Robust Appearance-based Sign Language Recognition

**Von der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften genehmigte Dissertation**

vorgelegt von

Morteza Zahedi, M.Sc.

aus

Gorgan, Iran

**Berichter: Universitätsprofessor Dr.-Ing. Hermann Ney
Universitätsprofessor Dr.-Ing. habil. Gerhard Rigoll**

Tag der mündlichen Prüfung: 21.09.2007

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

To my family

Abstract

In this work, we introduce a robust appearance-based sign language recognition system which is derived from a large vocabulary speech recognition system. The system employs a large variety of methods known from automatic speech recognition research for the modeling of temporal and language specific issues. The feature extraction part of the system is based on recent developments in image processing which model different aspects of the signs and accounts for visual variabilities in appearance. Different issues of appearance-based sign language recognition such as datasets, appearance-based features, geometric features, training, and recognition parts are investigated and analyzed.

We discuss the state of the art in sign language and gesture recognition. In contrast to the proposed system, most of the existing approaches use special data acquisition tools to collect the data of the signings. The systems which use this kind of data capturing tools are not useful in practical environments. Furthermore, the datasets created within their own group are not publicly available which makes it difficult to compare the results. To overcome these shortcomings and the problems of the existing approaches, our system is built to use video data only and evaluated on publicly available data. First, to overcome the scarceness of publicly available data and to remove the dependency on impractical data capturing devices, we use normal video files publicly available and create appropriate transcriptions of these files. Then, appearance-based features are extracted directly from the videos. To cope with the visual variability of the signs occurring in the image frames, pronunciation clustering, invariant distances, and different reduction methods are investigated.

Furthermore, geometric features capturing the configuration of the signers' hand are investigated improving the accuracy of the recognition system. The geometric features represent the position, the orientation and the configuration of the signers' dominant hand which plays a major role to convey the meaning of the signs.

Finally, it is described how to employ the introduced methods and how to combine the features to construct a robust sign language recognition system.

Zusammenfassung

In dieser Arbeit wird ein robustes, erscheinungsbasiertes Gebärdenspracherkennungssystem aufbauend auf einem Spracherkennungssystem für großes Vokabular vorgestellt. In diesem System werden viele Methoden aus der automatischen Spracherkennung zur Modellierung zeitlicher und sprachspezifischer Eigenheiten eingesetzt. Die Merkmalsextraktion in diesem System basiert auf neuen Entwicklungen der Bildverarbeitung, um unterschiedliche Aspekte der Gebärden und der visuellen Unterschiede in der Erscheinung zu modellieren. Verschiedene Sachverhalte der erscheinungsbasierten Gebärdenspracherkennung, wie z.B. Datensammlungen, erscheinungsbasierte Merkmale, geometrische Merkmale, Training und Erkennung werden untersucht und analysiert.

Außerdem wird der Stand der Forschung in der Gebärdensprach- und Gestenerkennung dargelegt. Im Gegensatz zum hier vorgestellten System bauen die meisten existierenden Ansätze auf spezielle Datenaufnahmetechniken, um die Gestendaten im Computer zu speichern. Systeme, die auf spezielle Datenaufnahmegeräte angewiesen sind, sind jedoch in praktischen Anwendungen oftmals nicht einsetzbar. Die Datensammlungen, die in den Systemen verwendet werden, sind oftmals von den publizierenden Gruppen erstellt worden und

sind nicht öffentlich verfügbar, was es schwierig bzw. unmöglich macht, die Ergebnisse zu vergleichen. Um diese Defizite zu bewältigen, werden in unserem System nur Videodaten verwendentet, und die Evaluation findet ausschließlich auf öffentlich verfügbaren Datensammlungen statt.

Um den Mangel an frei verfügbaren Daten zu reduzieren, und um auf unpraktische Datenaufnahmegeräte verzichten zu können, benutzen wir zunächst Videos aus öffentlich verfügbaren Quellen. Anschließend annotieren wir diese, um sie in unserem System zu trainieren und zu testen.

Um die großen visuellen Variabilitäten der Gebärden in den Videobildern zu modellieren verwenden wir Aussprachevarianten, invariante Distanzfunktionen und unterschiedliche Merkmalsextraktions- und Reduktionsverfahren.

Außerdem werden geometrische Merkmale, die die Handkonfiguration des Gebärdenden repräsentieren, benutzt, um die Genauigkeit des Erkennungssystems zu verbessern. Diese modellieren die Handposition, -orientierung und -konfiguration der dominanten Hand des Gebärdenden, die eine entscheidende Rolle für die Bedeutung einer Gebärde spielen.

Schlussendlich wird beschrieben, wie die vorgestellten Methoden benutzt und zu einem robusten Gebärdenspracherkennungssystem mit einer hohen Erkennungsrate kombiniert werden können.

Acknowledgments

Laudation to the God of majesty and glory! Obedience to him is a cause of approach and gratitude in increase of benefits. Every inhalation of the breath prolongs life and every expiration of it gladdens our nature; wherefore every breath confers two benefits and for every benefit gratitude is due.
Whose hand and tongue is capable
To fulfil the obligations of thanks to him?
– Sa'di (*1184–1283*) in Golestan

I would like to thank the people who have supported and accompanied me during my four-year stay in Germany to prepare this dissertation. I appreciate all suggestions, guiding, contributions, and even a "Hallo" or "Tschüss" from the people who joined and encouraged me to reach this goal.

First of all, I would like to thank Prof. Dr.-Ing. Hermann Ney, head of the Lehrstuhl für Informatik 6 of RWTH Aachen University, who supervised this thesis. His guiding and ideas not only opened a new window into new aspects of statistical pattern recognition for me, but also he added some constraints and restrictions into the method of thinking which have been very beneficial and accelerating my research process. Learning how to manage a research group to perform projects was one of the most important lessons I have made. His insist on introducing the state of the art at the beginning of the papers and presentations taught me how to make a scientific report or paper more understandable. He also supported me to attend at different conferences and workshops which have been very important and helpful to me.

I am very thankful to Prof. Dr.-Ing. Gerhard Rigoll who accepted to be co-referee of this thesis. His publications in the field of gesture recognition were very useful and beneficial.

I am also very thankful for my colleagues of the image processing group at the Lehrstuhl, especially Daniel Keysers and Thomas Deselaers as heads of the group and Philippe Dreuw as a colleague working on the same field of research. They have helped me a lot with their suggestions, ideas and contributions. Also participation at the workshops and conferences with them have been one of the most enjoyable times of my work.

My special thanks go to Jan Bungeroth and Shahram Khadivi, who have helped me to analyze the German environment and to handle life in Germany and who have spent their time for discussions about a variety of aspects and problems such as politics, shopping, travelling, etc.

The ones working at my office during the time, Jan Bungeroth, Daniel Keysers and Daniel Stein have made working time enjoyable. I thank my office-mates very much for a good advice when I was worrying for something and for saying a short sentence or for inviting me to drink a cup of tea or coffee.

I am very grateful to all my colleagues at the Lehrstuhl I6, working in the translation and speech recognition groups. Their talks, comments and discussions during my research especially at PhD seminars were very helpful and beneficial. Although I could not participate in all I6 affairs like "DVD Abend", Day tours, coffee breaks, parties and so on, participation at few of them was very enjoyable.

Many thanks go to my Persian friends and their families in Aachen due to their kindness to spend their time with me and my family to help each other to handle the life, which goes on very slowly. I appreciate their contributions and efforts to make enjoyable plans like trips

and parties at the weekends.

Finally, I am very thankful to my family, Farnaz and Tabassom due to their patience and bearing the difficulties of living abroad. Without their support and encouragement my efforts concluding in this dissertation would not have been possible.

This dissertation was written during my time as a researcher with the Lehrstuhl für Informatik 6 of RWTH Aachen University in Aachen, Germany. This work was partially funded by Iran Scholarship Office at MSRT (Ministry of Science, Research and Technology) and the Chair of Computer Science 6 (I6).

Contents

1 Introduction

In the name of God who owns soul and wisdom.
These are the best attributes of God.
– Ferdowsi (*935–1020*)

The study of interaction between human beings and computers is an interdisciplinary subject which is related to many fields of study and research of computer science. This field of study is named human-computer interaction or man-machine interaction. The researchers who work in this field employ different interfaces between the user and the computer including software or hardware components depending on the tasks performed by the machine. The researchers in the field of human-computer interaction, concerning new methodologies and techniques try to improve the interaction between users and computers by making computers more user-friendly and receptive to the user's needs.

Since deaf people and also people who hear hard have to communicate with hearing people in everyday life, a system which translates sign language into spoken language and vice versa is very helpful. Deaf people use sign language which is a visual form of communication including the combination of hand shapes, orientation and movement of the hands, arms or the body, and facial expressions instead of voice which is used in spoken languages for communication. In this particular case, human-computer interaction methodologies are used to facilitate communication between deaf people and hearing people. Furthermore, [Traxler 00] shows that the reading skills of the majority of the deaf society is poor. This makes automatic aids even more valuable.

Figure 1.1 shows the structure of a translation system. This dialog system can be used in a large variety of environments like shops, governmental offices and also for the communication between a deaf user and information systems like vendor machines or PCs.

This system includes, as shown in Figure 1.1, speech recognition and synthesis parts to communicate to hearing people. The progress of research in these fields is reported in [Gauvain & Lamel$^+$ 05]. In addition, the development of machine translation systems over the last decade is reported in [Ney 05a]. In particular, [Bungeroth & Ney 04] present a statistical framework for sign language translation. However, in this work we are only going to concentrate on the sign language recognition which is the key part for recording and

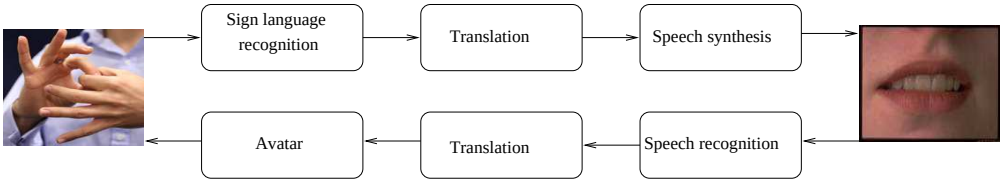


Figure 1.1: Translation machine for communication of deaf and hearing people. It can be used to translate sign language into spoken language and vice versa.

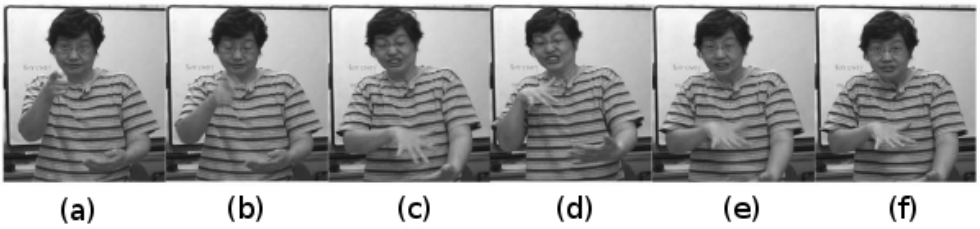


Figure 1.2: An example for American sign language; video frames of a sign language sentence translated into English as: "Are you studying hard?" [Ong & Ranganth 05].

recognition of signings.

1.1 Sign language

Sign languages are usually developed among deaf communities, which include friends and families of deaf people or people with hearing impairment. So, contrary to the popular belief, a universal sign language does not exist. Therefore sign languages like spoken languages are developed differently depending on the community and the region. Consequently, they vary from region to region and also they have their own grammar, e.g. there exists American sign language and German sign language. However, there is no relation between sign language in a particular region to the spoken language.

Sign language includes different components of visual actions of the signer made by using the hands, the face, and the torso, to convey his/her meaning.

Manual components. In sign language the main information is expressed by hand/arm gestures including position, orientation, configuration and movement of the hands. Figure 1.2 shows sequential frames of the sign language sentence "YOU STUDY". These gestures are named manual components of the signing and convey the main meaning of the sign.

Manual components are divided into two categories: glosses and classifiers. Glosses are the signs which are signed for a sign language word like car and house. Classifiers are used in American sign language to express movement, location, and appearance of a person or a subject. Signers use classifiers to express a sign language word by explaining where it is located and how it moves or what it looks like by its appearance.

