. There

is a disadvantage to this approach however: **poor expandability.** For every word, a separate model must be built. Building models often requires a lot of time and data, so expanding the vocabulary becomes costly and time-consuming. In addition, a large vocabulary of thousands of models will become difficult to search through efficiently. For these reasons, most current ASR methods focus on modeling phonemes instead of whole words. The advantage is that with only a limited number of phonemes nevertheless an infinite number of words can be formed. Once all phonemes in a language have been modeled, adding a new word to the vocabulary only requires adding a translation from a string of phonemes to this word to the lookup dictionary. Consequently, the phoneme level is the preferred level of modeling in current ASR.

Finally, a short note on **simultaneity**. Traditionally, speech is thought of as a sequential stream of sounds that represent meaning. The meaningful parts in speech, phonemes, all occur neatly in a row, or so it is often assumed. In sign languages, on the other hand, the meaningful parts (handshape, orientation etc.) occur simultaneously as well as sequentially. How to process these simultaneous 'phonemes' is a problem for ASLR. But notice that simultaneity can also yield problems for ASR: it is becoming clear that in speech, too, certain meaningful parts can occur simultaneously. For example the type of sound (e.g. an /a/) and its pitch. More on this will be said in section 4, in the discussion on the basic units of modeling.

3. Current results of automatic sign recognition methods

Having looked at some of the problems of automatic recognition of speech and sign language in general, we now turn to automatic sign recognition. What is the current

standard of ASLR research? At the moment, the most successful methods can recognize around 85-90% of the signs correctly, usually using small vocabularies (around 100 words is typical). Most studies use data from one signer only to train and test the method. (For comparison: Current state-of-the-art speech recognition systems, like automatic dictation, have a word-level accuracy of around 99% for one speaker and a vocabulary of thousands of words.[8]) Note that current ASLR methods are all trained and tested on their own data sets, which for the most are not publicly available. This makes it difficult to compare their results directly.

The recognition rate of current ASLR methods typically drops when a method is trained and/or tested with different signers. This is due to the inter-signer variability mentioned earlier. How well the different methods handle continuous signing is not always clear. Many projects use sign sentences as test material, but whether these are real, natural sentences or merely single signs put together is not always mentioned. The difference is important, because in a natural sentence, signs influence each other, whereas a combination of isolated signs does not suffer from this phenomenon.

Table 1 provides an overview of ASLR research. The overview is not exhaustive, but focuses on the most important results. Before turning to the evaluation of these methods, it is important to note that the criteria for successfulness can vary depending on the goal one has in mind. If the objective is to create a sign language user interface for the computer which can understand a limited range of fixed commands, such as ‘Open web browser’, ‘Exit’ etc., obviously a crude recognition method with a limited vocabulary is adequate. On the other hand, if the objective is to make the portable sign language interpreter mentioned in the introduction, a more sophisticated method is required. Such a

method must be capable of recognizing continuous signing, from any person, in any environment, and must have a large vocabulary. Therefore, a method which can only handle small vocabularies is not necessarily useless. However, since the ultimate goal of ASLR research is to recognize unconstrained, natural sign language, we will evaluate the various methods from that perspective.

Table 1: Overview of the most important automatic sign recognition methods.

