AUTOMATIC SIGN LANGUAGE RECOGNITION
INSPIRED BY HUMAN SIGN PERCEPTION

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.Ch.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 23 juni 2010 om 10.00
door

Gineke Anna TEN HOLT

doctorandus kunstmatige intelligentie
geboren te Groningen
Dit proefschrift is goedgekeurd door de promotoren:
Prof. dr. ir. M.J.T. Reinders
Prof. dr. H. de Ridder

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof. dr. ir. M.J.T. Reinders, Technische Universiteit Delft, promotor
Prof. dr. H. de Ridder, Technische Universiteit Delft, promotor
Prof. dr. K. Emmorey, San Diego State University, USA
Prof. dr. J.J. Koenderink, Technische Universiteit Delft
Prof. dr. ir. R.L. Lagendijk, Technische Universiteit Delft
Prof. dr. L.J.M. Rothkrantz, Technische Universiteit Delft
Dr. T.C. Andringa, Rijksuniversiteit Groningen

Prof. dr. C.M. Jonker (reservelid), Technische Universiteit Delft

Dr. A.J. van Doorn en Dr. E.A. Hendriks hebben als begeleiders in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

This work was partly supported by the VSB-fund: grant number 2003451, and by the ICT Delft Research Centre of Delft University of Technology.

This work was carried out in graduate school ASCI. ASCI dissertation series number 195.

Printed and cover design by Universal Press – Veenendaal – NL.
© 2010 Gineke ten Holt, all rights reserved.

cover: handshape ‘T’ with several depictions of the path the eyes follow when perceiving this shape.
“WE TAKE OUR SENSES FOR GRANTED”

OLIVER SACKS, *Musicophilia*
## CONTENTS

1. Introduction ......................................................................................... 1
   1.1 Background Information .............................................................. 3
       1.1.1 Sign Languages ................................................................. 3
       1.1.2 Automatic Classification ...................................................... 8
       1.1.3 Automatic Sign Language Recognition .................................. 8
   1.2 Relevant Literature ......................................................................... 10
   1.3 Summary of Results ....................................................................... 11

Part I Human Sign Language Recognition ................................................. 13

2. Which Fragments of a Sign enable its Recognition? ............................ 15
   2.1 Introduction ................................................................................. 16
   2.2 Method ....................................................................................... 17
       2.2.1 Material ............................................................................ 17
       2.2.2 Stimuli .............................................................................. 21
       2.2.3 Task .................................................................................. 22
       2.2.4 Subjects ............................................................................ 23
   2.3 Results ......................................................................................... 23
       2.3.1 Recognition Overall ............................................................ 23
       2.3.2 Recognition per Sign .......................................................... 25
       2.3.3 Learning Effect ................................................................... 25
       2.3.4 Influence of Segment Length ............................................... 29
       2.3.5 Influence of Subject Background ........................................... 29
       2.3.6 Influence of Different Rest Positions ..................................... 31
   2.4 Discussion ..................................................................................... 33
       2.4.1 Main Conclusions ................................................................ 33
       2.4.2 The Recognition Process ..................................................... 34

3. Signs in which Handshape and Hand Orientation are not or only partially visible: what is the Consequence for Lexical Recognition? .. 41
   3.1 Introduction ................................................................................. 42
   3.2 Method ....................................................................................... 44
# CONTENTS

3.2.1 Sign Set .................................................. 44  
3.2.2 Recording .................................................. 45  
3.2.3 Manipulations ............................................. 46  
3.2.4 Task ...................................................... 49  
3.2.5 Participants ............................................... 50  
3.3 Results ..................................................... 50  
3.3.1 Processing of Responses ................................. 50  
3.3.2 Recognition per Condition ............................... 52  
3.3.3 Recognition of Individual Signs ......................... 54  
3.3.4 Participants’ Backgrounds ............................... 55  
3.3.5 Influence of Sign Characteristics ....................... 56  
3.4 Discussion ................................................. 57  
3.4.1 Conditions ............................................... 57  
3.4.2 Individual Signs ......................................... 59  
3.4.3 General Conclusions and Future Directions .......... 61  

4. Sign Language Perception Research for Improving Automatic Sign Language Recognition ........................................... 69  
4.1 Introduction .................................................. 70  
4.1.1 Related Work ............................................. 71  
4.2 Information Distribution in the Time Domain ........... 71  
4.2.1 Experiment 1: Detecting a Sign ......................... 72  
4.2.2 Experiment 2: Recognizing a Sign ....................... 73  
4.2.3 Experiment 3: Recognition of Isolated Sign Phases .. 73  
4.2.4 Results .................................................. 73  
4.3 Informative Elements of Signs ............................. 75  
4.3.1 Experiment 4: Importance of Hand Configuration for Recognition .... 76  
4.3.2 Experiment 5: Allowed Variation within Elements .. 76  
4.4 Discussion ................................................. 80  
4.4.1 Practical Application ..................................... 80  
4.4.2 Implications for Sign Recognition Theory ............ 82  

Part II Applying Sign Perceptual Knowledge ................. 85  

5. Influence of Handshape Information on Sign Recognition ................................................. 87  
5.1 Introduction .................................................. 88  
5.2 Method ...................................................... 89  
5.2.1 Dataset .................................................. 89  
5.2.2 Sign Recognition Algorithm ......................... 89  
5.2.3 Handshape Representation ............................. 90  
5.2.4 Missing Values ......................................... 91  
5.2.5 Evaluation Method ....................................... 92
7.4.5 Future Directions ........................................ 136

8. Discussion .................................................... 141
  8.1 Application of Knowledge about Human Sign Recognition ............ 141
  8.2 Investigating Human Sign Recognition Processes ..................... 145
  8.3 Future Directions ........................................ 147

Appendices ....................................................... 151
  A. Dataset ...................................................... 153
    A.1 Dataset Recording ....................................... 153
    A.2 Extracted Features ....................................... 154
    A.3 Sign Vocabulary .......................................... 155
  B. Dynamic Time Warping using Multiple Features ....................... 159
    B.1 Introduction .............................................. 159
    B.2 Multi-dimensional Dynamic Time Warping ......................... 160
      B.2.1 Algorithm ............................................. 160
      B.2.2 DTW on Derivatives .................................. 163
      B.2.3 Dimension Selection .................................. 163
    B.3 Experimental Results ..................................... 164
      B.3.1 Dataset ................................................ 165
      B.3.2 Artificially Warped Signals ............................ 165
      B.3.3 Application in Gesture Classification ..................... 167
    B.4 Discussion ................................................. 169
  C. Recognition Algorithms .................................... 171
    C.1 Combined Discriminative Feature Detectors Algorithm ............ 171
      C.1.1 Feature Selection ..................................... 171
      C.1.2 Training of Feature Classifiers ........................ 172
      C.1.3 Calculation of SDT .................................... 175
      C.1.4 Testing ................................................. 175
      C.1.5 Calibration ............................................ 175
    C.2 Key Frame Algorithm ....................................... 176
      C.2.1 Calculation of Key Feature Vectors ....................... 176
      C.2.2 Feature Classifiers and Property Classifiers .............. 177
      C.2.3 SDT, Calibration and Testing ........................... 177
    C.3 Aspects Algorithm ......................................... 179
      C.3.1 Determining SDTs ....................................... 180
      C.3.2 Calibration and Testing ................................ 180

Bibliography ...................................................... 181
Machines play an increasingly important role in daily life. Automatic appliances are pervading every aspect of human life, from transport to cleaning to communication. It is no wonder that techniques for interacting with machines are constantly developing. From buttons to key boards to point-and-click-interfaces to touch screens; man-machine interaction tends to grow more natural and intuitive. Automatic recognition of speech and gestures is the next step towards natural interaction: to communicate with machines the way we communicate with each other. Recognition of gestures made with hand held devices containing motion sensors (such as mobile phones [37] and the wii controller [74]) is being utilized more and more nowadays. However, the most natural way of recognizing gestures would be vision-based, with no sensors or devices encumbering a person. Such vision-based gesture recognition is still a major challenge [32, 42, 46].

Automatic sign language recognition is part of the general area of automatic gesture recognition. Recognition of sign language signs is often used as a starting point for general gesture recognition, because sign language signs are as a rule more strictly defined than ordinary gestures, both with regard to meaning and with regard to (ideal) form. Still, sign language recognition is not easy to perform automatically. It is especially difficult when the process is vision-based, with no specialized devices, such as magnetic trackers or form-sensitive gloves, to gather the data. All information regarding the hands and other important parameters must then be extracted from the video using computer vision techniques, after which the information must be used in such a way that the sign can be recognized. Both the automatic image processing part and the pattern recognition part are challenging problems [64].

Similar problems do not seem to hold for human beings: they are capable of processing sign language quite effortlessly. This evokes the question of how exactly they achieve this, and whether similar techniques can be employed in automatic sign language recognition. In the field of audio processing, such human-inspired techniques have been applied successfully [38], for example in masking noise by exploiting the fact that in the human ear, frequencies adjacent to the perceived frequency are suppressed [75]. It is
1. Which elements of a sign are informative for human sign language recognition?
   (a) Which moments / phases in time?
   (b) Which properties / characteristics of the sign?
2. Are the same elements also informative for the automatic recognizer?
3. Can we improve the automatic recognizer by focusing on the informative elements?

Box 1: Research Questions

conceivable that information about human sign recognition could be used similarly to improve automatic sign language recognition.

There are many ways in which automatic sign recognition could benefit from knowledge about human sign recognition. Knowing where human signers focus their attention during sign recognition could help us use (limited) computer vision resources more efficiently; Learning which elements of signs humans use to distinguish between similar signs could help develop techniques to avoid confusions between such signs in automatic recognition; Learning more about which elements of a sign are most informative could make it possible to disregard certain elements in advance; And knowledge about allowed variation could facilitate better techniques for estimating parameters of a sign based on only a few examples, thus removing the need for large sets of training examples.

In this thesis, we investigate options for using knowledge about human recognition of sign language to improve an existing automatic sign language recognition system [56, 57]. There is not a large body of research concerning human processing of sign languages — not surprising, since sign languages have only been subject of serious linguistic study since circa 1960 [80]. Though interesting work is available (such as [19, 30, 44, 66, 68]), research is by no means exhaustive. For this reason, and to tailor the research to the requirements of our automatic recognition system, we conduct our own experiments on human sign language recognition. Using the results of this research, we investigate possible improvements to the recognizer.

Our focus has been on the information content of various sign elements — fragments in time, as well as aspects of a sign such as hand shape or location (see section 1.1.1). The main research questions are given in box 1. Part I of the thesis describes experiments with human signers investigating the relevance of several of such elements for lexical recognition (question 1). Part II describes experiments with the above-mentioned automatic recognizer aimed at testing possible improvements based on the information gathered in Part I (questions 2 and 3). All recognition is done on isolated signs.
1.1 Background Information

Because the subject of this thesis is interdisciplinary, several scientific areas are involved: research into sign languages, automatic classification, and automatic sign language recognition. In this section, we briefly explain the concepts of these areas that are most important with regard to the research described in this thesis. Table 1.1 lists important terms that are used throughout the thesis. Since their precise definition often differs between areas of research, the table lists a definition for each as it is used in the current thesis. Sign glosses are typeset in capitals.

1.1.1 Sign Languages

Sign languages are natural languages with the same power of expression as spoken languages. This fact was first acknowledged after Stokoe demonstrated that signs, like words, consist of combinations of smaller elements that are meaningless in themselves [80]. Since then, great advances have been made in the study of sign languages, but many topics are still subject of discussion [29, 52]. One of the areas under discussion is the phonology of sign languages. The term phonology is used here to refer to the study of the elements that are important for the meaning of a sign or word. Examples of sign phonemes are a certain location (e.g. on the chin) or a certain handshape (e.g. a fist). Sign phonemes can occur both simultaneously (a certain handshape at a certain location) and sequentially (a sequence of two handshapes, e.g. going from open hand to fist). The precise organization of sign phonemes is still topic of discussion [12, 29]. In Sign Language of the Netherlands (SLN), phonemes are divided into five phonological categories: general hand location, handshape, hand orientation, hand movement, and the non-manual component [73]. They are listed in table 1.2. We refer to these categories as sign aspects. A change in aspect value entails a change in the meaning of the sign. Figure 1.1 shows examples of pairs of signs that differ in one of the aspects.

The status of the non-manual component as a phonological category is sometimes disputed [29]. In this thesis, the non-manual component is mostly disregarded, because the automatic recognizer that is used only analyses the manual part of signs. Also, handshape and hand orientation are sometimes considered a single aspect ‘hand configuration’ [29]. Nonetheless, in describing signs, we use the five aspects as defined in [73].

Other sign characteristics, such as speed of movement, do not have the status of a phonological category, because varying such a characteristic does not influence the meaning of a sign. Such characteristics can have an adverbial, grammatical or prosodic function [52, 73]; however, since the research in this thesis is concerned with isolated signs only, this is not relevant here.
Table 1.1: Definition of concepts used in the thesis. Some definitions differ from the common interpretation of the term.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Signs</strong></td>
<td></td>
</tr>
<tr>
<td>gesture</td>
<td>Man-made hand movement intended to communicate sign.</td>
</tr>
<tr>
<td>sign</td>
<td>Sign language sign. Unless stated otherwise, from Sign Language of the Netherlands.</td>
</tr>
<tr>
<td>SIGN GLOSS</td>
<td>Representation of a sign in words. Printed in small caps or caps.</td>
</tr>
<tr>
<td>(sign) fragment</td>
<td>Period of time within a sign</td>
</tr>
<tr>
<td>(sign) aspect</td>
<td>One of the phonological categories</td>
</tr>
<tr>
<td>(sign) element</td>
<td>Part of a sign, either a fragment or an aspect</td>
</tr>
<tr>
<td>stroke</td>
<td>Expressive phase of a sign as defined by Kita et al. [51]</td>
</tr>
<tr>
<td>preparation</td>
<td>Sign phase from movement onset to stroke start</td>
</tr>
<tr>
<td>retraction</td>
<td>Sign phase from stroke end to movement end</td>
</tr>
<tr>
<td>transition</td>
<td>Sign fragment that could belong to either of two adjacent phases</td>
</tr>
<tr>
<td><strong>Human Sign Perception</strong></td>
<td></td>
</tr>
<tr>
<td>recognition</td>
<td>Identifying which sign has occurred. Unless stated otherwise, lexical recognition is meant.</td>
</tr>
<tr>
<td>lexical recognition</td>
<td>Identifying an isolated sign, without (sentence) context</td>
</tr>
<tr>
<td><strong>Automatic Classification</strong></td>
<td></td>
</tr>
<tr>
<td>(sign) property</td>
<td>Visually measurable characteristic of a sign, such as horizontal position of the hand. See appendix A for an overview of properties.</td>
</tr>
<tr>
<td>frame</td>
<td>One image from a sign video. Unless stated otherwise, sign videos contain 25 frames / sec.</td>
</tr>
<tr>
<td>feature</td>
<td>One property in one frame</td>
</tr>
<tr>
<td>feature representation</td>
<td>Matrix containing all features extracted from a sign (all properties in all frames)</td>
</tr>
<tr>
<td>detection</td>
<td>Concluding whether an event (feature, property) has occurred or not</td>
</tr>
<tr>
<td>(sign) recognition</td>
<td>Unless stated otherwise, refers to identifying a test sign as being an example of the target sign or not (= detection of the target sign)</td>
</tr>
</tbody>
</table>
Table 1.2: Phonological categories in Sign Language of the Netherlands, referred to as aspects in this thesis. The values of the aspects determine the meaning of a sign.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>handshape</td>
<td>shape of the hand(s)</td>
<td>flat hand with fingers spread</td>
</tr>
<tr>
<td>hand orientation</td>
<td>orientation of the hand</td>
<td>palm facing left, fingers upward</td>
</tr>
<tr>
<td></td>
<td>palm(s) and fingers</td>
<td>on the chest</td>
</tr>
<tr>
<td>location</td>
<td>general location in which</td>
<td>on the chest</td>
</tr>
<tr>
<td></td>
<td>the sign is (mainly) made</td>
<td></td>
</tr>
<tr>
<td>movement</td>
<td>movement path of the hand(s)</td>
<td>zig-zag from high to low</td>
</tr>
<tr>
<td>non-manual component</td>
<td>facial expression, mouth</td>
<td>pursed lips</td>
</tr>
<tr>
<td></td>
<td>shape and/or mouthing</td>
<td></td>
</tr>
</tbody>
</table>

Signs, like gestures in general, are movements executed over time. The movement can be divided into movement phases [48, 62]. Figure 1.2 illustrates this. From a rest position, a signer has to bring his hands into the desired location and configuration. This phase is generally called the preparation. The phase containing the movements, location changes etc. that make up the sign is called the stroke, and the phase in which the hands return to the rest position is called the retraction. In continuous signing, preparation and retraction may be merged into the preceding respectively next sign in the sequence. Movement phases in gestures and sign language signs are subject of discussion [48, 51, 62], but in general, the existence of the three phases described here is accepted.

There are several ways of gathering information about the way people process sign language. In this thesis, we adopt the strategy of performing recognition experiments with human signers. Signers are asked to watch sign language stimuli from which certain elements have been removed or obscured, and try to identify the sign. The recognition results yield information about the role the missing elements play in the sign recognition process.

Important considerations in such experiments are the choice of participants, the manner in which stimuli are presented, the interface with which the participants work, and the instructions they are given. Because there are few native signers in the Netherlands, it was not possible to include only participants with this background in the experiments, nonnative and late signers were also used. Information about the specific choices made for each experiment can be found in the corresponding chapters.
Figure 1.1: Illustration of sign aspects. The following pairs of signs each differ in one of the aspects. (a) DAY and TO-LIVE (handshape) (b) TO-BAKE and BAKER (orientation) (c) FEELINGS and DRY-OFF (location) (d) MINUTE and THUNDERSTORM (movement) (e) FEELINGS and FRIDAY (non-manual component). The first four are not strictly minimal pairs, since they also differ in non-manual component: the Dutch word is mouthed along with the sign.

Images used with the permission of the Royal Effatha-Guyot Group (Kentalis). Images b, c, and e have been adapted by the author.
1.1 Background Information

Figure 1.2: Main phases of a sign in time. More detailed (sub-)phases are sometimes suggested [51], but these main phases are generally not disputed. Note that no non-manual component is performed. This is the case for all signs in our dataset.

Figure 1.3: (a) Basic pattern recognition problem. Using two features (right hand horizontal position and right hand vertical position), two signs must be distinguished. Based on the training examples of the two signs (colored circles and squares), a decision boundary (black line) can be created to distinguish between the signs ONE and SHOWER. An unknown test sign (star) can then be classified. In this case, the test sign would be classified as ‘ONE’. Note that many different decision boundaries are possible here, including ones that would classify the test sign as ‘SHOWER’. Finding the optimal classifier based on a set of training data is a fundamental pattern recognition issue. (b) When more classes are involved, a multi-class classifier is needed. In this example, the test sign (star) would be classified as ‘STEWARDESS’. (c) When the goal is to distinguish a single sign (STEWARDESS) from everything else, this means that the sign must be detected in a large, unspecified group of movements. In such a case, only the target class can be modeled properly. This is called one-class classification: the decision boundary is created based solely on the examples of the target class.
**1.1.2 Automatic Classification**

This thesis is about automatic recognition of isolated sign language signs, which is a classification problem. Basically speaking, automatic classification entails assigning an unknown object to a certain class based on certain properties of the object. Which properties to use and how to achieve the classification are subjects in this field.

Usually, a training set of objects of known class is used to train a classifier. Such a training set can be used to select informative features and to determine a decision function based on these features. For example, from the training set it can be learned that the color of a signer’s hair is not a useful feature for classifying a sign, but the position of his right hand is. Furthermore, from the training examples it can be calculated which values of right hand position indicate which class: e.g. a high hand position indicates HAT, a low hand position indicates BELT. Part of the training data (the test set) is held back and used to evaluate the trained classifier.

Figure 1.3 shows three types of classifiers. In the figure, two features are used to describe the signs. In reality, many more features are usually employed. See [27] for more information about automatic classification.

With regard to the work described in this thesis, it is important to be aware of the difference between the classifiers shown in figure 1.3 (b) and (c). Classifier (b) assigns a test object to one of the available classes (ONE, SHOWER, or STEWARDESS). Classifier (c) assigns a test object to the class STEWARDESS or NOT-STEWARDESS. This type of classifier can be referred to as a ‘detector’, since it detects the presence of the sign STEWARDESS. It is this type of classifier that is used in the current thesis. For each sign in the dataset, a one-class classifier is built in the manner of figure 1.3 (c). The decision function is determined based solely on the target class examples. At the same time, the other signs in the dataset are used to model a non-target class, which is employed in the feature selection step that precedes the classification step, and also in the evaluation of the classifier. In this sense, the classifiers used in the thesis are hybrids of figure 1.3 (a) and (c).

**1.1.3 Automatic Sign Language Recognition**

The process of automatic sign language recognition (ASLR) consists of two basic steps: information gathering and classifier training (see figure 1.4). Both elements can be implemented in various ways. Information gathering can be achieved by using sign videos as input and employing computer vision techniques to extract relevant information from the video. This is the technique we employ. An alternative is working with sensors that can directly measure relevant qualities such as the location and shape of the hand. Examples are magnetic positional trackers and instrumented gloves. Sensors of this kind often encumber the signer.
1.1 Background Information

Figure 1.4: Basic steps in vision-based ASLR. The first step is extracting information from video material. In this step, information about sign languages can be used to determine which characteristics should be extracted. The second step is training a classifier using the feature representations of the signs. This step is in most cases purely data-driven.

Once information about a sign has been gathered, it can be used to train a classifier. In principle, any pattern recognition technique can be used for this purpose. The choice of technique depends on the requirements of the system, computing resources, and the availability of training data. If a high level of accuracy is not necessary, a simple classifier will suffice. If only a limited amount of training examples is available, it is better to choose a technique that does not require the estimation of many parameters.

For many pattern recognition algorithms, it is necessary to align the signals that are to be compared. As part of the work described in this thesis, a variant of the Dynamic Time Warping (DTW) algorithm [54] was developed to align signs. This DTW algorithm is described in appendix B. The DTW alignment compensates for variations in signing speed.

Many different ASLR algorithms have been developed over the years, for various situations with diverse requirements and different datasets [64, 97]. This makes it difficult to compare the techniques. This thesis focuses on the possibilities of using knowledge about human sign language recognition for the benefit of ASLR, not on the merits of various ASLR algorithms. For this reason, we use a single recognition algorithm to test the possibilities of applying knowledge. For a global description of the algorithm, see appendix C.1. Information about the performance of this recognizer compared to others in the field can be found in [57].

The recognition algorithm used in this thesis [56, 57] was developed to be employed in an interactive learning environment for young, hearing-impaired children to practice sign language vocabulary (part of the ELo project [78]). As a consequence, the recognizer had to fulfill certain requirements. These requirements are listed in table 1.3. The requirement of multiple users is an important one, since many current ASLR systems can handle intra-signer variation quite well (the variation displayed by a single signer), but have more difficulty handling inter-signer variation (variation across signers) [101]. To facilitate signer-independent recognition, a large number of different signers (75) was included in our dataset.

An second important requirement was that the recognizer should judge the correctness of a sign, rather than assign the most likely class to an unknown sign. This was because
Table 1.3: Requirements for the automatic recognizer used in this thesis. The requirements are a consequence of the application of the recognizer in an interactive vocabulary-learning environment for young, hearing-impaired children.

<table>
<thead>
<tr>
<th>Property</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>information gathering</td>
<td>video-based (non-intrusive)</td>
</tr>
<tr>
<td>response time</td>
<td>real-time</td>
</tr>
<tr>
<td>meant for multiple users?</td>
<td>yes</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>medium to large (100–200)</td>
</tr>
<tr>
<td>type of input</td>
<td>isolated signs</td>
</tr>
</tbody>
</table>

in the learning environment, a child would be requested to make a certain sign, and it was the task of the recognizer to judge whether the sign was indeed made correctly. As a consequence, the recognizer had to consist of a set of one-class classifiers, each of which could distinguish its target sign from everything else. For this reason, evaluations of the recognizer used in this thesis are always evaluations of a set of individual classifiers, one for each sign in the dataset.

1.2 Relevant Literature

Both the field of sign language (psycho-)linguistics and the field of automatic recognition of sign language are relatively new areas of research. Serious research on sign language recognition started after Stokoe published an influential paper arguing the status of sign languages as natural languages [80]. Automatic recognition of sign language started in the late 1980s [33, 82]. Early attempts at sign language recognition often used techniques employed in the field of speech recognition. In most recognition algorithms, only basic knowledge about sign language is used to decide which features to extract (step 1 in figure 1.4), e.g. in [13, 26]. The remainder of the process is usually entirely data-driven, exhibiting “difficulties with incorporating results from linguistic research into recognition systems” [28]. In part, this is due to the lack of extensive (psycho)linguistic information that could be applied: Jiang et al. [43] considered using knowledge about human signing movements to generate artificial training examples, but rejected the option because such information is not sufficiently available. Overviews of sign language recognition methods can be found in [64, 85, 97].

It is only recently that more efforts are made to integrate knowledge about sign languages into automatic sign language recognition. Derpanis et al. [24] developed a method for movement recognition based on Stokoe’s definition of movement primitives [81]. Von Agris et al. [96] and Zieren et al. [101] studied the variation in German sign language signs in order to create systems that could handle such variation. Bowden et al. [10] designed sign features and feature categories based on the linguistics of
1.3 Summary of Results

British sign language. Han et al. [36] integrated linguistic information about the structure of signs in their search for appropriate sub-word units in signs. Ding and Martinez [25] and Vogler and Metaxas [95] proposed a separate parallel handling of sign aspects based on phonological sign language models. In each of these cases, attempts are made to integrate existing (psycho)linguistic knowledge in automatic recognition.

This thesis proposes a closer integration between the study of sign languages and automatic recognition. Rather than relying on existing information alone, we propose an approach in which experiments into sign language recognition are performed in parallel with the development of a recognizer. In this way, experiments can be set up to address specific difficulties encountered by the automatic recognizer. Conversely, the automatic recognizer can be adapted to more fully utilize the information provided by such experiments, for example by adapting the architecture. The work in this thesis is a first step towards such an integrated approach.

1.3 Summary of Results

Part I of the thesis describes two experiments investigating the information content of sign elements. In chapter 2, the salience of the phases shown in figure 1.2 for recognition is tested [89]. It is shown that the central phase of a sign is most informative, often as informative as the entire sign. The other phases also contain useful information for recognition, especially the preparation and the transition from preparation to stroke. Chapter 3 presents experiments on the importance of the aspects handshape and hand orientation for recognition [88], two aspects that were not present as features in the original recognizer [59]. The aspect handshape proves more informative than hand orientation for human sign recognition, making this the most interesting feature to add to the recognizer. In general, the results of the experiments in Part I suggest there may be a correlation between the rarity of the aspect values of a sign and the ease with which a sign can be recognized. Because of the small vocabularies used, this correlation could not be confirmed statistically. The final chapter of Part I gives an overview of the information gained from human signer experiments [86].

In Part II, the information gathered in part I is used to alter the recognizer. In chapter 5, the effect of adding handshape to the features is investigated [87]. Several handshape representations are tested. Most representations give a slight improvement in recognition performance, but one causes the performance to decrease. This may be due to the fact that this representation does not contain hand orientation information, while the others do.

Chapter 6 describes the effect of using only part of a sign as input for the recognizer [90]. Using the same test fragments as in the human signer test, it is shown that for the recognizer, too, the stroke is the most informative part. The general trends are the same
between human signers and the recognizer, but there are small differences. In the last chapter, experiments using even more radical reductions on the number of input frames are described. Using only a few key frames to represent a sign causes a decrease of performance (which also happens for human signers [66]). More sophisticated techniques for choosing key frames and combining their information may improve the results of this approach.
Part I

HUMAN SIGN LANGUAGE RECOGNITION
Abstract

In sign language studies, it is generally assumed that a sign can be divided into several phases in time: preparation, stroke and retraction, and that the stroke contains all information. However, this has not been tested empirically. We present an experiment that investigates the distribution of information in a sign, to find out where the information truly resides. Signers were shown isolated fragments of Dutch sign language signs (citation form). They were then asked to identify the sign. The results show that the stroke alone performs as well as the entire sign. However, the preparation together with the transition from preparation to stroke gives equally good recognition, suggesting that most of a sign’s information is available early. Surprisingly, in many cases preparation alone and retraction alone also give quite good recognition (66% resp. 60%). The recognition pattern across signs gives an indication of the recognition strategy of signers.

The material in this chapter was previously published as:
2.1 Introduction

The study of signs and gestures is relatively young compared to that of spoken language. The fact that sign languages are fully fledged languages and not limited systems of pantomime was acknowledged less than fifty years ago, when the work of [80] first described the linguistic structure of American Sign Language. Since then, much knowledge has been gained on the phonology, morphology, and cognitive processes of visual languages [29, 52]. But there is no clear consensus yet on many important issues [91]. There are various theories on sign language phonology [12, 21, 22, 60, 71, 80, 92], the structure of sign and gesture phrases [47, 51, 62] and sign language processing [11, 30, 39].

It is generally assumed that a sign or gesture can be divided into different fragments in time, often called (in chronological order) preparation, stroke, and retraction. These terms originate in the study of co-speech gestures, but are used for sign language signs as well [51]. Both in sign and gesture studies, it is always assumed that the stroke is the important part (lexical or meaning-bearing part) [48, 62, 100]. This claim is generally not disputed, though there is in fact, to our knowledge, no empirical evidence to support it. Since one cannot make a sign or gesture without performing some sort of preparation and retraction as well, it is not obvious that the latter phases are unimportant in the recognition process. There have been sign recognition studies using the gating paradigm [19, 30, 35]. These studies showed that subjects can identify signs based on the first part. [35] reported that from the start of the sign on 47% is sufficient. [19] showed that with context, only the first 37% of the sign is needed. [30] found that without context, the first 34% of the entire sign is enough for recognition. However, they used a different definition of the start of the sign than the other two studies did. These experiments suggest that the stroke is not the only phase in a sign containing information useful for recognition.

In this paper, we present an experiment that investigates the distribution of information over a sign in time. We want to test the ‘meaning-bearing’ quality of various time fragments of a sign in an empirical manner. To achieve this, we tested lexical recognition of sign language signs based on isolated fragments of the signs. That is, subjects were shown only a fragment of a sign, for example the preparation, and were asked whether they could recognize the sign. Special attention was paid to the transition between preparation and the stroke, because the gating experiments mentioned above suggest that at the start of the stroke, much of the information necessary to identify a sign is already available.
2.2 Method

2.2.1 Material

As stimuli, we used segments (in time) of movies of signs from Sign Language of the Netherlands (SLN). We will first describe the sign movies, then explain the nature of the segments that were extracted from them.

Signs

We used twenty different signs in the experiment. The signs were taken from the standard SLN lexicon. No compound signs were used. Fifty-two sign movies were made with this set. Each movie showed a signer making a sign from the set once. The signer started in a certain rest position, and returned to the same position after the sign. Several signs were recorded multiple times with different rest positions. The standard rest position was two hands resting on the table.

There were four two-handed signs in the set of twenty. These were recorded with the standard rest position only. The two-handed signs were:

- AUTO - ‘car’
- PAARD - ‘horse’
- SOEP - ‘soup’
- TELEVISIE - ‘television’

The other sixteen signs in the set were one-handed. These were recorded three times each, with different rest positions. The rest positions to choose from were: the standard rest position, a position with the dominant hand on the face, one with the dominant hand on the body, and one with the dominant hand in the neutral space (the dominant hand is a person’s preferred hand; for most people, this is the right hand). The non-dominant hand was always on the table. The signs were each recorded with the standard rest position, with a rest position that resembled the location where the sign took place, and with a rest position that was distinct from the sign’s location.

For example, a one-handed sign with location ‘chin’ would be recorded with the standard rest position, with ‘one hand on the face’ as the resembling rest position, and with ‘one hand in the neutral space’ as the distinct rest position. The variety in rest positions was included to create different preparations and retractions for the same signs. The sixteen one-handed signs were:
In summary: there were three movies for each one-handed sign, one for each rest position it was recorded with. For sixteen signs, this gave a total of 48 movies. Together with the four movies of the two-handed signs, this yields a total of 52 sign movies. We shall refer to the movies as 'sign productions'. From these sign productions, segments were taken.

Recording

The signs were all made by the same signer, a right-handed female sign language instructor (late signer). She was seated at a table. The view was limited to her upper body and the table top. The background was blue, the tablecloth was white and the signer’s clothing was red, providing good contrast. Lighting was diffuse, so that there were no sharp shadows. The movies were recorded with one digital camera (frontal view) at 25 fps, resolution 720 x 576.

The signs were recorded without any non-manual component (the part of the sign not executed by the hand — e.g. an obligatory facial expression). The signer was instructed to keep a neutral expression at all times. Many SLN signs have a mouthed Dutch word as non-manual component. In this experiment we wanted to ensure that subjects watching the signs achieved recognition based on the hand actions they saw, not from lip-reading the mouthed Dutch word. We checked that none of the signs in the set belonged to a minimal pair which could only be disambiguated with the aid of the non-manual component. However, TELEVISIE and BROER do form minimal pairs with old (no longer used) SLN signs.

Segments

In general, a sign can be divided in time into several phases. Sequentially, the main ones are preparation, followed by stroke, followed by retraction.

- Preparation - the phase where the hands move from rest position to location, handshape and orientation of the sign. The start of the preparation is characterized as
2.2 Method

Figure 2.1: Components of a movement as defined by [51]. Phases between parentheses are optional. An expressive phase can either be a stroke with possible pre- and post-stroke holds, or an independent hold.

- Stroke - the phase where the hands perform the sign. The start of the stroke is characterized as the moment the hand changes direction and/or velocity and/or the handshape is fully formed.
- Retraction - the phase where the hands move from location, handshape and orientation of the sign back to the rest position. The start of the retraction is characterized by relaxation of the hand.