During signing a sentence continuously, the hand(s) need to move from the ending location of one sign to starting position of the next sign. Also hand configuration changes from ending hand configuration of one sign to the starting configuration of the next. This movement is named movement epenthesis and although it happens frequently between the signs, it does not belong to set of the components of sign language. Figure 1.2.b shows a frame of movement epenthesis which occurs between the sign "YOU" and the sign "STUDY".

Non-manual components. Although most of the researchers working in the field of sign language recognition have focused on the recognition of manual components or lexical meaning of sign gestures, sign language communication further involves non-manual signals conveyed through facial expressions such as eye gaze, movement of eyebrows, head movement and posture, and movements of signers' torso. For example, in Figure 1.2 the signer spreads

the lips wide in frame (c) and (d) in parallel of signing the manual component "STUDY". To build up a system which is able to understand sign language communication an analysis of non-manual signs is necessary as well as lexicons.

Grammar. The grammar of sign languages follows some general rules of spoken languages, however there is no special similarity between the grammar of sign language and the spoken language of a particular region. For example the order of the words in sign language does not obey the rules of the spoken languages. Also some words which occur in a sentence of spoken language may be ignored in the translation of the sign language.

Furthermore, there are some other systematic changes for the sign appearance during continuous signing which affect the sign meaning. These are particularly concerned by the sign language grammar and do not occur in the spoken language. As we have mentioned before signers use various parts of the body like hands, head, face and configuration of the entire body to sign a sentence. Therefore sub-lexical units can be signed simultaneously by different parts of the body. This is the main property of sign language which makes it different to the grammar of spoken languages.

Indexing is a unique property of sign language grammar. The signer defines persons and objects by classifiers or by spelling their name in sign language alphabet and locates them in signing space around himself. He/she uses them again simply by pointing to their location which was described before. This indexing is also used as a person agreement (first person, second person or third person). Furthermore when signing a verb, the direction of the movement indicates the subject and the object of the verb, by corresponding changes in the start and end location, and hand orientation. Different manners of signing a verb can deal with a number of agreements in order to show the number of the persons doing a verb, or the number of repeating a verb, or the number of its objects or subjects.

Emphatic inflections which are used for the purpose of emphasis, derivation of nouns from verbs, numerical incorporation and compound signs are the other aspects of sign language grammar which results in systematic changes.

1.2 Notation systems

Sign language is different from spoken language in its relation to writing. The Spoken languages are primarily sequential: that is, the majority of phonemes producing a word, the words constructing a sentence and the sentences making a text are produced in a sequence one after another. Consequently, spoken languages can be written in sequential form. However, as described before sign language consists of manual and non-manual signs which are signed simultaneously by the body, both hands, face or head of the signer. So, contrary to traditional writing systems, a notation system for sign language should be able to represent non-sequential aspects of the signs.

Although some particular notation systems have been invented and are used in some small communities, we are going to introduce six non-official notation systems which are usually used by researchers:

Glosses. Glosses are written sign language words as a graphic representation of a sign. One sign is replaced with one gloss which is its corresponding word in spoken language. Since

glosses are not complete translations of signs, non-manual signs and classifiers can be added to the main text as an additional transcription. Glosses are usually written in capital letters.

Stokoe system. The Stokoe system[1] is known by the name of William Stokoe, who created this notation system and therefore brought the sign language to the attention of linguists. This system is the first phonological symbol system to represent the component parts of American sign language. The original notation contains 55 symbols in three groups; location (tab), hand shape (dez), and movement (sig). These three groups of symbols occur simultaneously. The location and movement symbols were iconic while hand shapes were represented by units taken from the number system and the manual alphabet of American sign language. Various research teams made changes in this notation as they adapted it to their own situations, with no commonly accepted standard ever developed. Hence, the system should no longer be viewed as one single notation but as a whole family of related systems.

HamNoSys. The Hamburg Sign Language Notation System, HamNoSys [Prillwitz 89], is a "phonetic" transcription system, which is a more general form of Stokoe system. It includes about 200 symbols which are used to transcribe signs in four levels consisting of hand shape, hand configuration, location and movement. The symbols are based on iconic symbols to be easily recognizable. Facial expressions can be represented as well, but their development is not finished yet. HamNoSys is still being improved and extended as the need arises. Figure 1.3 shows some sign language sentences represented with HamNoSys symbols.

SignWritting. SignWriting[2] was developed by Valerie Sutton in 1974. This system is a adapted version from a movement writing system called Sutton "DanceWriting" capable of recording the movements of sign languages.

Using this notation system the first newspaper in history written in the movements of sign languages is published by Sutton. However Sutton did not know the languages she wrote. Therefore SignWriting has no connection with any other writing system because the movements are written down in a generic form, not based on a any prior knowledge of the sign languages being written. It only describes how the body looks as it moves. This means that SignWriting can write any sign language in the world, including detailed facial expressions and mimetic actions. Figure 1.4 presents a sign in Glosses, Stokoe, HamNoSys and SignWriting notation systems.

Liddell and Johnson. The Movement-Hold notation system has been created by Scott Liddell and Robert Johnson [Liddell 93]. It uses English words to describe the physical actions of signing in opposite to the Stokoe model which uses graphical symbols. The Movement-Hold model segments the signing into sequential devisions of movements and holds. Movements refer to the segments in which the aspects of a sign change, while holds are segments where all aspects are consistent at least for 0.1 second. These segments contain all the necessary information needed to describe a sign such as the shapes, locations, orientations of the hand and also non-manual components like facial expressions and torso configuration.

[1] `http://world.std.com/ mam/ASCII-Stokoe.txt`
[2] `http://www.signwriting.org`

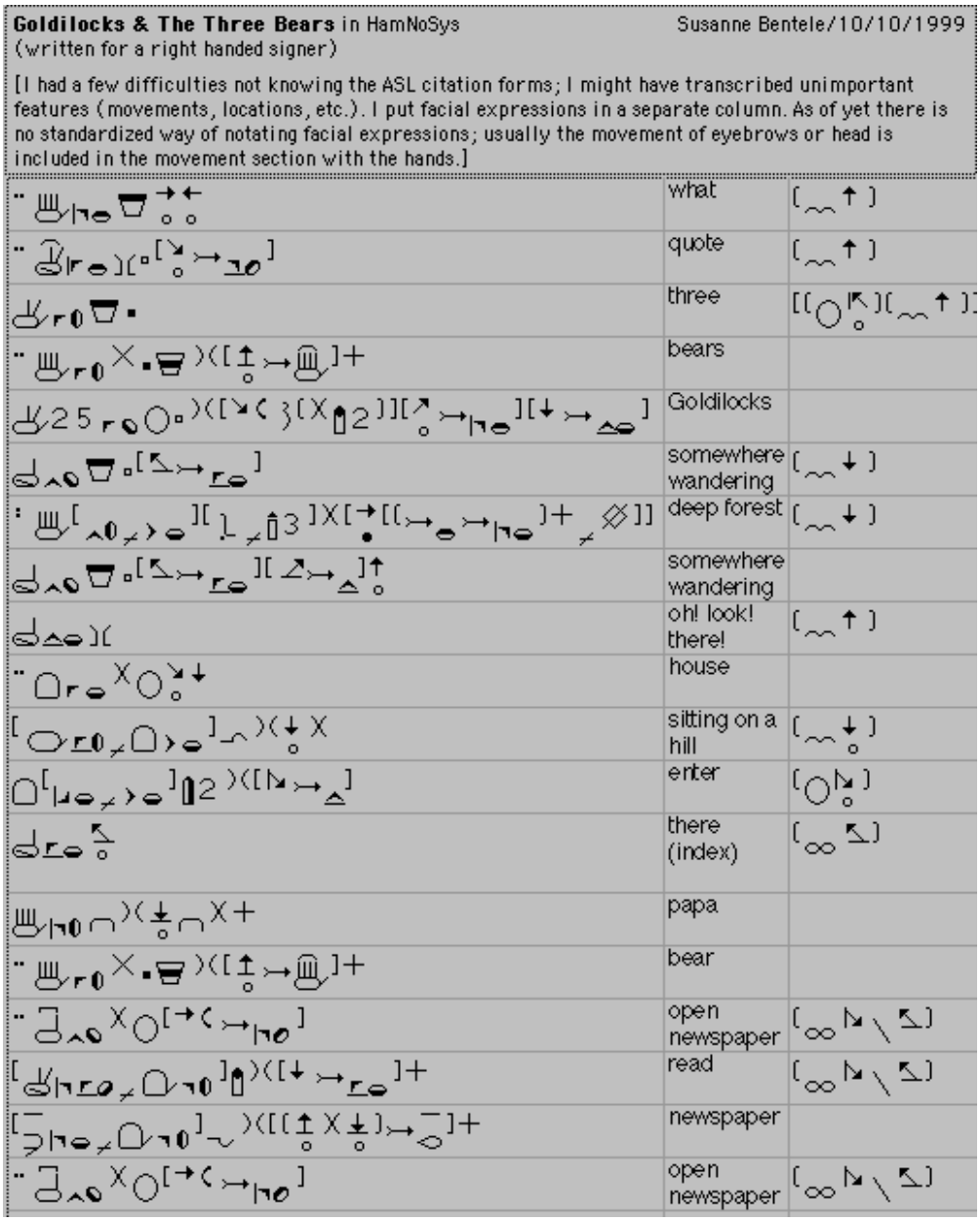


Figure 1.3: Sample of sentences represented in the HamNoSys notation system. The HamNoSys symbols, and other corresponding meaning in English and facial expressions are ordered from the left to the right column. This example is taken from `http://www.signwriting.org/forums/linguistics/ling007.html`.

SiGML. The Signing Gesture Markup Language, SiGML [Elliott & Glauert$^+$ 01] is based on HamNoSys symbols and represents signs in form of Extensible Markup Language (XML) – a simple but flexible format for the changes of structured and semi-structured data. Figure 1.5 shows a sentence in HamNosys which is translated into SiGML code.

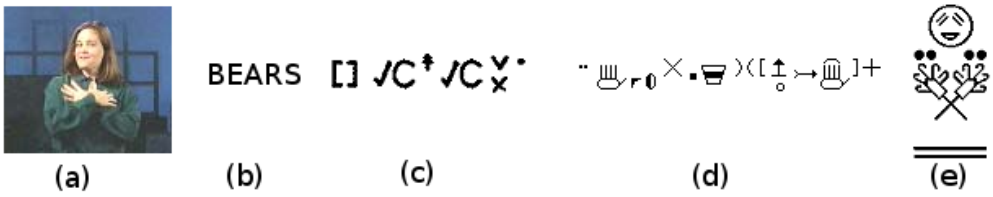
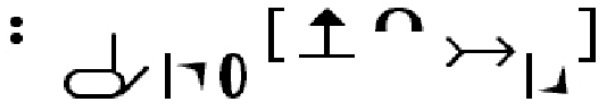


Figure 1.4: Sample of the sign "bears" presented in different systems; (a) shows the signing and (b), (c), (d) and (e) present it in Glosses, Stokoe, HamNoSys and SignWriting notation systems, respectively. This example is taken from `http://www.signwriting.org/forums/linguistics/ling001.html`.



```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE sigml SYSTEM "http://www.visicast.cmp.uea.ac.uk/sigml/sigml.dtd">
<sigml>
<hamgestural_sign gloss="going_to_DGS">
<sign_manual both_hands="true">
    <handconfig handshape="finger2" thumbpos="out"/>
    <handconfig extfidir="uo"  palmor="l"/>
    <par_motion>
        <directedmotion curve="u" direction="o"/>
        <tgt_motion>
            <changeposture/>
            <handconfig extfidir="do"/>
        </tgt_motion>
    </par_motion>
</sign_manual>
</hamgestural_sign>
</sigml>
```

Figure 1.5: Sample of a sign in HamNoSys notation shown in XML format of the SiGML system.

1.3 Sign language recognition

As mentioned before, a sign language recognition system is a key point of a communication system between deaf or hard hearing people and hearing people. It includes a hardware for data acquisition to extract the features of the signings, and a decision making system to recognize the sign language.