Modeling method	**Reference (sign language)**	**Recognition rate (vocabulary size)**	**Isolated or continuous**	**Remarks**	**Capturing method**
Two-stage neural network	Waldron & Kim, 1995 (American SL)	- 86% with back propagation - 84% with self-organizing (14)	Isolated	- Handshape, orientation, location and motion were used - Six signers for training and testing	One instrumented glove + Magnetic tracker
Machine learning techniques (instance based learning and decision tree building)	Kadous, 1996 (Australian SL)	- 80% for IBL - 40% for tree (95)	Isolated	Only a few characteristics of a sign are used: Hand position, finger bend and wrist rotation	One instrumented glove
Whole-word Hidden Markov Models (HMMs)	Grobel & Assam, 1997 (Dutch SL)	91.3% (262)	Isolated	Signer-dependent	Camera
HMMs	Liang & Ouhyoung, 1998 (Taiwanese S)	80.4% (250)	Continuous	Signer-dependent	One instrumented glove
HMMs	Starner et al., 1998 (American SL)	- 75% without constraints - 92% with constraints (40)	Continuous	Tested on 5-word sentences, both with and without constraints on word order	Camera
HMMs representing sub-units (see section 4)	Bauer & Hienz, 2001 (German SL)	81% (12)	Continuous	- No data on multiple signers available - Learning sub-units from the data tends to optimize for a certain data set instead of for the entire language	Camera
Simple Recurrent Network + HMMs	Fang et al., 2002 (Chinese SL)	92% (208)	Continuous		Instrumented gloves + Magnetic trackers
Whole word HMMs combined with efficiency techniques	Chen et al., 2003 (Chinese SL)	92% (5113)	Continuous	Accuracy drops to 84% when data from 6 instead of 1 signer is used	Instrumented gloves + Magnetic trackers
Parallel HMMs for handshape and movement (see section 4)	Vogler & Metaxas, 2004 (American SL)	96.1% (22)	Continuous	No data on multiple signers available	Instrumented gloves + Magnetic trackers
Whole word Markov chain models	Bowden et al., 2004 (British SL)	84% (49)	Continuous	No data on multiple signers available	Camera
Whole word HMMs	Zieren & Kraiss, 2004 (German SL)	98% (152)	Continuous	- No data on multiple signers available - Multiple hypotheses for position of hands and face are maintained and resolved with higher-level processing	Camera

The overview in table 1 shows that most sign recognition studies only use small vocabularies and a single signer. Some methods do not use handshape or hand orientation, or use a limited range of shapes only. It is striking that none of the current methods use the non-manual component of a sign (facial expression, mouth shape), although this is as much a 'phoneme' in sign languages as handshape or position, at least for Sign Language of the Netherlands and German Sign Language (Canzler and Ersayar, 2003; Schermer, Fortgens, Harder, and de Nobel, 1990). However, Canzler and Ersayar (2003) propose a technique for using facial features in ASLR.

Some methods capture their sign data with the aid of a digital camera; others use instrumented gloves and magnetic trackers. Instrumented gloves are gloves containing sensors which measure the angle of bend of a number of finger joints. With these data, approximate handshape can be calculated. Magnetic trackers are sensors that calculate their position with respect to a fixed source. This source can be worn on the body to give the position relative to the signer. These trackers are used to estimate hand position and motion. The advantage of capturing signs with gloves and trackers is that it is easier and faster to calculate handshape, position and motion from sensor data than from video images. If signs are captured on video, the hands and face must first be found and tracked in the image, and then handshape, palm orientation, position and motion must be determined from 2D image data. Both tasks are difficult to model computationally. The disadvantage of instrumented gloves, however, is that they do not provide data on the non-manual component of a sign. A camera usually films the entire upper body of the signer, and thus contains an image of the face as well as the hands. This is an advantage

of camera-based methods. Although no current method uses the non-manual component, techniques to automatically identify facial expressions are currently being developed in image processing research (e.g. Fasel, Stewart-Bartlett, Littelwort-Ford, and Movellan, 2002; Tian, 2004). If these techniques are applied to sign language images, they can be used to extract information about the non-manual component of a sign. But for glove-based methods, this is not an option.

Another matter that draws the attention is the fact that almost all recent methods rely on Hidden Markov Models (HMMs). What makes them so attractive? HMMs are statistical models. They can be trained to represent certain elements (words, phonemes) simply by providing them with relevant examples. There is no need to explicitly formulate rules that govern the behavior of these elements. Since it is often quite difficult to make explicit these abstract linguistic rules, it is a great advantage to model language elements without having to use rules. Moreover, since they are statistical in nature, HMMs can handle variation quite well. As long as the training data contain sufficient examples of all variants, the model can deal with each of them. And finally, HMMs can work in real time and are quite efficient considering the discriminative power they have.