Note that these terms are used as defined in [51], which is different from their use in general sign phonology literature. In continuous signing, the hands may not always start from and return to a rest position, and then some phases can disappear or merge. However, in this experiment, we used only isolated signs. The signer always started from a rest position and returned to this position after the sign had been performed.

More recent research [51, 62] distinguishes more detailed phases of a movement, such as dependent holds. [51] also state that a sign or gesture need not contain a movement. They propose the term expressive phase, and state that this phase either consists of the stroke plus pre- and post-stroke holds, or of an independent hold, a hold that is expressive all by itself. Figure 2.1 illustrates their theory. Parentheses mean that the phase is not compulsory. In our isolated signs, however, there is always a preparation and a retraction, though they may be short.

In our experiment, we focused on the three main phases of a sign: the preparation, the stroke and the retraction. As stroke we took what [51] call the expressive phase, that is, either the stroke plus pre- and post-stroke holds, or an independent hold. We also used their guidelines for determining the phases. Guidelines were based on visual properties (changes in velocity, direction of movement, etc). The preparation was defined as the time span between the moment the hands started moving from the rest position and the start of the stroke. The retraction was defined as the time span between the moment the hands started relaxing after the stroke and the moment they were back in the rest.
position.

First of all, we had to determine the borders of the three phases in our sign movies. To determine them as objectively as possible, we had three independent persons annotate them for all 52 sign movies.

The annotators had experience in working with sign language and sign phases, and used common definitions of the phase borders, based on kinematic properties. Still, they usually differed slightly in their exact placement of the borders. Their annotations are referred to as:

- **earliest** - the earliest time any annotator gave
- **latest** - the latest time any annotator gave
- **inbetween** - the time of the annotator in between

We are not only interested in the information content of preparation, stroke, and retraction, but also in the information contained in the transition from preparation to stroke. Here, transition means the time in which the movement goes from phase ‘preparation’ to phase ‘stroke’. In defining this transition, we utilized the amount of disagreement between border placements (difference between earliest and latest). We took this amount as an indication of the vagueness of the transition. Short, sharp transitions cause relatively small differences in annotation, while long, gradual transitions cause relatively large differences in annotation. For this reason, we used the border values in the definition of transition, interpreting the amount of disagreement as a measure of transition length. As a consequence, not all transitions had the same length.

The **transition**, then, was defined as the period from the earliest to the latest border between preparation and stroke, enlarged by 2 frames on either side (see figure 2.2). The two-frame margin is taken to ensure that the transition has a minimum length of 4 frames, and to ensure that preparation and stroke contain no trace of the transition. The average length of the transition was 6 frames ($SD = 1$). This was too short to use as a separate segment. Instead, we extracted preparation and stroke both with and without the transition included. The difference in result would give an indication of how informative the transition was.

The segments used in the experiment were thus defined as follows:

- **preparation (P)** - from the earliest start of the preparation, to the earliest start of the stroke minus 2 frames
- **stroke (S)** - from the latest start of the stroke plus 2 frames, to the earliest start of the retraction
- **retraction (R)** - from the latest start of the retraction, to the latest end of the retraction
Figure 2.2: Segments extracted from each sign movie. The borders between the phases were determined by three annotators, who often gave slightly different values (earliest, latest, inbetween). Their measurements are used in the definitions of the segments used in the experiment. A 2-frame margin is taken around the borders of the transition preparation-stroke, to ensure a minimum transition length of 4 frames.

- preparation + transition (P+) - from the earliest start of the preparation, to the latest start of the stroke plus 2 frames
- transition + stroke (S+) - from the earliest start of the stroke minus 2 frames, to the earliest start of the retraction

The segment definitions are illustrated in figure 2.2. From each sign production all five segments were extracted and used as an experimental condition. Moreover, the entire sign movie was also used, as a control condition. This yielded six segments for 52 sign movies, a total of 312 segments.

2.2.2 Stimuli

All stimuli were presented full screen on a 17 inch LCD computer screen. A stimulus was made from a segment by putting it three times in a row, with a second of gray in between. There was no gray before the first segment or after the last. The reason for the repetition was that some segments were quite short. Showing the segment several times compensated for this without making the stimulus unnatural by slowing it down.

Each subject was presented with all 312 stimuli. The stimuli were presented in a random order, which was different for each subject. Each subject was shown each stimulus only
once. The exception to the random order was the set of 52 stimuli that contained the entire sign movies. Because these were meant as a control condition (does the subject know the sign at all?), they were presented last. Within this group of entire signs, the order was again random.

2.2.3 Task

After a subject was shown a stimulus (i.e. a certain sign segment three times), the subject had to indicate whether s/he had recognized the sign or not by clicking a ‘yes’ or a ‘no’ button with the mouse. If the subject clicked ‘no’, s/he was immediately shown the next stimulus, without any button-clicks. If s/he clicked ‘yes’, s/he was presented a multiple choice list of 10 glosses, and asked to select the correct gloss by clicking with the mouse. There was also an option ‘Sign is not in list’, which the subject could choose if the sign s/he thought s/he had recognized was not in the list. (The correct sign was always present in the list, but of course, it was possible that a subject had made a mistake and thought s/he had seen a different sign which was not in the list.) After selecting a sign, or the ‘Sign is not in list’ option, the subject had to confirm her/his choice by clicking on a ‘Go ahead’ button. This immediately caused the next stimulus to be shown. Halfway through the complete task, there was a short break. For each action (clicking ‘yes/no’, selecting the answer and clicking ‘Go ahead’) the subject could take as much time as s/he wanted.

The glosses that served as alternatives in the multiple choice list were chosen as follows: first of all, a set of 40 extra signs was constructed which were chosen in such a way that for each of the 20 test signs, there was one sign that resembled it in handshape and another one that resembled it in either motion or location. These 40 signs, together with the 20 original test signs, formed the alternatives pool. For each choice that a subject had to make, the multiple choice list was constructed on the spot by pulling 9 random signs from the alternatives pool (checking that the correct answer was not among them), and inserting the correct answer at a random place. The limited number of alternatives ensured that a test sign would not occur more frequently than an alternative, which could be a hint for the correct answer.

All subjects received instructions to the task beforehand. The instructions were available both as text and as an SLN movie. After instruction, subjects could try the task briefly with 10 practice stimuli, drawn from signs that were not used in the actual experiment. If all went well, they could start the experiment.

In the instruction, subjects were informed that the movies they would see were always of an SLN sign, and that there were no compound signs in the set. They were also told that the correct sign was always present in the multiple choice list. Instruction stressed the fact that subjects should not guess the meaning of a sign, but only answer ‘yes’ when they could identify the sign with confidence.
The experiments were performed at different locations; at subjects' homes or offices, or at schools. The test program ran on a laptop, which was set up in a quiet place at the test location.

### 2.2.4 Subjects

Twenty-seven persons participated in the experiment. All of them were experienced SLN users. Thirteen were male, fourteen female. Their ages ranged from 19 to 67 years, average 31 ($SD = 11$). Eighteen were deaf, nine were hearing. Of the deaf, five had deaf parents, the other thirteen had hearing parents. Two hearing subjects had deaf parents, the other seven had hearing parents. Fifteen subjects lived in the northern region of the country, twelve lived in the western region (there are minor dialectic differences in SLN between regions).

9 subjects used SLN in their profession. 4 subjects indicated their use of SLN as 'always', 13 as 'daily', 9 as 'regularly' and 1 as 'sometimes'. The age to start using SLN ranged from 0 to 24 years, with an average of 9 ($SD = 9$).

### 2.3 Results

In this section, we present the results of the recognition experiment for the segments preparation (P), retraction (R), preparation + transition (P+), stroke (S) and stroke + transition (S+). The sixth segment in the experiment, the entire sign, is not presented. This segment was used as a comprehensibility test. If a person cannot recognize a certain sign when viewing the entire sign production, then it cannot be expected that he recognizes any segment from this sign production. Therefore, all segments that were extracted from this sign production were then removed from that person's results. The data presented here are from this corrected dataset. The total number of removed sign productions was 142 (of $27 \times 52 = 1404$). Most often (127 cases), the person did not recognize any production of a certain sign. Sometimes (15 cases) one production was recognized and another was not. In those cases, the sign was often mistaken for another sign (as opposed to not recognized at all).

#### 2.3.1 Recognition Overall

We calculated the recognition rates per sign per segment (taking sign productions of the same sign together). Figure 2.3 shows the distribution of recognition rates over the signs. It can be seen that recognition rates for P+, S, and S+ are near 100%. Values for P and R are clearly lower, though even for these segments, there are examples where recognition approaches 100%. Variation for P and R is much larger than for the other
Figure 2.3: Box plot of the medians and quartiles of the recognition rates per segment over all signs. R = retraction; P = preparation; P+ = preparation + transition; S = stroke; S+ = stroke + transition. In S+, there was 1 outlier, which is represented by ‘+’ (outliers are points that lie more than 1.5 times the interquartile range away from the first or third quartile). Number of subjects = 27, measurements per segment = 1262
three segments. Apparently, for P and R, recognition is sometimes possible, sometimes not. The more narrow distributions of P+, S, and S+ suggest that a sign can always be well recognized based on these segments.

A post-hoc test confirmed that our data can be separated into two groups, one formed by the segments P and R, and one formed by P+, S and S+. We will indicate these groups as ‘information-poor’ and ‘information-rich’ segments respectively.

### 2.3.2 Recognition per Sign

The variation in recognition rate between signs for the information-poor segments R and P compelled us to look at the recognition per sign. In figure 2.4, the recognition rates over all subjects are shown for each sign separately (again, sign productions of the same sign are taken together). The bars indicate recognition based on the five segments R, P, P+, S and S+. All signs can be recognized quite well by P+, S and S+. Some signs can even be recognized well by R and P (notably those in the last row of figure 2.4). In fact, for some signs there may be no significant difference between the recognition rates of information-poor and information-rich segments. We used the McNemar test for two related samples [76] to test for significant differences between various segments. In table 2.1, we listed which signs showed significant difference in recognition rate for which segments.

From the results in table 2.1, we see that for most signs, recognition rate of S is significantly better than that of P and R. Also, recognition rate of P+ is significantly better than that of P for most signs, and in many cases does not differ from the recognition rate of S. But there are exceptions to each of these general observations. Examination of the error data (the incorrect glosses chosen by subject) showed that in many cases subjects picked an answer sign which had one or two characteristics in common with the target sign (e.g. the same handshape and orientation). However, no characteristic or set of characteristics was exceptionally often equal.

### 2.3.3 Learning Effect

Even though we have dubbed them ‘information-poor’, the recognition rate of the segments P and R is quite high (average 66% and 60% respectively), as can be seen in figures 2.3 and 2.4. This could be due to an unwanted learning effect during the experiment. After all, subjects saw a total of 312 segments from a set of movies containing only 20 different signs. Subjects could have learned which signs were present in the dataset, which would make recognition easier. To investigate possible learning, we grouped the segments that subjects had seen into groups of twenty according to the serial number of showing. That is, for each subject, the first twenty segments s/he saw were placed in group one, the next twenty in group two, etcetera, regardless of the type of segment.
Figure 2.4: Recognition percentages per sign. The bars indicate recognition rates based on the various segments that were extracted from the signs, averaged over all subjects and rest positions.
Table 2.1: Overview of significance of differences between pairs of segments per sign. The leftmost column shows the segments that were compared. The next three columns indicate levels of significance of difference (over 99%, between 99% and 95%, less than or equal to 95%). Each sign of the set of 20 is listed in the appropriate column. ‘all other signs’ means that all signs not listed elsewhere along the row belonged in this column. The number between parentheses indicates their number.

All differences were tested using the McNemar test for two related samples. When there were not enough occurrences of a different response for the two segments compared, the binomial test was used. In some cases, there were not enough different responses even for this test (fewer than 5). These cases are listed in the rightmost column. They were all cases where recognition was near 100% for both segments.

<table>
<thead>
<tr>
<th>Segments compared</th>
<th>Significant difference ( p &lt; .01 )</th>
<th>Significant difference (.01 \leq p &lt; .05)</th>
<th>No significant difference ( p \geq .05 )</th>
<th>No test possible (fewer than 5 occurrences of a different response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>preparation &amp; stroke</td>
<td>all other signs (15)</td>
<td>KOORTS PAARD SOEP</td>
<td>TELEFOON KIJKEN</td>
<td></td>
</tr>
<tr>
<td>retraction &amp; stroke</td>
<td>all other signs (16)</td>
<td>KOORTS</td>
<td>BROER VIES TELEFOON</td>
<td></td>
</tr>
<tr>
<td>preparation &amp; preparation+</td>
<td>all other signs (16)</td>
<td>PAARD SOEP</td>
<td>KIJKEN TELEFOON</td>
<td></td>
</tr>
<tr>
<td>stroke &amp; preparation+</td>
<td>MIS TEKENEN ZAND</td>
<td>BAD BROER EUROPA KIJKEN KOORTS PAPA SCHEP</td>
<td>all other signs (10)</td>
<td></td>
</tr>
<tr>
<td>stroke &amp; stroke+</td>
<td>BROER MIS TEKENEN</td>
<td>BAD EUROPA KOORTS SCHEP ZAND</td>
<td>all other signs (12)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.5: Recognition percentages according to serial number of presentation. The data points represent recognition rates of the various segments as calculated from the first 20 segments presented to each test subject, the second 20, etcetera, over all subjects. The first group contains the stimuli presented as nos 1-20, the second group those presented as nos 21-40, etcetera. The last 52 stimuli presented were always the entire sign segments. These are not shown. Because the order of presentation was random, the precise stimuli in the groups were different for each subject.
Then, for all subjects together, the recognition rates within the groups were determined. That is, for all preparation-segments in group one for all subjects, what was the percentage that was recognized correctly? And for the retraction-segments in group one? And so forth. If it was indeed a learning effect that was responsible for the high recognition rate of some segments, one would expect low recognition rates for the first groups and a steady rise of the recognition rates with increasing group number.

The result of the analysis is shown in figure 2.5. First, even though the recognition rates go up during the trial, the recognition rate of P and R is already around 40% at the start and reaches 60% quickly. Secondly, the difference in recognition between the information-poor and information-rich segments is present right from the start. P+ is already distinct from P and close to S in the first group of 20. In general, the segments seem to reach an equilibrium after 100 fragments. Note that the entire sign movie was never present in these groups, since these were always shown at the end of the experiment (they account for the fragments with serial numbers 261-312, groups 14-16, which are not shown).

2.3.4 Influence of Segment Length

In our experiment, we wanted to study the information content of different fragments of a sign. However, there is a confounding variable present: the lengths of the segments. Different segments have different lengths as well as different contents. At a first glance, the length of a segment does not seem to play a significant role in the recognition of segments. But we have not researched the influence of segment length systematically. A few observations, however, show the following: on average, R is longer than P (13.1 frames vs 6.2 frames), but recognition is about the same. P+ is about as long as R (12.1 and 13.1 frames respectively), but recognition of the first is much better (90% vs 60% average over all sign productions and subjects). In fact, S is not much longer than R on average (15.3 frames), but recognition between these segments differs significantly. Also, we plotted the length of each segment versus the recognition rate of that segment over all sign productions and persons. If segments had the same length, they were taken together. Figure 2.6 shows the plots per segment type. There does not appear to be any clear dependency of recognition rate on length.

2.3.5 Influence of Subject Background

Over all sign productions and segments taken together, deaf subjects showed slightly better recognition than hearing subjects. Closer inspection showed that the difference was due to better performance by the deaf subjects on segments R and P (see figure 2.7). A $\chi^2$-test for nominal data [76] indicated that the difference was significant for both R ($\chi^2 = 44.3, p < .001$) and P ($\chi^2 = 83.8, p < .001$), as well as for recognition over
Figure 2.6: Plots of segment length vs recognition rates for different types of segment over all sign productions and subjects. No clear dependency of recognition rate on segment length appears present. The segment S+ is not shown, it presents a pattern almost identical to that of S.
2.3 RESULTS

Figure 2.7: Recognition of the various segments by deaf and hearing subjects. Number of subjects: deaf = 18, hearing = 9

all sign productions and segments ($\chi^2 = 88.4, p < .001$). There was no difference in recognition of S ($\chi^2 = 0.30, p < .70$) or S+ ($\chi^2 = 2.3, p < .20$). A test for P+ showed a possible significant difference ($\chi^2 = 5.7, p < .02$). It is conceivable that the deaf subjects were more proficient in SLN than the hearing, and thus capable of recognizing a sign on the basis of very little information. We examined several characteristics that can indicate proficiency, such as age to start learning SLN, hearing status of parent, frequency of use and self-assessment of signing competence. However, when subjects were grouped according to these characteristics, no statistical differences were found between the recognition scores of the groups. It is possible, though, that greater proficiency still is the reason for the higher scores of the deaf, but if so, then this proficiency is not reflected directly in any of the characteristics mentioned.

We also looked at other subject characteristics: age of the subjects, gender, and the region of the country they were from. No statistical differences were found between the recognition scores of any of these groups.

2.3.6 Influence of Different Rest Positions

The sixteen one-handed signs in the set were recorded with different rest positions, namely a neutral rest position, a position resembling the sign’s own location, and a
position distinct from it. This created preparations and retractions of different length and nature. Could the difference in recognition of P and R between deaf and hearing subjects be due to this difference? For instance, could deaf subjects be more adept at identifying signs when the rest position resembles the sign’s own location? We calculated recognition rates of R and P for each type of rest position for deaf and hearing subjects over all sign productions. The results are shown in figure 2.8. Clearly, recognition hardly varies between rest positions, both for deaf and for hearing subjects. The deaf subjects score higher for all rest positions, there are no rest positions for which they score especially well compared to hearing subjects.

Statistical tests were performed for all segments. No significant differences were found between any of the rest positions for any segment for either deaf or hearing, with the exception of the segment P+ in the hearing group, where a slight difference was found ($\chi^2 = 4.8, p < .05$) in favor of the neutral rest position. It appears that the type of rest position has little influence on recognition success.
2.4 Discussion

2.4.1 Main Conclusions

This experiment was conducted to learn more about the information content of the phases of a sign. We found that the stroke of a sign contains enough information to identify the sign, which has always been assumed, but had not yet been proven by experiment. However, we also found that other phases of a sign are informative as well. Recognition based on the information-poor segments P and R was much better than chance, with an average of 60% for R and 66% for P over all signs and subjects. Figure 2.4 shows that there are even signs, such as TELEFOON, for which all segments show near-perfect recognition. Thus although it appears true that the stroke contains enough useful information to identify a sign, P and R also contain information, enough to allow signers to identify signs in many cases. This is the case immediately from the start of a trial, before a subject has the opportunity to perhaps start learning the signs in the dataset. Apparently, no convincing learning effect is involved. It is interesting to note that P gave better recognition than R. This may have to do with the fact that in normal conversation, one often identifies a sign based on the beginning, but not based on the end (because then one has already seen the stroke).

The transition from preparation to stroke turned out to be very informative. P+ behaves similar to S and S+, whereas P gives lower and more variable recognition rates similar to those of R (see figure 2.3). Also, from table 2.1 we see that P+ gave significantly better recognition than P for 18 out of 19 signs (for 1 sign, calculation of significance was not possible). And recognition of P+ did not differ significantly from that of S for 7 out of 10 signs (for 10 signs, calculation of significance was not possible). This suggests that the transition from preparation to stroke often contains enough information to identify a sign. The stroke itself performs about equally well with or without the transition, suggesting that the part of the stroke after the transition contains about the same amount of useful information as the transition. These findings concur with the results of earlier gating experiments [19, 30, 35], which showed that the first 34%-50% of a sign (probably corresponding to preparation plus a bit of the stroke) are sufficient for recognition.

We can interpret these findings in the context of sign linguistic theory [29, 73, 80]. First, the stroke contains all information about the meaning of the sign because it is the ‘core’ of the sign, and preparation and retraction are physical necessities, needed to bring the hand(s) into resp. out of the required position, orientation and shape. Secondly, much of the information that is needed to identify a sign is available early in the sign. The parameters of a sign that carry meaning are the location, handshape, hand orientation, hand motion, and the non-manual component. During the preparation, the hands are taking on the correct shape and orientation. When the transition to the start of the stroke is reached, almost all parameters necessary to determine a sign’s meaning are already available: handshape, orientation and location can be observed. When the direction
of motion becomes clear, at the very beginning of the stroke, this is often enough to predict the rest of the motion and identify the sign. The transition, especially with the expansion of 2 frames in our experiment, probably shows the direction the motion will take, thus making it possible to recognize a sign immediately after the transition. This theory also offers a clue as to why recognition of preparation and retraction was so often successful. During preparation, the hands already start assuming the required shape and orientation, and they are traveling in the direction of the required location. Three of the five sign parameters are already present to some extent. The same is true for the retraction: here, hand configuration and location are present at the start and disappear while the hands relax from the sign to the rest position. Again, location, handshape and orientation can be inferred. In many cases, the partial information is enough to allow identification of the sign. This also explains why the different rest positions did not seem to influence recognition of P and R: the important part of these segments is the end resp. start, where the sign parameters are most developed. This is the time furthest from the moment where the hands are in the rest position, so the difference due to the varying rest positions will be minimal.

2.4.2 The Recognition Process

In this investigation, we aimed at gaining insight into sign recognition. Are there factors that facilitate recognition, especially recognition of information-poor segments? Are certain signs more easily recognized than others, and can we find regularities in their characteristics that can explain this? Images of the signs, as well as a short description, are added as an appendix.

**Handshape** Figure 2.4 shows that some signs can be recognized well based on P or R alone (the signs in the bottom row), while others are poorly recognized by these segments (the signs in the top row). We refer to these signs as ‘easy’ resp. ‘difficult’ signs. All ‘easy’ signs have difficult handshapes for which a subset of the fingers must be extended. In contrast, the ‘difficult’ signs — signs for which P and R do not score well — have unpronounced handshapes, such as fists or flat hands. By unpronounced we mean handshapes that can be made without much effort or control [1]. Difficult handshapes are handshapes which require more precise motor control. Perhaps for this reason, the unpronounced handshapes are very common in SLN [73]. The difficult signs all have such common handshapes. Only TELEVISIE and TEKENEN have somewhat pronounced shapes. For TEKENEN this is the ‘U’-handshape: first two fingers extended, touching. TELEVISIE has a shape with thumb and first two fingers open, but curled, not fully extended. The lack of extension could make this handshape less distinct than for example that of WC (an easily recognized sign in which thumb and first two fingers are extended).
Location For location, a similar trend can be seen: all signs that are relatively poorly recognized from P or R alone (signs in the top row of figure 2.4) are located in the neutral space (from SCHEP up to and including AUTO). The other signs are all located on the body or the face, except for PAARD (which is in the neutral space).

Motion As for motion: many signs that are poorly recognized from P or R contain a path movement, where the hand(s) move from one point to a distinctly different point in space over the course of a sign (location change). Most of the signs that are well recognized even from P or R contain only hand-internal movement — a change in handshape or orientation. Exceptions are KIJKEN and MIS, ‘easy’ signs which contain a path, and TELEVISIE, a ‘difficult’ sign which does not. It is possible that a sign containing a path movement is more difficult to recognize because two locations are needed to accurately identify it. Here, solely the beginning (preparation) nor solely the end (retraction) are sufficient to infer the sign’s motion. On the other hand, in signs with only hand-internal movement, the motion parameter is essentially the same at the start and at the end, making it easier to identify the sign when only one of these moments is accessible.

Orientation For orientation, no relationship between certain values and ease of recognition could be observed.

In general, it appears that the rarity of a feature value plays a role in the ease with which a sign can be recognized. There is no extensive data on frequency of feature values in SLN available. However, some information can be found in [92] and [73]. Signs that are processed well based on only the information-poor segments tend to have relatively rare handshapes and locations. The amount of change in parameters during the sign also appears to have influence: to identify signs in which the location changes, information about only the start or only the end of a sign does not seem to be sufficient. This is less problematic for signs with a rare handshape and/or location.

We speculate that in the recognition process, several sign parameters are processed in parallel. Together, they activate a group of possible candidates that shrinks as more information becomes available, similar to the cohort model of [61]. Rare handshapes and locations narrow down the possibilities quickly, which is why signs with these parameters are well recognized even from the information-poor segments. This concurs with the findings of [30]. They conducted a gating experiment in which subjects saw from the start on increasingly longer fragments of a sign, and were asked each time to name all signs they considered as possible identifications. They found that subjects listed a decreasing amount of possibilities when more information became available. In particular, some sign parameters (e.g. location) seemed to be resolved at a certain point, after which the subject would only lists signs with the same value for that feature (e.g. only signs with location ‘chin’). One can imagine that a rare feature value quickly brings down the number of candidates, after which little extra information is needed for a pos-
itive identification. A path movement makes recognition more difficult because both start and end location must be available before the path feature can be inferred. Signs with rare handshape or location suffer less from this difficulty, perhaps because the path feature is not always needed to positively identify the sign.

However, the set of data from which these observations stem is small, containing only 20 signs. Furthermore, many signs contain combinations of the factors that possibly facilitate recognition. Most signs that have an unpronounced handshape also have a common location. And finally, the non-manual component was excluded in this experiment. It is therefore not possible to accurately separate the effects various parameters and their values have on the recognition process based on our current data. More experiments, with a larger dataset and set up to separate the various factors, are necessary.

Acknowledgments

This research was partly funded by the VSB-fund, grant number 2003451, project “Een elektronische leeromgeving voor jonge kinderen met auditieve en communicatieve beperkingen voor het leren van Nederlandse Gebarentaal (ELo).”

The authors would like to thank Marielle Elzenaar and Jeroen Arendsen for their efforts in coding the movement phases, and Arend Harteveld for creating the interface used in the experiment.
Appendix 2.A: Description of Sign Dataset

Table 2.2 contains descriptions of the 20 signs that were used in the experiment. These sign descriptions are meant to give the reader a picture of the sign. They are not extensive phonological descriptions. The information was gathered from the digital SLN dictionary [34] and from signers.

Table 2.2: Sign descriptions

<table>
<thead>
<tr>
<th>Sign</th>
<th>SLN Hand-shape</th>
<th>Location</th>
<th>Orientation</th>
<th>Motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFDROGEN</td>
<td>5</td>
<td>shoulder, touching</td>
<td>palm toward signer</td>
<td>circular counterclockwise</td>
</tr>
<tr>
<td>to dry off</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTO</td>
<td>both hands S</td>
<td>neutral space</td>
<td>palms facing each other</td>
<td>up and down in alteration</td>
</tr>
<tr>
<td>BAD</td>
<td>AS</td>
<td>chest, touching</td>
<td>palm toward signer</td>
<td>up and down</td>
</tr>
<tr>
<td>BROER</td>
<td>V</td>
<td>front of shoulder, touching</td>
<td>palm down</td>
<td>tapping the body</td>
</tr>
<tr>
<td>EUROPA</td>
<td>E</td>
<td>neutral space</td>
<td>palm forward</td>
<td>horizontal arch forward</td>
</tr>
<tr>
<td>KIJKEN</td>
<td>V</td>
<td>in front of eyes</td>
<td>palm toward signer</td>
<td>forward</td>
</tr>
<tr>
<td>KIP</td>
<td>baby-snavel</td>
<td>in front of nose</td>
<td>palm forward</td>
<td>thumb and index finger open and close</td>
</tr>
<tr>
<td>KOORTS</td>
<td>B</td>
<td>cheek, touching</td>
<td>palm toward cheek</td>
<td>no motion</td>
</tr>
<tr>
<td>MAMA</td>
<td>B</td>
<td>cheek, touching</td>
<td>palm toward cheek</td>
<td>tapping the cheek</td>
</tr>
<tr>
<td>MIS</td>
<td>Y</td>
<td>in front of nose</td>
<td>palm sideways</td>
<td>forward</td>
</tr>
<tr>
<td>PAARD</td>
<td>both hands 1</td>
<td>neutral space</td>
<td>palms near facing each other</td>
<td>one finger tapping on the other from above</td>
</tr>
<tr>
<td>PAPA</td>
<td>1</td>
<td>chin, touching</td>
<td>palm toward signer</td>
<td>tapping the chin from the front</td>
</tr>
<tr>
<td>SCHEP</td>
<td>geld</td>
<td>neutral space</td>
<td>palm down</td>
<td>vertical arch down-and-up-again</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Sign</th>
<th>SLN Handshape</th>
<th>Location</th>
<th>Orientation</th>
<th>Motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOEP</td>
<td>both hands B</td>
<td>neutral space</td>
<td>one palm down, one facing the signer</td>
<td>dominant hand moves down along non-dominant arm</td>
</tr>
<tr>
<td>TEKENEN</td>
<td>U</td>
<td>neutral space</td>
<td>palm forward</td>
<td>zig zag downward</td>
</tr>
<tr>
<td>TELEFOON</td>
<td>Y</td>
<td>side of head</td>
<td>palm toward head</td>
<td>no motion</td>
</tr>
<tr>
<td>TELEVISIE</td>
<td>both hands X-tweenul</td>
<td>neutral space</td>
<td>palms facing each other</td>
<td>symmetric wrist rotation</td>
</tr>
<tr>
<td>VIES</td>
<td>Y</td>
<td>neutral space/side of neck</td>
<td>palm down</td>
<td>from neutral space to neck, thumb first</td>
</tr>
<tr>
<td>WC</td>
<td>W</td>
<td>neutral space</td>
<td>palm forward</td>
<td>fingers curl and extend</td>
</tr>
<tr>
<td>ZAND</td>
<td>geld</td>
<td>neutral space</td>
<td>palm down/sideways</td>
<td>upward</td>
</tr>
</tbody>
</table>

For examples of the handshapes mentioned, see [73]; the handshapes can also be found at research.tenholt.nl/handshapes.

Figure 2.9 shows depictions of the signs in the dataset. Perpendicular slashes through an arrow mean repeated movement. A cross means the hand touches the body. A cross followed by dots means the hand touches the body repeatedly. A square box means the hand closes. A square box followed by dots means the hand closes repeatedly. In sign ‘WC’ the hand repeatedly switches between the two handshapes shown, without other movement.
2.4 DISCUSSION

Figure 2.9: Depictions of the signs in the dataset
Signs in which Handshape and Hand Orientation are not or only partially visible: What is the Consequence for Lexical Recognition?

Abstract

We present an experiment about lexical recognition of human sign language signs in which the available perceptual information about handshape and hand orientation was manipulated. Stimuli were videos of signs from Sign Language of the Netherlands (SLN). The videos were processed to create four conditions: (1) one in which neither handshape nor hand orientation could be observed, (2) one in which hand orientation could be extracted but not handshape, (3) one in which an approximation of the handshape could be seen, and (4) one where the video was unmodified. In general, recognition of the signs was almost impossible in the first two conditions, while condition three showed a rise in recognition rate to ca 60%. However, some signs were recognized well even in conditions 1 and 2. Their success rate cannot be linked to a single sign property, but seems due to a combination of factors. In general, handshape information appears more salient for resolving the lexical meaning of a sign than hand orientation information.

The material in this chapter was previously published as:
**3.1 Introduction**

Research on the perception of sign languages is sparse compared to similar research for spoken languages. Since Stokoe [80] declared for the first time that sign languages are natural languages, there have been several studies on sign perception. But there are still many open questions about the processing of signs. We know from spoken language research that speech is highly redundant, and that this is necessary because the perceptual system tends to miss a lot of the presented information. Is this also the case for sign languages? Do signs contain the same high level of redundancy? And if so, are certain elements more important than others?

To answer such questions, we must gain insight into sign perception. A time-honoured method to do so is by selecting or removing parts of the stimulus material and observing the consequences for recognition. Several studies have been performed on sign languages using this method. Parish et al. [66] removed time fragments from a sign movie, reducing it to a set of keyframes. By choosing salient frames, the number of selected frames could be greatly reduced without affecting the intelligibility of the sign. ten Holt et al. [89] selected certain time fragments corresponding to phases within the sign and tested recognition based on these fragments alone. They found that various phases give accurate recognition, suggesting there is information redundancy in the time domain. Both Grosjean [35] and Emmorey and Corina [30] conducted gating experiments in which they tested how well participants could resolve a sign when they saw increasingly larger parts of the beginning of the sign. Grosjean reports that on average, participants could identify signs after seeing the first 51%. This pertains to the isolation point, the moment when a participant identifies the sign correctly and does not subsequently identify it differently. Emmorey and Corina [30] report isolation times of 34% of the entire sign, but their definition of the start of the sign differs from Grosjean’s definition. Emmorey and Corina [30] also report recognition times of 44% of the entire sign. The recognition point is defined as the moment a participant indicates an 80% confidence score for his/her identification of the sign (Grosjean does not report on the recognition point). In a recent study, Arendsen et al. [5] asked participants to watch a movie and respond as quickly as possible when they saw the beginning of a sign language sign. Participants needed only about a third of the sign to make this identification. These studies all suggest there is a reasonable amount of redundancy in signs in the time domain.

However, signs can be subdivided in time, but they can also be altered by removing or obscuring some of their formational parameters, i.e. movement, location and handshape [81], hand orientation [9], and the non-manual component. Though the organization of the parameters remains topic of discussion [29], it is generally agreed that these five parameters determine the meaning of a sign. The status of the non-manual component as a formational parameter remains unclear; see Emmorey [29] for a discussion. However, in Sign Language of the Netherlands (SLN), non-manual
components (more specifically *mouth patterns*) are considered to be a parameter with the same status as the other four.

It is remarkable that even though four out of five parameters have to do with the hands, studies indicate that an observer focuses mostly on the head and upper body of a signer [63]. The hands are apparently processed by the peripheral vision, which invokes the question of how much hand detail is actually perceived. [73].