Most researchers use special data acquisition tools like data gloves, colored gloves, location sensors, or wearable cameras in order to extract features of the signings. In contrast to the existing approaches, our system is designed to recognize sign language words and sentences by using simple appearance-based features and geometric features of the signers' dominant hand which are extracted directly from the frames captured by standard cameras. Therefore this system can be used rather easily in practical environments.

When using appearance-based features, the sign language recognition system has to overcome several problems like visual variability of utterances of each sign language word, different pronunciations, and large amount of features extracted directly from the image frames.

Concerning these issues, we are going to introduce an automatic sign language recognition (ASLR) system which is derived from a large vocabulary automatic speech recognition (ASR) system named "Sprint" [Sixtus & Molau$^+$ 00, Lööf & Bisani$^+$ 06]. Since speech and sign languages are sequences of features over the time, this system is hopefully able to use the insights gained in speech recognition research.

1.4 Organization of this document

This document is structured as follows: The following chapter contains a list of scientific goals which address the main contribution of this work. Then, we introduce state of the art for the researches concerning sign language translation and recognition. Specially we are going to focus on gesture and sign language recognition which show how the researches are developed in this field and list the results of these research groups. Chapter 4 introduces the databases produced in this work including structure of datasets and annotation files. Appearance-based features and geometric features of signers' dominant hand are explained in Chapter 5. Chapter 6, named automatic sign language recognition, includes a system overview and illustrates different methods which are employed in the system. The Bayes' decision rule and the hidden Markov model as base of the system and also other classification approaches like nearest neighbor classifier and leaving one out method for validation are illustrated in this chapter. Furthermore, principle component analysis and linear discriminant analysis for dimensionality reduction, three different ways employed for pronunciation clustering and also different methods like invariant distance functions for modeling of visual variability are included. The experimental results employing the presented methods and features are listed in Chapter 7. Finally, Chapter 8 concludes the document with an overview of the results achieved by employing the presented methods, including the future perspective of automatic sign language recognition systems.

2 Scientific Goals

The rich are ignorant of the nobility of science.
– Biruni (*973–1048*)

The main goal of this work is to build a system for robust appearance-based sign language recognition. An appearance-based sign language system contains different aspects and thus raises several problems. The contributions of this work are listed below:

- *Automatic sign language recognition system*
  Since speech and sign language are recorded as sequences of the temporal feature vectors, we developed an automatic sign language recognition system, based on a large vocabulary speech recognition system.

- *Database preparation*
  In the field of sign language recognition, most of the researchers construct their own databases which are not publicly available. In this work, well-annotated databases are produced to be used for sign language recognition. They are publicly available to other research groups.

- *Appearance-based features*
  Using appearance-based features which are extracted directly from the image frames recorded by normal stationary cameras results in a system which is easy to use in a practical environment. We avoid employing special acquisition devices to capture the signing features.

- *Modelling of variability*
  To model visual variability of different utterances of sign language words, invariant distances like the tangent distance (TD), the image distortion model (IDM) and the combination of them are employed to improve the accuracy of the system.

- *Geometric features*
  The dominant hand of the signer plays a main role for the meaning of the signing. The position, orientation, and configuration of the dominant hand are important features to be extracted from the video image frames.

- *Frame interpolation*
  The linear frame interpolation is used to help the tracking method employed to locate the dominant hand of the signer.

  As the frame rate of the video input is low comparing to the input of a speech recognition system, the frame interpolation gives the intermediate information of the image frames.

- *Pronunciation clustering*
  To investigate the influence of different pronunciation considerations, different pronunciation partitioning methods are implemented and compared with each other. The experiments show the Gaussian mixture densities help to cope with different pronunciations of the words. Also for this case, the nearest neighbor classifier is useful in the particular case of segmented sign language recognition.

- *Feature selection*
  The linear discriminant analysis (LDA) and principle component analysis (PCA) are employed to select the most important elements of the features. Also different combination methods are investigated to combine the features in a proper way.

3 State of the Art in Sign Language Recognition

Up from Earth's Center through the Seventh Gate
I rose, and on the Throne of Saturn sate,
And many Knots unraveled by the Road;
But not the Master-Knot of Human Fate.
– Avicenna (*980–1037*)

Sign language is studied in two major categories which are automatic translation and recognition. In this chapter first some important researches in the field of translation are going to be listed and then we are going to focus on the state of the art of sign language recognition systems.

Here a list of recent works on sign language translation is shown:

- [Sáfár & Marshall 02] working on the ViSiCAST project, suggest a rule-based approach for the automatic translation from English into British sign language. They discuss the language-processing component of an English to British sign language translation system focusing upon the inherent problems of the knowledge elicitation of sign language grammar within a HPSG[1] framework. They use an intermediate stage of semantic representation of the signs in the translation process.

- [Huenerfauth 04] presents a multi-path architecture for machine translation of an English text into American sign language animation. In contrary to traditional machine translation architectures, he proposes a new semantic representation using virtual reality 3D modeling software to produce the especially complex American sign language phenomena.

- Morrissey and Way [Morrissey & Way 05] investigate corpus-based methods for example-based sign language translation from English to the sign language of the Netherlands. Their work is based on the ECHO corpus [Crasborn & van der Kooij$^+$ 04] containing only a few hundred sentences of a fairy tale. Furthermore, it is stated while given the sparseness of the sign language data collections even highly established evaluation metrics are problematic in their usage.

- [Bauer & Nießen$^+$ 99] and [Bungeroth & Ney 04] have suggested a statistical machine translation system especially for the language pair of German sign language (DGS[2]) and German. Figure 3.1 shows the system configuration which is presented in [Bungeroth & Ney 04]. This system is trained with bilingual corpora and also the preparation of the corpus is presented. In the publication the preliminary results of the system on a small bilingual corpus are reported.

[1] Head-driven Phrase Structure Grammars [Pollard & Sag 87]
[2] Deutsche Gebärdensprache

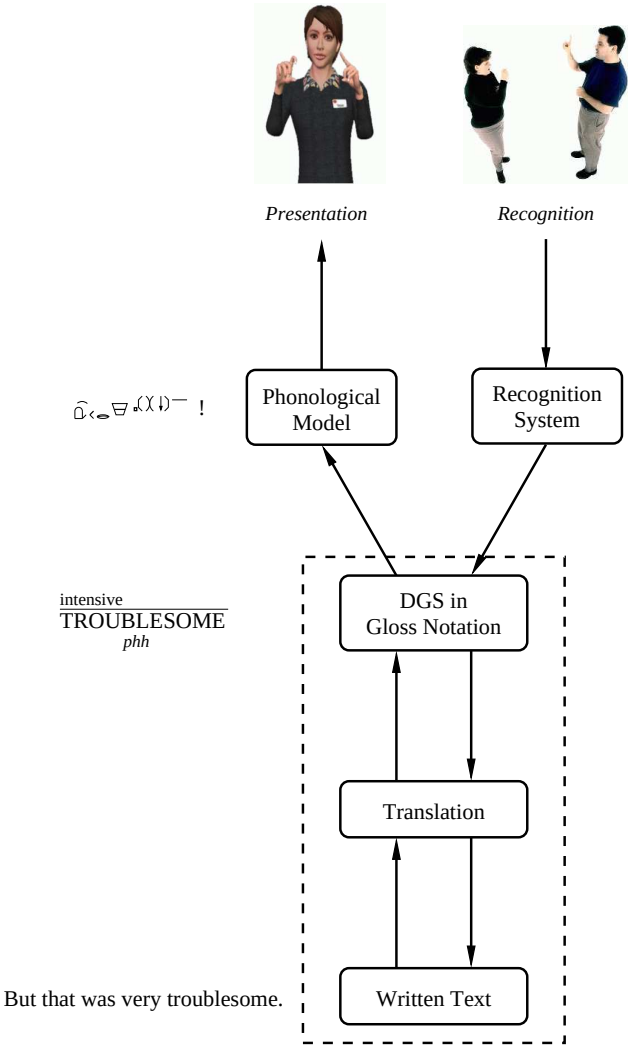


Figure 3.1: Automatic sign language translation system [Bungeroth & Ney 04].

Also, [Stein & Bungeroth+ 06] have presented a novel approach for the automatic translation of written German into German sign language. In this work, a phrase-based statistical machine translation system which is enhanced by pre- and post-processing steps is introduced. The best word error rate of 38.2% and the position-independent word error rate of 27.4% on a bilingual database focusing on the weather report domain of German TV is reported. A version of this database is going to be prepared in our group for sign language recognition [Bungeroth & Stein+ 06, Zahedi & Dreuw+ 06b].

- [Chiu & Wu+ 07] present a system for the language pair Chinese and Taiwanese sign language. They show that their optimizing method surpasses the IBM model 2 [Brown & Pietra+ 93]. They use a corpus of the language pair Chinese and Taiwanese sign language of about 2000 sentences.

Several studies on gesture and sign language recognition have been published. These publications can be separated into four categories according to the signs they try to recognize.

3.1 Static hand posture recognition

In the first category, researchers propose methods to recognize static hand postures or the sign language alphabet. They use images of the hands and extract feature vectors according to the static information of the hand shape. This approach cannot recognize the letters of the sign language alphabet which contains local movement made by the wrist, the knuckles, or the finger joints, as e.g. the sign for 'j' in American sign language (ASL).

[Cui & Swets$^+$ 95] present a self-organizing framework for learning and recognizing hand signs. The proposed framework selects automatically the most discriminating features (MDF) by using a multi-class, multivariate discriminant analysis which works with a small number of training samples. Their system can handle simple and complex backgrounds. A recognition rate of 96% is reported for recognition of 28 different hand signs which have not been used in the training phase.

[F. Quek 96] present an inductive learning system that is able to derive rules of disjunctive normal form formulate. The learning algorithm uses a feature vector which consists 28 features such as the area of the bounding box, the compactness of the hand, and the normalized moments. They have reported 94% recognition rate.

[Triesch & von der Malsburg 01] present a view-based system for a person-independent hand posture recognition in front of a complex background. They have employed an elastic graph matching (EGM) method, which has already been applied to object and face finding and recognition, to represent hand posture features as a graph. They have reported a recognition rate of 92.9% and 85.8% for 12 static hand postures signed by 19 persons against simple and complex backgrounds, respectively. Some image frames of this data set are shown in Figure 3.2.

In [Deselaers & Criminisi$^+$ 07], a new method is presented for localizing and recognizing hand poses and objects in real-time. In [Deselaers & Criminisi$^+$ 07] the task of simultaneously recognizing object classes, hand gestures, and detecting touch events is cast as a single classification problem. To achieve maximum class discrimination for a given image, a random forest algorithm is employed which adaptively selects and combines a minimal set of appearance, shape, and stereo features. This minimal set results in a good generalization as well as efficiency at run time.

3.2 Dynamic hand posture and sign language alphabet recognition

In the second category researchers collect sequential feature vectors of the gestures and, by using the dynamic information, it is possible to recognize letters with local movement as well. It is clear that in this case local movement and changes of the hand shape are important. In these approaches, only the movement due to changing hand postures is regarded, while path movement is ignored (movement made primarily with the shoulder or elbow).

A design for the recognition of 25 dynamic gestures from the international hand alphabet is reported in [Birk & Moeslund$^+$ 97]. As one can see in Figure 3.3, the image frames of the data set includes only the hand part of the signer. Therefore it is rather easy to segment the hand. They have employed the principle component analysis (PCA) to extract the features and an error rate of 99% is reported for off-line recognition of hand postures.

[Mehdi & Khan 02] use a sensor glove to extract the features from the signs of American sign language and employ artificial neural networks (ANN) to recognize 24 letters of the American sign language alphabet.

[Abe & Saito$^+$ 02] propose a virtual 3D interface system which allows a user to work in a three dimensional space and his commands, including hand movement and hand poses, are recorded by two cameras. An error rate of 93.1% for 15 dynamic hand postures is reported. Also, [Malassiottis & Aifanti$^+$ 02] present a system using a 3D sensor which generates a dense range image of the scene. In contrary to the systems which use color sensors, a robust segmentation of the hand under various illumination conditions is guaranteed.