But HMMs have disadvantages, too. An important disadvantage is that it requires a great amount of time and data to train a single model. The more variety the data contain, the more examples an HMM requires to represent all variants. This means that training HMMs for multiple signers will take more resources than training them for a single signer (more signers means more variation among signs). Furthermore, the tax on resources makes HMMs unattractive for modeling entire words. After all, this involves creating one or two models for thousands of words, which requires an enormous amount

of training data and time. Nonetheless, Chen et al. (2003) have invested this great amount of work to build whole word HMMs for Chinese SL. But their recognition rate drops significantly when they use six signers instead of one. To make their system user-independent, they will probably need even more data than the ca. 60,000 training examples they have used so far. Furthermore, the total amount of models required would probably make the model bank impossible to search through in finite time. In effect, true user-independence does not seem to be feasible for such a system. Another drawback of modeling whole words is that all the work that is done benefits one sign language only. Recognizing a different language in the same manner entails training new HMMs all over again.

Another problem has to do with the way HMMs calculate probabilities. HMMs do not process the input data as a whole, but rather piece by piece, with the pieces assumed to be statistically independent. Therefore, they calculate total probabilities by multiplying the probabilities of the pieces. Probabilities are fractions of 1, so the more probabilities are multiplied, the smaller the total probability becomes. This means that a sign that is produced more slowly than it ought to be according to the training database is assigned a considerable lower probability than a version that is produced in accordance with the training data. This inability to deal with temporal variation is a problem for HMMs.

For these reasons, many researchers consider modeling whole words with HMMs uneconomical. Some researchers are now trying to model parts of signs instead. Just like in spoken languages, there are only a limited number of meaningful sign parts, the ‘phonemes’ of sign language, with which an infinite number of signs can be created. If these sign parts are modeled, only a limited number of models are necessary. In such a

case, the resources HMMs require would not be a problem. Furthermore, it may be possible to use the modeled 'sign phonemes' to represent signs from different sign languages. However, since researchers who model sign parts have only tested their methods on small vocabularies, it is difficult to estimate how successful they will eventually be.

Not all researchers have abandoned the idea of modeling entire signs, though. Instead, Bowden, Windridge, Kadir, Zisserman, and Brady (2004) tried to find a more efficient alternative for HMMs. They chose to use Markov chain models. Markov chains can be regarded as more restricted versions of HMMs. In Bowden et al.'s approach, successful Markov chain models can be created from as little as one training example. Consequently, modeling with Markov chains is fast and easy. Under these circumstances, modeling entire words for a large vocabulary is unproblematic. On the other hand, it is uncertain how well Bowden et al.'s method can handle co-articulation effects and effects of adjacent signs (see section 2) in natural sentences. Moreover, like all methods that model entire words, theirs will still have to deal with the problems of storing and manipulating a huge model bank.

A final issue we wish to draw the attention to is the fact that current research focuses solely on context-independent signs, or lexical versions of context-dependent signs. Signs that vary with context, such as directional verbs, localized signs and personal pronouns, are only used in their dictionary form or are not used at all. This is understandable, because ASLR is still in its early stages, and context-dependent signs are harder to handle than signs that always have the same form. But ignoring context-dependency in

developing ASLR methods can pose serious problems later, if one wants to expand the method to deal with natural sign language. This is a topic of the next section.

4. Limitations of automatic sign recognition

The problems of ASLR discussed in the previous section show that the current methods are still far from perfect. Partly, this is because of difficulties that both speech recognition and sign recognition share, which were discussed in section 2. However, as ASLR research continues, it is becoming apparent that the recognition of sign languages has its own specific difficulties. These difficulties are due to characteristics of sign language that have no immediate counterpart in spoken languages, or are due to the differences between the visual and auditive modality. As such, solutions for these problems cannot be borrowed from automatic speech recognition. New solutions will have to be found. Some sign-specific problems were described in ten Holt (2004). These problems are discussed below.