It is possible to study the relative importance of the formational parameters by obscuring one or more of them and observing the effect on perception and recognition of signs. Poizner et al. [68] removed handshape, hand orientation and non-manual component by recording signs as point light displays. One of their experiments was a lexical recognition test resulting in a recognition accuracy of about 80%. However, in this test all signs had the same handshape, a handshape known to the participants. Tartter and Knowlton [84] investigated natural American Sign Language (ASL) conversation in point-light conditions, affixing twenty-seven lights to strategic body locations aimed at giving the observer at least some information on each of the four hand parameters. They found that natural conversation was possible under these conditions. In a subsequent study Tartter and Fischer [83] investigated the impact of each parameter separately in the same twenty-seven point-light condition. They tested how well signs that were identical except for the value of a single parameter could be distinguished, ensuring that context offered no clues. They found that signs only differing in handshape scored relatively poorly compared to signs that differed only in one of the other parameters. Apparently, handshape was not accurately perceived. Nevertheless, their previous study [84] had shown that natural conversation was possible. It seems that natural conversation does not require full access to the handshape parameter. All three studies indicate that recognition remains possible in situations where the representation of the signs has been severely diminished — when at least some information on each formational parameter remains.

The examples above show that sign languages contain a high degree of redundancy. Neither removal of time fragments nor removal/reduction of formational parameters renders a sign completely unintelligible. Signs sometimes remain remarkably understandable even when large elements are removed. This suggests that not every element of a sign is strictly necessary for recognition, and that some elements may be more informative than others (as is the case with various choices of keyframe in Parish et al. [66]). It is this difference in information content that we are concerned with in the current study. Focussing on this topic, we expect to gain a better understanding of sign language perception. This may result in possibilities to improve automatic sign language recognition systems, like the one proposed in [58, 78]. For optimal sign recognition, it is important to learn which elements of a sign contain essential information for lexical recognition.

In our experiment, we investigate the importance of the formational parameters handshape and hand orientation (sometimes considered a single parameter *hand*
configuration) for lexical recognition by removing this information in varying degrees from a movie of a sign and observing the effect on lexical recognition performance. There are four conditions in which a movie is shown. In the first condition, there is neither handshape nor hand orientation information, only location and movement can be observed. In the second condition, some information on hand orientation is available, too. In condition three, some handshape information and some hand orientation information can be observed along with location and motion, and in the fourth condition the sign is shown without modification (see table 3.2 for an overview of the conditions). The difference in recognition success between the conditions will tell us more about the importance of handshape and hand orientation for recognition.

3.2 Method

To test the influence of handshape and hand orientation information on recognition, we created sign movies that allowed a signer to see only part of this information. We worked with isolated signs (citation form). This removes any clues for recognition from context. In our case, this is desirable, because the automatic sign language recognition we work on is also meant for isolated signs [78]. Twenty different signs were recorded and four versions of each were created, which varied in the amount of available handshape/hand orientation information. The next section describes the sign set, subsequent sections discuss the versions, the manner in which they were created, and the setup of the experiment.

3.2.1 Sign Set

Twenty different signs were selected from the standard vocabulary of SLN. They are listed in table 3.1, showing SLN glosses and English translations. The English terms are used throughout the text. Illustrations and specifications are provided in appendix 3.4.3. There were several criteria for selection. Because the effect of handshape and hand orientation is the subject of investigation, the other parameters (location, motion, non-manual component, one- or two-handedness) were kept as equal as possible across the twenty signs. This was only valid for path motion, hand internal motion was not checked. The signs were recorded without any non-manual component, because we were interested in the information conveyed by the hands only. An extra criterion was therefore that all signs must be uniquely identifiable without using the non-manual component. All signs were one-handed, located in the neutral space, and had either no path motion or a path motion that was generally from high to low, not repeating. For some signs, this path was straight, for others arched or inclined (e.g. from top left to bottom right). Of the signs without a path motion, some signs were entirely motionless, some contained handshape or hand orientation changes (hand
internal movement). One sign was located near the temple, all the others were made in front of the signer’s body. In the end, not all signs were uniquely identifiable without the non-manual component. For these signs, the possible alternative meanings were listed in advance, and these were also considered correct identifications.

Table 3.1: Sign set used in the experiment.

<table>
<thead>
<tr>
<th>AFGEPRIJSD</th>
<th>- PRICED-DOWN</th>
<th>OKTOBER</th>
<th>- OCTOBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLEEN</td>
<td>- ALONE</td>
<td>OPNIEUW</td>
<td>- AGAIN</td>
</tr>
<tr>
<td>BLAUW</td>
<td>- BLUE</td>
<td>SCHADUW</td>
<td>- SHADOW</td>
</tr>
<tr>
<td>EEN</td>
<td>- ONE</td>
<td>SCHOOONMAKEN</td>
<td>- TO-CLEAN</td>
</tr>
<tr>
<td>FILM</td>
<td>- MOVIE</td>
<td>SECONDE</td>
<td>- SECOND</td>
</tr>
<tr>
<td>GEDEELD</td>
<td>- DIVIDED-BY</td>
<td>SNEEUW-</td>
<td>- SNOWDROP</td>
</tr>
<tr>
<td>DOOR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAMER</td>
<td>- HAMMER</td>
<td>SPONS</td>
<td>- SPONGE</td>
</tr>
<tr>
<td>LEUK</td>
<td>- NICE</td>
<td>STOPLICH</td>
<td>- TRAFFIC-LIGHT</td>
</tr>
<tr>
<td>LIMBURG</td>
<td>- LIMBOURG</td>
<td>TEKENEN</td>
<td>- TO-DRAW</td>
</tr>
<tr>
<td>NOG NIET</td>
<td>- NOT-YET</td>
<td>VLAG</td>
<td>- FLAG</td>
</tr>
</tbody>
</table>

3.2.2 Recording

The recorded signs were made by a right-handed female sign language instructor (late signer). She was seated at a table. The view was limited to her upper body and the table top. The background was black, the tablecloth was white and the signer’s clothing was turquoise, providing good contrast. Lighting was diffuse, so that there were no sharp shadows. The movies were recorded with two digital cameras (Allied Vision Technologies ‘Guppies’). One gave a frontal view, these images were the basic sign videos. The other recorded the signer at an angle. These obliquely recorded images (together with the frontal view images) were used for processing the basic videos to create conditions 1–4. The signs were recorded at 30 fps, resolution 640 x 480 pixels. As explained in the previous section, the signer did not perform the non-manual component, but kept a neutral expression. The signer kept her left hand (non-signing hand) underneath the table. The signer always started with her signing hand on the table, and put her hand on the table again after the sign. The average length of a sign movie was 3.15 seconds.
3.2.3 Manipulations

Participants saw sign movies in conditions which varied in the amount of handshape/hand orientation information that was available: without any handshape or hand orientation information, with hand orientation information but without handshape, with some handshape information and some hand orientation information, and ‘clean’ (without any manipulations). The hand orientation shown is the orientation of the palm, without the direction of the fingers. Table 3.2 gives an overview of the four conditions.

We altered the basic sign movies into movies showing the desired amount of handshape/hand orientation information by means of automatic image processing techniques. To this end, we employed the following manipulations and implementations. The first step was color segmentation using a model of skin color to locate the head and hand in the videos. Basically, this skin color model has a certain range of colors that are regarded as ‘skin color’, and it checks each pixel in each frame of a movie to see whether it is skin colored or not. Adjacent skin colored pixels are put together into ‘blobs’, and small blobs are ignored as noise. In this way, two blobs were found in each frame, one for the head and one for the visible hand. Once the locations of hand and head were known, all kinds of manipulations could be performed. We describe here the ones used to create conditions 1–3 of the experiment (condition 4 requires no processing). An example of the result is shown in figure 3.1.

Table 3.2: Availability of the five formational parameters in conditions 1–4 of the experiment and the appearance of the conditions in the sign movies. Loc = location, Mov = motion, Or = hand orientation, HS = Handshape, NMC = non-manual component.

<table>
<thead>
<tr>
<th>Cond</th>
<th>Loc</th>
<th>Mov</th>
<th>Or</th>
<th>HS</th>
<th>NMC</th>
<th>Appearance in movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>hand is completely covered with a pink circle</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>yes</td>
<td>partly</td>
<td>no</td>
<td>no</td>
<td>hand is completely covered, orientation of the palm is shown as a deformed circle</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>partly</td>
<td>partly</td>
<td>no</td>
<td>hand is completely covered, an approximation of the handshape silhouette is shown</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no manipulations</td>
</tr>
</tbody>
</table>
Condition 1  The first condition must not show any handshape or hand orientation information, only movement and location. This was achieved by covering the hand entirely with a circle that moves with the hand. First, the location of the center of the hand-blob (the average location of all hand-pixels) was calculated in each frame. Then all pixels in a circle around this center were colored pink, in such a way that the entire hand was covered. When the hand moved, the center of the circle moved with it. The same circle radius was used for all frames of all movies. If we had taken the smallest necessary size in each frame, or per movie, the circle would have become larger if the hand assumed a larger handshape. This would have given extra information we did not wish to provide. The circle’s radius was therefore kept fixed. The radius was determined experimentally at 100 pixels.

Condition 2  The second condition must show movement, location and hand orientation, but not handshape. To achieve this, an ellipse is used to represent the hand palm orientation. Imagine a paper circle taped onto the hand palm. When the palm cants or flips, the circle cants and flips with it. To an observer, this circle will look like an ellipse when the hand is canted. If the observer knows the shape is really a circle, s/he can quite easily interpret the deformation of the shape as a hand orientation [53]. In this way, palm orientation can be shown to the viewer without showing the hand. See the second row of figure 3.1 for an example. Note that the direction of the fingers is ignored. Also, there is no difference between the front and the back of the ellipse, so the difference between palm and back of the hand cannot be seen in this representation. The ellipse is therefore a simplified representation of the hand orientation parameter. To create this representation automatically was not trivial. First, the position of the hand pixels in 3D had to be known (because 3D information, such as depth, and orientation are directly linked to each other). This 3D position was obtained automatically by using two cameras. From the disparity (that is the difference in the image location of a certain point in 3D space) the depth of that point could be inferred (the human eyes use the same trick). The 3D position of each hand pixel was calculated in this manner. This gave a ‘cloud’ of pixels (a blob in 3D). Next, the main orientation of this cloud was found by calculating which plane fitted the cloud best (using Principle Component Analysis). A circle was then drawn in this plane, with its center at the central location of the cloud. This was the ‘paper circle taped to the hand’ representing the palm orientation. Projecting back to the 2D image, the circle became an ellipse, the shape (deformation) of which depended on the palm orientation. This ellipse was drawn in the image at the location of the hand. To avoid the hand showing behind the ellipse, the entire hand was first covered as in condition 1. The ellipse and circle moved together with the hand.

Condition 3  Condition three must show an approximation of the handshape without details. This was achieved by taking the silhouette of the hand in each frame and
Figure 3.1: Examples of the four manipulations on the sign FLAG. Three frames of the sign movie are shown. The top row shows condition one, without any handshape or hand orientation information available. The second row shows condition 2, the ellipse that indicates the palm orientation. Row three shows condition 3, the handshape approximation, and row four shows the unmodified condition 4.
simplifying the shape by means of a Fourier transform. Basically, this means that all small variations of the boundary are discarded, and a smooth, rounded shape is retained. We used the first ten Fourier components to construct the simplified shape. This new shape was drawn in the image at the location of the hand. To avoid the hand showing behind the shape, the entire hand was first covered as in condition 1. Again, shape and circle moved together with the hand.

In this way, the four manipulations listed in table 3.2 were obtained. The manipulations were performed on each of the twenty sign movies, resulting in 20 (signs) x 4 (manipulations) = 80 stimuli.

Error Handling in Condition 2  In certain cases, the hand orientation could not be calculated correctly. This was often the case for fist-like handshapes (e.g. ASL handshapes ‘A’, ‘S’, ‘O’). This was to be expected: a fist is basically a sphere, and a sphere has no best-fitting plane. In such cases, the orientation was adjusted by hand to fit the palm orientation. Also, the depth calculation was sometimes incorrect. This led to an incorrect calculation of hand orientation for one or a few frames. To remedy this, interpolation between reliable frames was used. In cases where the hand orientation was constant and interpolation gave incorrect results, the calculated orientation from the previous frame was used.

3.2.4 Task

The stimuli were shown to the participants on a 17-inch LCD computer screen. It was not possible for a participant to view a stimulus again or to slow it down. After the stimulus was shown, the participant saw a screen in which s/he had to choose between the following options: “I recognized the sign” and “I did not recognize the sign”. If s/he chose the second option, s/he could then click a ‘go ahead’ button. If s/he chose the first option, s/he had to type in the gloss of the sign s/he thought s/he had seen. After that, s/he could click the ‘go ahead’ button. Clicking the ‘go ahead’ button immediately caused the next stimulus to be shown. For each action (clicking and typing), the participant could take as much time as s/he wanted.

The stimuli were shown per condition: first all stimuli of condition 1 were shown, then all those of condition 2, etcetera. Within a condition, the order of the stimuli was random and different for each participant. Each individual stimulus was presented five times during a trial, for greater accuracy of the results. The only restriction on the randomness was that the same stimulus could not occur twice in a row. A trial took approximately 45 minutes. Halfway through (after completing condition 2) there was a short break. The experiments were performed at different locations: at participants’ homes or workplaces, or at schools. The test program ran on a laptop, which was set up in a quiet place at the test location.
All participants received instructions to the task beforehand. The instructions were available both as text and as an SLN movie, and contained examples of the manipulations. The examples used signs that were not in the experimental sign set. To aid comprehension of the ellipse in condition 2, a physical circle attached to a hand was also demonstrated during instruction. The physical example made clear how a circle in 3D would look like an ellipse in 2D on the screen, and how it represented hand orientation. After instruction, participants could try the task briefly with 10 practice stimuli, made from signs that were not used in the actual experiment. They then had the opportunity to ask questions. If all went well, they could start the experiment.

3.2.5 Participants

Twenty-seven persons participated in the experiment. All of them were experienced sign language users, either in SLN or in Signed Dutch. Since the task was lexical recognition, and there is not much difference in lexicon between SLN and Signed Dutch, the two were not treated differently in this experiment. Eighteen participants were female, nine male. Ages ranged from 17 to 51 years, average 30 (SD = 7.5). Seventeen were deaf, ten were hearing. Of the deaf, three had deaf parents, the other fourteen had hearing parents. All hearing participants had hearing parents.

Seventeen participants used signing in their profession: three as SLN interpreters, four as teachers of SLN, and ten in other capacities. One participant indicated her use of signing as ‘always’, twenty indicated theirs as ‘daily’, and six as ‘regularly’. The age to start using SLN/Signed Dutch ranged from 0 to 28 years, with an average of 11 (SD = 9.2).

3.3 Results

3.3.1 Processing of Responses

The fourth condition, in which the sign movies were unmodified, was used to check whether a participant was familiar with a sign. Upon evaluation of the responses, it became clear that a participant often gave a different gloss for a sign than the intended one (in about 30% of the presentations). Sometimes these were spelling errors or minor rephrasals (such as ‘seconds’ instead of ‘second’). These variations were corrected (about 2% of the presentations). More often, a participant responded with a gloss for a sign that was truly a different concept. Examples are ‘to color’ instead of ‘to draw’ (resemblance of movement and meaning) and ‘minute’ instead of ‘second’ (resemblance of handshape and meaning). These findings concur with the results of Tartter and Fischer [83], who found confusions to be made between minimal pairs of
signs (signs that only differ in one formational parameter) under normal conditions. Upon learning the correct meaning after the experiment, our participants often indicated that they knew the sign in a different form, or that they generally rely on the non-manual component to distinguish pairs such as ‘minute’ and ‘second’. Though the signs in question could be distinguished on the basis of the other parameters, participants were apparently used to relying on the non-manual component.

There were many cases of such personal ‘alternate meanings’ for a sign. In fact, the fraction of incorrect recognition in the unmodified condition was 0.42 (30% alternate meanings, 12% unrecognized signs). However, in this experiment our interest was not really in the exact meaning of a sign; rather we wanted to see how well a participant could identify the same sign in conditions 1–4. In fact, it does not really matter which meaning a participant ascribes to a sign as long as it is the same meaning in the various conditions. For that reason, a system of ‘personal meanings’ was adopted. A participant’s personal meaning for a sign is the meaning s/he attributes to the sign in condition four, provided s/he does this for at least three of the five repeated presentations. This guarantees that the participant is consistent in his/her interpretation of this sign, and also that s/he can only have one personal meaning for a sign. Using personal meanings, the definition for a ‘correct’ answer becomes the following: a meaning given for a stimulus is considered ‘correct’ if it equals the personal meaning for that participant for that sign, and ‘incorrect’ otherwise (even if the given meaning is the original ‘true’ meaning!). If instead all participant-sign combinations with an incorrect answer in condition four had been removed, then 187 participant-sign combinations would have had to be discarded (36% of the data).

There were two situations in which there was no personal meaning for a participant for a sign: when s/he had chosen “I do not know this sign” in condition four, and when s/he did not give the same meaning at least three times in condition four. These were signs completely unknown to the participant. The sign SNOWDROP was very obscure: it was unknown to 14 of the 27 participants. For this reason, this sign was excluded from the test results. For the other signs, if a sign was unknown to a certain participant, then all data for that sign were removed from the test results of this participant. In total, 52 participant-sign combinations were removed (10% of the data). The consequence was that the number of measurements per sign as well as per participant was no longer constant. The average number of signs per participant was 17 ($SD = 1.4$), and the average number of participants per sign was 24 ($SD = 3.8$).

Recognition scores in this experiment were measured as follows: participants saw each stimulus five times during a session. For each presentation, they gave an answer which was either correct or incorrect. The score on a stimulus was defined as the cumulative score on the five presentations, where ‘correct’ was counted as 1 and ‘incorrect’ as 0. Scores could therefore range from 0 to 5, six distinct values. However, consistency per stimulus was high: most scores were either 0 or 5.
3.3.2 Recognition per Condition

The recognition scores per condition are shown in figure 3.2. It shows the six possible scores and the percentage of occurrence of that score or lower, for each condition separately. Thus we see that for condition 1 and 2, the largest part of the scores (85%) is zero. The other scores occur in small amounts. For condition 3, most scores are either zero (about 30%) or five (about 40%). For condition 4, only scores of three or higher are possible; most are five. In all conditions, consistency is high: a participant most often
recognizes all five presentations of a stimulus, or none of them. The scores for condition 1 and 2 are practically equal: they are almost exclusively zero, which means there is hardly any recognition in these conditions. Scores for condition 3 are more varied, most of them are either 0 or 5, but some scores in between occur also. Apparently, some recognition is possible in this condition.

Figure 3.3: Scores of individual signs per condition averaged across participants. The four panels represent the four conditions. The horizontal axis shows the signs, the vertical axes the scores. For each sign, a boxplot is shown, with the thick grey line representing the median, the edges of the box the first and third quartile, and the ‘whiskers’ the minimum and maximum. Circles and stars are outliers and extreme values respectively. The black line connects the mean scores of the signs. The signs were grouped into three clusters on the basis of statistical differences, and ordered by median score within the groups. In condition 4, only scores of 3 or higher were possible.
3.3.3 Recognition of Individual Signs

Figure 3.3 shows the recognition scores per sign and condition. The scores are presented in two ways: (1) boxplots indicating the median plus the first and third quartiles, and (2) a black line connecting mean scores. This way of presenting was chosen because the mean scores are sensitive to outliers and the medians sometimes showed extreme scores. This particularly holds for condition 3, where most signs show a large spread in scores. Figure 3.2 suggests that this is mainly due to the fact that participants predominantly scored 0 and 5.

A two-way within-subjects repeated measures ANOVA was performed to investigate the impact of conditions 1 to 3 and the nineteen signs on the recognition scores. Condition 4 was left out because this condition was used to establish the set of signs most participants were familiar with. Because there were missing values (only five participants knew all nineteen signs, for the others one or more signs had to be removed), a missing values analysis was done in SPSS 14.0. This routine employs an expectation-maximization algorithm to replace the missing values by estimates on the basis of the existing distributions. Furthermore, Mauchly’s test showed that the data did not satisfy the sphericity assumption for condition \((\chi^2 = 37.89, p < 0.001)\) nor for sign \((\chi^2 = 295.14, p < 0.001)\). We therefore used the Geenhouse-Geisser correction on the degrees of freedom.

A significant effect was found for condition \((F_{1.1,27.9} = 330.31, p < 0.001)\). This could be attributed to condition 3 showing systematically higher scores than those in conditions 1 and 2. For all signs, behaviour in condition 2 is almost identical to that in condition 1, with the exception of BLUE, which yields some scores above zero in condition 2 but not in condition 1. Only four other signs show nonzero scores in both conditions 1 and 2. In this respect, the sign set is quite homogeneous. Post hoc tests show no significant difference in score between conditions 1 and 2 \((p < 0.12)\). The scores in both these conditions do differ significantly from the scores in condition 3 \((p < 0.001\) for both).

There was also a significant main effect of sign \((F_{7.6,190.0} = 12.77, p < 0.001)\). Post-hoc tests on the recognition scores for the signs suggest a clustering into three groups. These clusters can be found in figure 3.3, separated by vertical lines. In addition, table 3.3 shows an overview. Recognition scores of signs within a group do not differ significantly from each other. From figure 3.3 we see that group III contains the signs that show some recognition in conditions 1 and 2 and good scores in condition 3. These are the signs that score best over all conditions. Group I contains the signs with the lowest scores over all conditions. Group II contains the signs with mediocre scores.

There was also a significant interaction between condition and sign \((F_{11.7,293.1} = 8.65, p < 0.001)\). This is caused by the fact that (1) the recognition scores
Table 3.3: Three groups of signs based on common patterns of significant/non-significant differences among the signs. Signs within a group do not differ significantly from each other. Group I could be characterized as the signs generally worst recognized, group II as the signs moderately well recognized, and group III as the signs best recognized. '*' indicates the sign contains a path movement, '+' indicates it contained palm orientation change(s).

<table>
<thead>
<tr>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGAIN +</td>
<td>HAMMER</td>
<td>NICE *+</td>
</tr>
<tr>
<td>SHADOW *</td>
<td>TO-DRAW *</td>
<td>BLUE +</td>
</tr>
<tr>
<td>NOT-YET +</td>
<td>MOVIE +</td>
<td>FLAG +</td>
</tr>
<tr>
<td>PRIDED-DOWN *</td>
<td>SPONGE</td>
<td>OCTOBER *</td>
</tr>
<tr>
<td>LIMBOURG *</td>
<td>ALONE *</td>
<td>DIVIDED-BY *</td>
</tr>
<tr>
<td></td>
<td>ONE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TO-CLEAN *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRAFFIC-LIGHT *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SECOND</td>
<td></td>
</tr>
</tbody>
</table>

in groups I and II range from low to high for condition 3 but remain almost zero for conditions 1 and 2, and (2) the scores in group III range from low to high for conditions 1 and 2, but remain constantly high for condition 3.

3.3.4 Participants’ Backgrounds

The participants differed in various characteristics, such as age, hearing status, hearing status of parents, age to start learning signs, etc. To see whether such qualities affect performance, recognition scores were compared for the following characteristics: hearing status, parent hearing status, age, gender, age to start learning signs, frequency of use of signs, using SLN in profession, self-acclaimed level of signing, and region of residence (north or west area of the Netherlands). No differences in scores were found for any of these characteristics. As an example, we show the scores for deaf and hearing participants in figure 3.4.

The absence of any influence of the participants’ background was confirmed statistically by means of separate ANOVAs, with an overall significance level of 0.05 after Bonferroni correction. Furthermore, no factor had an interaction with either condition or sign. We therefore conclude that, in this experiment, background was of little importance.
3.3.5 Influence of Sign Characteristics

There were significant differences in scores among the individual signs. These differences are most pronounced in condition 3, since in conditions 1 and 2 most signs score close to zero. However, at least one sign (BLUE) did show an increase in score between condition 1 and 2. Certain sign characteristics may influence how well a sign can be recognized in a certain condition. They are analyzed below.

Movement and hand orientation

Ten of the signs in our dataset contained a path movement, the other nine did not. A two-way within-subjects repeated measures ANOVA was performed to compare recognition scores on signs with and without a path movement in conditions 1–3. The Greenhouse-Geisser correction was once again used. Signs containing a path performed significantly better ($F_{1,26} = 4.65, p = 0.041$). There was no interaction with condition ($F_{1.4,36.8} = 0.47, p = 0.565$). In principle, finer distinctions of movement

Figure 3.4: Recognition of Hearing (N = 10) and Deaf (N = 17) participants for conditions 1–4. The horizontal axis shows the conditions, the vertical axis the mean recognition scores. Error bars indicate the standard error of the mean. No difference in recognition scores is present between hearing and deaf participants ($F_{1,25} = 0.083, p = 1.000$).
could be made (for different types of path). However, because we used only twenty
signs, such groups would become too small to draw any general conclusions.

Six signs contained a change in hand orientation, either once or repeatedly. Again a
two-way within-subjects repeated measures ANOVA was performed to compare
recognition scores between signs with and without hand orientation changes, using
the Greenhouse-Geisser correction. No significant difference was found
\( F_{1.2,31.6} = 256.87, p < 0.001 \). However, there was an interaction between the presence
of hand orientation change and the experimental condition \( F_{1.5,39.1} = 8.37, p = 0.02 \).
Closer inspection showed that the groups behave the same in condition 1, but
somewhat differently in conditions 2 and 3 (see discussion).

### Handshape

The sign dataset contained a variation of handshapes. Table 3.4 lists them and gives
the equivalent ASL handshape

For depictions of the handshapes, see http://research.tenholt.nl/handshapes. It is
conceivable that the handshape parameter influences recognition in condition 3 (where
the simplified handshape is shown). For example, a more rare handshape may make it
easier to recognize the sign. For that reason we calculated the frequency of each
handshape in the SLN lexicon\(^1\). When a sign contained multiple handshapes (5 cases),
we used only the one that occurred first in the sign in time. However, the correlation
between handshape frequency and recognition of that handshape in condition 3 was
low \( r = 0.15, p = 0.66 \), so handshape frequency does not appear to facilitate
recognition.

We did the same for handshape groups. In SLN, handshapes are divided into eight
groups [73]. We calculated the frequencies of each group in the SLN lexicon and
compared them to the recognition scores in condition 3 for each group. Again, no
significant correlation was found \( r = -0.06, p = 0.89 \). Apparently, handshape
frequencies have no influence on recognition success.

### 3.4 Discussion

#### 3.4.1 Conditions

In this experiment, lexical recognition of SLN signs was tested in conditions in which
handshape and hand orientation information were present in varying degrees. In the
first condition, only location and motion could be observed. This condition is
comparable to that in the point-light displays of Poizner et al. [68], except that in our
Table 3.4: Handshapes occurring in the dataset (column 1), their frequencies in the SLN lexicon\(^1\) (column 2), the handshape groups these shapes belong to (column 3) and the frequencies of the groups in the SLN lexicon (column 4). When a sign contained multiple handshapes, only the one that occurs first in the sign in time was used. The equivalent ASL shapes (column 5) are taken from the overview at http://www.handspeak.com/abc.

<table>
<thead>
<tr>
<th>Handshape</th>
<th>HS Freq (%)</th>
<th>Handshape Group</th>
<th>HS Group Freq (%)</th>
<th>Equiv. ASL HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>one or two fingers extended</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>U</td>
<td>2</td>
<td>one or two fingers extended</td>
<td>23</td>
<td>U</td>
</tr>
<tr>
<td>S</td>
<td>7</td>
<td>fist</td>
<td>18</td>
<td>S</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>one or two fingers extended</td>
<td>23</td>
<td>L</td>
</tr>
<tr>
<td>baby-open snavel</td>
<td>1</td>
<td>thumb opposed to other fingers</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>three or more spread fingers</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>O</td>
<td>2</td>
<td>round closed shape</td>
<td>6</td>
<td>O</td>
</tr>
<tr>
<td>geld</td>
<td>5</td>
<td>fist</td>
<td>18</td>
<td>10-t</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>round open shape</td>
<td>9</td>
<td>C</td>
</tr>
<tr>
<td>B-nul</td>
<td>10</td>
<td>flat shape</td>
<td>24</td>
<td>5-thumb</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>flat shape</td>
<td>24</td>
<td>B</td>
</tr>
<tr>
<td>1v</td>
<td>1</td>
<td>one or two fingers extended</td>
<td>23</td>
<td>1-bent</td>
</tr>
</tbody>
</table>
experiment the arm was visible, possibly allowing a participant to deduce some information about hand posture. In the second condition, location, motion and some hand orientation information were available, and in condition 3, location, motion and approximate handshape (and through that some information about hand orientation, too). In the unmodified condition 4, everything was visible. Recognition scores were very low both in condition 1 and in condition 2. For condition 1, this was to be expected: with no information about three of the five sign parameters (handshape, hand orientation and non-manual component), the set of possible signs is simply too large to make recognition possible [61]. The hand orientation information revealed in condition 2 did not improve scores at all: there was no significant difference in scores between conditions 1 and 2. The data suggest that showing the hand orientation, at least in the present representation, does not lead to improved recognition compared to showing only position and motion. A possible explanation is that the hand orientation representation used in this experiment is limited: the difference between the palm and the back of the hand cannot be seen (the ellipse does not have a front and a back), implying that even after seeing the ellipse, two palm orientations remain possible. In addition, the finger orientation is not represented at all. Therefore, after seeing this orientation representation, not many signs can be excluded from the set of possible signs. During a session, participants often remarked that in condition 2, too many signs were still possible options for them to make a confident choice.

When handshape information was made available in condition 3, there was a significant rise in recognition rate. Apparently, seeing the handshape, even in a simplified form, made it possible to eliminate enough candidates to identify a sign in about 60% of the cases. This could be a consequence of the fact that there are many handshapes in SLN, which means that resolving handshape greatly reduces the possible sign candidates. Also, though the handshape approximation was shown as a silhouette, it was still possible to deduce some hand orientation information from the image, so that in this condition, both handshape and hand orientation could be observed to some degree. But this limited representation was not sufficient for reliable sign recognition: recognition in condition 3 was still significantly lower than that in the unmodified condition. In some cases, details that could not be observed in condition 3 are necessary for resolving a sign: e.g. it is sometimes necessary to see whether a handshape has one or two extended fingers to distinguish two signs. Also, when the handshape is a fist, it is difficult to deduce its orientation from its silhouette.

### 3.4.2 Individual Signs

In condition 1, there is little difference between the scores of individual signs. The same is true for conditions 2 and 4. However, there are signs (most of the signs in group III) that show some recognition in conditions 1 and 2. Most notably FLAG, OCTOBER and DIVIDED-BY. These signs all contain a rare motion pattern that can
already be observed in condition 1, though the path of OCTOBER is not as unique as the other two. This probably explains their good scores in the conditions in which other signs score almost exclusively zero. The sign BLUE is the only sign that shows an increase in score between conditions 1 and 2. The sign contains a characteristic flip of the palm, which becomes visible in condition 2, causing the rise in recognition rate. The other signs all score mostly zero in conditions 1 and 2.

Condition 3 is the only condition in which all signs show a variety of scores. Recognition improves for all signs compared to conditions 1 and 2. However, for signs in group I, this rise is mostly small, with median scores still below two. Signs in group III score almost perfectly in condition 3 (except DIVIDED-BY, which has more variable scores). Signs in group II are variable, showing mediocre averages and large spreads in scores. A few signs in group II (ONE, TO-CLEAN, TRAFFIC-LIGHT, SECOND) and NICE from group III show a large increase in recognition between conditions 1/2 and 3. These signs all contain handshape changes or finger orientation changes (which cannot be observed in condition 2, but become visible when the fingers become visible). Combinations of handshapes and typical hand orientation change patterns quickly narrow down the set of possible signs, making identification easier. ONE is an exception in that it contains neither handshape change nor hand orientation change. However, its lack of any movement is quite typical, and this together with the sign’s commonness may make it well recognizable when the shape is visible. Another exception is TO-CLEAN, which has a common handshape (ASL shape ‘S’) but still shows a large increase in recognition in condition 3. This specific case may be explained as follows: when only the path is visible, there are many possible signs (at least 5 SLN signs have the same path). TO-CLEAN is an infrequent word in Dutch compared to the alternatives (such as TO-DRAW in our setting)\textsuperscript{3} [7]. Assuming a similar situation for the SLN glosses, participants may be more inclined to enter the alternatives than the infrequent TO-CLEAN. However, once the handshape is visible, the other signs are no longer possible (since none have fist-like handshapes), and participants identify TO-CLEAN, causing the rise in recognition scores in condition 3. The signs that score low in condition 3 are generally signs with common handshapes (especially fist-like handshapes) that are difficult to interpret in silhouette. This difficulty of interpretation can be caused by the fact that the hand orientation is hard to make out (for e.g. a silhouetted fist), or that the fingers are invisible in silhouette (which happens when the fingers are pointed towards the camera).

Signs containing a path movement scored significantly better than signs without a path in all conditions. This is not surprising, since the identification of a specific path (visible in all conditions) narrows down the possible candidates more than the knowledge that the sign contains no path (the group of signs with no path is much larger than any group of signs sharing a specific path). In the same way, knowledge of a specific hand orientation change decreases the number of sign candidates more than knowing the sign contains no orientation change. However, the hand orientation is not visible in condition 1 and cannot always be deduced in condition 3. It is therefore not
surprising that signs containing hand orientation change tend to score better in condition 2 only.

From these observations, better general recognition seems linked to rarity of motion pattern and handshape, as well as to presence of handshape and hand orientation changes and path movement. However, none of these factors are reliable predictors for recognition scores. Some signs with hardly any of these characteristics are still recognized well (such as OCTOBER), whereas signs containing a number of them can score low (e.g. PRICED-DOWN). Such cases may be exceptions, explained by ad hoc reasoning (OCTOBER was well recognized because the experiment was performed in September/October?) or coincidences such as in the case of TO-CLEAN. Visibility issues also play a part: a rare handshape that cannot be seen clearly in condition 3 (which is the case for e.g. NOT-YET and PRICED-DOWN, where the fingers are pointed towards the camera) will not improve recognition, thus sabotaging any correlation between frequency of handshape and recognition score. However, the sign set is too small and the possible factors of influence too many to reliably point out signs which are exceptions. Similar research on larger sign sets is necessary to detect which sign characteristics are reliable predictors for recognition performance.