[Hernandez-Rebollar & Lindeman$^+$ 02] present a system for recognizing the 26 hand shapes of the American sign language alphabet and the two signs, "space", and "enter". Therefore, a user can form a sentence using a sequence of letters and spaces and submit it to a system. Special data-gloves are used to extract the features. The reported results show that 21 out of 26 letters have been recognized with 100% accuracy; the worst case, letter "U", achieved 78%.

The Chair of Technical Computer Science at the RWTH Aachen (LTI) has created the LTI-Gesture database, which contains 14 dynamic gestures, 140 training and 140 testing sequences [Akyol & Canzler$^+$ 00]. The videos have been recorded in a car using an infrared camera and the image frames contain the hand of the driver who signed 14 commands for a media player. An error rate of 4.3% has been achieved by the LTI group on this database [A.Pelkmann 99]. The best error rate of 1.4% is obtained on this database by using appearance-based features and modeling of visual variabili-

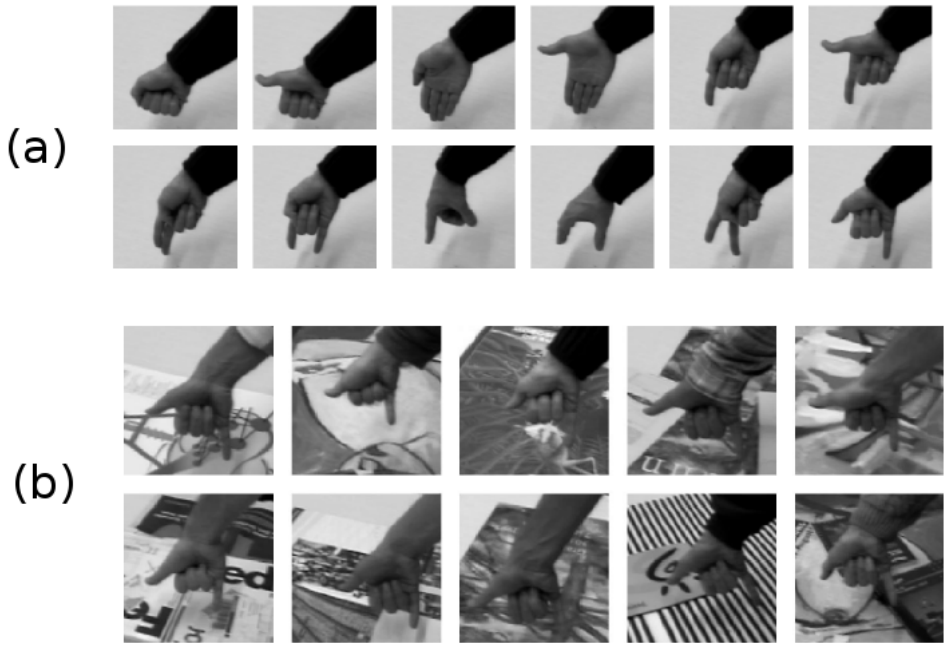


Figure 3.2: Example images from [Triesch & von der Malsburg 01]; (a) the 12 hand postures, (b) one sign performed against different complex backgrounds.

ties [Dreuw 05, Dreuw & Keysers+ 05, Dreuw & Deselaers+ 06b].

3.3 Gesture and isolated sign language recognition

The third category contains the studies that try to recognize gestures or isolated sign language words. In addition to the local movement of the hands, signing includes also the path movement of the hands. Therefore, most systems employ hand segmentation and hand tracking.

A system based on instance-based learning and decision tree learning is presented in [Kadous 96] which is able to recognize 95 isolated Australian sign language words with a recognition rate of 80%. A special data glove named Power-glove is used to extract the features from the signing. Using a similar method, a sign language recognition system is presented in [Matsuo & Igi+ 97] which is able to recognize 38 isolated signs from Japanese sign language. In [Matsuo & Igi+ 97], they use a pair of stereo cameras to collect the information of the 3D movements.

[Rigoll & Kosmala 97] have introduced a feature extraction method which aims at representing the features of the dynamics in the sequence of image frames, avoiding a mixture

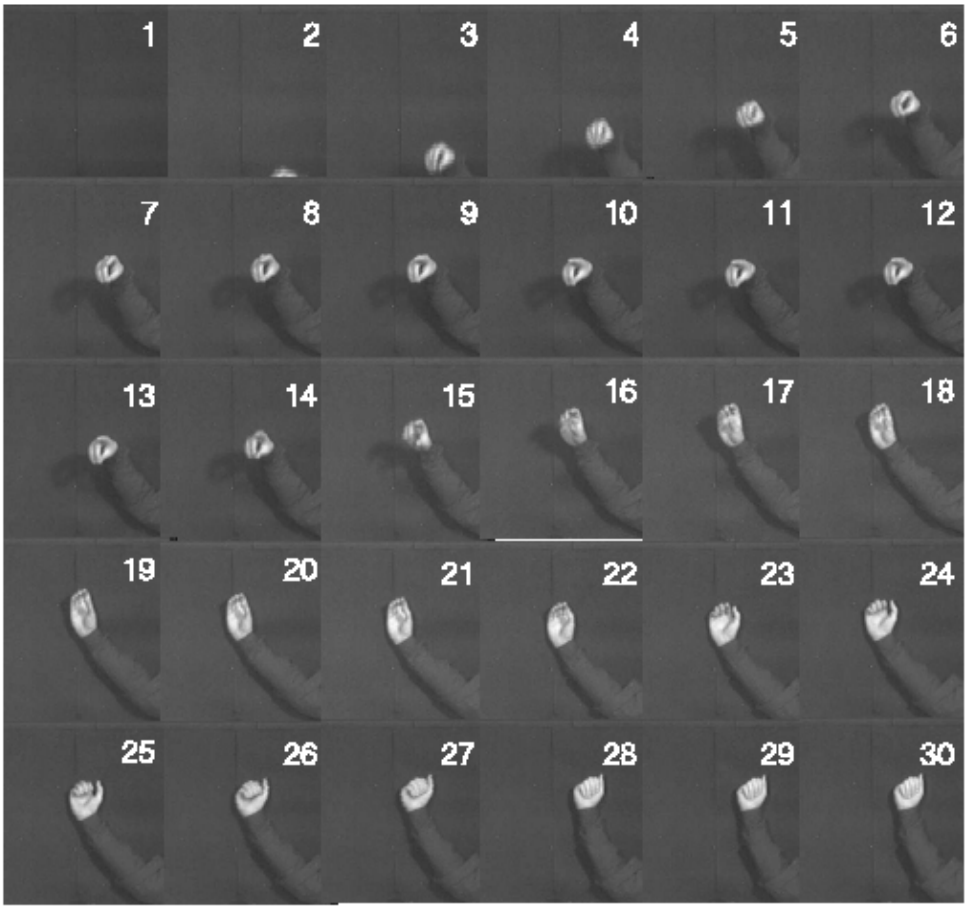


Figure 3.3: An example of sequential image frames in [Birk & Moeslund+ 97] which expresses "P", "C", and "A".

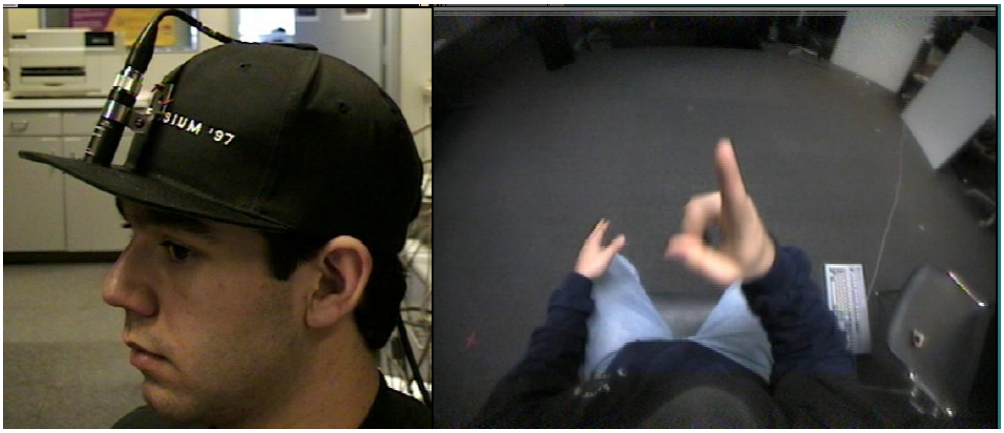


Figure 3.4: A wearable camera which is installed on to a cap worn by the signer is employed as video recording equipment in [Starner & Weaver$^+$ 98].

between spatial and sequential information, and robustness against variations of the gesture and of the person performing the gestures. They have presented an improved extension of this system in [Eickeler & Kosmala$^+$ 98], which is independent of the user position in the image frames, able to reject undefined gestures, and capable of continuous online recognition of gestures. Finally, in [Rigoll & Kosmala$^+$ 98], they present an advanced real-time system for the recognition of 24 complex dynamic gestures like "hand waving", "spin", "pointing" and "head movement". Their system is person and background independent and has a recognition rate of 92.9%.

3.4 Continuous sign language recognition

[Liang & Ouhyoung 98] present a system based on data-gloves which employs a hidden Markov model and uses the time-varying parameter threshold of the hand posture to indicate end-points of the signs in a stream of continuous Taiwanese sign language sentences. The error rate of 19.6% is reported in [Liang & Ouhyoung 98] for a database containing 250 signs.

C. Vogler and D. Metaxas in [Vogler & Metaxas 97, Vogler & Metaxas 99] present a framework for the recognition of isolated and continuous American sign language sentences from three dimensional data. A three dimensional tracking method using motion capturing sensors produces the data for the context-dependent hidden Markov models of the classifier. Also the geometric properties of the tracked hand of the signer are used to constrain the hidden Markov models. Using three dimensional features and context dependent hidden Markov models, yield a recognition rate of 89.91% which is reported for a data set containing 53 signs of American sign language.

T. Starner et al. [Starner & Weaver$^+$ 98] present a real-time hidden Markov model-based system to recognize continuous American sign language sentences. The features are extracted in two ways; a single desk mounted camera and a wearable camera installed on to a cap worn by the signer. Figure 3.4 shows a picture of their video recording tool which is shown in [Starner & Weaver$^+$ 98]. The recognition rate of 92% and 97% are reported for a data set including 40 signs, by using a desktop camera and a wearable camera, respectively.

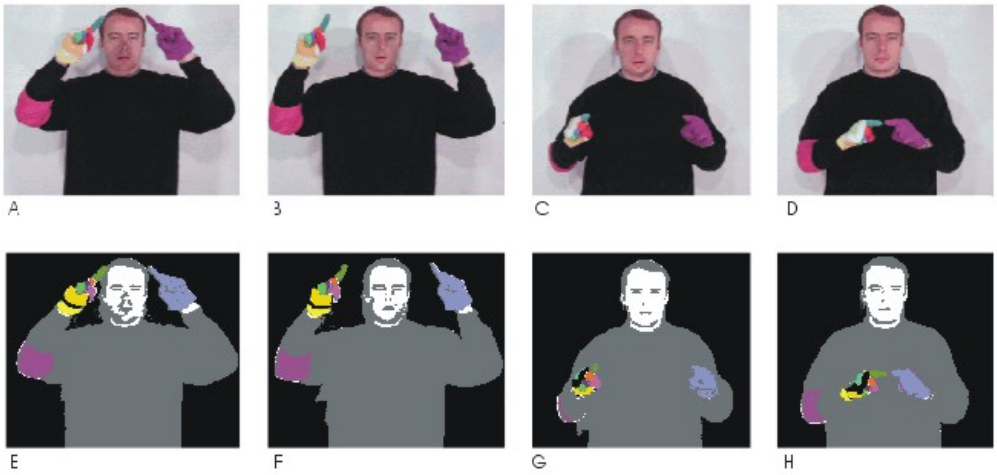


Figure 3.5: Sample frames from colored gloves used in [Bauer & Hienz+ 00b]. The dominant hand glove is painted with seven different colors to indicate the areas of five fingers, palm and back, and the non-dominant hand glove has another sole color.