Distinguishing gestures from signs. A problem that has hardly been addressed at all yet, is the blurring of signs and gestures in sign language. In spoken language, human gesturing does not interfere with the speech signal, since it is in a different modality. But in sign languages, gesture and speech share the same modality and blur (Liddel and Metzger, 1998; Taub, 2004[9]). This issue has not been addressed by ASLR researchers at all. Nevertheless, it could constitute a major difficulty.

Context-dependency. In the previous section, context-dependency was already mentioned as a problem area that ASLR research ignores for now. But if natural sign language is to be recognized, the problems of handling context-dependent signs must be

solved. These problems mostly pertain to the spatial character of sign language grammar. Consider the ASL sentence “Mother brings the food from the store to the kitchen”. TO-BRING is a directional verb. The direction denotes from where to where something is brought. To interpret this sentence, a recognition method must not only be able to identify the signs in the sentence, it must also take into account the direction in which TO-BRING was made. If it does not, it cannot determine what the starting point of the movement was, and what the end point. If both STORE and KITCHEN have been localized earlier, they will not be signed again. The sign TO-BRING will simply start at the reference location of STORE and end at the reference location of KITCHEN. If the automatic recognizer does not retain information about referent locations and the direction of the verb, it cannot interpret this sentence. At best it would yield the incomplete proposition “Mother brings the food from X to Y”, but crucial information would be lost.

The problem lies in the fact that a context-dependent sign takes on different forms in different grammatical contexts while still denoting the same concept. The sign TO-WALK still means ‘to walk’, whether it is made to the left, to the right, or in a zigzag movement. The extra information imparted by the direction of motion is grammatical information. But in other signs, direction of motion is a sign feature which discriminates two different signs. It is the context that offers the clue as to how the motion must be interpreted: as a sign characteristic, or as a grammatical feature. To process such a sign automatically, the context will have to be taken into account by the recognizer. There are similar phenomena in spoken languages, but context-dependency seems to be a more important factor in sign languages.

Basic unit of modeling. As mentioned in section 2, one must choose a level of modeling in automatic speech or sign recognition, such as word level or phoneme level. In current ASR research, the phoneme is the preferred level of modeling. The phoneme can be defined as the smallest contrastive unit in the sound system of a spoken language. For sign languages, however, it is not clear at all which part of a sign should be taken as the smallest contrastive units. In signs, several features carry meaning: handshape, palm orientation, position, motion and the non-manual component. If one of these elements changes, usually the meaning of the sign changes as well, so these features could be considered the basic units of sign language. But one cannot use the same techniques that ASR uses in spoken phoneme modeling to model these features, because spoken phonemes are largely sequential, whereas the features of sign language can be simultaneous as well as sequential. Although certain features can only occur in sequence (e.g., two different handshakes), the features of sign language often are simultaneous and occur together (e.g., a handshape, position and orientation).

The difficulty lies in the level of description for both types of language. In spoken languages, extensive phonetic and phonological research has resulted in several levels of description, including distinctive features (such as voicing, place of articulation and manner of articulation), phonemes (which are combinations of values of each of these features, such as “voiced, bilabial, stop,”) and syllables (sequences of phonemes, with extra information such as pitch), and words. In sign language, these levels are also distinguished (Perlmutter, 1992; Sandler, 1996; van der Hulst and Mills, 1996). But the phonetics and phonology of sign languages have been researched for a much shorter

period than those of spoken language. Consequently, ASLR research has not really exploited the results yet in deciding on the appropriate basic units.