3.4.3 General Conclusions and Future Directions

This experiment was conducted to learn more about the importance of the formational parameters handshape and hand orientation for lexical recognition of sign language signs. Results show that with only location and motion information, signs cannot be recognized, and that adding hand orientation information does not improve recognition. However, when reduced handshape information is added, recognition does improve, though it remains significantly lower than recognition of the unmodified representation. It is interesting to note that confusions and mis-identifications are made even in the unmodified condition. The same was found by Tartter and Fischer [83]. Apparently, minimal distinctions in the hand parameters of signs are often overlooked, since in normal conversation extra clues (context and/or the non-manual component) can be used to find the correct identification. It would be interesting to know if such confusions are also made by signers during production, and if so, whether an observer would notice.

Our findings for condition 1 form an interesting complement to the experiment of Poizner et al. [68]. Their point-light presentation of the signs was roughly equivalent to our condition 1, since only location and motion could be observed. Poizner et al. tested lexical recognition with handshape information provided separately, and found good recognition (ca. 80%). Our findings show that if the handshape information had not been provided, recognition would probably have been close to zero.

The background of the participants did not matter in the current experiment. The task
was performed equally well by deaf and hearing, late and early signers, and children of deaf and hearing parents. Nor did any other participant characteristic have a significant effect. In contrast, other studies have found differences in performance between late and early signers, e.g. Emmorey and Corina [30]. These studies, however, measured reaction time. For studies such as this one, where response time is not a factor, no data on the influence of participant background are available, since the only studies we know of use only native signers as participants.

As mentioned in the introduction, this sign perception research was conducted for the benefit of automatic sign language recognition. From this experiment, we conclude that automatic recognition has more to gain from a detailed representation of handshape than hand orientation (two features that are currently not often used). However, the results in condition 2 may be influenced by the representation used: it is possible that participants had trouble interpreting the hand orientation representation (the ellipse) correctly, though observation during the experiment suggested that participants understood the representation quite well.

Ideally, we would like to test each formational parameter separately: handshape, hand orientation, location, motion, and the non-manual component. However, it is difficult to separate location from motion, and handshape from hand orientation, since they are separate qualities of the same entity (the location resp. the hand). In the current form, we have found a way to separate hand orientation from handshape, by representing hand orientation with a different entity, and showing handshape as a silhouette. Other experiments in this area (e.g. showing only non-manual component, removing motion and location by showing the hand separately) would improve knowledge on sign perception and make better sign language recognition possible.

Notes

1Standaard Lexicon Nederlandse Gebarentaal, deel 1. Stichting Nederlandse Gebarentrum, 2002
2SPSS for Windows, rel. 14.0.0. Chicago:SPSS Inc., 2005

Acknowledgements

This research was partly funded by the VSB-fund; grant number 2003451: project “Een elektronische leeromgeving voor jonge kinderen met auditieve en communicatieve beperkingen voor het leren van Nederlandse Gebarentaal (ELo)” and the ICT Delft
Research Centre of Delft University of Technology. The authors would like to thank Marielle Elzenaar for signing the stimuli, Arend Harteveld for creating the interface used in the experiment, and André Redert for his help with the 3D calculations.
Appendix 3.A: Description of Sign Dataset

Descriptions of the twenty signs used in the experiment. Figure 3.5 shows pictures of the signs. Sign p was later removed from the set because it was unknown to more than half the participants. Perpendicular slashes through an arrow mean the movement is repeated. In sign c, the hand flips and changes shape, without other movement. In sign q, the hand repeatedly switches between the two handshapes shown, without other movement. In sign r, the square C-shape means the hand opens once. Images are used with the permission of the Royal Effatha Guyot group (Kentalis). Images a, g, i–m and p–t have been modified by the authors.
Figure 3.5: Sign Images
Table 3.5: Sign descriptions

<table>
<thead>
<tr>
<th>Sign</th>
<th>Hand-shape SLN</th>
<th>Hand-shape ASL</th>
<th>Location</th>
<th>Movement</th>
<th>Orientation Palm</th>
<th>Hand Internal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>path?</td>
<td>direction</td>
<td>type</td>
</tr>
<tr>
<td>PRICED-DOWN</td>
<td>baby-snavel QQ</td>
<td>beak → Q</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>straight</td>
</tr>
<tr>
<td>ALONE</td>
<td>1</td>
<td>1</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>straight</td>
</tr>
<tr>
<td>BLUE</td>
<td>B → L</td>
<td>B → L</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ONE</td>
<td>1</td>
<td>1</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MOVIE</td>
<td>geld</td>
<td>10-t</td>
<td>temple</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DIVIDED-BY</td>
<td>1-v</td>
<td>1-bent</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>arched</td>
</tr>
<tr>
<td>HAMMER</td>
<td>S</td>
<td>S</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NICE</td>
<td>baby-snavel Q</td>
<td>beak Q</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>straight</td>
</tr>
<tr>
<td>LIMBOURG</td>
<td>L</td>
<td>L</td>
<td>neutral space</td>
<td>y</td>
<td>down right</td>
<td>straight</td>
</tr>
<tr>
<td>NOT-YET</td>
<td>L</td>
<td>L</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OCTOBER</td>
<td>C</td>
<td>C</td>
<td>neutral space</td>
<td>y</td>
<td>down left</td>
<td>straight</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Sign</th>
<th>Hand-shape SLN</th>
<th>Hand-shape ASL</th>
<th>Location</th>
<th>Movement</th>
<th>Orientation Palm</th>
<th>Hand Internal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>path?</td>
<td>direction</td>
<td>type</td>
</tr>
<tr>
<td>AGAIN</td>
<td>U</td>
<td>U</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SHADOW</td>
<td>5</td>
<td>5</td>
<td>neutral space</td>
<td>y</td>
<td>down left</td>
<td>straight, slow</td>
</tr>
<tr>
<td>TO-CLEAN</td>
<td>S</td>
<td>S</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>wiggly</td>
</tr>
<tr>
<td>SECOND</td>
<td>1</td>
<td>1</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SNOW-DROP</td>
<td>bloem</td>
<td>5-close</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SPONGE</td>
<td>5↔S</td>
<td>5↔S</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TRAFFIC-LIGHT</td>
<td>O→5</td>
<td>S→5</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>straight</td>
</tr>
<tr>
<td>TO-DRAW</td>
<td>U</td>
<td>U</td>
<td>neutral space</td>
<td>y</td>
<td>down</td>
<td>wiggly</td>
</tr>
<tr>
<td>FLAG</td>
<td>B</td>
<td>B</td>
<td>neutral space</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Abstract

Current automatic sign language recognition (ASLR) seldom uses perceptual knowledge about the recognition of sign language. Using such knowledge can improve ASLR because it can give an indication which elements or phases of a sign are important for its meaning. Also, the current generation of data-driven ASLR methods has shortcomings which may not be solvable without the use of knowledge on human sign language processing. Handling variation in the precise execution of signs is an example of such shortcomings: data-driven methods (which include almost all current methods) have difficulty recognizing signs that deviate too much from the examples that were used to train the method. Insight into human sign processing is needed to solve these problems. Perceptual research on sign language can provide such insights. This paper discusses knowledge derived from a set of sign perception experiments, and the application of such knowledge in ASLR. Among the findings are the facts that not all phases and elements of a sign are equally informative, that defining the ‘correct’ form for a sign is not trivial, and that statistical ASLR methods do not necessarily arrive at sign representations that resemble those of human beings. Apparently, current ASLR methods are quite different from human observers: their method of learning gives them different sign definitions, they regard each moment and element of a sign as equally important and they employ a single definition of ‘correct’ for all circumstances. If the object is for an ASLR method to handle natural sign language, then the insights from sign perception research must be integrated into ASLR.
4.1 Introduction

The current tendency in man-machine interaction is toward more and more natural methods of interfacing. It is not surprising, therefore, that automatic gesture processing is receiving a lot of attention. Most of this attention is focused on technical issues. However, to process natural gestures, attention must also be paid to the nature of human communication. The importance of this element depends on the application for which the gesture processing is to be used. In natural language processing or language teaching environments, the human criteria for communication must be adapted, because ultimately the machine’s goal is human communication. Current gesture processing largely ignores this issue, concentrating mostly on data-driven methods. This can lead to undesirable results, an example of which is the outlier problem: a gesture executed in a strange, but nevertheless correct way, will most often be rejected by a statistical gesture processing method. Such a method learns the definition of the gesture from example data. If the strange variant was not in its training data, then it will reject it henceforth. ‘Common’ and ‘correct’ are largely equivalent for such methods. For humans, the notion “uncommon but correct” is conceivable. Apparently, humans employ different criteria for judging correctness than mere frequency of occurrence. If a machine algorithm is to process natural gestures, it must adapt similar methods and cannot rely on data-driven methods alone. More knowledge on human gesture processing must be gathered, and ways of using this knowledge in automatic gesture processing must be developed. This not only could solve problems such as the one described above, it may also benefit gesture processing in other ways, such as reducing computational load and removing the need for large training sets.

Many gesture processing techniques are first tried out in the field of sign language recognition, since sign language signs are gestures that are both complex and well-defined. Automatic sign language recognition (ASLR) is also an interesting field in its own right, since it can help create better communication aids for the deaf and interactive tools for sign language learners. The research described in this paper is all done on Sign Language of the Netherlands (SLN). However, many of the principles can be extended to gestures in general. Current ASLR, like other gesture processing areas, mostly focuses on problems of information acquisition (how to obtain information from image streams or instrumented gloves) and data-driven classification (which method to use, how to construct training sets, etc.). However, for successful
processing of natural sign language, other issues must also be addressed, such as: which elements of a sign are informative? Are all parts of a sign equally important, and if not, where should we focus our attention and resources? And how much variation is allowed in the execution of a sign? This paper discusses the impact of such issues on ASLR. A set of previously performed sign language perception experiments is presented, their results are compared and several practical applications are proposed. The impact of the insights gained on sign language processing theory is also discussed.

4.1.1 Related Work

ASLR research is comparatively young: most research has been done since the 1990s. The majority of these studies either give no theoretical background to motivate their choice of features to represent a sign, or limit themselves to mentioning that handshape, hand orientation, hand motion and hand location determine the meaning of sign language signs and then extract (a subset of) features representing these qualities [8, 10]. Studies using instrumented gloves to collect data (e.g. [55, 99]) mostly use all sensors available. Studies using computer vision often limit the sign representation to location and motion plus some basic shape characteristics, because handshape and hand orientation are hard to extract (e.g. [79, 101]). Or they choose to use rotation- and scale invariant features [40]. Lee et al. [55] model handshape, hand orientation and motion separately, but do not describe how they arrive at values for these elements. The same is true for Bowden et al. [10], who categorize motion, handshape, hand configuration and location into distinct categories, but do not describe how they arrive at the categories. Derpanis et al.[23, 24] and Vogler and Metaxas [94, 95] are among the few who give a motivation for their sign representations, which (partly) reflect existing sign phonological models. On the whole, little use is made of theoretical knowledge in the field of ASLR. An extensive overview of ASLR research is given in [64].

4.2 Information Distribution in the Time Domain

Three experiments were performed to gain insight into the distribution of information over a sign in time. A sign (or gesture) can be divided into certain phases in time. The three main phases are (in chronological order) preparation, stroke, and retraction [51, 62]. These terms originated in the study of co-speech gestures, but are used for sign language signs as well. In figure 4.1 the phases are shown schematically in white, gray and black. Preparation is the phase in which the hands are moved to the location and configuration necessary to perform the intended action (e.g. if the intended action is at eye-height and the hands are at the table, the hands must first be lifted). The stroke is the phase containing the desired action, and the retraction contains the
relaxation back to the rest position or to the next movement. In fluent signing, the retraction of a sign can merge with the preparation of the next sign. In the experiments described here, however, all stimuli were isolated signs starting from and ending in a certain rest position, so all three phases were always present. In the experiments it was investigated which moments in a sign facilitate human sign detection and sign recognition. The first two experiments investigated how early in a sign recognition can take place, to learn how much of a sign is actually needed. The third experiment investigated the information content of sign phases in isolation.

4.2.1 Experiment 1: Detecting a Sign

This experiment was set up to investigate at which moment of a movement signers realize that it is a sign, and not a different movement. Participants saw a movie and were asked to press a button as soon as they saw a sign beginning. The movies contained either a sign, or a non-communicative movement (a fidget) followed by a sign, or two fidgets. The fidgets were distractors that made sure participants could not
simply react to any movement. Thirty-two signs and nine fidgets were used to make one hundred and twelve stimuli. Sixteen signers and eight non-signers participated in the experiment. A more extensive description can be found in [5].

4.2.2 Experiment 2: Recognizing a Sign

In this experiment, the same setup and stimuli were used as in experiment 1, but this time the participants were asked to press the button as soon as they recognized a sign, and were then asked to identify the sign. The purpose was to determine at which moment in a sign signers can identify it. Thirty-two signers participated (no non-signers were included since they would not be capable of identifying a sign). See [6] for more details.

4.2.3 Experiment 3: Recognition of Isolated Sign Phases

This experiment was set up to investigate the information content of the different phases of a sign. In the experiment signers were asked to view isolated fragments of a sign movie and try to identify the sign. The fragments corresponded to the phases preparation, stroke and retraction. Because the boundaries between these sign phases are not sharp, three independent annotators were asked to determine them. Using their annotations (shown as vertical bars in figure 4.1), the segments preparation, stroke, and retraction were extracted from sign movies as depicted in figure 4.1. Because previous research indicated that the transition between preparation and stroke may be especially salient [5, 30], both preparation and stroke were extracted with and without this region included. The difference between the scores of these pairs would reveal the information content of the transition region. All segments were extracted from twenty different signs (a subset of the material used in experiments 1 and 2), and twenty-seven signers participated in the experiment. See [89] for more details.

4.2.4 Results

Here, the combined information gathered in experiments 1–3 for the overlapping stimuli (fifty-two stimuli from twenty different signs) is discussed. Figure 4.2 shows the time lines of ten example stimuli as horizontal bars. Descriptions of the signs and details on phase border determination can be found in [89]. The vertical black ticks indicate the phase borders (with transition areas between the end of the preparation and the start of the stroke, and between the end of the stroke and the start of the retraction). On top of the time line, histograms of the detection and recognition times from experiments 1 and 2 are shown in white and black respectively. The histograms were calculated by dividing each sign into a constant number of bins, regardless of
Figure 4.2: Overview of recognition scores and detection and recognition times for ten signs (experiments 1–3). The horizontal bars depict the time lines of the signs. The vertical black ticks indicate the phases borders (with transition areas between the end of the preparation and the start of the stroke, and between the end of the stroke and the start of the retraction). The four main phases are indicated by shades of gray. The segment P is put below the segment P+ because they partly overlap. On top of the time line, normalized histograms of the detection and recognition times from experiments 1 and 2 are shown in white and black respectively. Below the time line, white and black triangles indicate the median detection and recognition time respectively. The recognition percentages of the phases in experiment 3 are shown as bar graphs to the left of each time line. The segment S+ is not shown, since it gave results identical to those of S.

sign length. Each histogram was normalized separately (divided by the maximum), so that all values lie between zero and one. The triangular marks below the time line indicate the median detection and recognition times, also in white and black
respectively. The four bars to the left of the time line show the recognition percentages of the four fragments from experiment 3. A t-test confirmed that detection takes place earlier than recognition. Averaged over all overlapping stimuli, detection takes place after 28% of the entire sign, recognition after 34%. Both detection and recognition were positively correlated to both the start of the transition and the start of the stroke (determined as shown in figure 4.1). Detection happened on average after 1% of the stroke and recognition after 18%. Apparently, detection and recognition are possible on the basis of just the first part of the sign — not even the entire stroke-phase is necessary. The results of experiment 3 were in accordance with these findings. The stroke in isolation gives recognition scores not significantly different from 100%, and adding the transition to the preparation increases recognition scores to the level of the stroke, confirming that the first part of the sign contains enough useful information to make recognition possible. There was no correlation between the length of the preparation (without transition) and the recognition times, nor between the recognition score of the preparation and the recognition times. Preparation length and preparation recognition score had no correlation with detection times either. The graphs in figure 4.2 show that there can be large differences between individual signs regarding the usefulness of the information contained in preparation, retraction and transition. The detection and recognition times of individual signs show less variation.

4.3 Informative Elements of Signs

Signs can not only be subdivided in time, they can also be studied in terms of their informative elements. Signs are linguistically equivalent with words, and like words, they consist of smaller elements that determine their meaning. In words, these elements are called phonemes. In sign languages, the corresponding elements are not clearly described yet. There is considerable discussion on which elements would play the role of sign language phonemes, and what their organization should be [29]. However, it is generally assumed that hand location, hand orientation, hand motion and handshape are important for the meaning of a sign, since pairs of signs (minimal pairs) can be found that differ only in one of these elements, and yet have different meaning. Other qualities of signs, such as speed of movement or scale, are generally not considered important for determining meaning. It is not clear, however, if each of these informative elements is equally salient. Nor is it clear how much variation is allowed within the elements. For example, if a position value is described as “in front of the eyes”, how large is this area exactly? Is the allowed variation the same for all position values, or are some values vaguer or more ‘lenient’? And how is this for the other elements?

The experiments described below were performed to increase knowledge on these matters. In the first, handshape and hand orientation were separated using image
processing techniques, to investigate their salience for sign recognition. The second experiment investigated how much the informative elements of a sign can be changed before it becomes unacceptable as an example of that sign.

4.3.1 **Experiment 4: Importance of Hand Configuration for Recognition**

Signers were asked to view sign movies in which the hand configuration was obscured to varying degrees. The term “hand configuration” is used for the combination of handshape and hand orientation. They were then asked to try to identify the sign. In condition 1 (C1) the hand was entirely covered, so that only location and motion could be observed. In condition 2 (C2), a representation of palm orientation was shown on top of the covered hand. In condition 3 (C3), a crude version of the handshape silhouette was shown on top of the covered hand. Figure 4.3 (a) shows an example of a stimulus sign in conditions 1–3. Condition 4 (C4) was unmodified and used as a control condition. Twenty signs were used, in four conditions, giving a total of eighty stimuli. Participants saw each stimulus 5 times and scored between 0 and 5, depending on how many showings they identified correctly. Twenty-seven signers participated. The experiment was described more extensively in [88]. The results are shown in figure 4.4. In C1 most signs cannot be recognized. C2 hardly differs from C1. In C3, scores vary greatly, both between signs and within a sign (between participants). An analysis of variance confirmed that there was no difference in score between C1 and C2, and that C3 gave higher scores, though still not as high as in the unmodified condition (not shown). We conclude that location and motion alone (C1) are in general not enough for successful sign recognition. Adding hand orientation information (C2) does not improve recognition scores, but adding crude handshape information (C3) does, although the amount with which the scores improve varies between signs. A few signs can be recognized well even in C1 and C2.

4.3.2 **Experiment 5: Allowed Variation within Elements**

This experiment was designed to learn more about which variations are acceptable in signs. To this end, four signs were recorded in a number of different variations. Variations included changing the direction of movement, the scale of movement, the speed of movement, the handshape, the hand orientation, etc. An extensive description can be found in [2, 4]. There were one hundred and thirty-one stimuli in total. Figure 4.3 shows an example. Signers were asked to view the movies and indicate on a five point scale how acceptable the sign was. Twenty-six signers participated.

The results show that participants vary widely in the strictness of their judgments, with some participants rejecting far more variants than others. However, the order of
Figure 4.3: (a) Example of stimuli in experiment 4. Two frames of the sign FLAG are shown in conditions 1 (left column), 2 (center column) and 3 (right column). (b) Example of stimuli in experiment 5. Two frames of the sign SAW are shown: in the ‘normal’ form (left column), with modified movement direction (center column) and with modified hand orientation (right column).
Figure 4.4: Results of experiment 4. Box plots of recognition scores (0–5) per sign in conditions 1 (no hand configuration), 2 (hand orientation visible) and 3 (crude handshape visible). The black line indicates the mean score, the thick gray bar within the box is the median. Circles and asterisks are outliers and extreme values respectively.

Figure 4.5: Results of experiment 5. The bars show acceptability of different types of variation in a sign. Temporal means that the timing of elements was changed, spatial means changes in location, direction, etc. Spatio-temporal means variations that combine the two types.
acceptability for the movies tends to be highly correlated between participants (using Kendall’s tau). Closer inspection shows that acceptability ratings also vary according to the type of variation, as shown in figure 4.5. Temporal variations (changing the speed of the movement or inserting holds) appear to be acceptable. Spatial variations, on the other hand (e.g. changing the location) are generally unacceptable. Variations combining the two are only slightly more acceptable than spatial variations. Apparently, temporal variations do not affect acceptability much, whereas spatial variations quickly cause a sign to be considered unacceptable. However, iconicity is a complicating factor here. All four test signs contained iconicity, i.e. they mimicked the object or action they represented. Some spatial variations damaged the iconicity, and it may be for this reason that those variations were unacceptable. Temporal variations hardly ever damage iconicity. Also, because of iconicity some variations were acceptable for certain signs but not for others. For example, tilting the hand would be acceptable in the sign ‘to saw’, but not for the sign ‘to carry a tray’. In the first case, it is a conceivable modification of the action the sign depicts: sawing at an angle. In the second case, it is not: tilting the tray would cause the contents to fall off. Through iconicity, the meaning of a sign has an influence on which variations are acceptable. This complicates the formational rules for signs.

Human and Machine Judgment Compared

The human judgments from experiment 5 were compared to the judgments of three different ASLR algorithms [57]. These algorithms were trained to assess the correctness of a sign. They were trained on 120 signs, executed in regular form, from 75 different signers. The algorithms learned the definition of one sign by comparing the 75 examples of the sign to the 119 x 75 other examples. In this way, they learned which feature values represented ‘normal’ variation of the target sign, and which indicated a truly different sign. The algorithms were tested on a subset of the sign variants created for experiment 5: only three of the four signs were used, and all variants that were invisible to the machines were removed (such as handshape variations, which would ‘look’ identical to the machines because they did not extract detailed handshape information). The correctness values given to these sign variants by an algorithm were regarded as machine acceptability ratings. The variants were ranked according to these ratings, and the resulting rankings were compared to the rankings human signers had made in experiment 5 (see [3] for more details). Somewhat surprisingly, the rank correlation (Kendall’s tau) between the machine and human rankings was low (average 0.30). In contrast, the correlation among human rankings was generally high (average 0.68), as was the correlation among the three machine rankings (average 0.61). Apparently, training on a set of ‘normal’ signs does not enable a machine algorithm to extract an accurate (human) model of what a ‘correct’ sign is. Also, the ASLR method that gave the highest average correlation was not the one with the best classification performance. This means that classification performance on regular signs is apparently
not a reliable predictor for success in acceptability assessment of deviant sign forms. Table 4.1 shows the human-machine correlations per individual sign. The degree of correlation seems to be related to both the algorithm and the sign it is used on, suggesting that there is not one algorithm that always gives the most human-like ratings, but that this differs per sign.

**Table 4.1**: Correlations between acceptability judgments of three ASLR methods with the average human acceptability ratings, for the signs CURTAIN, OVEN and (HAND)SAW. The rightmost column gives the average correlation for the three signs. For comparison, the bottom row shows the correlation of human acceptability judgments with the average human judgment (average of the correlations of individual humans with the average human). It is clear that the human-human correlation is higher than the human-machine correlation for all three signs. The human-machine correlations vary both with the type of ASLR algorithm and with the specific sign involved. This may be because some algorithms are weaker in recognizing certain types of sign than others. But more complex factors, such as iconicity, may also play a part in explaining the lack of correlation for some algorithms and signs (e.g., CDFD on sign OVEN).

<table>
<thead>
<tr>
<th>Recognition Method</th>
<th>CURTAIN</th>
<th>OVEN</th>
<th>SAW</th>
<th>Average of three signs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDTW</td>
<td>0.49</td>
<td>0.10</td>
<td>0.34</td>
<td>0.31</td>
</tr>
<tr>
<td>CDFD</td>
<td>0.40</td>
<td>0.02</td>
<td>0.41</td>
<td>0.28</td>
</tr>
<tr>
<td>QDFFM</td>
<td>0.33</td>
<td>0.14</td>
<td>0.43</td>
<td>0.30</td>
</tr>
<tr>
<td>Humans</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

### 4.4 Discussion

#### 4.4.1 Practical Application

The knowledge derived from the presented experiments can be given practical application in ASLR or gesture processing. For example, the time domain experiments reveal that it is not necessary to process the entire sign. Processing only the first half, or only the stroke phase, would give enough information to identify a sign. The discarded parts would still contain useful, if redundant information (segments P and R do enable some recognition), and using more information is often desirable. However, this should be weighed against the increase in computational costs and the extra variation that such parts can contain. For example, preparation and retraction of a sign can be very different for the same sign depending on where the hands are located before the sign and have to go afterward. Such variable phases are hard to model.
statistically, and in the end they would probably prove unhelpful for classifying the sign. In which case it would be better to disregard them from the start.

Focusing on certain sign phases does of course presuppose the ability to detect such phases automatically in the data stream. Unfortunately, automatic detection of movement phases is not a trivial matter [100]. A practical solution would be to annotate the phase borders of a single example per sign by hand, and to find the corresponding phase borders in other examples of the same sign through a synchronizing algorithm such as dynamic time warping [54].

In the domain of the informative elements both experiments have a fairly direct application. Experiment 4 showed that handshape is more salient for sign recognition than hand orientation. So if an algorithm using only the location and motion elements (most current algorithms) is to be extended, it would seem most beneficial to add the handshape element to the representation. Although in practice, detecting the hand orientation may be necessary in order to correctly classify a handshape. The participants in experiment 4 may also have derived some notion of orientation in condition 3, even though they only saw the handshape silhouette. Experiment 5 yielded that temporal variations in signs are relatively harmless, whereas spatial variations more quickly cause a sign to become unacceptable. This shows us that in ASLR, modeling temporal patterns is probably not useful. Such patterns can still be typical, of course. However, they are apparently not necessary: a sign example not following the pattern is still a valid version of the sign, as demonstrated in experiment 5, so an ASLR algorithm should not reject it. This is an example of how perceptual knowledge can be used for the correct handling of atypical examples (outliers), something which is hard to achieve for a data-driven method.

Experiment 5 and the human-machine comparison described in section 4.3.2 also yielded insights that are not immediately applicable, but nevertheless important for ASLR. One striking matter is the fact that determining the “ground truth” for ASLR (which variants should the algorithm recognize as a certain sign?) is not trivial: human observers vary in their determination of what can still be called a ‘correct’ version of a sign, with some being more strict, others more lenient. Therefore, choosing a strictness level seems necessary before variants can be judged as ‘correct’ or ‘incorrect’. Furthermore, the amount of variation that is allowed for a certain sign element appears to depend on a number of factors. The type of variation is important (temporal vs. spatial), and possibly even the exact nature of the element. Then there is the presence of iconicity: if a variation damages the iconicity, it immediately seems to become unacceptable. Knowledge on these factors could enable us to construct rules for allowed variation in every situation. However, to do this, more research is necessary. Current ASLR methods try to learn such rules from training data, but do not arrive at the same criteria as human beings. Training the methods on more deviating sign variants may improve their performance. However, constructing such variants and determining a ground truth (which variants are ‘correct’) remains a problem.
4.4.2 Implications for Sign Recognition Theory

In experiments 2 through 4 the recognition of sign language signs under difficult circumstances was investigated — information was sparse or the response had to be as quick as possible. The results differ for individual signs, with some signs being recognized earlier (experiment 2), or yielding higher scores in the information-sparse conditions (i.e. segments P and R in experiment 3, and conditions 1 and 2 in experiment 4). No single sign characteristic explains these differences, i.e. there is no significant correlation between better score and rarity of handshape, better score and presence of a path movement, etc. However, in the experiments with sparse information, the well-recognized signs tend to contain more ‘restrictive’ elements than the poorly-recognized signs. We define restrictive elements as elements that only occur in a relatively small group of signs. For example: a rare handshape that only occurs in a small number of signs is a restrictive element. Presence of a path movement is also a restrictive element: the group of signs with no path movement is larger than any group of signs sharing a specific path movement. The presence of such elements is beneficial for recognition, because they limit the set of possible signs (assuming a cohort model of sign recognition [61]). The following elements seem to function as restrictive elements: rare handshapes, path movements, a change in handshape or hand orientation, and repetition. Well-recognized signs appear to contain more of these elements than poorly-recognized signs. However, the large number of factors and small number of signs in the dataset makes it impossible to test this statistically. More research with a larger sign set, designed to separate out these factors, is necessary.

In the experiments that focus on a quick response, an analysis of variance showed that the presence of a rare handshape causes quicker recognition. The other restrictive elements listed above give no effect, except for the presence of a hand orientation change which actually causes recognition to be slower. This may be due to the difference in tasks: when timing is an issue, a stimulus that is high in information-density may cause a delay, because more information needs to be processed. When timing is not important but information is sparse or obscured, high density of information becomes a benefit, because a small part of the stimulus may still be very informative. Another possibility is that recognizing patterns for which several frames of a movie are necessary (such as path movements and handshape/hand orientation changes) takes time, whereas spatial patterns (such as handshapes and locations) can be observed at a glance. This would explain why the restrictive elements that are patterns in time do not cause quicker recognition in the experiments where a fast response is needed, while one of the spatial patterns does have an effect. Once again, more data are needed to test the influence of various elements on recognition performance.

Studying the data from several perceptual experiments has given us more knowledge on the nature of sign language recognition. We have gained a fairly good insight into
the importance of different time phases within a sign, with the stroke being the most important phase and the beginning of the stroke a very salient moment. The matter of informative elements within a sign remains more difficult, mainly because of the large number of factors that appears to play a role. Some straightforward conclusions can be drawn about the salience of the handshape element and the relative insignificance of timing within a sign. However, to really gain insight into allowed variation, more research is necessary. Since this is an area in which ASLR can profit substantially, solving the problem of intra- and inter-signer variation, it would certainly be worthwhile.

**Acknowledgments**

The research discussed in this paper was made possible by the Delft ICT Research Centre and the VSB-fund, grant number 2003451. The authors would also like to thank the Dutch Foundation for the Deaf and Hard-of-hearing Child and the Royal Auris Group, who were partners in this project.
Part II

APPLYING SIGN PERCEPTUAL KNOWLEDGE
Influence of Handshape Information on Sign Recognition

Abstract

Research on automatic sign language recognition (ASLR) has mostly been conducted from a machine learning perspective. We propose to implement results from human sign recognition studies in ASLR. In a previous study it was found that handshape is important for human sign recognition. The current paper describes the implementation of this conclusion: using handshape in ASLR. Handshape information in three different representations is added to an existing ASLR system. The results show that recognition improves, except for one representation. This refutes the idea that extra (handshape) information will always improve recognition. Results also vary per sign: some sign classifiers improve greatly, others are unaffected, and rare cases even show decreased performance. Adapting classifiers to specific sign types could be the key for future ASLR.

The material in this chapter was previously published as:
5.1 Introduction

With imaging hardware such as cameras becoming cheaper and more advanced, the interest in ambient intelligence and natural, multi-modal (i.e. speech and gesture) interfaces has grown over the last decade. As a consequence, the study of automatic sign and gesture recognition has been gaining attention steadily. Sign language signs (in their citation form) are more formal and more strictly defined than general gestures (e.g. co-speech or emblematic gestures). As such, recognition of (isolated) sign language signs is often used as a starting point for general gesture recognition methods, but it has a merit of its own, too: automatic sign language recognition techniques can facilitate communication among the deaf, and also between deaf and hearing people.

Automatic sign language recognition (ASLR) has been studied since the early 1990s — see [64] for an overview of ASLR methods, and [97] for an impression of the current state of the art. Focus has mostly been on methods for capturing a sign, on the choice of feature representation of a sign and on the pattern recognition algorithms to recognize or detect a sign. Derpanis et al. [23] and Vogler and Metaxas [95] are among the few who try to use sign linguistic information in their ASLR approaches. In general, little attention has been paid to the nature of signs and sign processing: issues such as “What are the important characteristics of a sign?”, “Are all parts of a sign equally informative?” and “How much and what kind of variation is allowed in a sign?”.

In previous work, we addressed such questions, performing recognition experiments with human signers to gain insight into the human sign recognition process [88, 89]. In [86], it was discussed how the results of these studies could be applied in automatic sign language recognition. One of the conclusions was that handshapes, even in a simplified form, may be helpful for automatic sign language recognition. Because of the environment our recognition system is used in, a program for training sign language vocabulary, it is aimed specifically at recognition of isolated, citation-form signs. We therefore do not address issues of continuous sign language recognition here.

This paper describes the results of an experimental study on the implications of adding simplified handshape information to an existing ASLR system. Though handshape is an integral part of a sign language sign, our ASLR system can function reasonably well without it [57]. For this reason, we do not adopt prior assumptions about the merits of adding handshape for this system. During the study, it was noticed that results can vary substantially depending on which subdivision of the data is made to form training- and test sets. Therefore, the experiments described here were performed with several different subdivisions of the dataset and results were evaluated together. It was found that handshape information can improve recognition performance, but that results depend on the handshape representation that is used. A second interesting finding is that individual sign classifiers respond differently to the addition of
5.2 Method

5.2.1 Dataset

The dataset for training and testing consisted of 91 isolated signs from the standard lexicon of Sign Language of the Netherlands Standaard Lexicon Nederlandse Gebarentaal, deel 1. Stichting Nederlandse Gebarencentrum, 2002. Each sign was recorded from 75 different persons (mostly non-signers). Signs were captured using two synchronized Allied Vision Technologies ‘Guppy’ cameras, at 25 frames per second, resolution 640 x 480 pixels. Examples of the signs in the dataset can be viewed online

http://research.tenholt.nl/GW09.

5.2.2 Sign Recognition Algorithm

For this experiment, the ASLR method described in [57, 59] was used. An overview of its architecture is shown in figure 5.1. Stereo cameras are used to capture the sign. A
skin colour model is then applied to the images, so that head and hands can be found and tracked. From the hand location in the stereo images, the 3D hand positions can be calculated, as well as several other characteristics, such as velocity, acceleration, and direction of motion (see [59] for the complete list). These feature types are extracted for each frame in a sign recording, resulting in a \([\text{number of frames} \times \text{number of feature types}]\) matrix, which is the feature representation of a sign recording (in figure 5.6, examples of such matrices can be seen). Each combination of feature type and frame is henceforth considered one feature. The feature values are smoothed in the time dimension using a 3-frame median filter (for outlier removal), followed by an 11-frame Gaussian filter (for further smoothing). Dynamic time warping is used to align sign examples.