B. Bauer et al. [Bauer & Hienz+ 00b, Bauer & Hienz 00a] present a video-based recognition system for continuous German sign language sentences. Their approach employs a hidden Markov model for each sign. Geometric features from the signers' hand and fingers are extracted from the frames recorded by a color video camera. The signer wears simple colored cotton gloves shown in Figure 3.5 which make a segmentation of the hand and fingers rather easy. The recognition rate of 94.0% and 91.8% is achieved for different data sets consisting of 52 and 97 signs. By adding a language model, the recognition rate is improved to 95.4% and 93.2%, respectively.

R. Bowden et al. [Bowden & Windridge+ 04] divide the process in to a two-stage classification procedure. In the first stage, a high level description of the hand shape and motion is extracted from the video. This description is based upon sign language linguistic aspects and describes the actions at a conceptual level. Then, the second stage of the classifier models the temporal transitions of individual signs employing a hidden Markov model which is combined with the independent component analysis (ICA). They report a recognition rate of 97.67% for a data set including 43 signs, where the classifier is trained by single instance samples.

Large vocabulary sign language recognition has its own problems and some research groups have focused on these. For example [Fang & Gao+ 04] propose a fuzzy decision tree with heterogeneous classifiers to reduce the recognition time of a large vocabulary sign language recognition without a loss of accuracy. Since the time of the recognition process is an important aspect in the huge search space due to a variety of classes, their method on a large vocabulary database of 5113 signs reduces the recognition time by eleven percent and also improves the recognition rate from 82.9% of a self-organizing feature maps/hidden Markov model (SOFM/HMM) [Fang & Gao+ 04] classifier to 83.7%.

Also, [Vogler & Metaxas 99] deal with the scalability of the vocabulary size of data sets. In contrary to spoken languages, in sign language phonemes may occur simultaneously and the number of possible combinations of phonemes increases when the vocabulary size of the data set increases. They propose parallel hidden Markov models to consider the different

combinations at the training stage. The parallel hidden Markov models are employed to model the parallel processes independently.

Non-manual signals have attracted the interest of some research groups to recognize movement and posture of the signer's torso, head movements, facial expressions like eye gaze, movement of eyebrows and lip movement.

[Erdem & Sclaroff 02] describe a system for detection of head movements. Relevant head gestures in American sign language videos are labeled by a system which analyzes the length and the frequency of the motion signal's peaks and valleys. This system is particularly used to detect two important types of periodic head gestures in sign language communication: "head nods" and "head shakes".

H. Kashima et al [Kashima & Hongo$^+$ 01] propose an iris detection method which is adaptable to different face and eye movements. This system segments the face region by using the color differences from the standard skin color. Then, the eye and mouth regions are extracted by using a hybrid template matching and finally the iris regions are recognized by using the saturation and the brightness from the extracted eye regions. They report a recognition rate of about 98% and 96% for the eye region and iris detection, respectively.

U. Canzler and T. Dziurzyk [Canzler & Dziurzyk 02] have presented a system for analyzing automatically the lip movements using point distribution models and active shape models. C. Vogler and S. Goldenstein [Vogler & Goldenstein 05] present a 3D deformable model tracking system which is able to recognize dynamic facial expressions without training.

To extract facial features it is necessary to track the head of the signer which should handle the problems like occlusion of the signer's hands and the face. Some approaches to track the head are reported in [Cascia & Sclaroff$^+$ 00, Akyol & Zieren 02].

Most researches of all four categories use special data acquisition tools like data gloves, colored gloves, location sensors, or wearable cameras to extract the features. Some researches of the first and second category use simple stationary cameras [Triesch & von der Malsburg 01, Birk & Moeslund$^+$ 97] without any special data acquisition tools. However, their images contain only the hand and skin color segmentation allows them to perform a perfect segmentation. In the third and fourth categories due to the occlusion between the hands and the head of the signer, segmentation which is based on skin color is very difficult. Instead of gloves, some researchers use different methods. For example in [Starner & Weaver$^+$ 98] the camera is placed above and in front of the signer. Then in the images captured by this camera the occlusion between the hands and the head of the signer is decreased. In another case the camera is installed on a hat which the signer wears to avoid the capturing of the face of the signer in the images. These methods or special tools may be difficult to use in practical situations.

Automatic sign language recognition is an area of high practical relevance since sign language often is the only way of communication for deaf people. In contrast to existing approaches, our system is designed to recognize sign language words and sentences using appearance-based features extracted directly from the frames captured by standard cameras. Appearance-based approaches offer some immediate advantages to automatic sign language recognition in contrast to systems which require special data acquisition tools. In particular, they may be used in a "real-world" situation where no special data recording equipment is feasible. Using a laptop with standard cameras placed in fixed positions, for example on a table, this system could be easily used in shops, offices and other public places.

4 Data Sets

What is easiest and most useful in arithmetic? "restoring", referring to the process of moving a subtracted quantity to the other side of an equation; and "comparing", referring to subtracting equal quantities from both sides of an equation.

– Khwarazmi (*790–840*)

All systems for automatic sign language translation and recognition, in particular statistical systems, rely on adequately sized corpora. They need a suitable corpus of sign language data to train with. Unfortunately, most of the currently available corpora are too small or too general to suit the mentioned tasks. Furthermore, most researchers in the field of sign language recognition work on the databases created within their own group. These databases are not publicly available. Therefore, there is no way to compare different approaches of sign language recognition systems. Also, some other databases have been created by linguistic research groups. These databases have not been produced for sign language recognition purpose; i.e. no suitable transcription is available or sign language words occurring in the database have not been repeated frequently to be useful for training. To use this data for the training or for performance evaluation in sign language recognition systems, the necessary transcriptions have to be created which is a costly process and requires a lot of human work.

In this chapter we introduce two sets of databases which have various properties. The RWTH-BOSTON databases including RWTH-BOSTON-10 [Zahedi & Keysers$^+$ 05b], RWTH-BOSTON-50 [Zahedi & Keysers$^+$ 05c, Zahedi & Keysers$^+$ 05a] and RWTH-BOSTON-104 [Zahedi & Dreuw$^+$ 06a], and the RWTH-ECHO databases including RWTH-ECHO-BSL, RWTH-ECHO-SSL and RWTH-ECHO-NGT [Zahedi & Dreuw$^+$ 06b] which are subsets of the databases which are created by linguistics. They are adapted to be used for the training and the evaluation of recognition systems.

For storing and processing sign language, a textual representation of the signs is needed. While there are several notation systems covering different linguistic aspects, we focus on the so-called gloss notation. Glosses are widely used for transcribing sign language video sequences; they are a form of semantic representation for sign language.

In our work, a gloss is a word describing the content of a sign written with capital letters. Additional markings are used for representing the facial expressions and other non-manual markings [Bungeroth & Stein$^+$ 06]. The manual annotation of sign language videos is a difficult task, so notation variations within one corpus are often a common problem. To avoid this, we follow the specifications of the Aachener Glossenumschrift [DESIRE 04] for transcription of RWTH-ECHO databases. In this section, we introduce the databases which are used in this work.

4.1 RWTH-BOSTON databases

The National Center for Sign Language and Gesture Resources of the Boston University published a database of ASL sentences [Neidle & Kegl$^+$ 00]. Although this database has not been produced primarily for image processing research, it consists of 201 annotated video streams of ASL sentences and these video streams can be used for sign language recognition.

In the RWTH-BOSTON databases, there are three signers: one male and two female signers. All of the signers are dressed differently and the brightness of their clothes is different.

The signing is captured simultaneously by four standard stationary cameras where three of them are black and white and one is a color camera. Two black and white cameras, placed towards the signer's face, form a stereo pair (Figure 4.1.a and b) and another camera is installed on the side of the signer (Figure 4.1.c). The color camera is placed between the stereo camera pair and is zoomed to capture only the face of the signer (Figure 4.1.d). The movies published on the Internet are at 30 frames per second and the size of the frames is 312×242 pixels[1] (Figure 4.3). We are going to use the published video streams at the same frame rate but we are going to use only the upper center part of the size 195×165 pixels since the lower part of the frames show some information about the frame such as the date and the time of recording the video. Also, the left and right border of the frames are unused.

4.1.1 RWTH-BOSTON-10

In the first step of this work, we have created a small database including segmented sign language words to investigate the results of using appearance-based features in a sign language recognition system. To create the RWTH-BOSTON-10 database for American sign language word recognition, we have extracted 110 utterances of 10 words from the original database as listed in Table 4.1. These utterances have been segmented manually. Table 4.1 lists the sign language words, number of corresponding utterances and also minimum, maximum and average number of image frames occurring in the utterances for each sign language word. As one can see, the number of utterances and also the length of the utterances are not distributed uniformly. The frequency of utterances for different lengths is shown in Figure 4.2.

We use the frames captured by two of the four cameras, consisting of one camera of the stereo camera pair in front of the signer and the other lateral. Using both of the stereo cameras and the color camera may be useful in stereo and facial expression recognition, respectively. Both of the used cameras are in fixed positions and capture the videos in a controlled environment simultaneously. In Figure 4.3 the signers and the views of the cameras are shown.

This database is too small to be separated into training and evaluation sections, and it is therefore used by the leaving one out method in the experiments. Although this database is a very small snapshot from the original database, it is a good database to do preliminary experiments for testing features and methods.

4.1.2 RWTH-BOSTON-50

The RWTH-BOSTON-50 database contains 483 utterances of 50 words from the original database which have been segmented within our group manually. This database includes approximately all words which occur at least two times in the original database. The sign

[1] `http://www.bu.edu/asllrp/ncslgr.html`

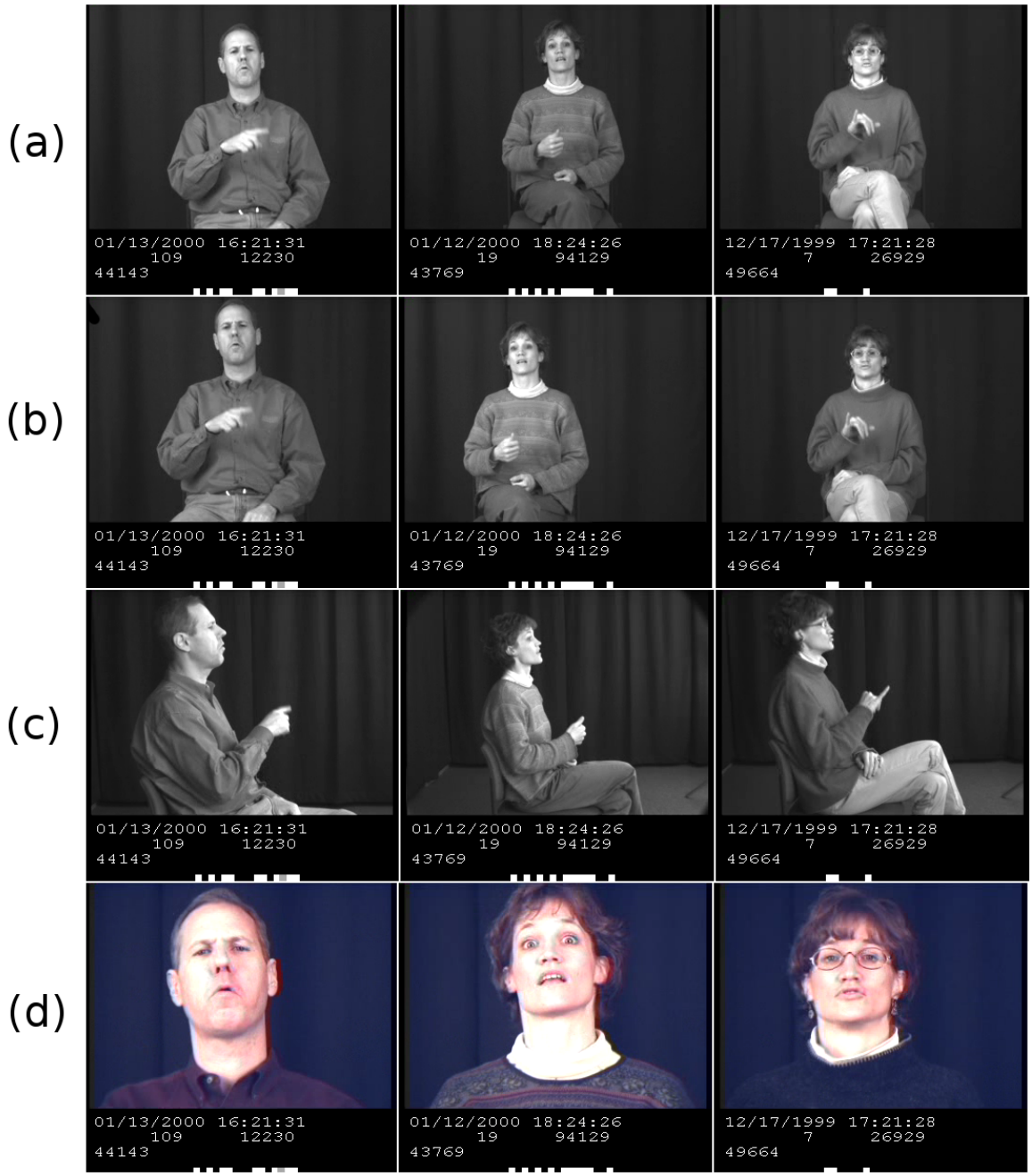


Figure 4.1: Sample frames from the original Boston database: three signers viewed from the (a and b) pair of stereo cameras, (c) side camera and (d) a camera zoomed to capture the signer's face perspectives.