Current ASR methods often use the sequentiality of phonemes. The property of sequentiality makes it difficult for ASLR to use existing ASR techniques. For instance, in ASR, an HMM assumes that all data at any one time must be ascribed to a single model. Because of the assumption that only one phoneme can be present at a given moment, only one phoneme model should explain the data. But for a sign, the observations at any given moment in time might have to be ascribed to several models if features are used as basic units: one to explain the handshape-part, one for the motion-part, etc. For such a task, most methods borrowed from ASR are ill-equipped. Either the methods must be adapted to allow for simultaneous basic units, or different basic units must be chosen.

Vogler and Metaxas (2004) try to develop methods for simultaneous language units. They propose parallel HMMs to model the parallel units of sign language. Their method has only been tested for parallel modeling of movements and handshapes, though. Their results were a 4% improvement from using ordinary HMMs (Vogler and Metaxas, 1997, 2004).

Others tried to find different basic units in sign language, in particular units that are guaranteed to be sequential. An example of different language atoms is the sign 'phoneme' system of Movements and Holds proposed by Liddell and Johnson (1989). This system divides signs into parts where there is no change in sign features, and parts where there is. These are the basic units. Each unit has a bundle of features associated with it, in which handshape, orientation, movement and location are stored. Vogler and Metaxas (1999) tried to use these 'phonemes' as basic units for automatic sign

recognition. Bauer and Hienz (2001) tried yet another approach, modeling self-organizing 'sub-units'. These are sign parts that have no linguistic justification but are learned from the data itself as useful basic units.

In conclusion, there are different ways in which signs can be divided into parts. In deciding which basic unit is best, linguistic arguments as well as practical arguments can be considered. Using sequential units is convenient because this allows ASR methods based on sequentiality to be used. On the other hand, perhaps ASR will have to abandon the notion of sequentiality, too, since certain meaningful features in speech (e.g., suprasegmental properties such as pitch) occur simultaneous to other features as well. The approaches that model parts of signs have not been tested sufficiently to allow any conclusions about their success to be drawn. At this point, a question that arises is whether recognition models should restrict themselves to only one unit of representation. Can language perhaps be recognized best by modeling various levels (parts of words, words, sentences) all at the same time? This will be discussed further in section 5.

Transitions between signs. When two signs are made in succession on different locations, an extra movement has to be inserted to transport the hand from the first location to the second. This phenomenon is often referred to as *movement epenthesis.* The extra movement has no meaning, but is a consequence of the fact that a person cannot move his hand instantaneously from one location to another. Note that this is not the same as co-articulation. The latter refers to the changing of sounds/signs when they overlap in time. With epenthesis, an *extra* movement is inserted. Movement epenthesis can be solved in ASLR by modeling the epenthesis movements explicitly. There are not many possible trajectories, so this is can be done (Vogler and Metaxas, 1997). However,

usually the hands already start shifting location in anticipation of the next location before the sign is finished. This contextual influence is a problem for ASLR, since it causes the position feature of a sign to change. In speech, epenthesis is not considered a major problem. This means that there are no solutions for epenthesis that can be borrowed from ASR.

Repetition. Repetition occurs when a sign, in its lexical form, consists of a circular or to-and-fro movement that is made a couple of times. An example is the Sign Language of the Netherlands word PRATEN (to talk). PRATEN is a two-handed sign. Both hands assume the '1' handshape and move forward from the mouth and back to the mouth in alternation. Evidence suggests that in these signs, the number of repetitions is unimportant: in an experiment, Ten Holt (2004) found that in a laboratory situation, a signer, when asked to make the Dutch sign PRATEN ten times in a row, varied the number of repetitions from two-and-a-half to five (each back or forth was counted as one repetition). This kind of repetition must be distinguished from certain forms of added repetition (adding repetition to a sign that has none in its lexical form) or altered repetition that are used for denoting aspect in sign languages such as ASL. In these cases, the sign itself has none or a different kind of movement, and a distinct type of repeated movement is added to convey the extra grammatical meaning (Klima and Bellugi, 1979). The signs we are concerned with here are signs that in their lexical form already contain a repetitive movement. In these signs, variations in the number of repetitions and precise end position were observed (Ten Holt, 2004), which did not result in any difference in meaning. The significance for such signs seems to lie in the mere presence of repetition, not in the exact number of repetitions.