With the feature representations, classifiers can be trained. For each sign in the set, a separate classifier is trained to distinguish that sign from all other signs. Examples from the target class are used as target training examples, examples of all the other classes are taken together as one large non-target class. For training, the Combined Discriminative Feature Detectors (CDFD) method is used (see [57] for details). In short, training consists of a feature selection step (based on a feature’s usefulness in separating the target class from the non-target class), followed by a training of feature detectors for the selected features. Unknown signs are recognized by first applying the feature detectors of the selected features, and then combining their outputs to determine if the sign is recognized or not. The sign is recognized if a certain fraction of the selected feature detectors gives a positive response. This fraction is determined in the training phase. The sensitivity of the feature detectors can be varied to create stricter or more lenient versions of the classifiers.

5.2.3 Handshape Representation

The system described in [59] was trained without the use of handshape information. To add handshape information, first a representation must be chosen. There are several possible representations of handshape [15]. Because our system is required to work real time, the method must be fast. In combination with the fact that the dataset contains no zoom images of the hands, this means that fitting detailed 3D hand models is not an option. Speed and robustness considerations caused us to select a relatively simple but efficient shape representation. This representation is formed by taking the bounding box of a hand in a certain frame from the image (the bounding box is always aligned with the main axes of the frame), segmenting the subimage into hand and non-hand pixels, dividing the subimage into a fixed number of squares, and calculating the ratio hand/non-hand pixels for each square (see figure 5.2 (a)). Together, these ratios form a feature representation of the handshape, which is translation and scale invariant, but not rotation invariant. This means that the representation carries some information about hand orientation as well as shape (the
Figure 5.2: Procedures for generating handshape representations. (a) Ratios of 4 x 4 grid (R44). The bounding box of the hand is lifted from the frame and divided into subregions. The image is then segmented with the skin color model, and the ratio skin/background is calculated for each subregion. Together, the ratios form the shape representation of the hand. (b) Ratios of 2 x 2 grid (R22). Same method as for (a). The third representation (INV) is made by taking the first seven invariant moments [41] of the raw hand blob shown in (c).

same handshape under a different orientation results in a different feature description). The number of squares used in the method can be varied. We chose a 4 x 4 grid (R44), resulting in 16 values for each hand, so 32 extra feature types in total.

For comparison, two alternative methods are tried as well: a more crude version of the above one in which a 2 x 2 grid is used (R22), and the invariant moments of the hand blob (INV), a representation which is translation, scale and rotation invariant [41]. For the latter, the first seven moments are taken for each hand. These representations result in 8 resp. 14 extra feature types in total. Figure 5.2 (b) and (c) refer to the alternative representations.

5.2.4 Missing Values

Signs where the hands overlap with each other or with the face form a problem for the handshape representation, because the skin segmentation results in a merged blob, the shape of which is not necessarily informative or representative for the shape(s) of the hand(s) involved. It was decided to extract no handshape features in these cases. As a result, the dataset is no longer complete — some feature representations contain missing values for certain handshape features. If the features were missing for only a few frames (maximally 5), the values were interpolated using a Gaussian filter in the
time dimension (integrated into the filtering step). If feature values were missing for only a few examples in the dataset (maximally 10%), average values (calculated over the other examples) were imputed.

Remaining missing values in the dataset are handled as follows: if missing values remain for a feature in examples of the target class of a classifier (e.g. in the examples of the sign ‘SAW’ when training the classifier for ‘SAW’), then this feature is rejected in the feature selection stage. Missing values in examples of the non-target class are ignored. During testing, missing values are ignored: the fraction of detected features is calculated over the existing features only.

5.2.5 Evaluation Method

To investigate the effects of handshape information, classifiers are trained and tested on the dataset with and without handshape features added, and the results are compared. Each classifier is trained and tested according to a five-fold cross-validation scheme. The cross-validation is repeated 10 times for different random orderings of the data. For each repetition and each cross-validation fold, four versions of the classifiers (with the three handshape representations and without handshape features) are trained and tested.

5.3 Results

5.3.1 Recognition Results

The results of the classifiers are shown as Receiver Operating Characteristic (ROC) curves (plots of the true positive rate of a classifier against the false positive rate, for a number of strictness settings). Figure 5.3 shows the average ROC curves over all signs, cross-validation folds and —repetitions for the original classifiers (No HS) and the three classifiers that result from training with various handshape representations. R44 produces the best curve, and R22 gives improvement as well. Adding INV representations actually worsens classification performance. To investigate whether this trend is significant, a one-way, repeated measures ANOVA was performed on the areas under the curve (AUCs) of the ROCs of all classifiers, cross-validation folds and —repetitions. There was a significant effect for handshape representation \( F_{3,13647} = 546.20, p < 0.0001 \). Post hoc tests, using the Bonferroni correction for multiple comparisons, showed that the four methods all differed significantly from each other (all \( p < 0.0001 \)), with R44 giving the best performance, followed by R22, No HS, and INV — the last handshape representation made performance significantly worse. The explanation may be that the INV representation is more sensitive to noise
5.3 RESULTS

Figure 5.3: Average ROC of recognition results for all four conditions over all signs, cross-validation folds and cross-validation repetitions \(N = 4550\). No HS = no handshape representation, INV = invariant moments, R22 = ratios of 2 x 2 grid, R44 = ratios of 4 x 4 grid.

than the other two: small changes in the skin segmentation can cause differences in the details of the hand blob shape, and thus in the higher-order shape moments. The ratio-representations are more robust against this type of noise, since the variation averages out over a subregion.

The same tests were performed for partial AUCs (AUCs for the false positive (FP) range \([0 – 0.1]\) of the ROC). The partial AUC (pAUC) is interesting because the operating point chosen for a classifier will most often lie in this range — a classifier with an FP rate of more than 10% is generally not acceptable in practice. The results for pAUCs were similar to the results for AUCs \(F_{3,13647} = 859.57, p < 0.0001\), all pairwise comparisons significant at the 99.99% level).

5.3.2 Individual Signs

Experiments with human signers showed that there are usually great differences between results for individual signs. In this ASLR experiment, the same effect occurs. First of all, the ROCs of individual signs can differ substantially (see figure 5.4). This is to be expected — the characteristics of some signs will make them more easily recognizable than others. However, there are also differences in the effects of adding handshape information. With addition of R44 features, most signs (75%) improve, but for some signs performance does not change, and for a few, there is even a decrease in performance. Paired t-tests per sign (using cross-validation folds as repetitions) were
performed to investigate which sign classifiers benefit from the addition of R44 handshape information. Two measurements were used: the AUC and the partial AUC.

Table 5.1 summarizes the findings. For the more interesting partial AUC only one sign shows a decrease in performance. This sign, BIRD, is made by flapping the arms and hands, which will cause a lot of variation in the precise handshapes and orientations. This, combined with the fact that the number of repetitive ‘flaps’ is arbitrary, makes the handshape characteristics of the sign quite variable, and thus possibly confusing rather than informative. For twenty-two signs, there is no significant difference. This is partially due to a ceiling effect: four signs already show near-perfect recognition without handshape. Another cause are the missing values: for three signs, the hands overlap for almost the entire duration of the sign, so that nearly all handshape features are missing for all examples. In these cases, it is logical that handshape information has no influence. The rest were often signs which, like BIRD, contained a lot of handshape variation, or signs with fist-like handshapes. Figure 5.5 shows the average partial AUC for all signs both with and without R44.

---

**Figure 5.4:** ROCs for No HS (dashed) and R44 (solid), for different signs and cross-validation folds. Within each column, the difference between versions of the same sign when trained and tested with a different subdivision of the data can be seen. Within each row, differences between individual signs can be observed. (b) and (c) show cases where handshape information is not helpful.
Table 5.1: Classification performance per sign after adding handshape information in representation R44. Significant increase or decrease in performance was determined through paired t-tests per sign over all cross-validation folds and repetitions, $\alpha = 0.05, df = 49.$

<table>
<thead>
<tr>
<th>Measurement</th>
<th># Increase</th>
<th># No Difference</th>
<th># Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>69 (76%)</td>
<td>19 (21%)</td>
<td>3 (3%)</td>
</tr>
<tr>
<td>partial AUC [0 - 0.1]</td>
<td>68 (75%)</td>
<td>22 (24%)</td>
<td>1 (1%)</td>
</tr>
</tbody>
</table>

Figure 5.5: Average partial AUCs (FP range [0 – 0.1]) of all signs with and without R44. Averages are taken over all cross-validation folds and repetitions. The error bars represent the standard error of the mean.

5.3.3 Feature Selection

Feature selection plays an important role in the CDFD-classifier. Features that are not selected will not have any influence on classification. Since features are selected based on their ability to separate the target class from other movements, studying the feature selection can tell us more about which handshapes are informative for ASLR. Figure 5.6 shows the feature selections for a few classifiers trained with R44. The selections are averaged over all cross-validation folds and repetitions (for each cross-validation fold, a different classifier with a different feature selection is trained). The gray level of
Figure 5.6: Average feature selections for four signs over all cross-validation folds and repetitions for signs with R44. Grey level indicates how often a feature was selected (white = always). Feature types 1–25 are the original, non-handshape feature types, types 26–41 are R44 of the right hand, 42–57 are R44 of the left hand. BIRD and FROG are two-handed signs, WAITER and MICROPHONE are one-handed.

A feature indicates how often it was selected, with white representing selection always. For classifier BIRD, the sign that drops in performance when R44 features are added, handshape features are selected. Apparently, these features appear informative during training, but prove confusing for the test examples (an indication that our feature selection method might be improved). For classifier WAITER, many features are selected, and this classifier shows a clear increase in performance. Classifier MICROPHONE is an example of what happens when almost all handshape features are missing (in MICROPHONE, the hand is in front of the mouth, so there is facial overlap for almost the entire sign). FROG, finally, is an example of a classifier for which the feature selection varies (like for BIRD), but for which it has no significant effect on the recognition performance.

To investigate the influence of handshape feature selection (R44) on performance, the average increase in pAUC and the average number of selected handshape features were calculated for each sign. Then the linear correlation between the two was
Figure 5.7: Scatterplot of the average number of selected R44 handshape features versus the average increase in partial AUC (averages calculated per sign).

Calculated. There is a slight positive correlation, which is not significant for $\alpha = 0.01$ ($\rho = 0.21, p = 0.023$). Figure 5.7 shows the scatterplot of the two variables. It mostly shows that when handshape features are selected, they are generally not harmful, though exceptions (such as BIRD) exist.

5.4 Discussion

5.4.1 The merits of adding handshape

From the results of this experiment, it is clear that adding handshape information is in general beneficial for ASLR. However, this is not automatically true for every handshape representation. For the classification method described here, using invariant moments of the handshape blob decreased performance, probably because of its greater sensitivity to noise in combination with the low resolution of the handshape images. The more robust R22 and R44 representations did increase recognition performance. A second explanation for their success may be the fact that the ratio-representations intrinsically carry some hand orientation information as well as shape information, because they are sensitive to rotation. The INV representation is rotation invariant, which may be a disadvantage in the current situation: the orientation of the hand can be useful information for detecting a sign. But rotation sensitivity can be a disadvantage, too. Signs that show no increase or even a decrease in performance with R44 information added, are in many cases signs with much translational and rotational hand movement combined (such as BIRD or ELEPHANT). For these signs, individual examples will differ in the rotation of the handshape at different moments in the sign (because individual signers perform the combinations of
translational and rotational movements in different tempos). A rotation-invariant representation would not reflect these variations (as long as the rotations were not out-of-plane). The ratio-representations, on the other hand, may reflect so much of this noisiness that the information is no longer useful for classification, or even becomes confusing.

It is not possible to give characteristics that clearly connect the signs that did not benefit from handshape information. The analysis is made difficult because there are three possible causes: a ceiling effect (the classifier can hardly improve any more), a missing values problem (there are hardly any handshape features present), or the sign is truly 'handshape-indifferent'. Removing the first two categories leaves only about fifteen signs which do not improve. These tend to contain the large variability in handshape and orientation mentioned above, or to have fist-like handshapes, which occur commonly and thus may not be helpful in distinguishing signs. However, the set is too small to test the significance of said tendencies. The set of signs that does benefit from handshape information contains all kinds of shapes, including fists, but no signs with great variability in path and orientation (like the ones mentioned above). The results demonstrate that variability in performance between individual signs occurs in ASLR as well as in human signer experiments. Future work should consider adapting recognition algorithms for individual signs, or perhaps for sign types or handshape types.

5.4.2 Applying human recognition insights to machine recognition

This study was set up to implement our previous findings from human sign recognition research [86] in ASLR. Direct comparison between the results here and the outcome of our human signer experiments is not possible, because different sign sets had to be used. However, certain tendencies can be compared. Firstly, in [88], humans could only recognize a few signs based on location and motion alone. In ASLR, quite reasonable performance can be achieved under these conditions (see figure 5.3, ‘No HS’, which is basically the same condition). Of course, human beings are not used to seeing signs that way, whereas an automatic algorithm is trained to perform under certain circumstances. Secondly, in [89], it was demonstrated that the first and last part of a sign (in time) are not necessary for recognition; the central (‘stroke’) part suffices. The ASLR algorithm also disregards the start and end of a sign. The feature selection graphs show that selected features always come from the stroke part of the signs (see figure 5.6). So there are both differences and similarities between human and automatic recognition.

An important difference between human and automatic sign processing is the following: in experiments with human signers, more information never caused a deterioration in performance. But the performance of an ASLR algorithm can drop when information is added: when the INV handshape representation was used,
performance of our CDFD classifiers decreased. An ASLR algorithm has the capability
to find information in any feature that it is provided with. This capability probably
allows it to perform so much better than human signers in recognizing signs based on
location and motion alone. But the downside becomes clear when a noisy or
uninformative feature is present. Even with the precaution of feature selection, meant
to weed out uninformative features, the system can still over-train (try to train on
noise), causing a deterioration of performance.

In conclusion, we can say that it is possible to apply results from human signer studies
in ASLR — adding simplified handshape information proved beneficial for ASLR, as
predicted. But the problem described above shows that such applications must be
considered carefully.

Notes

1Standaard Lexicon Nederlandse Gebarentaal, deel 1. Stichting Nederlandse Gebarencentrum, 2002
2SPSS for Windows, rel. 14.0.0. Chicago:SPSS Inc., 2005
3CELEX Dutch database (Release N32) [On-line], 1990. Available:

Acknowledgments

This research was partly funded by the VSB-fund; grant number 2003451: project ELo,
and the ICT Delft Research Center of Delft University of Technology.
Abstract

Human perception of sign language can serve as inspiration for the improvement of automatic recognition systems. Experiments with human signers show that sign language signs contain redundancy over time. In this paper, experiments are conducted to investigate whether comparable redundancies also exist for an automatic sign language recognition system. Such redundancies could be exploited, for example by reserving more processing resources for the more informative phases of a sign, or by discarding uninformative phases. In the experiments, an automatic system is trained and tested on isolated fragments of sign language signs. The stimuli used were similar to those of the human signer experiments, allowing us to compare the results. The experiments show that redundancy over time exists for the automatic recognizer. The central phase of a sign is the most informative phase, and the first half of a sign is sufficient to achieve a recognition performance similar to that of the entire sign. These findings concur with the results of the human signer studies. However, there are differences as well, most notably the fact that human signers score better on the early phases of a sign than the automatic system. The results can be used to improve the automatic recognizer, by using only the most informative phases of a sign as input.

At the time of printing, the material in this chapter was provisionally accepted as a paper by ACM Transactions on Applied Perception, May 2010.
6.1 Introduction

The study of sign language is quite young compared to that of spoken language. Serious research did not begin until Stokoe [80] published his influential work, proving that sign languages were fully-fledged natural languages, rather than gesture systems and pantomime, as was generally believed until then. Most sign language research since then has been complex, fundamental research describing vocabulary, grammar, and basic building blocks of signs, work that is still in progress today [29, 72]. The field of automatic sign language recognition (ASLR) has developed somewhat independently from the linguistic research. The earliest attempts mostly focused on translating speech recognition techniques to the field of ASLR [14, 79, 98], using only minimal knowledge about sign language. It is only recently that more interaction between linguistic research and ASLR has emerged. Derpanis et al. [24] developed a method for movement recognition based on Stokoe’s definition of movement primitives [81]. Von Agris et al. [96] and Zieren and Kraiss [101] studied the variation in German sign language signs in order to create systems that could handle such variation. Bowden et al. [10] designed sign features and feature categories based on the linguistics of British sign language.

We are developing an ASLR system [57] as part of an interactive vocabulary learning environment. At the same time, we perform experiments in sign language perception and processing. The knowledge gained from these experiments is used as inspiration for finding areas in which the automatic recognizer can be improved, and for developing techniques that can be used to achieve such improvements. A set of experiments into the distribution of sign information over time revealed that for human signers, signs contain redundant information. Ten Holt et al. [89] found that certain fragments of a sign in isolation give recognition performance equal to that of the entire sign. In other words, parts of the sign are apparently redundant. In Arendsen et al. [6], it was shown that signers do not need to see the entire sign to recognize it. In many cases, the first third of the sign already suffices. This concurs with the findings of Emmorey and Corina [30] and Grosjean [35], who also found that recognition occurs during the first half of the sign. The experiments all suggest that for human signers, signs contain a high level of redundancy over time.

In this paper, it is investigated whether such redundancy also exists for our automatic recognition system. If this is the case, then the efficiency of the system could be improved by ignoring redundant or less informative parts of a sign. This would bring down processing times, enabling a quicker response and freeing resources for other tasks, such as more detailed feature extraction. Ignoring unimportant parts could also improve the accuracy of the system, because noise will be discarded along with the (redundant) information. For example, the experiment described in ten Holt et al. [89] showed that the central phase of the sign is the most informative one for human signers. If the same is true for the automatic recognizer, then instead of processing the
6.2 Experimental Setup

In both studies described here, the automatic sign language recognizer was trained and tested on a certain subset of frames from videos of sign language signs. This section describes the general setup: the dataset that was used, the procedures that were followed, and the details of the automatic recognizer. The specific frame selection for each study is described in the next sections.
6.2.1 Dataset

The dataset used in this experiment consisted of 91 isolated signs from the standard lexicon of Sign Language of the Netherlands (SLN) [34]. Examples of the signs can be viewed online at visionlab.tudelft.nl/users/gineke-ten-holt. No compound signs were used. Each sign was performed by 76 different persons. They were mostly non-signers, copying the signs from examples that were played for them on a monitor. Inspection by a sign language instructor revealed that their variations and aberrations resembled those of regular beginning signers. Recordings were made with a neutral background and diffuse lighting, using a calibrated stereo camera (Allied Vision Technologies ‘Guppies’) at 25 frames per second. Figure 6.1 shows the recording setup. The resolution of the images was 640 * 480 pixels. Start and end of a sign were indicated by the hands being on the table.

6.2.2 Sign Phases

Sign language signs and gestures can be divided into three basic movement phases: preparation, stroke and retraction. There is some discussion about the exact number and structure of (sub)phases [51, 62], but in general, the three main phases are not disputed. They are shown schematically in figure 6.2. The preparation corresponds to the phase in which the hands move towards the intended location and configuration for the sign, the stroke contains the desired movements and configuration changes, and the retraction contains the relaxation from the final location and hand configuration back towards the rest position or towards the next sign. In continuous signing, preparation and retraction can merge or even disappear, but since our recognizer only works with isolated signs [77], this is not an issue at the moment.

6.2.3 Locating the Stroke Phase

In both studies described here, information about the start and end of the stroke of a sign is relevant. However, finding the stroke automatically in a gesture is not a trivial task [100]. For that reason, starts and ends of the stroke are determined as follows for all sign videos in the dataset: For one example of each sign (the model sign), an annotation of start- and end frame is made by hand by one annotator. The annotator belongs to the group of annotators described in Arendsen et al. [5], annotating another dataset. Analysis of the data from Arendsen et al. showed that the annotator was reliable for stroke annotations (less than 0.2 frames average diversion from the mean of the three annotators, $SD = 1$). Once the stroke annotations for the model are known, the corresponding frames in all other examples can be found automatically by aligning the examples with the model through Dynamic Time Warping (DTW) [54] based on
Table 6.1: Properties extracted from the sign videos. All properties are extracted for each frame in a video. The number between brackets is the number of properties that falls under each description. Total number of properties: 57.

<table>
<thead>
<tr>
<th>Location</th>
<th>Motion</th>
<th>Handshape</th>
</tr>
</thead>
<tbody>
<tr>
<td>left/right hand position x, y, z (6)</td>
<td>left/right hand velocity (2)</td>
<td>left hand shape descriptors (16)</td>
</tr>
<tr>
<td>distance between hands in x, y, z (3)</td>
<td>left/right hand acceleration (2)</td>
<td>right hand shape descriptors (16)</td>
</tr>
<tr>
<td></td>
<td>left/right hand motion direction in x, y, z (6)</td>
<td>left/right hand size derivative (2)</td>
</tr>
<tr>
<td></td>
<td>left/right hand curvature (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left/right hand curvature derivative (2)</td>
<td></td>
</tr>
</tbody>
</table>

automatically extracted features. The model itself is subsequently excluded from the dataset.

6.2.4 Feature Representation

The first step in the automatic sign recognition process is the extraction of features from the sign videos through automatic image processing techniques. This feature extraction is described in detail in Lichtenauer et al. [59]. Summarizing, the head and hands are detected in the videos using a skin model and a tracking algorithm. When the location of the head and hands are known, a number of properties can be extracted, such as position of the hands relative to the head, velocity and acceleration of the hands, direction of hand motion, and distance between the hands. Stereo cameras facilitate the calculation of hand positions in 3D. In addition to the location and motion properties, handshape descriptors are extracted (see ten Holt et al. [87] for details). Table 6.1 lists the properties. These properties are extracted for each frame in a video.

To allow comparison between sign examples, the examples must be aligned in time. This is achieved by aligning all of them with one model per sign using DTW, transforming them to fixed-length examples (this fixed length being the length of the model). Each combination of a frame and a property is considered one feature, which means that each sign video is represented by a $[\text{number of model sign frames} \times \text{number of properties}]$ matrix of feature values. The feature representations are used as input for the automatic recognizer.
6.2.5 Automatic Recognition System

The automatic recognizer used here is built for the task of two–class classification. That is, its task is to distinguish a specific sign (target sign) from “everything else” rather than choose between a set of signs and pick the most likely. See [57, 59] for details. This means that a separate classifier is trained for each sign, aimed at distinguishing its target sign from all other movements. “All other movements” are represented by all other signs, put together in a single non–target class.

The automatic recognizer uses the Combined Discriminative Feature Detectors (CDFD) method described in [57]. First, automatic feature selection is performed, during which only those features most useful for distinguishing between target class and non-target class are retained. Next, a simple feature classifier is built for each selected feature. This basically entails determining a set of feature values that are considered indicative of the target class. For example, if “right hand X-position at frame 40” were a selected feature, then the feature classifier would contain the range of acceptable values for this feature as determined from the target class training examples, e.g. from “10 cm to the left of the nose” to “10 cm to the right of the nose”. The output of such a feature classifier is a simple TRUE or FALSE. The detectors are combined as follows: if the number of TRUEs surpasses a certain threshold, the sign is accepted, otherwise, it is rejected. This sign detection threshold is determined during training, by registering how many TRUEs each training example produces and taking the median.

6.2.6 Classifier Evaluation

A two–class classifier can use strict or more lenient settings. For more lenient settings, more target class signs will be accepted (true positives), but more non-target class signs will be accepted as well (false positives). The operating point chosen for such a classifier depends on its application: is it better to avoid false positives (incorrect signs being accepted) or false negatives (correct signs being rejected)? To determine the overall performance of a two–class classifier, it can be tested using a variety of strictness settings. For each setting, a true positive (TP) and false positive (FP) rate are determined. Plotting these against each other creates a curve, the Receiver Operating Characteristic (ROC) curve. Figure 6.3 shows an example of such a curve. The area under this curve (AUC) is an indicator of a classifier’s performance over a range of different settings (0 ≤ AUC ≤ 1). Because the first part of the curve is usually the most interesting (settings for which the false positive rate is higher than 0.1 will hardly ever be used in practical situations), we use the partial AUC (pAUC) calculated over the false positive range [0 – 0.1] as the performance measure for our classifiers (0 ≤ pAUC ≤ 0.1).
6.2.7 Testing Procedure

For each sign in the dataset, the classification performance is established. The classification task is to determine, for a given sign, whether that sign belongs to the target class or not. For each sign, five classifiers are trained and tested in a five-fold crossvalidation scheme [27] to reduce the influence a particular choice of training and test set might have on the performance. This means that the dataset is divided five times into different training and test sets. For each division (crossvalidation fold), a classifier is trained on the training set and evaluated on the unseen test set. The test set always contains unseen non-target signs (non-target signs of which there were no examples in the training set). Testing entails setting the classifier to a number of strictness levels, and testing for each level which test signs are recognized and which are rejected. This gives a TP and FP rate for each strictness setting. Together, they form an ROC curve. The pAUC of this curve is the performance measure for the classifier. For each crossvalidation fold, a pAUC is calculated in this way, resulting in five performance estimates per sign.

6.3 Study 1: Using Isolated Fragments of a Sign

In this study, the automatic recognizer is tested on isolated parts of sign language signs. The parts correspond to basic movement phases in the sign. The same experiment was performed with human signers in ten Holt et al. [89], though with a different dataset (using videos of a professional signer). A description of the fragments is given below. The results are given in section 6.3.2, and in section 6.3.2 a comparison is made between the recognizer and the human signers.

6.3.1 Selection Procedure

The goal of this study is to investigate the relevance of the sign phases preparation, stroke, and retraction for classification. Using the annotations of the start and the end of the stroke, these phases can be extracted from the sign videos. The start of the sign video is considered the start of the preparation, the end of the video is used as the end of the retraction. However, the boundaries between phases are not sharp. The part of the sign that could belong to either phase is called a ‘transition’ here. Previous studies concerning the moment of sign recognition [5, 6, 30] suggest that the transition from preparation to stroke is a salient region. To test the relevance of this region for classification, the phases preparation and stroke are extracted from a sign both with and without the transition (see figure 6.2). In the human signer study [89], the difference between the boundary assessments of three sign annotators was used to determine the transition. Because only one annotator is used in this study, the same
Figure 6.2: Illustration of the sign fragments used in the experiments. All fragment are based on the annotation of the start and end of the stroke phase. The transition between phases is defined as from 3 frames outside to 2 frames inside the stroke.
procedure cannot be followed. Instead, the average length of the transition from ten Holt et al. is used for all transitions. This average length is six frames, so the transition is defined as the period from three frames outside the stroke to two frames inside it. Figure 6.2 illustrates the fragment definitions. The transition preparation–stroke is also extracted separately as an extra fragment. This fragment was not used in [89], because it would have been too short for human signers to work with (ca. 0.25 sec.). Finally, the entire sign is also used, as a control condition, so the fragments extracted for the first study are:

- prep — preparation without transition
- pretrans — preparation with transition
- trans — transition on its own
- stroke — stroke without transitions
- transstroke — stroke with transition preparation–stroke included
- retr — retraction without transition
- all frames — the entire sign

6.3.2 Results and Discussion

Overall Results

Figure 6.3 shows the average ROC curves for the different fragments, taken over all signs and crossvalidation steps. Only the upper left quadrant of the figure is shown, since this is the most interesting area. Performance based on the preparation is worst, though better than chance (the ROC for chance performance is a straight line from (0,0) to (1,1), which would fall outside the figure). The retraction gives slightly better performance. The transition appears to perform equally well with and without the preparation included. The stroke-fragments seem to give performance equal to that of the entire sign. The inset of figure 6.3 shows the curves of these three fragments in more detail.

Main Effects

To determine the statistical significance of the test results, a two-way ANOVA has been performed, with fragment type and sign as the independent variables, and the pAUC as the dependent variable. The number of repetitions was five (the pAUCs resulting from the five cross-validation steps). The results show a significant main effect for both fragment type ($F(7, 2912) = 1122.12, p < 0.001$) and sign.
Figure 6.3: Average ROCs for the sign fragments, taken over all signs and crossvalidation steps. Error bars represent the standard error of the mean. As a performance measure, the area under the curve is calculated for the false positive range [0 - 0.1] (dotted line). The inset shows the upper left corner in more detail.
Post-hoc tests (using Tukey’s Honestly Significant Difference (hsd) to correct for multiple comparisons) show which fragment types differ significantly from each other ($\alpha = 0.05$). The results are listed in table 6.2. They confirm what is shown in figure 6.3: the stroke-fragments give performance equal to that of the fragment ‘all frames’, and the fragments preptrans and trans do not differ significantly from each other. Post-hoc tests have also been performed to investigate which signs differ significantly (using Tukey’s hsd). The results are too extensive to show here, but figure 6.4 illustrates the distribution of the sign scores. Not all signs can be recognized equally well, but there does not seem to be any clustering of signs into groups.

Interaction

The ANOVA also shows a significant interaction effect between signs and fragment types ($F(630, 2912) = 5.51, p < 0.001$). This may be explained in part by a ceiling effect: all signs show high scores for the stroke-fragments (stroke and transstroke) and the entire sign, whereas their scores on the other fragments differ. Another possible explanation is the difference between the relative salience of preparation and retraction: for some signs (60%), the preparation gives lower scores than the retraction, whereas for others (40%), the reverse is true. Figure 6.5 shows a few typical examples of different fragment-score patterns that occur within the sign set. To eliminate the overall differences between signs, the scores of the fragments are taken relative to the score of the fragment ‘all frames’. The first sign scores low on all non-stroke fragments. It is followed by two signs for which the preparation scores better than the retraction, two signs for which the retraction scores better than the preparation, and two signs for which all fragments give high scores. The last sign shown is MICROPHONE, the only sign in the set that contains no movement during the stroke. The signs are marked in figure 6.4.

This experiment was not explicitly set up to provide insights into why certain signs give better performance for certain fragments. Nevertheless, we can offer some speculation on the topic. Signs that score poorly on all non-stroke fragments, such as PLANT, seem to be signs that are not very strictly defined. For example, PLANT consists of drawing a few leaf-shapes in the air with thumb and forefinger, with both hands alternating. The hand that starts, the number of ‘leaves’, the height at which to start and finish, all vary between signers. For this reason, the characteristics of the start and end of the stroke will vary greatly between examples, making it hard to create good classifiers based on the non-stroke fragments.

Signs that score better for the preparation than for the retraction, such as SHADOW and BAR (OF CAGE), tend to have the following in common: at the start of the stroke, the hands are in a relatively rare (distal) position, but at the end, they are in the neutral space, the area in front of the body in which many signs are made. The reverse is true for signs that score better on the retraction, such as FLAT and QUILL. These signs tend
Table 6.2: Fragment groups based on post-hoc tests (using Tukey’s hsd). Recognition scores of fragments within a group do not differ significantly from each other. Recognition scores of fragments from two different groups always differ significantly from each other ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>all frames</td>
<td>transition</td>
<td>retraction</td>
<td>preparation</td>
</tr>
<tr>
<td>stroke</td>
<td>preparation + transi-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transition + stroke</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FLAT scores particularly badly for the preparation, perhaps because the start position of the sign is close to the rest position, making the preparation short and uneventful. Note that shortness alone would not explain the low recognition scores: the fragment transition, which is only six frames long, performs well in most cases.

In signs that score well for all fragments, such as BELT and SHOWER, both the start and end position are often unusual. For example, BELT starts and finishes low (at waist height), SHOWER starts and finishes high and to the right (at shower head height). We did not observe similar trends for other sign parameters, such as handshape, orientation, or motion in our dataset. It is interesting to note that the sign with the motionless stroke (MICROPHONE) scored among the highest overall. This may have to do with the fact that the stroke does not contain any extra information compared to the end of the preparation or the start of the retraction. All parameters present in the stroke are thus also present in the preparation and retraction.

Comparison to Human Signers

It is interesting to compare the results of this study to those of the fragment recognition experiment performed with human signers [89]. Figure 6.6 compares the results of both experiments. In both cases, the recognition scores are taken relative to the score on the entire sign. For the human signer data, this is trivial, since recognition of the entire sign was always 100% in the human experiment. This is because in the human experiment, signs for which the fragment ‘all frames’ was not recognized were considered unknown to the participant, and all fragments extracted from such a sign were removed from the participant’s test results. Computing the relative scores of the fragments makes it possible to compare the results of the human signers to those of the recognizer. The relative scores are averaged per sign, and the distribution of sign scores drawn as a box plot.

The general trends are the same in both box plots: performance of the stroke-fragments does not differ much from that of the entire sign. Preparation and retraction give lower
Figure 6.4: Mean and standard error of the pAUC of all signs used in study 1, calculated over all fragments and cross-validation steps ($N = 35$). There are significant differences between signs, but there is no clear clustering. The marked signs are example signs discussed in section 6.3.2 and shown in figure 6.5.
Figure 6.5: Ratios pAUC of fragment / pAUC of ‘all frames’ for a selection of signs, averaged over five crossvalidation steps. Error bars represent the standard error of the mean. The scores for fragments stroke and transstroke are fairly constant. Fragments preparation and retraction show the greatest variation.
6.3 STUDY 1: USING ISOLATED FRAGMENTS OF A SIGN

Figure 6.6: (a) Box plot of the ratios fragment score / score on ‘all frames’, of human signers. $N = 20.$ (source: [89]) (b) Box plot of the ratios pAUC of fragment / pAUC of ‘all frames’, of the automatic recognizer. The plots were made by first averaging the scores per sign, then creating a box plot of the sign scores. This is the same procedure that was followed in ten Holt et al. [89]. The red line represents the median, the box represents the middle 50 percentile, the lines indicate the minimum and maximum (excluding outliers). Dots represent outliers. The notches indicate the 95% confidence interval. $N = 91.$ Note that because of the averaging step, this interval is estimated more conservatively than in the ANOVA.
scores. The most striking difference is in the scores of the preparation and preparation + transition. Though the relationship of these fragments is the same, their scores are lower for the automatic recognizer than for the human signers. It seems that human signers have an advantage over the machine when it comes to recognizing a sign based on the preparation-fragments. The explanation may be that human signers have much experience with trying to identify a sign based on its preparation and transition: in normal conversation, one tries to identify a sign as soon as possible. This could lead to an increased ability to recognize a sign based on the beginning. For the retraction, no such skill is necessary: recognizing based on the retraction is never a necessity, since by then, the stroke has already been observed. This would explain why there is little difference between the retraction-scores of human signers and the automatic recognizer. The second study takes a closer look at recognition based on the beginning of a sign.