Table 4.1: List of the words, number of utterances and the minimum, maximum and average number of the image frames for each sign language word in the RWTH-BOSTON-10 database.

Word	Number of utterances	Number of image frames		
		Minimum	Maximum	Average
CAN	17	3	14	6.94
BUY	15	4	7	5.33
CAR	15	5	17	7.8
BOOK	13	5	9	7.15
HOUSE	11	15	21	17.45
WHAT	10	6	19	11.2
POSS (Possession)	9	5	9	6.88
WOMAN	8	6	10	7.75
IX “far” (Pointing far)	7	8	17	13
BREAK-DOWN	5	14	18	16
Overall	110	3	21	9.15

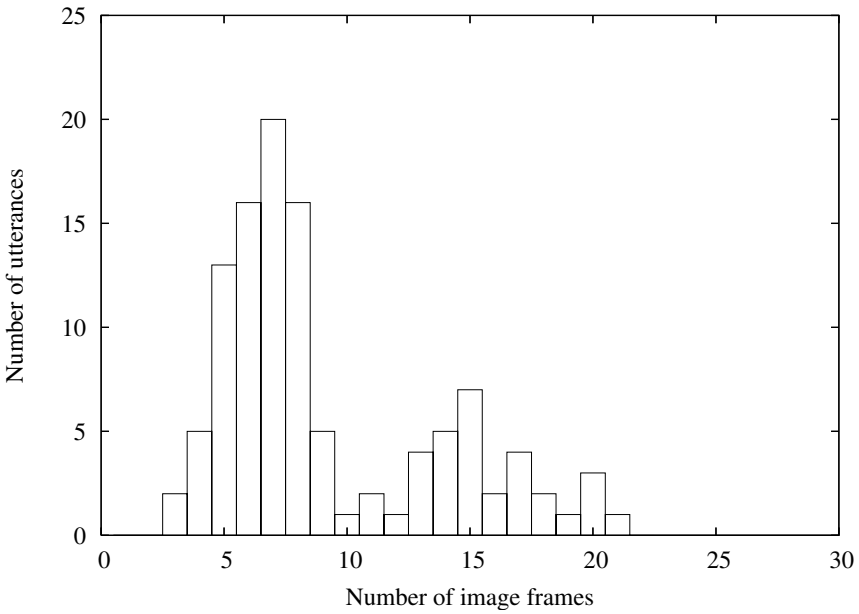


Figure 4.2: Frequency of utterances with different lengths in the RWTH-BOSTON-10.

Figure 4.3: The sample frames used in the RWTH-BOSTON databases.

Table 4.2: Length of utterances in the RWTH-BOSTON-50 database.

	Minimum	Maximum	Average
Number of image frames	2	29	9.14

language words along with the number of utterances of each word are listed here:

IX_i (37), BUY (31), WHO (25), GIVE (24), WHAT (24), BOOK (23), FUTURE (21), CAN (19), CAR (19), GO (19), VISIT (18), LOVE (16), ARRIVE (15), HOUSE (12), IX_i"far" (12), POSS (12), SOMETHING/ONE (12), YESTERDAY (12), SHOULD (10), IX-1p (8), WOMAN (8), BOX (7), FINISH (7), NEW (7), NOT (7), HAVE (6), LIKE (6), BLAME (6), BREAK-DOWN (5), PREFER (5), READ (4),COAT (3), CORN (3), LEAVE (3), MAN (3), PEOPLE (3),THINK (3), VEGETABLE (3) VIDEOTAPE (3), BROTHER (2), CANDY (2), FRIEND (2), GROUP (2), HOMEWORK (2), KNOW (2),LEG (2), MOVIE (2), STUDENT (2), TOY (2), WRITE (2).

As one can see, although some words occur only 2 times, there are some words occurring even more than 20 times. However, there is not enough data to separate the database into training and evaluation sets and the leaving one out method is employed for this database as well. More information about the RWTH-BOSTON-50 database like the minimum, maximum and average number of the image frames for 483 utterances of the database are presented in Table 4.2. Also the histogram for the number of utterances with different lengths in the RWTH-BOSTON-50 is shown in Figure 4.4.

4.1.3 RWTH-BOSTON-104

To use the Boston database for ASL sentence recognition, we have separated the recordings into a training and evaluation set. To optimize the parameters of the system, the training set has been further split into separate training and development parts. To optimize the parameters in the training process, the system is trained by using the training set and evaluated

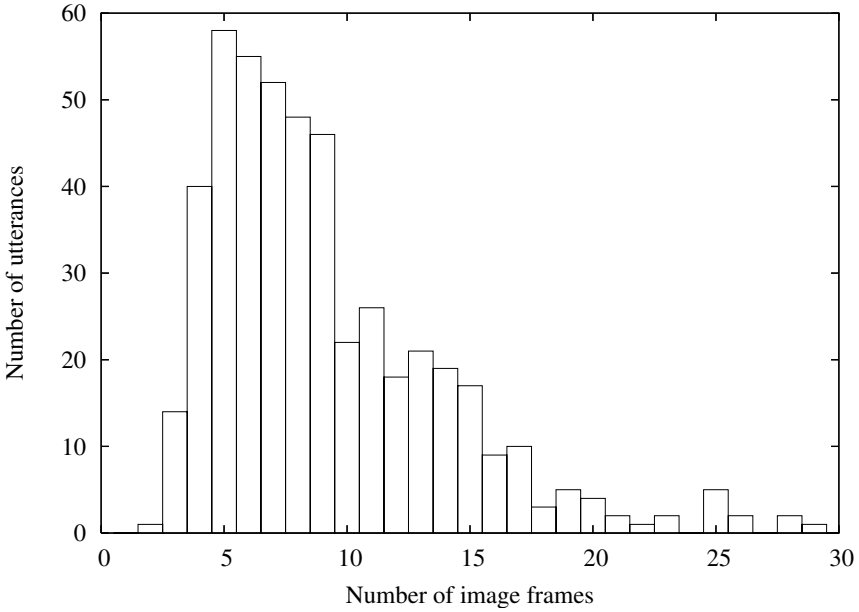


Figure 4.4: Frequency of utterances with different lengths in the RWTH-BOSTON-50.

using the development set. When the parameter tuning has been finished, the training data and development data had been used to train one model using the optimized parameters. This model has been then evaluated on the so-far unseen test set. Corpus statistics for this database are shown in Table 4.3 which include number of sentences, running words, unique words, singletons, and out-of-vocabulary (OOV) words in the each part. Singletons are the words occurring only once in the set. The out-of-vocabulary words are the words which occur only in the evaluation set, i.e. there is no visual model for them in training set and they cannot be therefore recognized correctly in the evaluation process.

Table 4.4 gives example sentences which are shown in the gloss notation. Also English translation of sentences are shown for each sentence.

Table 4.3: Corpus statistics for the RWTH-BOSTON-104 database.

	Training set		Evaluation set
	Training	Development	
Number of sentences	131	30	40
Number of running words	568	142	178
Vocabulary size	102	64	65
Number of singletons	37	38	9
Number of OOV words	–	0	1

Table 4.4: Example sentences of the RWTH-BOSTON-104 which are presented in the gloss notation; the English translation is also provided for the reader.

ASL	JOHN LOVE MARY
English	John loves Mary.
ASL	MARY VEGETABLE KNOW IX LIKE CORN
English	Mary knows that, as for vegetables, he likes corn.
ASL	JOHN FISH WONT EAT BUT CAN EAT CHICKEN
English	John will not eat fish but eats chicken.

4.2 ECHO databases

The ECHO database[2] consists of three corpora in British sign language (BSL), Swedish sign language (SSL) and the sign language of the Netherlands (NGT). All three corpora include the videos from sign narrations of the same five fable stories, a small lexicon and interviews with the signers. In addition, there is sign language poetry provided in BSL and NGT. Figures 4.5, 4.6 and 4.7 show sample image frames. The corpora have been annotated linguistically and include sign language and spoken language transcription in English. In addition, SSL and NGT sections include Swedish and Dutch transcription, respectively.

These videos have been transcribed by using the ELAN software and the transcription includes word and sentence boundaries necessary for the sign language recognition. The annotations are stored as EAF files, i.e. an XML format used by the ELAN software. This allows several annotations on different tiers on the same time line to represent different aspects of the signing like the movement of the right hand, the left hand, or the facial expressions.

To use the ECHO databases in the field of sign language recognition, we have chosen some parts of the five fable stories of the original database and have created a database for each of the subcorpora. The missed part includes some very long sign language sentences where the annotations do not fit to the recorded video and also some video parts recorded to introduce the lexicon of the dataset. We named the created databases RWTH-ECHO-BSL, RWTH-ECHO-SSL, RWTH-ECHO-NGT, which have been all extracted from the original ECHO databases [Zahedi & Dreuw$^+$ 06b].

Although the data has been recorded in a completely controlled environment with constant background, it is currently very hard to use these three databases for sign language recognition: The number of singletons and the vocabulary size are too high in relation to the total number of utterances. To reduce the data sparseness, we have decided to split the corpus into training and testing data only, i.e. for these corpora no development sets have been specified. Furthermore, the test set has been selected to have no out-of-vocabulary words, i.e. each word in the test set is at least once in the respective training set. The training corpora consists of the sentences and their segmented words (word boundaries), but evaluation contains only sentences.

[2] `http://www.let.ru.nl/sign-lang/echo`

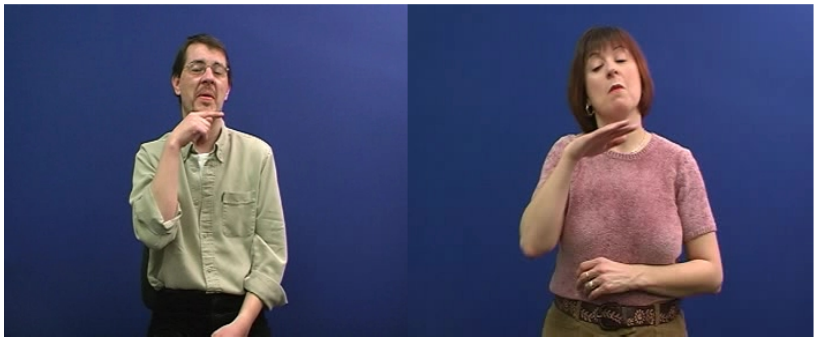


Figure 4.5: Sample frames from the RWTH-ECHO-BSL databases.

Table 4.5: Corpus statistics for the RWTH-ECHO-BSL database.

	Training set	Evaluation set
Number of sentences	205	57
Number of running words	2620	241
Vocabulary size	532	97
Number of singletons	340	56

Table 4.6: Language model statistics for the RWTH-ECHO-BSL database.

LM type	Log-likelihood	Perplexity
Zerogram	1864.72	534
Unigram	1318.97	84.85
Bigram	1263.44	70.38
Trigram	1259.62	69.48

Table 4.7: Information of the recordings from different signers (RWTH-ECHO-BSL).