If this is indeed the case, though, it creates a problem for ASLR. In automatic speech recognition, models generally assume that a word consists of a certain number of elements (sounds), in a certain order, which must all be there, without any extra sounds. It is clear from the discussion above that such constraints cannot be used for a language that has words containing repetition. This was shown in ten Holt (2004). Consider the interjection *well well well*, the closest thing to a repetitive word that we could find in spoken language. An HMM representing *well well* does not consider the expression *well well well* as fitting because there are too many elements in the data (the third *well* has no place in the model). For similar reasons, a long model cannot explain a short sequence. Intuitively, a cyclic model (a model allowing repetition of parts of it) would be necessary for a cyclic sign. HMMs can handle repetitions in the data. However, whether they can be adapted to represent repetitive signs, or whether a new kind of model needs to be developed, remains to be seen.[10]

5. Future Directions

Looking back on the problems of ASLR – context-dependency, distinguishing signs from gestures, determining the basic units of modeling, movement epenthesis and repetitive signs – it becomes apparent that they all have a certain central feature in common. They are problematic because they all display a mismatch between physical form and meaning. Usually, a spoken word or a sign has a certain recognizable form with which a certain meaning is associated. Some forms have multiple meanings (homonyms); some meanings have multiple forms (synonyms). But in the sign language phenomena discussed in section 4, there are forms that have no meaning (movement epenthesis and the extra

repetitions of repetitive signs), and meanings that have no form (in context-dependent signs, meaning is provided by context, and not (solely) by form). This mismatch between form and meaning might very well be the reason why these phenomena are so difficult to deal with in automatic recognition. And although gestures are not entirely meaningless, they are not symbols in the same sense that signs are symbols. So they, too, can be considered to be forms that have no fixed and well-circumscribed meaning. Given these phenomena, a recognizer cannot assume that all elements in its data carry meaning, and at the same time must allow for the possibility that there are aspects of meaning that have no counterpart in the data. This is a problem for all methods that process the data exclusively bottom-up (i.e., on the basis of the input data only, without using prior knowledge and expectations).

One possible solution is to use top-down processing as well in the recognition process, that is, to use what has been recognized already to predict what will probably be encountered next. This provides clues for choosing between ambiguous physical forms. For example, in spoken language, when the sounds /h/, /E/, and /l/ have been recognized already, but the fourth is ambiguous and could be interpreted as /m/ or as /b/, top-down processing can help. Because *helm* is an existing word, and *helb* is not, the /m/ is more appropriate in the local context. As a result, the ambiguous sound can best be interpreted as /m/. This process is sometimes called ‘bootstrapping’: pulling yourself up by your own bootlaces. It is paradoxical, because recognition results (the already recognized sequence /hEl/) are used to form a prediction of what is to come, and this prediction is used to achieve further recognition results (e.g., with respect to the /m/ that was previously not

recognized yet), which can then be used to form further or improved predictions, which can again help recognize yet other parts of the data, etcetera.

Top-down processing is essential for human language processing. It is impossible to recognize speech in everyday, noisy environments without the aid of higher-level knowledge (such as the fact that *helm* is a word and *helb* is not, for example). Because ASLR is a young field of research, it has not yet felt the need for top-down processing. However, we argue that the phenomena mentioned in the previous section are typical examples of physical forms that require higher level knowledge to be successfully mapped onto their meanings. We therefore believe that top-down processing should be integrated in the process of automatic sign recognition. Local context must be utilized to enable automatic recognition of sign language utterances. Bauer, Hienz, and Kraiss (2000) and Vogler and Metaxas (1997) already move in this direction with their use of bigram language models. Higher-level information can also be used in HMMs, in the form of ad hoc constraints that are built in by hand. These can improve recognition results, but only as long as the practice situation corresponds to the previously built-in constraints.