6.4 Study 2: Using Increasingly Larger Start Fragments

This study tests the automatic recognizer with increasingly larger start fragments of signs. The goal is to determine at which point during a sign recognition becomes possible. ‘Recognition’ is defined here as a performance score equal to 95% of the score of the entire sign. The study is compared to several human signer studies in which similar experiments are performed [6, 30, 35]. In this study, transitions are not used. The annotated start frame of the stroke is used as the end of the preparation and the beginning of the stroke.

6.4.1 Selection Procedure

The first \( N \) frames of each sign are selected, for various values of \( N \). The frames are taken from the fixed-length versions of the signs, that is, after synchronization with the model sign. The fragment sizes used are \([10, 12, 14, \ldots, 50]\) and 60. This gives a total of twenty-two fragment sizes. The automatic recognizer is trained and tested with each fragment size.

6.4.2 Results and Discussion

Results

Figure 6.7 (a) shows the general trend of the recognition scores. It shows the average ROC curves for a subset of fragment sizes. Clearly, the first 10 frames are not informative enough: the recognizer gives a performance equal to chance. Adding ten
more frames makes some recognition possible. Recognition scores continue to increase until fragment size 40, after which little further improvement is made. This is not surprising: the stroke starts on average at frame 27 ($SD = 3.2$) and the average stroke length is 28 frames ($SD = 9.4$), which means that the first 40 frames encompass a large part of the stroke in most cases. Study 1 showed that the stroke in isolation gives performance equal to that of the entire sign, so once the start fragment includes most of the stroke, little improvement can be achieved by adding extra frames.

A two-way ANOVA has been performed on the recognition scores, using fragment size and sign as the independent variables, and the pAUC as the dependent variable. The number of repetitions was five (the pAUCs resulting from the five cross-validation steps). Because we are interested in the general trends, only fragment sizes 10, 20, 30, 40, 50 and 60 were used in the test. Results show a main effect of fragment size ($F(5, 2184) = 7359.71, p < 0.001$), as well as sign ($F(90, 2184) = 18.02, p < 0.001$), and there is also an interaction effect between them ($F(450, 2184) = 5.11, p < 0.001$).

Post-hoc tests (using Tukey’s hsd) show that all fragment sizes except 50 and 60 differ significantly from each other ($\alpha = 0.05$). The sign scores show a similar distribution as in study 1 (see figure 6.4). The interaction effect appears due to the fact that for some signs recognition starts to increase earlier than it does for other signs (see figure 6.7 (c)).

Figure 6.7 (b) shows the recognition results for all fragment sizes. The result is an S-curve, which is steepest around 25 frames. This coincides with the average start frame of the stroke, which is 27 ($SD = 3.2$). The dotted line indicates the threshold that is used for recognition here: a pAUC of 0.089, which is 95% of the average pAUC of the fragment ‘all frames’. This threshold is reached at fragment length 36, which is 41% of the average sign length, and includes 32% of the average stroke length. For more detailed results, the moment of recognition is also determined for each sign separately — that is, the moment the sign’s pAUC reaches 95% of the sign’s ‘all frames’ pAUC.

The average fragment length for which this happens is 36 ($SD = 6.4$). Relating to the lengths of the individual signs, the average recognition point is at 41% of the sign ($SD = 0.07$), and includes 31% of the stroke ($SD = 0.22$).

Figure 6.7 (c) shows the recognition results for four of the signs that were discussed in section 6.3.2. The slopes of the S-curves are similar, but for signs that score well for the preparation (such as SHOWER and BELT), the curve is moved to the left compared to signs that do not score well for the preparation (PLANT and QUILL), indicating that recognition starts to occur earlier. This explains the interaction effect between sign and fragment length. Note that the length of the preparation alone is not decisive for the recognition performance: the preparation of QUILL is almost as long as that of SHOWER, but the recognition scores are different. Apparently, the information present in a fragment is more important for recognition than its length, something that was noted in study 1 as well.
**Figure 6.7:** (a) Average ROC curve for a subset of fragment sizes (over all signs and crossvalidation steps). Curve 50 is almost entirely covered by curve 60. Error bars indicate the standard error of the mean \( (N = 455) \). The pAUC is calculated for the false positive range [0 - 0.1] (dotted line). (b) Average pAUC for all fragment sizes (over all signs and crossvalidation steps). Error bars indicate the standard error of the mean \( (N = 455) \). The dotted line indicates pAUC = 0.089, which is 95% of the average recognition based on 'all frames'. (c) Average pAUC for four individual signs also shown in figure 6.5. Error bars indicate the standard error of the mean \( (N = 5) \). Stars indicate the moment the stroke starts. This graph illustrates the interaction that exists between fragment length and sign: some signs already show high scores for short fragments, notably signs for which the preparation gives good scores. For others, scores only start increasing at a greater fragment length.
Comparison to Human Signers

There are several human signer studies that attempt to establish the moment (confident) sign recognition happens. Grosjean [35] performed a gating experiment, in which human signers saw increasingly larger start fragments of a sign and were asked to identify the sign, giving a confidence score to their identifications each time. He determined the recognition point (the moment of confident sign recognition) at 47% of the sign, counting from the moment of sign onset. Clark and Grosjean [19] found that with context, only 37% of the sign is needed, again counting from sign onset. Emmorey and Corina [30] reported recognition times of 34% of the sign, counting from the moment that the hands entered signing space. Arendsen et al. [6] performed an experiment in which signers were asked to indicate the moment they started to recognize a sign, for videos containing both signs and other movements. They found a median response time of 13 frames (538 ms), counting from the moment the hands started moving. Using only the videos containing signs, the recognition time was calculated as 34% of the sign [86], again counting from movement onset. It is clear that precise comparison of these values with the machine recognition times is hard because of the differences in reference points. The recognition scores in study 2 are calculated from the start of the sign video, which may include a short period of inaction before movement onset. Nonetheless, the recognition time established for the automatic recognizer, 41% of a sign, falls within the interval of values reported for human signers (34%–47%). If we use the recognition time counted from the start of the stroke in the comparison, the machine value (31% of the stroke) is similar to the values reported by Emmorey and Corina [30] and ten Holt et al. [86] (34%). In general, it can be concluded that both human signers and the machine do not need the complete sign in order to recognize it, less than half the sign is sufficient.

6.5 Discussion

The goal of this study was to investigate whether redundancy exists in sign language signs over time for our automatic recognizer. The results of recognition experiments with human signers were used as an inspiration to locate such redundancies. The results show that there is redundancy over time for the automatic recognizer, and that the redundancies resemble those of the human signers. In the first study, it was shown that using only the central phase of a sign gives equally good recognition as when the entire sign is used. Preparation and retraction give lower scores, but do contain useful information. The second study showed that the first half of the sign contains sufficient information to achieve automatic recognition of a certain level (95% of the level of the entire sign). The results can be used to improve the automatic recognition system, making it faster and less perceptible to noise by reducing the number of frames used for recognition. For example, the recognizer could use only the stroke as input, instead
of the entire sign, without any loss of recognition performance. This would make the recognition process faster, because fewer frames would need to be processed and processing could commence before the end of the sign. It would also discard any noise contained in the preparation and retraction phases. Another option would be to use only the first half of the sign as input, which would also increase processing speed and enable a quick response, which is often important in human-computer interaction applications. Processing fewer frames would also free resources for more sophisticated feature extraction, such as more detailed handshape representations.
Using Sparse Sign Representations

Abstract

This paper presents a set of experiments in automatic sign language recognition investigating the effects of using sparse sign representations. Two algorithms are proposed. In the first, a key frame representation is used to represent the signs. The key frames are selected using knowledge about the sign. At the same time, the requirement is added that a sign property is only considered correct if it is correct in all key frames. The second algorithm is a variant of the key frame algorithm: the same key frame representation is used, with the same correctness requirement for properties. Properties are then divided into groups called aspects, which represent linguistically relevant characteristics, and a sign is considered correct only if it is correct in all aspects. Both algorithms are tested and compared to a classifier that uses the entire stroke of a sign as input, and give lower performance. The reason may be that the combination method, both for key frames and for aspects, is quite strict: if one element is incorrect, the entire combination is rejected. This causes individual features to have a large impact on the recognition of the entire sign. Since features can always be erroneous even in a correct sign, this may cause the deterioration of recognition performance. Using different combination methods, with more robust handling of incorrect feature values, is a possible solution.
7.1 Introduction

Automatic sign language recognition is a relatively new area of research that has applications in a variety of areas, such as human-computer interaction, facilitating communication between deaf and hearing, and sign language learning. We develop an automatic sign language recognizer as part of an interactive environment for young deaf children to practice sign language vocabulary [78]. The basic recognizer is described in [59]. Using results from human signer experiments as inspiration, the recognizer was further developed to use more features [87] and fewer input frames [90]. In this paper, the possibilities of using sparse sign representations are investigated.

In [90] it was shown that not all frames of a sign are essential for recognition: when only the frames of the stroke (the central part of the sign) were used as input for the CDFD recognizer, recognition performance stayed the same. A sign could therefore be represented by the stroke-phase only. However, adjacent frames within the stroke are highly correlated for many properties, such as hand position and handshape, since the human hands cannot move instantly from one location or configuration to another. This raises the question whether a sign could be represented by a subset of stroke frames, and what influence such a sparse representation would have on the recognition performance. For human signers, signs can remain intelligible when the frame rate is reduced, although performance does decrease and other factors, such as the choice and resolution of the frames, influence the results [16, 44, 66]. It would be interesting to see the effect of using fewer frames on the automatic recognizer.

Besides reducing the number of frames used to represent a sign, it may be worthwhile to investigate the effect of imposing requirements on the relationships between the chosen frames. In the CDFD recognizer, all features are considered independent (a feature is a certain sign property at a certain frame). Recognition of the sign is achieved when the number of correct features surpasses a certain threshold (the Sign Detection Threshold or SDT). This could lead to undesirable results. Figure 7.1 gives an example: suppose there is a sign that consists of multiple, repeated circles in front of the body, followed by a rising hand. A sign consisting of the same circular movement, but followed by a lowered hand, could be recognized as correct if the circular part contains enough features to surpass the SDT, regardless of the incorrect latter part. This does not seem to be the best way to represent a sign. It would be better to include a requirement that all moments within a sign must be reasonably correct.

For comparable reasons, requirements could be imposed on the relationships between properties of a sign: without such requirements, a sign could be recognized based on certain properties alone, regardless of other, incorrect properties. For example, in the CDFD recognizer, a handshape is represented by sixteen properties, while a location is represented by only three properties (X, Y and Z). This means that theoretically, the
7.2 Method

7.2.1 Dataset

The dataset used in the experiment consists of 91 signs from the standard lexicon of Sign Language of the Netherlands (SLN). Each sign was recorded from 75 different persons. For each video, features were extracted automatically. Fifty-seven properties were extracted from each frame, resulting in a \([\text{nr of frames} \times 57]\) matrix of features for each sign video. For an overview of the extracted properties, see [57] or appendix X.

The recognizer used in this study judges the correctness of a sign. This means that it consists of a set of two-class classifiers, one for each sign in the dataset, rather than a multi-class classifier. For each two-class classifier, the goal is to distinguish its target sign from any other movement. Here, “any other movement” is represented by all non-target signs in the dataset. Thus, the dataset is divided into target sign examples and non-target sign examples for each sign classifier separately. Per classifier, all sign examples are aligned with a single target class example (the model sign) through Dynamic Time Warping (DTW) [54], thus turning them into fixed-length examples (the fixed length being the length of the model sign). The model sign is not used in the subsequent training and testing of the classifier. For details about the DTW alignment, see [57].

For the recognition methods used in this paper, it is necessary to determine the start and end of a sign’s stroke (the stroke is the central part of a sign or gesture [51]). To achieve this, annotations of these moments are made by hand for the model signs. The corresponding stroke frames of other examples of a sign can then be found by aligning
Figure 7.1: Example of a weakness in the CDFD recognizer. Consider an artificial sign X that consists of a repetitive circular motion in front of the body (a) followed by lifting the hand to a high position (b). If someone would perform a sign Y which consists of the same circular motion (a) followed by lowering the hand (c), this would not be considered a correct version of X by most people. However, the automatic recognizer could find it correct if the number of (correct) features in the first part of the sign were enough to surpass the sign detection threshold. For example, if part (a) contains 800 features, and part (b)/(c) contains 200, and the sign detection threshold is determined at 0.8 (80% of the features must be correct), then the threshold can be reached if all features of part (a) are correct, even if all features of the latter part are incorrect. The sign Y would then be recognized as correct on the merits of part (a) alone.
7.2 Method

a sign with the model through DTW [54]. For details about the annotation procedure, see [90].

7.2.2 Original Recognition Algorithm

The sign recognition algorithm used in this experiment is based on the CDFD algorithm described in [57]. In the CDFD algorithm, a separate classifier is created for each sign. The classifier is trained using a set of training examples, both of the target class and of the non-target class, and tested on a separate test set. Training starts with an automatic feature selection step, in which only those features are retained that are useful for distinguishing the target-class from the non-target class.

$$\text{SelFeat} = \{(f, p)_1, (f, p)_2, \ldots, (f, p)_{N_sf}\}$$

where $\text{SelFeat}$ denotes the set of selected features, $N_{sf}$ denotes the number of selected features. For the selected features, feature classifiers are created. The feature classifiers give an output of TRUE or FALSE to indicate whether a feature was correct or not:

$$\text{FC}_{(f,p)}(e) \rightarrow \{T, F\}$$

For example $e$, feature classifier $\text{FC}$ for property $p$ at frame $f$ gives TRUE or FALSE (0 or 1).

A Sign Detection Threshold (SDT) is determined by calculating the median number of selected features that are positively detected in the target training examples (relative to the total number of selected features):

$$\text{DR}(e) = \frac{\sum_{\forall (f,p) \in \text{SelFeat}} \text{FC}_{(f,p)}(e) == T}{N_{sf}}$$

$$\text{SDT} = \text{median}(\text{DR}_e)_{\forall e \in TTr}$$

where $\text{DR}$ denotes detection ratio, $TTr$ denotes the set of target training examples. During testing, an example $e$ is recognized as correct if it scores above the SDT. That is, only test examples for which enough selected features are correct are accepted:

$$\text{rec}(e) = (\text{DR}(e) \geq \text{SDT})$$

For an overview of the CDFD algorithm, see Appendix C.1.
7.2.3 Key Frame Recognition Algorithm

The key frame (KF) algorithm is largely the same as the CDFD algorithm, except that there is no automatic feature selection step. Instead, features are selected using pre-existing knowledge. Certain frames (key frames) are chosen:

\[ KF = \{ kf_1, kf_2, \ldots, kf_{N_{kf}} \} \]

where \( N_{kf} \) denotes the number of key frames.

All properties in the key frames are selected. Feature classifiers are built for all selected features in the same manner as in the CDFD algorithm. Property classifiers are built by combining the feature classifier outputs per property over all key frames to produce one output. Combining is done through logical AND:

\[ PC_p(e) = \text{AND} \left( FC_{(f,p)}(e) \right) \]

In other words, a property must be correct at all key moments in a sign before it is classified as correct. The output of the property classifiers is used to determine the SDT, in the same manner as in the CDFD algorithm:

\[ PDR(e) = \frac{\sum_{p=1}^{P} PC_p(e) == T}{P} \]

\[ SDT = \text{median}(PDR(e)) \]

where \( PDR \) denotes the property detection ratio, \( P \) denotes the number of properties, \( TTr \) denotes the set of target training examples.

Recognition is achieved by comparing the ratio of recognized properties to the SDT:

\[ \text{rec}(e) = (PDR(e) \geq SDT) \]

Appendix C.2 gives an overview of the KF algorithm.

Key Feature Vector Calculation Procedure

The choice of key frames is determined in the following way: Start and end of the stroke are always selected, plus a number of frames equally spaced in between. The number of additional frames can be varied.

The feature vectors extracted from the key frames could be used directly as key feature vectors, but this would give poor results if one such frame were mis-tracked or otherwise faulty. To avoid this, a set of frames (three or five) is selected around each
Figure 7.2: Calculation of key feature vector representation of a sign. The number of extra key frames between start and end of the stroke can be varied, as can the number of frames used to create each key frame (3 or 5) and the combination method (average or median).

key frame. Features are extracted for all frames in the set, and the average or median is calculated for each set (calculated per property). The resulting key feature vectors are used to represent the sign:

\[ KFV_k = \frac{1}{k+n} \sum_{i=k-n}^{k+n} Comb(FV_i) \]

where \( KFV_k \) denotes the key feature vector at key frame \( k \), \( FV_i \) denotes the feature vector at frame \( i \), \( n \) denotes the number of extra frames around the key frame in either direction (1 or 2), \( Comb \) denotes the frame combination (average or median). Figure 7.2 illustrates the key feature vector calculation procedure. In the experiments, various combinations of the number of selected frames, the number of frames taken around each selected frame, and the combination method are investigated.
7.2.4 Aspect Recognition Algorithm

The aspects algorithm uses the same key feature value representation for a sign as the key frame algorithm. Feature classifiers and property classifiers are created in the same manner. However, instead of determining one SDT on all properties, properties are divided into groups called aspects:

\[ P_a = \{ p_1, p_2, \ldots, p_{N_a} \} \]

where \( P_a \) denotes the set of properties belonging to aspect \( a \), \( N_a \) denotes the number of properties in aspect \( a \). A separate SDT is calculated for each aspect:

\[ PDR_a(e) = \frac{\sum_{p \in P_a} PC_p(e) == T}{N_a} \]

\[ SDT_a = \text{median}_{\forall e \in T_{Tr}}(PDR_a(e)) \]

where \( PDR_a \) denotes the property detection ratio for aspect \( a \). Recognition is achieved by checking whether all aspect SDTs are surpassed:

\[ rec(e) = \text{AND}_{\forall a}(PDR_a(e) \geq SDT_a) \]

That is, a sign is required to be correct (enough) in all aspects before it is considered correct as a whole.

Properties are grouped into aspects that correspond to linguistically relevant sign characteristics, such as handshape, hand location and hand motion. In the current study, a simple division of properties into seven aspects is used. The aspects are listed in table 7.1. Appendix C.3 describes the aspects algorithm. In the algorithm, knowledge about the relevance of aspects for individual signs can be utilized. In the CDFD algorithm, all properties are always used for all signs, and automatic feature selection is used to exclude irrelevant properties. In the aspects algorithm, properties can be excluded based on knowledge about their relevance to specific signs. In this paper, we test a simple application of such knowledge: excluding aspects pertaining to the non-dominant hand for one-handed signs.

7.2.5 Performance Measurement

The testing procedure is the same as in [90]. In short, classifiers are trained and tested for every sign in a five-fold crossvalidation scheme [27], to reduce the influence that a particular choice of training and test set might have on the performance. This means that the dataset is divided five times into different training and test sets. For each division (crossvalidation fold), a classifier is trained on the training set and evaluated
Table 7.1: Organization of the 57 sign properties into aspects. Each aspect represents a linguistically relevant characteristic of a sign. The aspect ‘hand configuration’ contains the feature representation of the hand’s configuration, in which both handshape and hand orientation information is present to some extent. A prefix ‘d’ indicates a derivative. For more information about the properties, see [57] or appendix A.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Feature Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>left hand position</td>
<td>left hand X, Y, Z</td>
</tr>
<tr>
<td></td>
<td>dLeftHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>ddLeftHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>LeftHandMoveAngle3D</td>
</tr>
<tr>
<td>left hand motion</td>
<td>(sidewayness, upwardness, forwardness)</td>
</tr>
<tr>
<td></td>
<td>LeftHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dLeftHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dLeftHandSize</td>
</tr>
<tr>
<td>left hand configuration</td>
<td>16 shape descriptors</td>
</tr>
<tr>
<td>right hand position</td>
<td>right hand X, Y, Z</td>
</tr>
<tr>
<td></td>
<td>dRightHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>ddRightHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>RightHandMoveAngle3D</td>
</tr>
<tr>
<td>right hand motion</td>
<td>(sidewayness, upwardness, forwardness)</td>
</tr>
<tr>
<td></td>
<td>RightHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dRightHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dRightHandSize</td>
</tr>
<tr>
<td>right hand configuration</td>
<td>16 shape descriptors</td>
</tr>
<tr>
<td>hand relationship</td>
<td>LeftRightDistance</td>
</tr>
<tr>
<td></td>
<td>(horizontal, vertical, depth)</td>
</tr>
</tbody>
</table>
Table 7.2: Average pAUCs of various key frame representations, with the standard deviation between brackets. Averages are taken over all signs and crossvalidation steps. The score of the CDFD algorithm on all stroke frames is included for comparison.

| # KF | 3 Consecutive | | | 5 Consecutive | | |
|------|---------------|------|------|---------------|------|
|      | avg | med | avg | med | avg | med | avg | med |
| 3    | 0.0855 (0.0133) | 0.0846 (0.0145) | | | 0.0867 (0.0129) | 0.0856 (0.0135) | | |
| 4    | 0.0880 (0.0115) | 0.0870 (0.0126) | | | 0.0889 (0.0108) | 0.0878 (0.0118) | | |
| 5    | 0.0876 (0.0132) | 0.0852 (0.0142) | | | 0.0876 (0.0125) | 0.0866 (0.0132) | | |
| 6    | 0.0865 (0.0125) | 0.0852 (0.0141) | | | 0.0878 (0.0118) | 0.0861 (0.0133) | | |

CDFD algorithm on all stroke frames 0.0937 (0.0075)

7.3 Results

7.3.1 Key Frame Representation

Representations with small numbers of key frames were tested with several numbers of consecutive frames and combination methods, to investigate the effect of varying these parameters. The results are listed in table 7.2. As a combination method, the average seems to work better than the median. Taking five consecutive frames gives better results than using three. Scores do not increase monotonously with the number of key frames used: using four key frames gives better performance than using five or six. All representations give a performance lower than that of the CDFD method using all stroke frames. Average ROC curves for the representations are included in appendix 7.4.5.

To investigate the effect of using larger numbers of key frames, the KF algorithm was
7.3 Results

tested using 8, 10, ..., 28 key frames. The key feature value computation method of using five consecutive frames and taking the average is used for all of these cases, since this configuration gave the best results in the first study.

Figure 7.3 shows the average pAUCs for each number of key frames used. For numbers of key frames larger than six, the pAUC decreases. This probably has to do with the fact that key frames are combined through logical AND. If a property is incorrect in one of the key frames, the entire property is invalidated. As a consequence, the sign has a larger chance of being rejected. When more key frames are used, the chances increase that a property will be incorrect in one of them, due to a tracking error or because the feature classifiers are not perfect. As a result, there will be a greater chance of target signs being rejected, causing a lower true positive rate. The false positive rate will not decrease much, since most non-target signs were already being rejected anyway. Together, this causes a decrease in pAUC.

For small numbers of key frames (between three and six), the pAUC varies. In these cases, the increased information provided by the extra frames has a beneficial effect on recognition that compensates for the increased chances of a property failing. For larger numbers of key frames, the amount of extra information is relatively small, and the problem outlined above seems to become dominant.

7.3.2 Aspects Method

The aspects method is applied using the key frame representation that gives the best results: four key frames, each calculated by combining five consecutive frames and taking the average. Two versions of the method are tested. In the first, all seven aspects are used for every sign. In the second, all aspects pertaining to the non-dominant hand are ignored for one-handed signs. These are aspects 1–3 (all signers in the dataset are right-handed). The results are shown in figure 7.4. Figure 7.4 (a) shows the average ROC curves for both versions and compares them to the key frame method and the original CDFD method using the entire stroke. Using aspects causes a decrease in performance compared to the key frame method. Ignoring selected aspects improves the results slightly. Figure 7.4 (b) shows the partial AUCs for the same methods.

Introducing aspects causes a deterioration of performance. The explanation may be that certain aspects contain only a few features. This means that for these aspects, the SDT is determined over a small set of features, so that it can only assume a limited set of values. For example, for three features, the SDT can only be 0, 0.33, 0.67 or 1.0. Intermediate levels are not possible. The SDT is determined on the target training examples, on examples for which most features are usually correct. The median ratio will thus most often be 1.0, not 0.67 or lower values. If the aspect SDT is set at 1.0, a test example’s aspect will only be accepted if all features of the test example are correct. Any incorrectness in the features will invalidate the aspect, and through that, the entire
Figure 7.3: Partial AUCs for increasing numbers of key frames used in the KF algorithm. The average pAUC is plotted against the number of key frames used. Error bars indicate the standard error of the mean ($N = 455$).
Figure 7.4: Results of the aspects method. Error bars indicate the standard error of the mean ($N = 455$). (a) shows the average ROC curve over all signs and cross-validation steps for the aspects method with and without selection of aspects per sign, compared to the original CDFD method and the best key frame method (stroke 4 con 5 avg). (b) shows the average pAUC for the same methods.
test example. This would also explain why the pAUC rises slightly when certain aspects are ignored: with fewer aspects, the chance of an aspect failing and invalidating the sign becomes smaller.

7.4 Discussion

7.4.1 Effects of Requirements on Relationships between Features

This paper describes a first exploration of possibilities for classifying signs based on combinations of features rather than individual features. To do this, key moments in a sign are selected and requirements are introduced with regard to the relationship between the key moments of a sign (KF algorithm). In a second algorithm, additional requirements are formulated for the relationships between various properties of a sign (aspects algorithm). The results show that introducing requirements on these relationships causes performance to deteriorate. This deterioration is probably caused by the strictness of the requirements that are used: in both algorithms, properties are only classified as correct if they are correct in all key feature vectors. In addition, in the aspects algorithm a sign is only classified as correct if it is correct in all (relevant) aspects. A complicating factor in the aspects algorithm is the fact that some aspects contain only a few properties. This has the effect that such an aspect often fails if a single property is incorrect. In short, the strict combination methods leave little room for error: a single incorrectness can have large consequences, even cause an entire example to be rejected. This is undesirable, since errors can always occur in the feature representation of a sign due to tracking errors or other problems. Also, the feature classifiers are not perfect: even with perfect tracking, it can occur that a target sign contains a feature value (an outlier) that is rejected by the feature classifier. The method of taking multiple frames around a key frame and averaging or taking the median, which is meant to counter such errors, apparently does not suffice to eliminate such problematic values.

In the original CDFD method, there is a margin of error for incorrect feature values, because only a certain percentage of the selected features has to be correct for the sign to be recognized. Each individual feature is of limited importance because hundreds of features are used to calculate whether the SDT is surpassed or not, which determines the correctness of the sign. In the key frame algorithm, the number of features used to calculate whether the SDT is reached is reduced to 57, since frames are combined per property. This means that each individual feature has a larger effect on whether the SDT is reached or not than in the CDFD algorithm, and errors in the features have a greater impact on recognition. This impact becomes even greater in the aspects algorithm, because there, the number of features used to calculate whether an aspect SDT is reached is even smaller. And because the aspects are combined through logical
AND, each aspect is decisive for whether the sign is recognized or not. This means that certain individual features have a large impact on whether a sign is classified as correct or not, and any errors in such features directly affect the classification of the sign.

7.4.2 Alternatives to the Current Requirements

From the results of the experiments, it seems that the requirements imposed were too strict, and did not leave enough margin for errors in the feature values. Apparently, incorrect feature values occur frequently, whether this is due to feature extraction errors or imperfect feature classifiers. Therefore, more tolerance for such errors must be built into the classification method than is currently the case. There are various ways in which this could be accomplished: classifiers could be built directly for a combination of features, rather than first building separate feature classifiers and then combining the outputs. This would give a better modeling of the feature combination, and would reduce the chance of the entire combination being rejected because one feature was incorrect. This principle could be applied both to the combination of key frames and to the combination of features into aspects. Also, rather than just thresholding feature values and combining the binary outputs, the deviation of a feature from the threshold could be retained and used in the combination. For example, the deviations could be averaged and compared to a common threshold.

7.4.3 Comparison to Human Signers

It is interesting to compare the results of the key frame algorithm to experiments in which the recognition of human signers is measured when low frame rates are used. Unfortunately, most studies reporting such experiments vary both the frame rate and the spatial resolution, which makes comparison more difficult [16, 44]. Furthermore, in these experiments sign videos are sub-sampled to a certain frame rate (fps), whereas in this paper, absolute numbers of frames are used as a measure of sub-sampling. Finally, in this paper, only the stroke is sub-sampled, whereas in the human signer experiments, videos are often sub-sampled in total (including preparation and retraction). An interesting study by Parish et al. [66] reports that for constantly sub-sampled signs, recognition probability is around 0.5 for 5 fps (50% of the stimuli are identified correctly). Since most signs are 1–2 seconds long, this would correspond to a representation using 5–10 key frames in our experiment, for which we found a pAUC of around 0.88. Scores in Parish et al. increase to around 0.9 for 30 fps, which would be regarded as normal presentation. The pAUC of the CFD method is 0.93. An alternative for comparison is calculating the error of the KF algorithm at a single operating point. At an FP rate of 0.1, this error is 0.099 for representations using 5 and 10 key frames. This means the recognition probability is about 0.9. Precise comparison is difficult because of the differences between both studies. Recognition performance
of human signers decreases compared to the performance of ‘normal’ presentation when a subsampled sign is used. The same happens in the KF algorithm. In most cases, the performance of human signers increases when more frames are used, whereas the performance of the KF algorithm decreases with more key frames. However, there are instances in [66] where performance decreases for a higher frame rate, so apparently, extra frames can be confusing rather than enlightening even for a human signer.

7.4.4 Sparseness and Underspecification

It may be argued that a small number of frames cannot represent a sign well enough to convey all its details. Especially details of motion, such as arched movement or ‘wiggles’, are lost. However, this may not necessarily be a problem. Most sign parameters can be inferred using only a few frames, because they are relatively static: handshape, general sign location, orientation, even the non-manual component to some extent. The most difficult parameter to infer at a low frame rate is motion. Human recognition of sub-sampled signs was shown to be better when frames were chosen at the extremes of the motion path [66]. Possibly, human signers infer a motion path by interpolating between frames. A similar technique could be applied in automatic recognition, perhaps using knowledge to supply the recognizer with alternative interpolation patterns (straight, arched, wiggly) and checking whether the application of any of these patterns provides a match with a possible sign.

However, there are cases in which details of the motion path are essential. For example when the recognizer must be able to distinguish between signs whose only difference lies in these details. Or when the correctness of a sign has to be checked, to give feedback to a sign language learner. In the first case, the details are the only thing distinguishing two signs, in the second, the presence or absence of the details determine whether a sign is made correctly or not. In these cases, a sign representation with only a few frames cannot be used; too much motion detail would be lost. Separating the motion parameter from the other sign parameters may be an interesting alternative. A representation using only a few frames could then be used for the non-motion parameters, freeing resources for e.g. a more complex handshape recognition algorithm. For the motion parameter, only the hand centers would have to be tracked, which could be done at the highest frame rate to conserve as much path detail as possible. Vogler and Metaxas [95] investigated this type of approach in automatic sign language recognition using Hidden Markov Models (HMMs).

7.4.5 Future Directions

In this paper, simple methods were used for choosing key frames, organizing properties into aspects, and selecting aspects for signs. Alternative, more sophisticated
methods are possible for each of these. It would be interesting to select key frames that correspond to the extremes of the motion in the stroke, rather than simply sub-sampling the stroke. The activity index described in [66] could serve as an indicator of such interesting points. Using the extremes of a motion path may be a better representation of a sign than sub-sampling, because the other positional values can be interpolated better, and because handshapes tend to be fully formed at such extremes, whereas during motion, they may be in the process of changing. Different definitions of aspects could be used based on more knowledge about sign language. For example, if research showed different levels of tolerance for directional properties and non-directional properties, then separating such properties into different aspects would be useful: separate levels of strictness could then be used. And information about which aspects are relevant could be listed for each sign separately, and updated depending on the task (for some recognition tasks, checking a few aspects may be sufficient).

Including automatic feature selection is another possibility for improvement. For example, several possible key frames could be selected, using knowledge or a measure for frame suitability (such as the activity index described in [66]). Then, the combination that is most effective in distinguishing target class from non-target class could be chosen as key frames.

Finally, it would be interesting to conduct separate experiments investigating the effect of reducing frame rate and the effect of requirements on feature relationships. A separate experiment into frame rate reduction would give us information about the redundancy of frames: does performance always increase with frame rate? If so, at which frame rate does it stop increasing significantly? Does an informed choice of key frames give better performance than sub-sampling for the same frame rate? Once this is established, various methods of combining key frames and/or aspects can be investigated separately.
Appendix 7.A : ROC curves of various key frame representations

Figure 7.5 shows the average ROC curves for various key feature vector configurations. “kf N con M O” means the method used N key frames, M frames per key frame, and combination method O. The curve ‘CDFD on stroke’ indicates the results of the CDFD method using all stroke frames as input. In (a), different numbers of key frames are compared. Each key feature vector was calculated by taking the average over five consecutive frames. (b) shows the same using the median to combine consecutive frames. In (c), different numbers of consecutive frames and combination methods are compared when four key frames are used. All curves are averaged over all signs and cross-validation steps.
Figure 7.5: Average ROC curves for various key feature vector configurations. Error bars indicate the standard error of the mean ($N = 455$).
8.1 Application of Knowledge about Human Sign Recognition

The work in this thesis centers around the question: “Can knowledge about human sign language recognition be used to improve vision-based automatic recognition of sign language signs?”. From the experiments in the second part of the thesis, it becomes clear that, to a certain extent, this is possible — at least, for the recognition system that was experimented with. Adjustments and additions to the recognition system based on the information gained in human signer experiments had a positive effect on recognition performance. In chapter 5, the addition of a handshape feature led to a small increase in recognition scores. Chapter 6 showed that the amount of input data can be reduced without causing a noticeable decrease in recognition scores. Chapter 7 showed that reducing the number of input frames to only a few key frames causes a drop in recognition performance, which also happens for human signers in [66].