Signers	Number of segments	Duration (seconds)
Male signer	489	565.015
Female signer	754	907.163
Sum	1243	1472.18

4.2.1 RWTH-ECHO-BSL

The RWTH-ECHO-BSL database is signed by two signers (one male and one female) and the number of recordings and the duration of signing are shown in Table 4.7. Figure 4.5 shows the perspective of the signers. Statistics of the corpus is shown in Table 4.5.

Also, the perplexity and log-likelihood of zerogram, unigram, bigram and trigram language models are shown in Table 4.6. The perplexity and log-likelihood of the language model are defined in Section 6. Although the perplexity of the zerogram language model is expected to be equal to the vocabulary size, however in this context the words defined for “SILENCE” and “UNKNOWN”s in the “Sprint” framework causes the perplexity to be equal to the vocabulary size plus two. Also unigram and trigram language models show how strong structures exist in the text of the database. A small perplexity corresponds to strong language model restrictions.

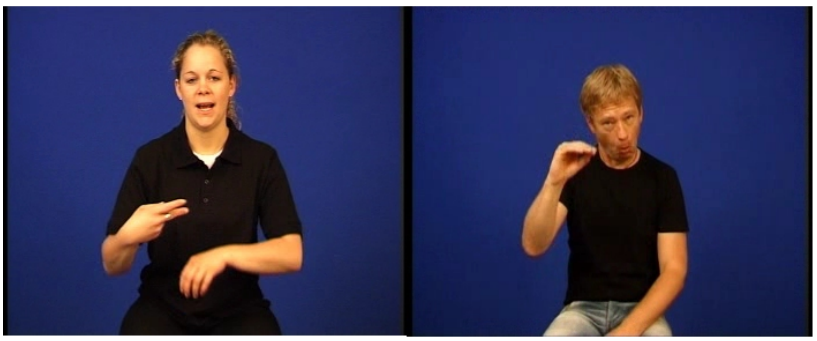


Figure 4.6: Sample frames from the RWTH-ECHO-SSL databases.

Table 4.8: Corpus statistics for the RWTH-ECHO-SSL database.

	Training set	Evaluation set
Number of sentences	136	23
Number of running words	2988	129
Vocabulary size	519	70
Number of singletons	279	44

Table 4.9: Language model statistics for the RWTH-ECHO-SSL database.

LM type	Log-likelihood	Perplexity
Zerogram	919.59	521
Unigram	713.52	128.24
Bigram	644.84	80.37
Trigram	646.1	81.06

Table 4.10: Information of the recordings from different signers (RWTH-ECHO-SSL).

Signers	Number of segments	Duration (seconds)
Male signer	631	799.36
Female signer	468	580.37
Sum	1099	1379.73

4.2.2 RWTH-ECHO-SSL

The RWTH-ECHO-SSL database is signed by one male and one female signer. Table 4.10 shows the number of recordings and the signing duration of each signer. The signers of the corpus are shown in Figure 4.6. Statistics of the corpus is shown in Table 4.8.

Table 4.9 shows the perplexity and the log-likelihood of the different language models. As one can see, comparing to the RWTH-ECHO-BSL database, the RWTH-ECHO-SSL database includes less running words in the test set and also unique words. The RWTH-ECHO-BSL has a stronger language model restrictions, i.e. it has a smaller bigram and trigram language model perplexity.

Figure 4.7: Sample frames from the RWTH-ECHO-NGT databases.

Table 4.11: Corpus statistics for the RWTH-ECHO-NGT database.

	Training set	Evaluation set
Number of sentences	188	53
Number of running words	2450	197
Vocabulary size	468	77
Number of singletons	268	40

Table 4.12: Language model statistics for the RWTH-ECHO-NGT database.

LM type	Log-likelihood	Perplexity
Zerogram	1513.57	470
Unigram	1079.89	80.62
Bigram	1002.78	58.92
Trigram	1005.45	59.57

Table 4.13: Information of the recordings from different signers (RWTH-ECHO-NGT).

Signers	Number of segments	Duration (seconds)
Male signer 1	428	482.27
Male signer 2	304	377.69
Female signer	552	808.22
Sum	1284	1668.18

4.2.3 RWTH-ECHO-NGT

The RWTH-ECHO-NGT database is signed by three signers (two males and one female). The duration and the number of segments signed by these signers is shown in Table 4.13. Sample frames of the database from the signers are shown in Figure 4.7. Table 4.11 and 4.12 show the statistics of the corpus and the perplexity of the database for the different language models.

5 Features

For beyond these colors and these perfumes, these are other colors in the heart and the soul.

– Rumi *(1207–1273)*

It is very common to extract special features like the exact position of the hands or the head of the signer in the signing space by using 3D models, stereo cameras, data-gloves, sensors, or colored gloves. These advanced acquisition tools are necessary to extract complicated features like fingertips or angels between the fingers, but it makes these systems very difficult to use in practical environments like offices, shops and in other everyday life situations.

[Gavrila 99] presents a survey on the visual analysis of human movement. He reviews the development of the research in this area and discusses real-time capture, transfer and processing of images on widely available low-cost hardware platforms, virtual reality, surveillance systems and advanced user interfaces. It is a survey on 2D approaches with and without explicit shape models and also 3D approaches.

In contrary to the systems which focus only on the recognition of the lexical by analyzing the hand movements, [Ong & Ranganth 05] deal with the different aspects of signing which have received little attention by the researchers. In this survey, building a signer independent recognition system and addressing the more difficult aspects of signing, such as grammatical inflections and mimic signs, which are two aspects of the gesture and sign language recognition, are discussed. The systems which are introduced in this survey paper use different data acquisition tools to record the different aspects of signings.

Most of the features used in the existing systems for sign language recognition focus only on one aspect of the signing like hand movements or facial expressions. We are going to introduce a sign language recognition system using appearance-based features which include whole information of the image frames and also the geometric features of the signers' dominant hand which plays the main role in the signings. To extract the features no special data acquisition tool is employed. Image processing methods are performed on the original image which is captured by normal stationary cameras. Using a laptop with standard cameras placed in fixed positions, for example on a table, this system could be used easily in shops, offices and other public places.

5.1 Appearance-based image features

Appearance-based features including the original image and its transformations like downscaling, thresholding, filtering, etc. are used successfully for optical character recognition (OCR) [Keysers & Deselaers$^+$ 04, Keysers & Gollan$^+$ 04, Keysers & Deselaers$^+$ 07], medical image processing [Keysers & Deselaers$^+$ 07, Deselaers & Müller$^+$ 07] and object recognition [Deselaers & Keysers$^+$ 04, Deselaers & Keysers$^+$ 05a, Deselaers & Keysers$^+$ 05b, Deselaers & Hegerath$^+$ 06, Hegerath & Deselaers$^+$ 06]. This encourages us to use this kind

Figure 5.1: Example of the features: original image (left), intensity thresholded image (center), and down-scaled image (right).

of features for gesture recognition [Dreuw & Deselaers+ 06b] and sign language recognition [Zahedi & Keysers+ 05b, Zahedi & Keysers+ 05a, Zahedi & Dreuw+ 06a] as well. The appearance-based features including the sequence of whole image frames contain all informations like hand and head movements and facial expressions conveying the different simultaneous aspects of signing. To extract the appearance-based features we do not rely on complex preprocessing of the video signal. Furthermore, the system using only these features works without any segmentation or tracking of the hands. Because we do not rely on an intermediate segmentation step, the recognition can be expected to be more robust in cases where tracking and segmentation are difficult.

The definition of the features is based on basic methods of image processing. These features are directly extracted from the image frames. We denote by $X_t(i,j)$ the pixel intensity at position (i,j) in the frame t.

5.1.1 Original image

We can transfer the image matrix of the size $I \times J$ to a vector x_t and use it as a feature vector. In databases like the RWTH-BOSTON databases, where additional appropriate cameras with different views are available we can simply concatenate the image frames of the different cameras to collect more information from the signer at a certain time.

5.1.2 Down-scaled original image

Although the feature vector containing the whole information of image frames includes all information, the size of the feature vector is too big and the sign language process therefore needs a huge amount of memory and takes a long time. Furthermore a large feature vector needs large databases with more training data to train several parameters which is a problematic issue of sign language recognition. The problem with a high dimensional feature vector is going to be described in detail in Section 6.5.

A Gaussian function using intensity of neighboring pixels of a mapping pixel is used to scale the original images down. This Gaussian filter smoothes the image and weights the intensity information of the neighboring pixels in contrast to the down-scaling methods which are based on a linear interpolation.

5.1.3 Intensity or skin color thresholding

If image frames include body parts like head and hands of the signer, a skin color thresholding ignores the background and useless information. Also in gray valued images, skin parts are usually lighter than others like for example the image frames of the RWTH-BOSTON databases. Intensity thresholding removes almost everything except the hands and the head of the signer. As the thresholding is not a perfect segmentation, we cannot rely on it confidentially for tracking the hands. The output of this thresholding consists of the two hands, face and some parts of the signer's clothes. Intensity thresholding is formulated by

$$x_t(i,j) = \begin{cases} X_t(i,j) & : \quad X_t(i,j) > \Theta \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.1}$$

where x_t is the feature vector at the time t with the brightness threshold Θ. In Figure 5.1 an original image frame, the intensity thresholded image frame and its down-scaled frame created by a Gaussian function are shown.

To extract skin color parts of colored image frames, a skin color based model is employed which is based on the Compaq Cambridge Research Lab image-database presented in [Jones & Rehg 98] and [Jones & Rehg 02]. They use a dataset of nearly one billion labelled pixels to generate color probability histograms. This dataset includes 3077 pictures containing masked skin regions and 6286 pictures not containing skin.

The Bayes' formula is used to calculate the probability s of a specified color c being the skin color with the following formula

$$p(s|c) = \frac{p(c|s) \cdot p(s)}{p(c|s) \cdot p(s) + p(c|\bar{s}) \cdot p(\bar{s})} \tag{5.2}$$

where $p(s)$ and $p(\bar{s}) = 1 - p(s)$ are the prior skin and the non-skin probability calculated over all labeled pixels of the dataset and c is a representation of the RGB color of the image at a specific position (i,j). The skin color model of the Compaq Cambridge Research Lab is used to calculate the probabilities $p(c|s)$ and $p(c|\bar{s})$.

The images thresholded with the skin color probability are calculated by:

$$x(i,j) = \begin{cases} X(i,j) & : \quad S(i,j) > T_p \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.3}$$

where S is the skin probability image which is created according to its skin probability map. The image x includes the pixels with their corresponding skin color value thresholded with value of T_p.

Smoothing the skin color probability maps by applying a Gaussian filter improves the skin color segmentation by keeping the continuous skin parts and by removing the gaps and noises. Also instead of a sharp thresholding, sigmoid functions can be used for segmenting the skin color regions.

$$x(i,j) = \begin{cases} \frac{1}{1+exp(-\alpha \cdot (S(i,j)-T_p))} & : \quad S(i,j) > T_p \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.4}$$

Figure 5.2 shows some sigmoid functions which are used for the skin color segmentation. Figure 5.3 shows the difference between the normal skin color probability map and a Gaussian filtered skin color probability map with their thresholding results. The original image

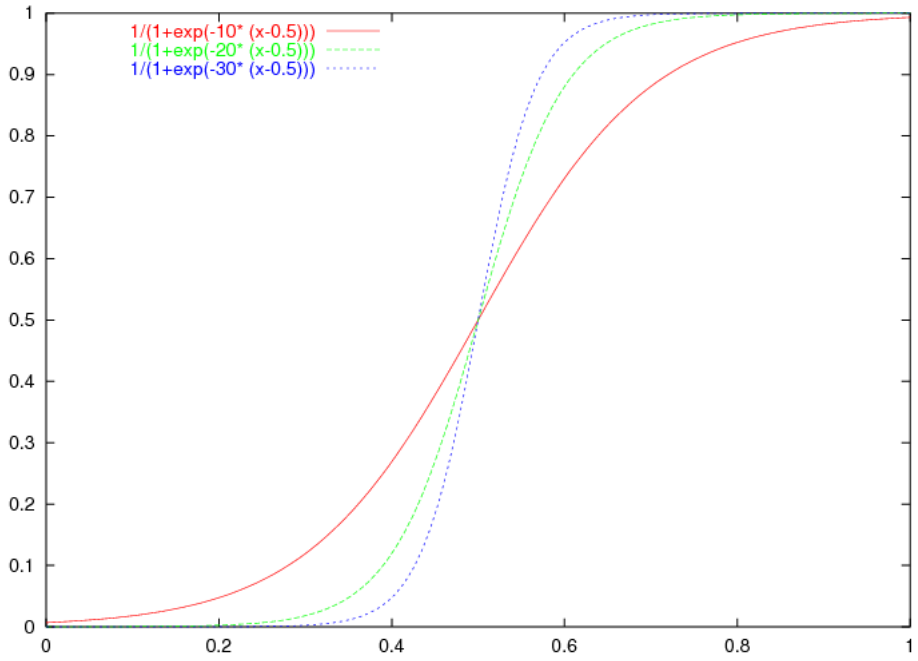


Figure 5.2: Example of three different sigmoid functions with $\alpha = 10, 20, 30$ and $T_p = 0.5$ which improve the segmentation of the original images with skin color probability maps.

thresholded by a skin color probability map which is smoothed by using a sigmoid function (Figure 5.3.c) includes less artifacts and gaps comparing to the fixed thresholding which uses a non-smoothed skin color probability map (Figure 5.3.b).