When humans process language, they maintain hypotheses at many different levels of representation simultaneously: about phoneme identities, possible syllables, possible words, possible sentences, plausible meanings, etcetera. Maybe automatic speech or sign recognition will need to do the same: maintain multiple hypotheses at many different representational levels, in effect using more than the immediate, local context. It may be the only way for a sign recognizer to handle context-dependency, epenthesis, gestures, repetition, incomplete information, obscuring, and other difficulties with the same ease

and accuracy as humans. Perhaps this is what is required if the palm-top sign language interpreter is ever going to be realized.

Acknowledgements

The authors would like to thank an anonymous reviewer for his/her valuable comments and suggestions.

Notes

[1] See also: http://www.visicast.sys.uea.ac.uk/Public.html
[2] See: http://asl.cs.depaul.edu/project_info.html
[3] See: http://www.vcom3d.com/
[4] See: http://www.gebarennet.nl/
[5] See also: http://sy.jdl.ac.cn/en/synthesis.asp
[6] Actually, there is a third process, translation, needed to convert sign language to spoken language, since sign languages are natural languages and not signed encodings of spoken languages.
[7] The term 'word' is used throughout this article to mean either a spoken or a signed word.
[8] See: http://www.scansoft.com/naturallyspeaking/standard/ for a state-of-the-art ASR system.
[9] See also: http://research.communication.utexas.edu/isgs/Contributions/Taub/taub.html
[10] Current ASLR research does not mention the problem of repetition. Whether researchers do not use repetitive signs, or whether their methods can handle the problem, is not clear. It is not mentioned in any of the studies we are acquainted with.

References

Bauer, B., and Kraiss, K.-F. 2002. Video-Based Sign Recognition using Self-Organizing Subunits. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002), Québec City, Canada, August 11-15 (on CD-ROM)*, ed. R. Kasturi, D. Laurendeau, and C. Suen. IEEE Computer Society.

Bauer, B., Hienz, H., and Kraiss, K.-F. 2000. Video-Based Continuous Sign Language Recognition Using Statistical Methods. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00), Sept 3-7.* Volume 2: 2463.

Bowden, R., Windridge, D., Kadir, T., Zisserman, A., and Brady, M. 2004. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In *Proceedings of the 8th European Conference on Computer Vision (ECCV'04), Prague, Czech Republic, May 11-14*, ed. T. Pajdla, and J. Matas, volume 1:391- 401. Heidelberg: Springer-Verlag.

Canzler, U., and Ersayar, T. 2003. Manual and Facial Features Combination for Videobased Sign Language Recognition. In *Proceedings of the 7th International Student Conference on Electrical Engineering (POSTER 2003), Czech Technical University, Prague, May 22*, IC8.

Chen, Y., Gao, W., Fang, G., Yang C., and Wang, Z. 2003. CSLDS: Chinese Sign Language Dialog System. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'03)*, *Nice, France, October 17*, 236-238. IEEE Computer Society.

Cox, S.J., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., and Abbott, S. 2002. TESSA, a system to aid communication with deaf people. In *Proceedings of the 5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2002), Edinburgh, Scotland, July 8-10*, 205-212.

Fang, G., Gao, W., Chen, X., Wang, C., and Ma, J. 2002. Signer-independent continuous sign language recognition using SRN/HMM. In *Gesture and Sign Language in Human-Computer Interaction: Proceedings of the International Gesture Workshop*

(GW2001), London, UK, April 18-20. ed. I. Wachsmuth and T. Sowa, 2298: 76-85. Heidelberg: Springer-Verlag.

Fasel, I., Stewart-Bartlett, M., Littelwort-Ford, G., and Movellan, J.R. 2002. Real time fully automatic coding of facial expressions from video. In *Proceedings of the 9th Symposium on Neural Computation, California Institute of Technology, May.*

Grobel, K., and Assam, M. 1997. Isolated Sign Language Recognition Using Hidden Markov Models. In *Proceedings of the IEEE International Conference on Systems, man and Cybernetics, Orlando, Florida.* 162-167.