On the other hand, the experiments in part II also show that not all results found for human signers can be used successfully in the automatic recognition system. In chapter 2, the fragments ‘preparation’ and ‘preparation + transition’ gave good scores for human signers. The second fragment even gave performance equal to that of the stroke (for which recognition results were almost equal to those for the complete sign). This is not the case in the automatic system, as shown in chapter 6: when tested with the same fragments, the automatic system scored lower than the human signers, and its scores for the fragment ‘preparation + transition’ were significantly lower than those for the stroke. In chapter 5, it was shown that not all handshape information is necessarily beneficial: one of the representations tested caused recognition scores to drop. Finally, chapter 7 showed that the automatic system is not as robust as human signers: adding more key frames can have a negative effect on performance. For human signers, the opposite effect is generally found [18, 66].
It seems that knowledge about human sign recognition can be used in general terms to find inspiration for improving automatic sign language recognition, for example by directing towards informative elements of a sign. However, detailed results about human recognition are not necessarily transferable to the field of automatic recognition. There are several areas in which human and automatic sign recognition differ. Some are of a general nature, some are specific to the type of recognizer used in this thesis. These differences complicate the comparison between human and machine recognition, and may explain why some techniques can be transferred between the two fields while others cannot. In the subsequent sections, the most relevant differences are discussed.

**Data Acquisition** There are differences in data acquisition between human signers and automatic recognition systems. Human vision is very sophisticated [65], and in general, computer vision techniques cannot yet achieve the same standards of robustness and detail [31], especially in real-time settings. This means that in a real-time sign recognition system such as ours, sign characteristics will often be extracted using a simplified representation. For example, the handshape representation used in the CDFD recognizer is quite crude. A human signer, on the other hand, will usually have access to the full, detailed handshape. This difference means that certain effects found for human sign recognition may not occur in automatic recognition. For example, we found that for human signers, handshape is a salient feature and increases the recognition performance substantially (see chapter 3). Because of the difference in handshape representation, however, the handshape feature of the automatic recognizer contains less detail, and is therefore less salient. This is a possible explanation for the fact that the performance of the automatic recognizer increases only slightly when a handshape feature is added (see chapter 5).

**Processing Time** In the experiments described in Part I, the human signers must perform all sign processing during the playing time of the stimulus, plus possibly a short period afterwards while they can still remember the stimulus. This means that human signers have less processing time available for a short sign (fragment) or a quickly made sign (fragment) than for a long or slow sign (fragment). This is not the case for an automatic recognizer: such a system generally records a sign and then processes it, which means processing can take as long as necessary. Processing time is only limited by the requirements defined for the system, e.g. that a response must be given within a second, to adhere to the requirement of real-time functioning. For the CDFD recognizer, processing time was sufficiently short that this requirement did not need to be enforced (see [56] for details).

This difference may cause certain effects found for human signers to be absent in an automatic recognizer. For example, a fragment that is too short for a human signer may be perfectly usable for an automatic recognition system, because for the system,
processing time available for a fragment is not limited by the fragment length. This is illustrated in chapter 6: the automatic system could use the fragment ‘transition’, which was too short for human signers to work with, and achieve good results.

**Recognition, Detection, and Checking Correctness** Another factor that complicates the comparison between human and machine results is the difference in tasks executed by the human signers in Part I of the thesis, and the **CDFD** classifier. There are different types of classification, as explained in the introduction. The task given to human signers in Part I is a recognition task: a multi-class classification task in which they attempt to assign a test sign to a class (figure 1.3 (b) in chapter 1), where their entire vocabulary forms the set of possible classes, and where they have the option of replying ‘not recognized’ if the classification is not clear enough.

The **CDFD** recognizer consists of a set of classifiers. Each classifier is ideally meant to be a ‘correctness checker’: given a test sign, it evaluates whether this test sign is a correct version of its target sign or not (figure 1.3 (c) in chapter 1). This is a form of detection in which an event must be detected in a large, unspecified group of events. Such a task would normally be a one-class classification task: the target class must be distinguished from ‘everything else’, but ‘everything else’ cannot be properly modeled. Only the target class can be properly modeled, which is the definition of a one-class classification problem. As a consequence, human judgment must be used in choosing the features to represent the objects. Furthermore, evaluation is difficult, because it is not possible to determine whether a set of test objects is representative of the non-target class.

In practice, the **CDFD** classifiers are not pure one-class classifiers: the other classes in the dataset are used to model the non-target class. The modeled non-target class is employed in selecting informative features to represent the objects, and in the evaluation of the classifiers. Thus, the feature representation and performance of the classifiers do not truly reflect the ability of a classifier to distinguish its target sign from ‘everything else’, but rather the ability to distinguish its target sign from the other signs in the given dataset. However, in training the feature classifiers, only the target class is employed. This makes a **CDFD** classifier a hybrid between one-class and two-class classification (figures 1.3 (c) and (a) in chapter 1).

Summarizing, human signers perform multi-class classification, while the automatic recognizer executes a form of one-class classification. This difference in recognition task influences how well information gained in one task can be applied in the other task. For example, human signers cannot recognize a sign using only the properties location and motion (see chapter 3). Does this mean that the **CDFD** classifier cannot do so either? In fact, the **CDFD** classifier performs quite well using almost exclusively location and motion features [59]. But this is evaluated using only 119 classes to model the non-target class, whereas human signers must decide between all signs in their
vocabulary. If the dataset used by CDFD contained many signs that differ only in handshape, its performance would probably be much lower. Such influences make it difficult to compare human and machine performance.

Or suppose there is a sign with a unique handshape value. For a recognition task (which sign is it?), observing this handshape would suffice\(^1\). For the task of checking correctness however (is it a correct version of sign \(a\)?), all other relevant properties must be checked and found correct as well, before the sign can be deemed correct. Which properties are relevant in this sense is not a trivial problem, and may differ per sign [2]. Therefore, a conclusion drawn from a recognition experiment (handshape suffices) is not necessarily true for a correctness-checking system as well (all properties must be checked). This does not strictly apply to the CDFD recognizer used in this thesis, because it does not truly perform correctness judgments, but rather a hybrid form of one-class classification on a dataset. But it does illustrate how a conclusion from one situation may not hold in another situation due to differences in the task that is performed.

**Differences in Processing Techniques** From the experiments in part I of the thesis, we know that modifications of input may have an effect on the recognition success of human signers. However, the details of human sign processing are largely unknown; modeling human language processing in general is a complex matter [45], and research into the cognitive processes involved in sign language processes is still ongoing [29]. Without knowing the details of human sign processing, it is difficult to predict whether a certain stimulus-response relationship observed for human signers will occur in an automatic recognizer as well.

For example, suppose a human signer can recognize the sign **BELT** based on the location property only. If this is the case because **BELT** is simply the only sign with this specific location value, then we can expect to find the same effect in our CDFD recognizer. But the human signer may use more sophisticated techniques in the recognition process. Suppose there are more signs with these location values, but they are all very infrequent signs (obscure words). The the sign **BELT** may be selected as a result of its higher prior probability. If this is the cognitive process underlying human recognition, then the automatic recognizer cannot be expected to produce the same results, since it does not utilize information about prior sign probabilities. Thus, it would not be able to distinguish between **BELT** and the other signs when using only the location property. This would explain why applying broad, general knowledge is more successful than applying specific observations: a more general observation is less likely to be connected to a specific detail of the recognition process, and therefore more likely to be applicable in a different context.

\(^1\)Note that if it were not guaranteed that the stimulus was a sign, but could also be nonsense, then all other properties would have to be checked. But in that case the handshape value would not truly be unique: there could be many nonsense movements using the same handshape
The above suggests that it would be interesting to attempt to construct a cognitive model of human sign processing; to understand not only what happens, but also why it happens. Gaining insight into the human sign recognition process could open up new avenues of improvement for automatic recognition. For example, including information about prior sign probabilities in the recognition system and using the information to choose between possible signs when a stimulus is ambiguous. At the same time, problems encountered by the automatic recognizer would indicate new directions for the cognitive research: how do human signers overcome such problems? This kind of mutual exchange could stimulate both fields. The work in this thesis can be regarded as a first step in such a process.

8.2 Investigating Human Sign Recognition Processes

The previous discussion makes it clear that there are several factors complicating the application of knowledge about human sign recognition to automatic systems. Looking at the effects of modifying stimuli and trying to apply this knowledge is an interesting first step, but as we have seen, it is most useful for detecting broad trends and pointing out interesting areas. To benefit more fully from knowledge about human sign recognition, it is necessary to learn more about the recognition process. Creating a cognitive model of human sign recognition would not only allow us to use more detailed knowledge about human recognition, but could also inspire new approaches for automatic sign recognition architectures.

Let us look at a possible model for lexical sign recognition. Emmorey and Corina [30] conducted a gating experiment in which signers saw increasingly larger parts of a sign and were asked each time to guess the sign. From their responses, it could be deduced at which point sign aspects were resolved: from that point onward, a subject produced only responses with that particular aspect value (e.g. a certain handshape). The results suggest that signers resolve sign aspects in parallel, using the resolved aspect values to eliminate candidate signs, until only one possibility is left (similar to the cohort model of [61]); alternatively, a choice could be made when only a few candidates are left, based on several parameters such as similarity to a prototype, prior probability, etc. This recognition model would explain why signs with a rare value for one of the aspects are recognized more quickly or easily (as observed in part I of the thesis): the rare value, once it is resolved, leaves only a few possible candidate signs.

Such a model of sign recognition could be used to create a recognizer built on the same principles. For example, a recognizer consisting of parallel aspect detectors, each of which continuously updates the possible aspect values (see figure 8.1). Values for which the probability falls and remains below a certain threshold could be eliminated. At each moment, a set of possible sign identifications can be determined by taking all signs in the dictionary that consist of a combination of possible aspect values. As the
Figure 8.1: Example of a recognition system inspired by human sign recognition research. Each sign aspect is processed separately. All possible values for the aspects receive initial probabilities. With each frame that is processed, the probabilities are updated based on the information extracted from the frame. Aspect values for which the probability drops below a certain threshold (e.g. 40%) are no longer considered possible. After each update, a set of possible signs is generated by selecting all signs from the dictionary that consist of a combination of possible aspect values. As long as this set consists of more than one sign, processing is continued. When only one sign remains possible, a decision is achieved. Alternatives and fine-tunings are possible in all steps of the process. For example: the aspect value probabilities could be updated after a set of frames rather than after each frame, or probability adjustments could be weighted according to their time of occurrence (so that one ‘bad’ measurement cannot easily invalidate a value that has had a high probability for many previous frames).
number of possible values for each aspect decreases when more information becomes available, the number of possible candidates decreases as well, until only one is left, which would be the recognized sign. Or the choice could be made when there are only a few possible candidates left. To select between the candidates, the probabilities of the aspect values involved could be used (the sign with the most probable aspect values is chosen), as well as other information, such as the a priori probabilities of the signs or the context-dependent probabilities of the signs given earlier recognized signs. Inspiration for such techniques can be found through human sign recognition research. The parallel handling of aspects would have the added benefit of allowing different frame rates to be used for different aspects. To retain all details of a motion path, a high frame rate is necessary. Other aspects, such as handshape, do not change much over the course of a sign: a (non-compound) sign generally contains only one or two handshapes, and in most cases, handshape pairs have specific qualities (e.g. open-closed) [73]. Such aspects could be processed using a lower frame rate. This would have the benefit of more processing time and resources being available for each frame, which would make it possible to use more complex image processing techniques, improving the feature extraction and aspect recognition.

Of course, this is only a simple example. Both the cognitive model and the recognition system are likely to be more complex than the picture sketched here. But it illustrates the way information about the cognitive processes underlying sign recognition could give rise to new recognition systems. In such systems, knowledge about human sign recognition could be integrated more easily. For example, research might show that human signers recognize handshape mainly based on certain moments in time, for example the start and end of the stroke, when the hands are relatively still. Such knowledge could be easily integrated into a system that has a parallel-processing architecture similar to human processing. Due to its construction, the information could not be applied in the CDFD-recognizer used in this thesis, because this recognizer expects the same number of properties for each frame. Thus it would not be possible to use a small number of frames for one aspect, and a large number of frames for another aspect. To summarize: in our opinion, designing systems with an architecture that resembles the structure of human sign recognition is a promising direction for the integration of human knowledge into automatic sign language recognition.

8.3 Future Directions

It would be interesting to develop an automatic sign recognition system with an architecture based on a cognitive model of human sign language processing. As mentioned above, such a process could be iterative, with knowledge about human sign processing inspiring designs for the automatic recognizer, and difficulties in automatic recognition inspiring new experiments for human signers. The work described in this
Discussion

thesis forms a tentative start to such a process. Difficulties in applying the results of human signer experiments suggest that changing the architecture of the CDFD recognizer might be an interesting next step.

Another opportunity for improvement is the dataset used for the automatic recognizer (see appendix A). The current dataset of 91 signs, though average-sized in the field of ASLR, is small with regard to the human sign vocabulary (around 5,000 in the standard lexicon[34]). This influences the quality of the classifiers trained with it, and also complicates comparison between human and machine results. More variation in the dataset would be beneficial. The current dataset consists of signs that are all made from the same rest position, which makes it difficult to evaluate techniques aimed at handling different preparations and retractions. In chapter 2, it was shown that different rest positions do not appear to influence the recognition performance for humans. Using different rest positions in the dataset would enable us to investigate how robust automatic recognition algorithms are in this respect. Small variations of the signs themselves were included in the dataset, but not in a structured way. Research such as [2] can show us which types of variation are possible, which would allow us to include such variations structurally.

Obviously, improvements in image processing techniques would also be beneficial for automatic recognition. Currently, certain tasks that human beings perform effortlessly still pose a problem to most computer vision systems, such as ignoring background noise, seeing depth, attaining high spatial resolution, etcetera. Solving these problems would improve the accuracy of the features extracted for automatic recognition, and thus make better classification possible. It would also allow an automatic recognizer to retain representations of a sign that are closer to the representations humans use than is currently possible.

Of course, applying human techniques is not the ultimate answer to all problems. In some cases, human signers may handle difficulties in such a way that reproducing the technique in the automatic system would require adding an entire inner world representation, or training the system for twelve years. In such cases, where the human-inspired solution is undesirable, it may have to be bypassed in favor of a more practical machine learning technique. Also, there are areas in which automatic systems have an advantage over human signers, such as processing (clock) speed, memorization and storage capacity. These advantages may allow them to use techniques that are not possible for human beings. An example would be the Monte Carlo method, a method that is used widely in current game-playing AI [17], and which consists of evaluating moves by playing many random games quickly and choosing the move that leads more often to a win. Human beings could not do this in their head, nevertheless, the technique works quite well for a computer with enough processing power behind it. The advantages of automatic processing should not be disregarded, and solving the problem of automatic sign recognition should not entail building an entire human being. For automatic sign language recognition, the human
way is not necessarily the only and final solution. But a closer integration of human knowledge can be a big step forward in creating better automatic sign language recognition.
This appendix describes the dataset that was used for all automatic sign language recognition experiments reported in the thesis. For the human signer experiments, different datasets were used. These datasets are described in the corresponding chapters. The datasets did not overlap. The reason for using different datasets lies in the different requirements posed on the data by each experiment. For the automatic recognizer vocabulary, several requirements were a consequence of the application of the recognizer in a learning environment for young, hearing-impaired children. For example, the concepts in the vocabulary had to be conveyable using a picture. For the human signer experiments, different demands could be imposed on the vocabulary: for example, in chapter 3, all signs were required to either contain no path movement, or to contain a downward path movement.

### A.1 Dataset Recording

The dataset used in this thesis consists of 91 isolated signs from the standard lexicon of Sign Language of the Netherlands (SLN) [34]. Examples of the signs can be viewed online at research.tenholt.nl/GW09.

Each sign was performed by 76 different volunteers. They were mostly non-signers, copying the signs from examples that were played for them on a monitor. Multiple examples of each sign were available, by different signers, to ensure some variation in the signs made by the volunteers. Inspection by a sign language instructor revealed that the variations and aberrations displayed by the volunteers resembled those of regular beginning signers.

The non-manual components of the signs were not recorded. In Sign Language of the Netherlands, almost all signs possess a non-manual component, which in many cases consists of (part of) the Dutch word mouthed along with the hand movements. The status of such mouthings as true phonemes can be questioned, since they often
Figure A.1: Setup in which the dataset was recorded. Most of the signers in the dataset were inexperienced. An example video was visible to them on the monitor, recorded from experienced signers.

disappear when a sign is used in a sentence. Also, such mouthings would give a person watching the video a strong hint to the identity of the sign, even if they did not know any sign language. To avoid including such information, the signs were recorded with no manual component.

Recordings were made with a neutral background and diffuse lighting, using a calibrated stereo camera (Allied Vision Technologies ‘Guppies’) at 25 frames per second. Figure A.1 shows the recording setup. The resolution of the images was $640 \times 480$ pixels. Start and end of a sign were indicated by the hands being on the table.

A.2 Extracted Features

The properties that were extracted from the sign videos are listed in table A.1. All properties except for the ones representing handshape were already used in [59]. The handshape representation properties were added in the course of the work described in this thesis (see chapter 5).
A.3 Sign Vocabulary

Table A.1: Properties extracted for the sign videos.

<table>
<thead>
<tr>
<th>Property</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>left/right hand co-ordinates</td>
<td>( \mathbf{h}<em>{l/r}(t) = [X</em>{l/r}(t), Y_{l/r}(t), Z_{l/r}(t)] )</td>
</tr>
<tr>
<td>left/right hand motion</td>
<td>( \tilde{\mathbf{h}}_{l/r}(t) = S{</td>
</tr>
<tr>
<td>left/right hand acceleration</td>
<td>( \tilde{\tilde{\mathbf{h}}}_{l/r}(t) = S{</td>
</tr>
<tr>
<td>left/right sideways orientation</td>
<td>( \theta_{SI/r}(t) = \arcsin(dX_{l/r}(s)/ds) )</td>
</tr>
<tr>
<td>left/right upward orientation</td>
<td>( \theta_{UI/r}(t) = \arcsin(dY_{l/r}(s)/ds) )</td>
</tr>
<tr>
<td>left/right forward orientation</td>
<td>( \theta_{FI/r}(t) = \arcsin(dZ_{l/r}(s)/ds) )</td>
</tr>
<tr>
<td>left/right hand motion curvature</td>
<td>( \tilde{k}<em>{l/r}(t) = S{\kappa</em>{l/r}(t), c_k} )</td>
</tr>
<tr>
<td>left/right hand motion curvature change</td>
<td>( \tilde{\kappa}<em>{l/r}(t) = S{d\tilde{k}</em>{l/r}(t)/dt, c_k} )</td>
</tr>
<tr>
<td>left/right hand size change</td>
<td>( \tilde{B}<em>{l/r}(t) = S{dB</em>{l/r}(t)/dt, c_B} )</td>
</tr>
<tr>
<td>hand position difference</td>
<td>difference between hand co-ordinates</td>
</tr>
<tr>
<td>left/right hand shape descriptor</td>
<td>shape representation consisting of 16 parameters per hand</td>
</tr>
</tbody>
</table>

A.3 Sign Vocabulary

Table A.2 lists the signs that were used for training and testing the automatic recognizers used in this thesis. The set is a subset of the vocabulary of ELo, an electronic learning environment for young hard-of-hearing children [78]. The ELo vocabulary only includes concepts that can be conveyed using a picture were eligible to be included. In this thesis, only the non-compound signs from the ELo vocabulary were used, because these signs are more like two signs performed in succession than a single sign. Because we look, among other things, at time courses of signs, it is preferable not to use signs that, in some respects, are actually two signs. For example in determining a sign’s stroke: a compound sign would appear to have two strokes with an epenthesis movement connecting them. For this reason, compound signs were excluded. Videos of the signs can be viewed at research.tenholt.nl/GW09.
Table A.2: Sign dataset used for automatic recognition.

<table>
<thead>
<tr>
<th>English gloss</th>
<th>Dutch gloss</th>
<th>English gloss</th>
<th>Dutch gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACROPAT</td>
<td>ACROBAAT</td>
<td>FRAME</td>
<td>LIJST SCHILDERIJDING</td>
</tr>
<tr>
<td>AIRPLANE</td>
<td>VLIEGTUIG</td>
<td>FRIES</td>
<td>PATAT</td>
</tr>
<tr>
<td>ANT</td>
<td>MIER</td>
<td>FOG</td>
<td>KIKKER</td>
</tr>
<tr>
<td>ARROW</td>
<td>PIJL</td>
<td>GHOST</td>
<td>SPOOK</td>
</tr>
<tr>
<td>AXE</td>
<td>BIJL</td>
<td>GLOBE</td>
<td>WERELDBOL</td>
</tr>
<tr>
<td>BAG</td>
<td>ZAK</td>
<td>GUITAR</td>
<td>GITTAAR</td>
</tr>
<tr>
<td>BANANA</td>
<td>BANAAN</td>
<td>HAIR DRYER</td>
<td>FHN</td>
</tr>
<tr>
<td>BAR (OF CAGE)</td>
<td>TRALIE</td>
<td>HAMMER</td>
<td>HAMER</td>
</tr>
<tr>
<td>BASIN</td>
<td>TEIL</td>
<td>HOOP</td>
<td>HOPEL</td>
</tr>
<tr>
<td>BEAK</td>
<td>SNAVEL</td>
<td>LABEL</td>
<td>ETIKET</td>
</tr>
<tr>
<td>BELT</td>
<td>RIEM</td>
<td>LADDER</td>
<td>LADDER</td>
</tr>
<tr>
<td>BIRD</td>
<td>VOGEL</td>
<td>LETTER</td>
<td>BRIEF</td>
</tr>
<tr>
<td>BOOK</td>
<td>SCHRIFT</td>
<td>LIGHTNING</td>
<td>BLIKSEM</td>
</tr>
<tr>
<td>BRUSH</td>
<td>BORSTEL</td>
<td>LION</td>
<td>LEEUW</td>
</tr>
<tr>
<td>CABIN</td>
<td>HUTJE</td>
<td>MAIL BOX</td>
<td>BRIEVENBUS</td>
</tr>
<tr>
<td>CALF</td>
<td>KALF</td>
<td>MAYONAISE</td>
<td>MAYONAISE</td>
</tr>
<tr>
<td>CAT</td>
<td>POES</td>
<td>METAL</td>
<td>METAAL</td>
</tr>
<tr>
<td>CIRCLE</td>
<td>CIRKEL</td>
<td>MICROPHONE</td>
<td>MICROFOON</td>
</tr>
<tr>
<td>CIRCUS</td>
<td>CIRCUS</td>
<td>NAIL POLISH</td>
<td>NAGELLAK</td>
</tr>
<tr>
<td>CLOUD</td>
<td>WOLK</td>
<td>NEWSPAPER</td>
<td>KRANT</td>
</tr>
<tr>
<td>COOK</td>
<td>KOK</td>
<td>OVEN</td>
<td>OVEN</td>
</tr>
<tr>
<td>COW</td>
<td>KOE</td>
<td>PEOPLE</td>
<td>MENSEN</td>
</tr>
<tr>
<td>CROSS</td>
<td>KRUIS</td>
<td>PIANO</td>
<td>PIANO</td>
</tr>
<tr>
<td>CROWN</td>
<td>KROON</td>
<td>PLANT</td>
<td>PLANT</td>
</tr>
<tr>
<td>CURTAIN</td>
<td>GORDIJN</td>
<td>PLATFORM</td>
<td>PERRON</td>
</tr>
<tr>
<td>DESK</td>
<td>BALIE</td>
<td>PLUG</td>
<td>STEKKER</td>
</tr>
<tr>
<td>DRESS</td>
<td>JURK</td>
<td>POND</td>
<td>VIJVER</td>
</tr>
<tr>
<td>DRUM</td>
<td>TROMMEL</td>
<td>POSTAGE STAMP</td>
<td>POSTZEGEL</td>
</tr>
<tr>
<td>ELEPHANT</td>
<td>OLIFANT</td>
<td>POSTCARD</td>
<td>KAART</td>
</tr>
<tr>
<td>FAIRY</td>
<td>FEE</td>
<td>PRINCESS</td>
<td>PRINSES</td>
</tr>
<tr>
<td>FEET</td>
<td>VOETEN</td>
<td>QUILL</td>
<td>STEKEK</td>
</tr>
<tr>
<td>FLAG</td>
<td>VLAG</td>
<td>SAW</td>
<td>ZAAG</td>
</tr>
<tr>
<td>FLASHLIGHT</td>
<td>ZAKLAMP</td>
<td>SHADOW</td>
<td>SCHADUW</td>
</tr>
<tr>
<td>FLAT</td>
<td>FLAT</td>
<td>SHOWER</td>
<td>DOUCHE</td>
</tr>
<tr>
<td>FLATIRON</td>
<td>STRIJKIJZER</td>
<td>SPARROW</td>
<td>MUS</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>English gloss</th>
<th>Dutch gloss</th>
<th>English gloss</th>
<th>Dutch gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQUARE (OUT-SIDE)</td>
<td>PLEIN</td>
<td>TWINE</td>
<td>DRAAD</td>
</tr>
<tr>
<td>STACK</td>
<td>STAPEL</td>
<td>WAITER</td>
<td>OBER</td>
</tr>
<tr>
<td>STAMP</td>
<td>STEMPEL</td>
<td>WATCH</td>
<td>HORLOGE</td>
</tr>
<tr>
<td>STEWARDESS</td>
<td>STEWARDESS</td>
<td>WAVE</td>
<td>GOLF</td>
</tr>
<tr>
<td>STOOL</td>
<td>KRUK</td>
<td>WOOD</td>
<td>HOUT</td>
</tr>
<tr>
<td>SUBMARINE</td>
<td>DUIKBOOT</td>
<td>WORM</td>
<td>WORM</td>
</tr>
</tbody>
</table>
Abstract

In this paper, we present an algorithm for Dynamic Time Warping on multi-dimensional time series (MDDTW). The algorithm is compared to ordinary DTW, where only a single dimension is used for aligning the series. Both one-dimensional and multi-dimensional DTW are also tested when derivatives instead of feature values are used for calculating the warp. MDDTW performed best in finding a known ground truth under noisy conditions. The algorithms were also used to perform simple classification of a set of 121 gestures. MDDTW performed as well as or better than any single dimension in all tasks. In general, DTW on feature derivatives always gave better results than DTW on feature values.

B.1 Introduction

There are various problem areas where signals need to be synchronised. When two time signals are compared, or when a pattern is sought in a larger stream of data, one of the signals may be warped in a non-linear way by shrinking or expanding along its time axis. Simple point-to-point comparison then gives unrealistic results, because one might be comparing different relative parts of the same signal/pattern. In these cases, some sort of synchronisation is needed.

Dynamic Time Warping (DTW) [54] has long been used to find the optimal alignment between two signals. The DTW algorithm calculates the distance between each possible pair of points out of two signals. It uses these distances to calculate a cumulative distance matrix and finds the least expensive path through this matrix. This path represents the ideal warp - the synchronisation of the two signals which causes the
distance between their synchronised points to be minimised. Usually, the signals are normalised and smoothed before the distances between points are calculated.

DTW has been used in various fields, such as speech recognition [69], data mining [49], and movement recognition [20, 33]. Previous work in the field of DTW mainly focussed on speeding up the algorithm, the complexity of which is quadratic in the length of the series. Examples are applying constraints to DTW [50], approximation of the algorithm [70] and lower bounding techniques [49]. Keogh et al. [50] proposed a form of DTW called Derivative DTW (DDTW). Here, the distances calculated are not between points, but between their first-order derivatives. In this way, synchronising is done based on shape characteristics (slopes, peaks) rather than simple values. Most work, however, only considered one-dimensional series.

We work on gesture recognition. Gestures are recorded by cameras and multiple features are extracted at each time instance, giving us multi-dimensional time series. We therefore investigated techniques for the synchronisation of such series.

Multi-dimensional (time) series are series in which multiple measurements are made simultaneously. Such series have an K-dimensional vector of values for each (time) instance of the series. They can be synchronised by simply picking one dimension to perform DTW with and warping the entire signal according to the warp found in this dimension. However, in many cases, all dimensions will contain information needed for synchronisation. We therefore propose multi-dimensional DTW (MDDTW) for synchronising such series. An extension of DTW into 2 dimensions was proposed by [93], but not systematically tested. In [33] a type of MD-DTW is described, but only used for fixed-length series. In the next sections, we explain the MDDTW algorithm and test it on our dataset, which consists of gestures on which multiple features were extracted for each time instance. We compare MDDTW to regular 1D DTW on various dimensions of the signals. We also looked at using derivatives instead of the values themselves (DDTW) and at combining both approaches in MDDTW.

**B.2 Multi-dimensional Dynamic Time Warping**

**B.2.1 Algorithm**

Multi-dimensional series consist of a number of measurements made at each instance. The number of measurements is the dimensionality of the series, the number of time instances its length. Note that multi-dimensional series need not be time signals, any situation in which several measurements are made simultaneously depending on one variable gives a multi-dimensional series. In this paper, we assume that measurements are stored in a matrix, in which columns are features and rows are time instances.
Let $A, B$ be two signals of dimension $K$ and length $M, N$ respectively.

1. Normalise each dimension of $A$ and $B$ separately to $\mu = 0$ and $\sigma = 1$
2. If desired, smooth each dimension with a Gaussian filter
3. Fill the $M$ by $N$ distance matrix $\text{Dist}$:
   
   $$\text{Dist}(A, B) = \sum_{j=k(l)}^{K} |A(i,k) - B(j,k)|$$
4. Use this distance matrix to find the best synchronisation with the regular DTW algorithm

**Figure B.1:** The MDDTW algorithm

Take two series $A$ and $B$. DTW involves the creation of a matrix in which the distance between every possible combination of time instances $A(i) \leftrightarrow B(j)$ is stored. In 1D DTW, this matrix is created by taking the absolute or the squared distance between each combination of points. For MDDTW, the distance between each pair of K-dimensional points must be calculated. This distance can be any $p$-Norm. We use the 1-Norm, or the sum of the absolute differences in all dimensions. To combine different dimensions in this way, it is necessary to normalise each dimension to a mean of zero and a variance of 1. For this, the dimensions must be comparable. If one dimension contains real-valued measurements and one is binary, a more sophisticated distance measure must be found. An overview of the algorithm is given in Figure B.1.

The benefits of MDDTW can be seen when multi-dimensional signals are considered that have synchronisation information distributed over different dimensions. Take the artificial 2D signal shown in figure B.2 (a). It is clear that for the first half (in time) of the signal, dimension 1 is useful for finding the correct synchronisation, whereas dimension 2 is uninformative. The converse is true for the second half of the signal. If we were to perform 1D DTW on this signal using dimension 1, the result would be as shown in figure B.2 (b). The second half of the signal is uniformly synchronised, since there is no information for 1D DTW to work with. But it can be seen that for dimension 2, this is not the ideal synchronisation. 1D DTW on dimension 2 gives a similar (but converse) result.

MDDTW takes both dimensions into account in finding the optimal synchronisation. The result is a synchronisation that is as ideal as possible for both dimensions, as shown in figure B.2 (c). This is the advantage of MDDTW over regular DTW.

Though the signal in figure B.2 is artificial, the situation it depicts is by no means unrealistic. Figure B.3 shows a similar situation from a real-world multi-dimensional...
Figure B.2: The necessity for MDDTW. (a) shows two artificial 2D time series of equal mean and variance. Dimension 1 is shown in column 1, dimension 2 in column 2. The signals contain synchronisation information in both dimensions. If 1D DTW is performed using the first dimension, the result is suboptimal for dimension 2, as illustrated in (b). Note that the peaks and valleys in dimension 2 are not aligned properly. 1D DTW on dimension 2 gives a similar suboptimal match for dimension 1. MDDTW takes both dimensions into account and finds the best synchronisation (c).
signal: x- and y-co-ordinates of the right hand for the Dutch sign for acrobat. Here, the y-co-ordinate is informative for the first part of the signal, and the x-co-ordinate for the second part. To get both peaks properly matched, MDDTW is necessary.

B.2.2 DTW on Derivatives

In [50], it is argued that for synchronising shape-characteristics of signals (such as peaks and slopes), it is beneficial to take the first derivative of the signal and perform DTW on this. Since we want to perform such shape-matching in each of our dimensions, we considered this option for the MDDTW-algorithm. In our case, this meant taking the first-order derivative in each dimension separately. This gives us information about the slopes and peaks in each of our dimensions. The signals were first smoothed in each dimension with a Gaussian filter ($\sigma = 5$) to diminish noise effects. Then an approximation of the derivative was taken in each dimension using the filter

$$
der(a(t)) = \frac{a(t+1) - a(t-1)}{2}$$

Now we can perform MDDTW either on the signals themselves or on their derivatives. As a third option, we took the first-order derivatives and added them to the signal as extra dimensions, doubling the dimensionality of the signal. In this setting, both the signal values themselves and the derivatives are taken into consideration when searching for the ideal warp. We tested our algorithm in all three settings. They are denoted as S(signal), D(erivative) and SD(signal+derivative). In each case, the signals were smoothed and normalised before DTW was commenced.

B.2.3 Dimension Selection

To compare signals and various dimensions, it is necessary to normalise the signals. However, this is a problem if a signal contains only noise. For example, position data gathered on the non-dominant hand in a one-handed gesture. Normalisation will enlarge the noise in this dimension to the same proportions as the informative data in other dimensions (such as right hand positions). The noise will then heavily influence the synchronisation. This is not what we want. We therefore perform dimension selection before commencing normalisation and synchronisation.