Huang, X., Acero, A., and Hon, H.-W. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall PTR.

Kadous, M.W. 1996. Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In *Proceedings of the Workshop on the Integration of Gesture in Language and Speech (WIGLS'96), Delaware, USA, October 7-8*, 165-174. Wilmington, DE.

Liang, R.-H., and Ouhyoung, M. 1998. 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.* 558-565.

Liddel, S.K., and Metzger, M. 1998. Gesture in Sign Language Discourse. *Journal of Pragmatics*, 30: 657-697.

Liddell, S.K., and Johnson, R.E. 1989. American Sign Language: The phonological base. *Sign Language Studies*, 64: 195-277.

Lincoln, M., Cox, S.J., and Nakisa, M. To appear. The development and evaluation of a speech to sign translation system to assist transactions. *International Journal of Human-Computer Interaction*.

Perlmutter, D. 1992. Sonority and syllable structure in American Sign Language. *Linguistic Inquiry*, 23: 407-422.

Poizner, H., Klima, E.S., and Bellugi, U. 1987. *What the hands reveal about the brain*. Cambridge, Massachussetts: MIT Press.

Sandler, W. 1996. Phonological features and feature classes: The case of movements in sign language. In *Issues in the phonology of sign language. Theme issue of Lingua*, ed. H.G. van der Hulst and A. Mills. 98: 197-220.

Schermer, G.M., Fortgens, C., Harder, R., de Nobel, E. 1991. *De Nederlandse Gebarentaal*. Twello: Van Tricht.

Starner, T., Weaver, J., and Pentland, A. 1998. Real-time American Sign Language using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12): 1371-1375.

Taub, S.F., Galvan, D., and Pinar, P. 2004. The Encoding of Spatial Information in Speech/Gesture. Paper presented at Theoretical Issues in Sign Language Research (TISLR) 8, Barcelona, Spain.

Ten Holt, G. 2004. *The Eye of the Beholder: Automatic Recognition of Dutch Sign Language*. MA Thesis Artificial Intelligence, University of Groningen.

Tian, Y. 2004. Evaluation of Face Resolution for Expression Analysis. In *Proceedings of the IEEE Workshop on Face Processing in Videos (FPIV'04), Washington D.C, June 28.*

Van der Hulst, H., and Mills, A. 1996. Issues in sign linguistics: phonetics, phonology and morpho-syntax. In *Issues in the phonology of sign language. Theme issue of Lingua*, ed. H.G. van der Hulst and A. Mills. 98: 3-18.

Vogler, C., and Metaxas, D. 2004. Handshapes and movements: Multiple-channel ASL recognition. In *Lecture notes in computer science*, 2915: 247-258. Heidelberg: Springer-Verlag.

Vogler, C., and Metaxas, D. 1999. Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes. In *Gesture-Based Communication in Human-Computer Interaction: Proceedings of the International Gesture Workshop (GW'99), Gif-sur-Yvette, France, March 17-19*, ed. A. Braffort, R. Gherbi, S. Gibet, J. Richardson and D. Teil, 211-226. Heidelberg: Springer-Verlag.

Vogler, C., and Metaxas, D. 1997. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SM'97), Orlando, Fl, October 12-15*, 156-161. IEEE Computer Society.

Waldron, M.B., and Kim, S. 1995. Isolated ASL sign recognition system for deaf persons. In *IEEE Transactions on Rehabilitation Engineering*, 3(3):261-271.

Zieren, J., and Kraiss, K.-F. 2004. Non-Intrusive Sign Language Recognition for Human-Computer Interaction. Paper presented at the *9th IFAC/IFIP/IFORS/IEA Symposium of Analysis, Design, and Evaluation of Human-Machine Systems, Atlanta, Georgia, September 7-9*.