In dimension selection, the variance in each dimension of a multi-dimensional signal is calculated. If the variance falls below a certain threshold, the dimension is not taken into account in the synchronisation process, nor in the calculation of feature distances. In this way, dimensions with hardly any variance, that probably consist of noise, are disregarded. The variance threshold was empirically determined for our dataset by
calculating the variance of known noise-dimensions (inert hands). It is also possible to weight dimensions rather than simply select or discard them.

In the next section, we give a number of tests in which we compared MDDTW in various settings (S,D,SD) to regular DTW (1D DTW) in the same settings on various dimensions. For all tests, the same filters for smoothing and derivatives were used.

B.3 Experimental Results

We tested the MDDTW algorithm in different ways. First, we wanted to assess the accuracy of various DTW algorithms on a known ground truth. For this purpose, we created artificial warpings on a number of signals and stored the warps. We then applied various versions of DTW to the warped signals and compared the synchronisations calculated to the stored true synchronisations.

Secondly, we tested our algorithm in the domain of gesture recognition. We performed simple classification on a set of 121 gestures by comparing unknown examples to class prototypes. We used the synchronisation found by the various algorithms to determine at which time-points features should be compared. Better synchronisation should give
more appropriate feature comparison and therefore higher classification scores. Data for all tests was retrieved from our gesture dataset, which is described below.

**B.3.1 Dataset**

In our experiments, we used a dataset of gestures. The gestures are signs from the standard vocabulary of Sign Language of the Netherlands. The gestures were recorded with 2 stereo cameras and a number of features were automatically extracted from each pair of frames, resulting in a multi-dimensional signal. The extracted features were: 3D positions (x,y,z), relative to the head, of the left and right hand. Each gesture was stored in a NrOfFrames x 6 matrix. The gestures varied in length.

Our dataset consists of 121 different gestures. Each gesture was recorded from 67 different persons (all right-handed), giving us 67 examples for each of the 121 classes (for 9 classes, there were only 66 examples). In addition, there is a set of prototypes consisting of one example per class. These examples were hand-picked on the grounds of being a correct version of the class they represented. They were not optimised with respect to any of the classification methods mentioned below.

Many gestures are one-handed, in which case the left hand is inert. In the two-handed gestures, the left hand in most cases copies the right. For this reason, we tried only right hand features for 1D DTW, since in almost all cases the left hand would either be less informative (inert) or equally informative (copying).

**B.3.2 Artificially Warped Signals**

We created a warp as follows: we took a multi-dimensional signal (a gesture) and copied it. In the copy, we chose a random anchor point. This point was shifted 20 time instances to the left or right (direction was also chosen randomly). The adjacent points were shifted a fraction of 20 points, the fraction decreased with their increased distance to the anchor point. This created a localised time-warp. The warp was the same in all dimensions. After warping the time axis, the values in each dimension were re-interpolated to the original time axis. This gave us a warped signal whose correct warping to the original signal was known. We will refer to the artificially warped signal as the distorted signal.

After warping, the original and the distorted signal were normalised. Random noise was then added to both to create some differences in y-values. The variance of the noise was a given fraction of the variance of the signal. Figure B.4 shows an example of an artificial warping (zero noise).

We first smoothed both signals in each dimension with a Gaussian filter ($\sigma = 5$), just as we would do for a real signal. We then applied various versions of DTW to the original
and distorted signals and stored the calculated warps. We calculated the goodness of a warp as follows: let the original signal be \( s \), the distorted signal \( s' \). Let \( GT \) denote the ground truth warp, and \( W \) the warp found by the algorithm. Let \( W : s(i) \rightarrow s'(j) \) denote that \( W \) warps \( s(i) \) onto \( s'(j) \). Then the average aberration \( Ab \) is given by

\[
Ab = \frac{\sum_{i=1}^{M} |j - j'|}{M}
\]

where \( M \) is the length of the warp \( W, GT : s(i) \rightarrow s'(j) \) and \( W : s(i) \rightarrow s'(j') \).

We performed this operation on one random example of each class in our set, and took the median of the errors of all 121 examples. We repeated this 20 times. The means and standard deviations over these 20 runs are given in Table B.1.

It can be seen that MDDTW performs better than any single-dimension DTW variant. When there is no noise, there is no significant difference between the algorithms. Adding more noise enlarges the difference between MDDTW and the one-dimensional variants. More noise makes the correct warp more difficult to find, because even the correct warp will display differences in feature values. MDDTW can make a profit, because it has multiple dimensions in which the same warp is sought, whereas noise is different in each dimension, cancelling out to some extent.
Table B.1: Average aberrations of the ground truth in frames of the calculated warp for various types of DTW. 1D RHX means one-dimensional DTW using the right hand x-co-ordinate, 1D RHY using the right hand y-co-ordinate, etc. The aberration was calculated 20 times for the entire set. The values shown are the means and standard deviations over 20 runs. Noise level indicates the ratio noise variance / signal variance. Each type of DTW was performed both on the feature values and the feature derivatives. MDDTW was also performed on both combined.

<table>
<thead>
<tr>
<th>DTW on</th>
<th>noise level</th>
<th>1D RHX</th>
<th>1D RHY</th>
<th>1D RHZ</th>
<th>MDDTW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>0</td>
<td>1.25 (0.03)</td>
<td>1.24 (0.03)</td>
<td>1.23 (0.03)</td>
<td>1.25 (0.03)</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>1.75 (0.05)</td>
<td>1.96 (0.05)</td>
<td>1.78 (0.05)</td>
<td>1.37 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2.28 (0.06)</td>
<td>2.73 (0.09)</td>
<td>2.56 (0.10)</td>
<td>1.75 (0.05)</td>
</tr>
<tr>
<td>Deriv</td>
<td>0</td>
<td>1.31 (0.04)</td>
<td>1.29 (0.04)</td>
<td>1.27 (0.03)</td>
<td>1.24 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>1.60 (0.05)</td>
<td>1.75 (0.05)</td>
<td>1.64 (0.04)</td>
<td>1.33 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1.94 (0.07)</td>
<td>2.30 (0.09)</td>
<td>2.15 (0.07)</td>
<td>1.61 (0.04)</td>
</tr>
<tr>
<td>Both</td>
<td>0</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>1.25 (0.03)</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>1.31 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>1.58 (0.06)</td>
</tr>
</tbody>
</table>

DDTW appears to improve the results for conditions with more noise, not for noiseless conditions. The reason may be that taking the derivative will introduce some noise in the noiseless condition. Note that for MDDTW without noise, DDTW does not cause performance to drop.

B.3.3 Application in Gesture Classification

In our domain, we want to classify gestures by comparing their feature values. MDDTW should help us find the correct correspondences of time points between different gesture examples, so that the appropriate features will be compared.

To test the merits of MDDTW over ordinary DTW in this respect, we executed a few simple classification tasks using various versions of DTW for synchronisation. The tasks are entirely equal in all other respects.

Nearest Neighbour Classification

One basic way of testing classification performance is the nearest neighbour (NN) scheme. We used the prototypes of our dataset as the training set (one prototype per class). All other gestures were used as the test set. In the test, we simply warped each test gesture on each prototype. We then calculated the feature distance between test
Table B.2: Classification error of the Nearest Neighbour algorithm using various DTW variants and conditions for synchronisation. 1D RHX means one-dimensional DTW using the right hand x-co-ordinate, RHY using the right hand y-co-ordinate, etc. Each was tested 20 times on random samples of the data. The values shown are the means and standard deviations over 20 runs. The minimum error is shown in bold. It was significantly lower than the other errors (one-sided T-test, $p < 0.05$).

<table>
<thead>
<tr>
<th>DTW type</th>
<th>signal</th>
<th>derivative</th>
<th>signal + derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D RHX</td>
<td>0.791 (0.006)</td>
<td>0.722 (0.007)</td>
<td>n.a.</td>
</tr>
<tr>
<td>1D RHY</td>
<td>0.791 (0.006)</td>
<td>0.699 (0.006)</td>
<td>n.a.</td>
</tr>
<tr>
<td>1D RHZ</td>
<td>0.834 (0.004)</td>
<td>0.759 (0.005)</td>
<td>n.a.</td>
</tr>
<tr>
<td>MDDTW</td>
<td>0.748 (0.006)</td>
<td><strong>0.690 (0.006)</strong></td>
<td>0.699 (0.006)</td>
</tr>
</tbody>
</table>

gesture and prototype in the following way: let $A$ be the test gesture, $B$ the prototype, and $W$ the calculated warp. Then

$$Dist(A, B) = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=k(l)}^{k(K)} |A(i, j) - B(i', j)|$$

where $N$ is the length of $A$, $K$ contains the dimensions that are selected for both $A$ and $B$, e.g. [1,3,5] (see section 2.3), and $W : A(i) \rightarrow B(i')$. A test gesture was given the class of the prototype to which it had the smallest distance.

Our test set consisted of 8098 examples in 121 classes. We took 6000 random samples from this set and computed the error of the nearest neighbour classification for several different methods of synchronisation. This process was repeated 20 times. The mean error and standard deviation over these 20 runs are given in table B.2. The errors in table B.2 are large, which is to be expected for a NN with only one training example per class. However, we are not interested in classification performance so much as in comparing classification performance when various forms of DTW are used to achieve synchronisation. Table B.2 shows that when synchronisation is calculated with MDDTW, performance is better than with DTW synchronisation based on a single dimension. The best performance outright is given by MDDTW on the first derivatives of the dimensions. It performs significantly better than any other DTW variant. For each DTW variant, the Derivative condition performed better than the Signal or SD condition. All significances were calculated with a one-sided T-test, $p \leq 0.05$.

Warp Distances as Dissimilarity Features

In the previous section, we used the distance between a gesture example and a prototype, calculated for various synchronisations, to determine the class of the
example. A more sophisticated method is to use the distances of an example to each prototype as (new) dissimilarity features of the example [67] and represent each example by its pattern of distances to all prototypes. This gives us a different dataset: we still have 67 examples for each of the 121 classes, but now, each example has 121 features, namely, its distances to each of the prototypes. The values of these features will differ for different synchronisations. Therefore, the examples have different feature values for each DTW variant.

When we regard the dissimilarities as ordinary features, we can train and test ordinary classifiers on our new dataset. Again, we are not interested in the classification performance in itself, but in differences due to various DTW conditions. The performance of a few classifiers for different DTW conditions was evaluated. Each classifier was trained and tested 20 times on random partitionings of the dataset. The ratio of training set and test set was 3:1. Mean and variance of the classification error over these 20 runs are shown in table 3.

The linear density-based classifier gives the better performance, but we are more interested in the relative performance for the various forms of DTW and MDDTW. MDDTW outperforms 1D DTW when the x- or z-co-ordinate are used for synchronisation. For 1D DTW on the y-co-ordinate, performance is equal with MDDTW both for the signal and for the derivative condition. Possibly the y-co-ordinate is the most informative dimension for most gestures in our dataset when it comes to synchronisation. But there are exceptions, gestures which hardly vary in y-co-ordinate. When tested on these classes only, MDDTW performs better than DTW on the y-co-ordinate (results not shown). But the exceptions are few, so the average classification performance over the entire set of gestures is not influenced much.

Best performance outright is given by the MDDTW and y-co-ordinate derivatives. There is no significant difference between these two for either classifier. For MDDTW and for each DTW variant the derivative condition performed significantly better than the signal condition. All significances were tested with a one-sided T-test, p < .05.

### B.4 Discussion

We discussed a novel technique for synchronising multi-dimensional signals. The MDDTW algorithm is an extension of the regular DTW algorithm that takes all dimensions into account when finding the optimal synchronisation between two signals. We tested MDDTW against one-dimensional DTW, and also looked at the performance of both algorithms when first-order derivatives were used instead of feature values.

When testing against a known ground truth, MDDTW, as expected, showed an advantage over 1D DTW when more noise was added. In various classification tasks,
MDDTW usually outperformed 1D DTW on all dimensions. Sometimes, the performance of the best 1D DTW dimension equalled that of MDDTW. However, since there are cases for which the best dimension does not give good results, and since MDDTW performance is always equal to or better than that of a single dimension, using MDDTW is preferable.

MDDTW has the disadvantage of being more expensive than DTW in its calculation of the distance matrix. The extra processing time is linear in the number of dimensions. For large or high-dimensional datasets, it may therefore be worth the effort to discover the best single dimension and use 1D DTW. The best dimension in our dataset was not the one with the highest variance, so it is not immediately clear how it should be found. Empirical testing on a training set is probably the best way. For smaller or low-dimensional datasets, MDDTW is a better option.

For each single dimension and for MDDTW, DDTW gave the better performance. For our gesture dataset, top performance was given by MDDTW on derivatives, sometimes equalled by DTW on the y-co-ordinate derivatives. We therefore conclude that for our dataset, MDDTW on derivatives is the best synchronisation method.
C

RECOGNITION ALGORITHMS

C.1 Combined Discriminative Feature Detectors
Algorithm

The CDFD algorithm is the basic recognition algorithm used in this thesis [56, 57]. The recognizer consists of a set of classifiers, one for each sign in the dataset. Each classifier is individually trained and evaluated.

One CDFD sign classifier is a hybrid of one-class and two-class classification. In feature selection and evaluation, the non-target class is modeled: all non-target signs in the dataset are used as a representation of the non-target class. In training and calibrating the feature classifiers and in determining the sign detection threshold, only the target class is used. The various phases of the algorithm are explained below.

C.1.1 Feature Selection

The feature selection procedure is illustrated in figure C.1. For each feature (a property in a certain frame), a distribution is compiled of target class and non-target class values. Only features that can be used to distinguish between target and non-target class are selected. The criterion is that no more than 25% of the non-target class distribution may overlap the middle 50% of the target class distribution:

\[
F_{f,p}^{Neg}(t_{0.75}) - F_{f,p}^{Neg}(t_{0.25}) \leq F_{f,p}^{Neg}(t_{1.0}) \begin{cases} 
\text{select } (f, p) & \text{if } < 0.25 \\
\text{select } (f, p) & \text{if } \geq 0.25
\end{cases}
\]

where \( F_{f,p}^{Neg} \) denotes the cumulative non-target class distribution of property \( p \) in frame \( f \), and \( t_a \) denotes the feature value for which \( F_{f,p}^{Pos}(t_a) = a \).
C.1.2 Training of Feature Classifiers

For the selected features, feature classifiers are trained as illustrated in figure C.2. For all training examples, the distance to the mean is calculated (a). A threshold on the distance to mean is set so that 90% of the training example values fall within the threshold (2). This threshold is used to make a simple binary-output classifier (3):

\[ x' = |x - \mu| \]

Threshold = \( x' \) for which \( F_{f,p}^{\text{Neg}}(x') = 0.9 \)
Figure C.2: Feature classifier training in CDFD
**Figure C.3:** Setting of Sign Detection Threshold in CDFD

**Figure C.4:** Testing in CDFD
C.1.3 Calculation of SDT

After feature selection and the training of feature classifiers, it must be determined which fraction of the selected features must be detected before the sign is recognized. Figure C.3 illustrates the procedure that is used. The features of the target training examples are classified using the feature classifiers. For each example, the ratio of detected features to selected features is calculated. The median ratio is then taken over all training examples. This is the value that is used as the Sign Detection Threshold or SDT. Note that when the SDT is first set, 50% of the target training examples will be rejected. However, after the SDT is set, the feature classifiers are calibrated. Using the SDT, the feature classifier thresholds are widened until 95% of the examples is recognized (see section C.1.5).

C.1.4 Testing

The test procedure is illustrated in figure C.4. The selected features of a test example are classified using the feature classifiers. If the resulting ratio of detected features to number of selected features is higher than the Sign Detection Threshold, the example is recognized, otherwise, it is rejected.

C.1.5 Calibration

After the SDT is set, the feature classifiers are calibrated in the manner illustrated in figure C.5. All positive training examples are tested using the feature classifiers and the SDT (see C.1.4). As long as the percentage of recognized examples is not (close enough
to 95%, the feature classifier thresholds are adjusted and the examples tested again. See [56, 59] for more details about the calibration procedure.

### C.2 Key Frame Algorithm

The Key Frame (KF) algorithm is introduced in chapter 7 as an alternative to CDFD. In the KF algorithm, only certain frames are used to represent the sign. These frames are chosen using pre-existing knowledge. There is no automatic feature selection (though such a step could be added in the future). An extra layer of property classifiers is added, classifying properties as a whole based on properties in multiple frames. Using the output of the property classifiers, a SDT is set. From there on, the procedure is identical to CDFD. The steps of the KF algorithm are described below.

#### C.2.1 Calculation of Key Feature Vectors

In the KF algorithm, key feature vectors are used to represent a sign. The vectors are calculated as follows: certain frames of the sign are selected as key frames: the first and last frame of the stroke, and one or more frames equally spaced in between (subsampling of the stroke). Around each key frame, a number of consecutive frames is selected (3 or 5). For these frames, features are extracted. The extracted feature vectors are combined by taking the median or the average per property over consecutive frames. The result is one key feature vector per key frame. Figure 7.2 in chapter 7 illustrates the procedure.
C.2.2 Feature Classifiers and Property Classifiers

Figure C.6 illustrates how features and properties are classified in the KF algorithm. Feature classifiers are trained in the same manner as for CDFD (see section C.1.2), one classifier for each property in each key feature vector. Then, properties are classified by combining key feature vectors per property through logical AND. The result is a classification per property.

C.2.3 SDT, Calibration and Testing

After property classification, the KF algorithm is identical to the CDFD algorithm, except that where the latter uses the outputs of the feature detectors to determine the SDT and classify an example, KF uses the outputs of the property classifiers. The SDT calculation procedure is illustrated in figure C.7. The testing procedure is similar (compare figure C.4).
Figure C.7: Determination of SDT in KF

Figure C.8: Determination of SDT in aspects algorithm
C.3 Aspects Algorithm

The aspects algorithm is a variant of the KF algorithm. It uses the same key feature vector representation and trains and classifies features and properties in the same manner as KF. However, instead of recognizing the sign directly based on the result of the property classifiers, properties are divided into groups called aspects. An aspect corresponds to a linguistically relevant sign characteristic, such as location or handshape. Each aspect of a sign is recognized separately, and the sign is only considered correct if all aspects are recognized. A selection of relevant aspects can be made per sign if desired. The steps of the aspects algorithm are explained below.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Feature Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>left hand position</td>
<td>left hand Y, Y, Z</td>
</tr>
<tr>
<td></td>
<td>dLeftHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>ddLeftHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>LeftHandMoveAngle3D</td>
</tr>
<tr>
<td>left hand motion</td>
<td>(sidewayness,upwardness,forwardness)</td>
</tr>
<tr>
<td></td>
<td>LeftHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dLeftHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dLeftHandSize</td>
</tr>
<tr>
<td>left hand configuration</td>
<td>16 shape descriptors</td>
</tr>
<tr>
<td>right hand position</td>
<td>right hand Y, Y, Z</td>
</tr>
<tr>
<td></td>
<td>dRightHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>ddRightHandPos 3D magnitude</td>
</tr>
<tr>
<td></td>
<td>RightHandMoveAngle3D</td>
</tr>
<tr>
<td>right hand motion</td>
<td>(sidewayness,upwardness,forwardness)</td>
</tr>
<tr>
<td></td>
<td>RightHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dRightHandCurvature</td>
</tr>
<tr>
<td></td>
<td>dRightHandSize</td>
</tr>
<tr>
<td>right hand configuration</td>
<td>16 shape descriptors</td>
</tr>
<tr>
<td>hand relationship</td>
<td>LeftRightDistance</td>
</tr>
<tr>
<td></td>
<td>(horizontal, vertical, depth)</td>
</tr>
</tbody>
</table>
C.3.1 Determining SDTs

The first part of the aspects algorithm, up to and including property classification, is identical to KF. But instead of determining one SDT, properties are grouped into aspects according to table C.1, and a separate SDT is determined for each aspect. Figure C.8 illustrates this procedure.

C.3.2 Calibration and Testing

Once the SDTs are determined, a sign can be tested. The test procedure, up to and including property classification, is identical to KF. After property classification, the properties are divided into aspects, then the property classification outputs are averaged (again per aspect) and compared to the corresponding aspect SDT. If the aspect scores above the aspect SDT, the aspect is recognized. A sign is only recognized if all aspects are recognized.

The calibration procedure is the same as in CDFD. Basically, the aspects algorithm divides the data into aspects and performs KF for each aspect separately, combining the results of the aspect recognizers through logical AND to classify the sign.
BIBLIOGRAPHY


[80] W. C. Stokoe. *Sign language structure: An outline of the visual communication systems*


Automatic sign language recognition is a relatively new field of research (since ca. 1990). Its objectives are to automatically analyze sign language utterances. There are several issues within the research area that merit investigation: how to capture the utterances (cameras, magnetic sensors, instrumented gloves), how to extract interesting information from the captured data, and how to classify signs or sentences automatically using the extracted information. These issues are of an immediate and basic nature, and must be solved before any automatic recognition of sign language can be achieved. But other issues, pertaining to the nature of sign language and human recognition, are no less interesting: which elements of a sign are important for the meaning of an utterance? How do consecutive signs influence one another? Why are certain types of variation unimportant while others change the meaning of the sign? Automatic sign language recognition has, until recently, mostly focused on the first set of issues. In this thesis, we attempt to integrate knowledge about sign languages and human sign recognition into the automatic sign recognition process. Research on the (psycho)linguistics of sign languages is itself quite young (since ca. 1960), and many questions as yet unanswered. For this reason, we conduct our own studies of human sign language recognition. The knowledge gained from these experiments is applied in an existing automatic sign language recognition system.

The thesis is divided into two parts: the first part describes the experiments conducted with human signers, the second part describes experiments investigating the possibilities of integrating such knowledge in the automatic recognizer. This recognizer is meant to be used in an interactive environment for young children to practice sign language vocabulary. For this reason, it is vision-based (which is unobtrusive), and only handles isolated signs.

The experiments in part I of the thesis investigate the information content of various sign elements: fragments of a sign in time (chapter 2), and the sign aspects handshape and hand orientation (chapter 3). In time, the central phase of a sign is the most informative one, equally informative to the entire sign. Recognition based on other phases is also possible to a certain extent, and the transition from the preparation phase to the central phase appears to be a salient moment. As for the aspects, the aspect handshape proves more useful for recognition than hand orientation. Chapter 4 gives an overview of the human recognition research and discusses possibilities for application.
In part II, the possibilities of utilizing the results of part I in the recognition system are investigated. Chapter 5 describes the addition of the handshape feature to the system (which chapter 3 showed to be the most interesting feature to add). Adding handshape gives a small improvement in the recognition performance. In chapter 6, the salience of the sign fragments used in chapter 2 for the automatic recognizer is investigated. The central phase proves to be the most informative one, as it was for human signers. Chapter 7 describes experiments in which a small set of frames is used to represent a sign. The results show a deterioration in recognition performance. Strict demands on the correctness of the remaining frames are probably partly responsible for the performance decrease.

In conclusion, we can say that applying human knowledge in automatic sign language recognition is a complex task. Conclusions about human sign recognition do not necessarily hold for the automatic recognizer as well. The most important obstacles for utilizing information successfully seem to be: 1) data acquisition: computer vision is not as accomplished as human observers in capturing the complex, dynamic hand and face motions that form sign language. This means that information that is present in a sign movement for a human being may not be (correctly) observed by an automatic vision analysis system. Thus, the data that humans work with is not necessarily identical to the data the recognizer works with, and this may cause techniques that are successful for human signers to fail in the automatic system. And 2) differences in basic system architecture. Research into human sign recognition is still ongoing, there is no clear model of human sign recognition yet. This makes it more difficult to translate observations from human sign recognition to the automatic recognizer: human signers may use techniques that are not compatible with the current architecture of the recognizer. For example: human signers may process aspects independently. If the recognition system processes all data as a single stream, then such a technique cannot be implemented. A more thorough understanding of human sign recognition, more sophisticated computer vision techniques, and a close co-operation between the fields of automatic sign language recognition and human sign perception, seems the best way to overcome these obstacles.
SAMENVATTING

Automatische gebarentaalherkenning is een relatief nieuw onderzoeksgebied (sinds ca. 1990). Het doel van automatische gebaarherkenning is het automatisch analyseren van uitingen in gebarentaal. Er zijn meerdere interessante kwesties in dit onderzoeksgebied: hoe kan men de gebarentaaluiting het beste vastleggen — camera’s, magnetische sensoren, handschoenen met sensoren? Hoe halen we interessante informatie uit de verzamelde data? En hoe kunnen we een gebaar automatisch classificeren op basis van deze informatie? Dit zijn zeer directe, basale problemen, die opgelost moeten worden voordat er berhaupt gebaarherkenning bedreven kan worden. Maar er zijn andere kwesties die evenzeer van belang zijn voor herkenning, en die betrekking hebben op de aard van de gebarentaal: welke onderdelen van een gebaar zijn belangrijk voor de betekenis? Hoe benvloeden gebaren elkaar als ze na elkaar gemaakt worden? Welke variaties in gebaren zijn belangrijk voor de betekenis, en waarom?

Tot voor kort heeft de automatische gebarentaalherkenning zich voornamelijk beziggehouden met de eerste set vragen, de basale kwesties. In dit proefschrift proberen wij kennis over gebarentaal en menselijke gebaarherkenning de integreren in de automatische gebaarherkenning. Gebarentaal is een redelijk nieuw onderzoeksgebied (sinds ca. 1960), waarin nog vele vragen open zijn. Omdat de bestaande kennis verre van volledig is, hebben wij besloten zelf onderzoek te doen naar menselijke gebaarherkenning. Vervolgens wordt gekeken hoe de kennis die in deze experimenten opgedaan wordt, gebruikt kan worden in een bestaande automatische gebaarherkenner.

Dit proefschrift bestaat uit twee delen: in het eerste deel wordt het onderzoek naar menselijke gebaarherkenning beschreven, het tweede deel beschrijft onderzoeken naar de mogelijkheden om de opgedane kennis is de automatische herkenner toe te passen. Deze herkenner is onderdeel van een interactief leersysteem waarin jonge kinderen hun gebarentaalvocabulaire kunnen oefenen. Vanwege deze toepassing is de herkenner visueel (data wordt verzameld met camera’s, er is geen hinderlijke hardware nodig), en werkt hij alleen met individuele gebaren (geen zinnen).

In de experimenten uit deel I van het proefschrift wordt onderzocht hoeveel nuttige informatie onderdelen van een gebaar bevatten. Er wordt gekeken naar gebaardelen (fases) in de tijd (hoofdstuk 3) en naar de gebaaraspecten handvorm en handoriëntatie (hoofdstuk 5). In de tijd blijkt dat de centrale fase van een gebaar het meest informatief
is, even informatief als het gehele gebaar. Op basis van andere fases is eveneens een zekere mate van herkenning mogelijk. De overgang van de fase preparatie naar de centrale fase lijkt een informatief moment. Op het gebied van de gebaaraspecten blijkt dat handvorminformatie nuttiger is voor de herkenning dan handoriëntatieinformatie. In hoofdstuk 4 wordt een overzicht gegeven van het onderzoek naar menselijke gebaarherkenning, en worden de toepassingsmogelijkheden besproken.

Deel II van het proefschrift gaat over manieren waarop de kennis uit deel I toegepast zou kunnen worden in de automatische herkenner. Hoofdstuk 5 beschrijft de tovoeging van een handvormfeature in de herkenner, het feature dat in hoofdstuk 3 het meest informatief bleek. Deze tovoeging heeft een kleine toename in herkenningsscores tot gevolg. In hoofdstuk 6 wordt onderzocht hoe informatief de testfragmenten uit hoofdstuk 2 zijn voor de automatische herkenner. De centrale fase blijkt voor de machine het meest informatief, net als voor de mens. Hoofdstuk 7 beschrijft een aantal experimenten waarin een gebaar gerepresenteerd wordt met een klein aantal frames. Deze aanpak heeft een daling in de herkenningsscores tot gevolg. Deze daling is waarschijnlijk deels te wijten aan de strikte eisen die worden gesteld aan de correctheid van de gebruikte frames.

In conclusie kunnen we stellen dat het toepassen van menselijke kennis in automatische gebaarherkenning geen eenvoudige taak is. Conclusies die getrokken worden over menselijke gebaarherkenning gelden niet noodzakelijkerwijze voor een automatisch systeem. De belangrijkste problemen voor het gebruiken van menselijke kennis lijken te liggen op de volgende gebieden: 1) het verkrijgen van de data: computer vision is niet zo vaardig als de mens in het analyseren van complex, dynamische hand- en gezichtsbewegingen waaruit gebarentaal bestaat. Dit betekent dat de computer niet noodzakelijkerwijze kan ‘zien’ wat de mens ziet; informatie waarover de mens beschikt is wellicht niet toegankelijk voor de computer. Mensen en machine beschikken mogelijk over verschillende data, en dat kan tot gevolg hebben dat een herkennings techniek die effect sorteert voor de mens, niet goed werkt voor de machine. En 2) verschillen in systeemontwerp. Onderzoek naar menselijke gebaarherkenning is verre van compleet, en er zijn nog geen algemeen geaccepteerde modellen voor menselijke gebaarherkenning. Dit maakt het lastig menselijke technieken te vertalen naar het gebied van de automatische herkenning: wellicht gebruiken menselijke herkenners technieken die niet uitvoerbaar zijn in de automatische herkenner. Bijvoorbeeld, menselijke gebaarders verwerken gebaaraspecten misschien parallel. Deze techniek zou niet toegepast kunnen worden in een herkenningsysteem dat alle data als n stroom verwerkt. Een beter begrip van het menselijk gebaarherkenningsproces, meer ontwikkelde computer vision, en nauwe samenwerking tussen de vakgebieden van automatische gebaarherkenning en menselijke gebaarperceptie lijkt de beste manier te zijn om de problemen te overwinnen.
Acknowledgments

Many people helped me in the course of my work as a PhD. student. First of all, there are my fellow PhD. students in the project, Jeroen Lichtenauer and Jeroen Arendsen. Two very different characters, but both great colleagues and friends. I still remember the three of us going out in Maastricht together, having a great time. Thank you both for your inspiring co-operation in the ELo project. I would further like to thank Marielle Elzenaar for her assistance in all matters concerning Sign Language of the Netherlands, and for the way she cheerfully agreed to be recorded many times performing signs in all sorts of ways. I also want to thank Arend Harteveld, who sadly passed away, for creating the various interfaces needed in the experiments.

The work described here was done in co-operation with the Dutch Foundation for the Deaf and Hearing-impaired Child, and the Royal Auris Group. I want to thank all members of these two institutions for the very pleasant co-operation, particularly Gerard Spaai, Connie Fortgens, Elmar Wenners, and Michel Esveld. Without volunteers, this work would not have been possible. Thus, I want to thank all the signers who participated in my (sometimes tedious) experiments, as well as all the people who volunteered an hour of their time to be recorded as part of the main dataset.

My time in the Information and Communication Theory Group of Delft University was very pleasant, thanks to the great people who work(ed) there. I want to thank all of them, but a few merit a separate mention. Shortly after I started, our room was joined by Yunlei Li, whose warm friendship has been a highlight of my time in the group. I also want to thank her for being my first Chinese mentor and drilling me relentlessly on the (sometimes difficult) pronunciation of the sounds of Mandarin. Robbert Eggermont has been another great friend in the group, helping me out with many computer mishaps as well as with recording the main dataset. I would also like to thank David Tax, for his inspiring brainstorms on matters concerning pattern recognition, and for having the patience to let me try to explain our quaint algorithm to him time and time again. First Anja van den Berg, and later Saskia Peters, have been very helpful in all formal matters surrounding the life of a PhD. student. Berend Berendsen and Vikas Gupta have been excellent roommates during the later part of my time in the group.

Many people provided my with advice or inspiration at conferences, at meetings, or through e-mail contact. I would like to thank Els van der Kooij, Onno Crasborn, Inge
Zwitserlood, Victoria Nyst, Ulrich von Agris, Roland Pfau, and any others I may have forgotten.
I thank Harold Nefs for providing me with the eye-gaze tracking data used on the cover.
I have been fortunate to have four very active and interested supervisors. Ans Koenderink-van Doorn has been a great mentor, reading all my work conscientiously and relentlessly forcing me to examine, at every step, the scientific correctness of my work. Huib de Ridder has many times inspired new directions for my work with his creative suggestions. With his sharp insights, Marcel Reinders has often caused me to re-examine my work, while the calm and realistic insertions of Emile Hendriks would keep us from going too far off course.
I want to thank Herman Hiddema, for brainstorming with me many times on many subjects, and for many other things as well. And I thank my family and friends, who have always supported me and were prepared to listen to long explanations of what it was I was actually doing.
Gineke ten Holt was born on May 20, 1979 in Groningen. She graduated with honours from the Praedinius Gymnasium in Groningen in 1997. She went on to study Physics at the University of Amsterdam in 1997, and obtained her first-year degree with honours. Based on this degree, she was admitted to Artificial Intelligence at the University of Groningen in 1998. In 2004, she obtained her Master’s Degree in Artificial Intelligence with honours. The subject of her MA thesis was automatic recognition of Sign Language of the Netherlands.

Early 2005, she started as a PhD. student in the Information and Communication Theory Group at Delft University of Technology. She worked in a project called ELo: an electronic learning environment for young, hearing-impaired children. The aim of this project was to create a vision-based automatic sign language recognizer that could be used in an interactive environment in which young children could practice their sign language vocabulary. In this project, three PhD. students were employed to investigate both the human side of the issue (human sign language recognition and processing) and the computer side (computer vision and automatic pattern recognition). Ten Holt’s task within the project was to integrate the two sides and improve the automatic recognizer using the knowledge gained from human signer experiments. The project was conducted in co-operation with experts in the field of sign language education: the Dutch Foundation for the Deaf and Hard-of-hearing Child, and the Royal Auris Group.


10. Gineke A. ten Holt, Andrea J. van Doorn, Huib de Ridder, Marcel J. T. Reinders, and
Emile A. Hendriks. Signs in which handshape and hand orientation are either not visible or are only partially visible: What is the consequence for lexical recognition? *Sign Language Studies*, 10(1):5–35, Fall 2009.