Other alternative algorithms have been suggested to segment skin color parts in e.g. [Raja & McKenna$^+$ 98], [Sigal & Sclaroff$^+$ 00] or [Zhu & Yang$^+$ 00], too. However, we have

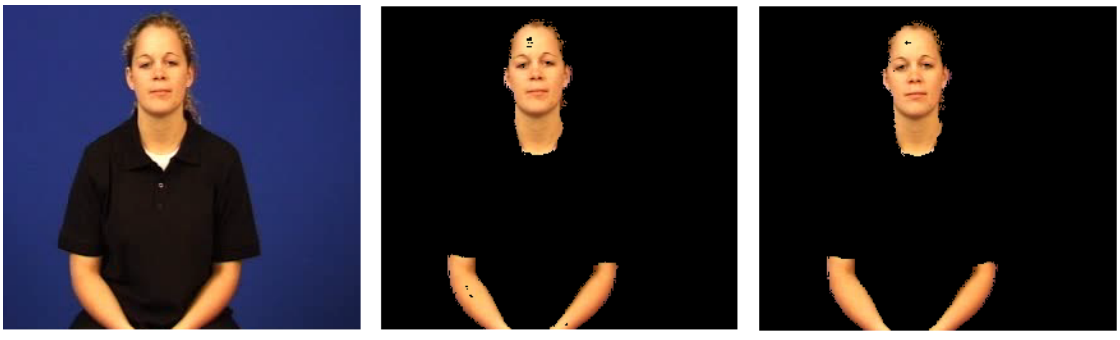


Figure 5.3: Skin color thresholding: (a) original image frame, (b) the image thresholded with the skin color probability map, (c) the image thresholded with the skin color probability map smoothed by a sigmoid function.

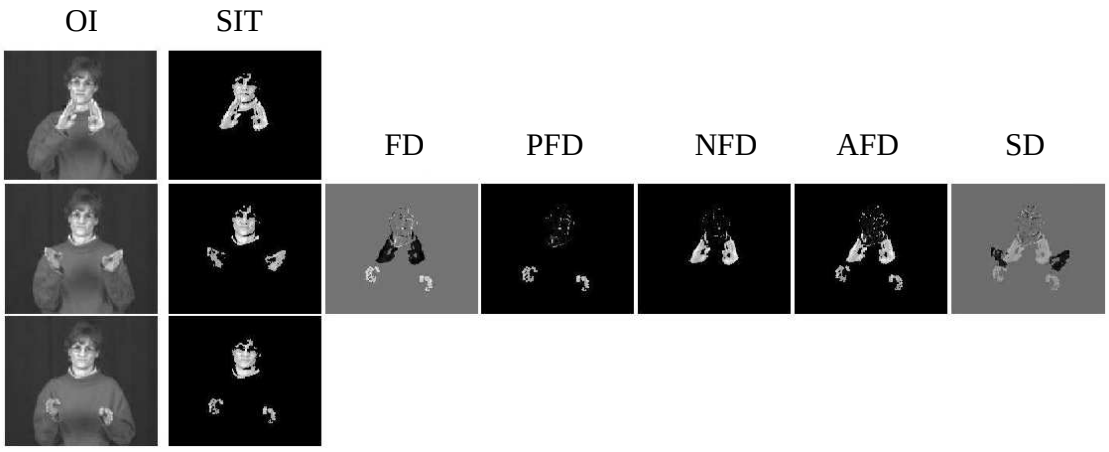


Figure 5.4: Examples for the original image frames, skin intensity thresholding and temporal derivative features.

employed the introduced method which is publicly available with a complete dataset.

5.1.4 Temporal derivative features

The temporal derivative features which represent the changes of the image frames through the time are used to emphasis the motion of the signs and gestures. Image frames resulting from comparing the current image frame to the predecessors and successors shows the motion information of the image parts like the change of position, the velocity or the acceleration of the changes. In this section, different temporal derivative features are going to be introduced.

First temporal derivative (FD). This feature is related to the variety of the motion change and measures the change rate between the successor frame and the predecessor frame.

$$x_t(i,j) = X_{t+1}(i,j) - X_{t-1}(i,j) \tag{5.5}$$

Positive first temporal derivative (PFD). This feature consists of positive members of the FD feature vector. In the databases like the RWTH-BOSTON databases the PFD feature vector posses information of the image pixels that do not belong to the skin intensity values of the predecessor frame. But in the successor frame the pixel values are in the skin intensity range (e.g. a moving hand or head).

$$x_t(i,j) = \begin{cases} X_{t+1}(i,j) - X_{t-1}(i,j) & : \quad X_{t+1}(i,j) - X_{t-1}(i,j) > 0 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.6}$$

Negative first temporal derivative (NFD). In contrast to the PFD feature vector, the NFD feature vector at the time t indicates that the intensity of the pixel is decreasing. This feature contains information of the image pixels belonging to the skin intensity range of the predecessor frame, but in the successor frame the hands or the face of the signer leave that region and these pixel values do not belong to the set of the skin intensity values.

$$x_t(i,j) = \begin{cases} X_{t+1}(i,j) - X_{t-1}(i,j) & : \quad X_{t+1}(i,j) - X_{t-1}(i,j) < 0 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.7}$$

(a) (b) (c)

Figure 5.5: Examples for the gradient images: original image frame (a) and image frames transformed by horizontal (b) and by vertical (c) Sobel filters.

Absolute first temporal derivative (AFD). This feature consists of the combined information of the PFD and NFD feature vectors by using the absolute value of the temporal difference images.

$$x_t(i,j) = |X_{t+1}(i,j) - X_{t-1}(i,j)| \tag{5.8}$$

Second temporal derivative (SD). The information related to the acceleration of the changes or the movements can be found in the SD feature vector.

$$x_t(i,j) = X_{t+1}(i,j) - 2 \cdot X_t(i,j) + X_{t-1}(i,j) \tag{5.9}$$

We apply the skin intensity thresholding (SIT) to the original frames and then extract the temporal derivative feature vectors. Some examples of the temporal derivative features are shown in Figure 5.4.

The feature vectors defined above can be concatenated to the original image frame to provide new feature vectors containing static and dynamic information of signs and gestures.

5.1.5 Gradient images

The edges and changes of the pixel intensities contain some information about the details of the body parts appearance such as the position, the orientation and the configuration of the fingers or face components. The derivatives of the image values with respect to the image coordinates are used to form the feature vectors. These gradient images are extracted from the original images or down-scaled images employing different filters like Sobel filters, Laplace filters, Gaussian filters, etc. These filters work as a kernel operator which enhance the change of the brightness of each pixel compared to the neighboring pixels. In Figure 5.5 an original image and its transformations by horizontal and by vertical Sobel filters are presented.

Sobel filters. The Sobel filters are used in different directions to calculate the gradient of the image intensity at each point. The result gives the rate of the change in that direction. This implies that employing a Sobel filter in a direction, the result of the Sobel operator in a region of the image without any change in that direction is a zero value and at a point on an edge is a non-zero value.

Here are the Sobel filters pointing in four directions and the resulting image frames after employing the filters are shown in Figure 5.6.

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

Horizontal (H) Vertical (V) Diagonal left (DL) Diagonal right (DR)

Figure 5.6: The Sobel filters and the sample resulting image frames employing them.

5×9 7×13 11×21 15×29

Figure 5.7: Examples for the resulting image frames from the larger horizontal gradient filters.

9×5 13×7 21×11 29×15

Figure 5.8: The sample image frames resulting from the vertical gradient filters in larger scales.

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0.1004 & -0.0234 & 0.1004 \\ -0.0234 & -0.3079 & -0.0234 \\ 0.1004 & -0.0234 & 0.1004 \end{bmatrix}$$

Laplace Gaussian

Figure 5.9: The Laplace filter and the Gaussian filter with the size of 3×3 and the sample resulting image frames.

3×3 15×15 21×21 27×27

Figure 5.10: Examples for the resulting image frames from the Gaussian filters in different sizes.

Laplace filter (L). Convolving a 3×3 Laplace filter, subtracts the brightness values of the neighboring pixels from the central pixel. If we use the Laplace filter in a region of the image frame that is uniform in brightness, the result is a reducement of the grey level to zero, but when there exists a discontinuity within the neighborhood, the result of the Laplace filter is a non-zero value which appears in the form of a point, line or edge.

In Figure 5.9, The Laplace filter and the resulting image frame after employing the filter is presented.

Gaussian filters (G). The Gaussian filter like the Laplace filter is based on a second derivative of the image values with respect to the horizontal and vertical axis. The basic Gaussian filter which is usually used for image processing is shown in Figure 5.9. In Figure 5.10 four Gaussian filters in different sizes, which are used in this work, with the image frames resulting from the filters are shown.

Scaled gradient filters. To consider the changes of the pixel intensities in a region larger than 3×3, the Gaussian function is used to make gradient filters. Then we concatenate the

$$\begin{bmatrix} 0 & -0.5 & 0 \\ -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \\ 0 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 & 0 & 1 & 0 \\ -0.5 & -2 & 0 & 2 & 0.5 \\ 0 & -1 & 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 & -0.5 \\ 0 & 0 & -1 & -2 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} -0.5 & 0 & 0 & 0 & 0 \\ 0 & -2 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

(a) (b) (c) (d)

Figure 5.11: The PLUS filters and the resulting image frames after convolving the filters: (a) horizontal PLUS (HP), (b) vertical PLUS (VP), (c) diagonal left PLUS (DLP), (d) diagonal right PLUS (DRP).

$$\begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix} \quad \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix} \quad \begin{bmatrix} 0 & -3 & -10 \\ 3 & 0 & -3 \\ 10 & 3 & 0 \end{bmatrix} \quad \begin{bmatrix} -10 & -3 & 0 \\ -3 & 0 & 3 \\ 0 & 3 & 10 \end{bmatrix}$$

(a) (b) (c) (d)

Figure 5.12: The Jähne filters and the resulting image frames using the filters: (a) horizontal Jähne (HJ), (b) vertical Jähne (VJ), (c) diagonal left Jähne (DLJ), (d) diagonal right Jähne (DRJ).

Gaussian filters to produce larger gradient filters. In Figure 5.7 and 5.8, the resulting image frames employing horizontal and vertical gradient filters in larger sizes are shown.

PLUS filters. Here, we are going to introduce the PLUS filters which are derived from the second derivative in the gradient direction (SDGD) and the linear Laplace filter. The resulting filter which is called PLUS (PLUS = Laplace + SDGD) is an edge detector filter that finds curved edges more accurately than its constituents[Verbeek & van Vliet 94]. In Figure 5.11 the PLUS filters pointing in four directions are presented.

Jähne filters. For a horizontal Sobel derivative operator, there exists a correlation caused by the cross smoothing in the x direction and for a vertical Sobel filter as well.

Figure 5.13: Sample frames tracking the dominant hand.

[Jähne & Scharr+ 99] introduce a new filter which is optimized for rotational invariance. They show that the noise is substantially better suppressed. In Figure 5.12, the Jähne filters pointing in four directions are presented with the resulting image frames.

5.2 Geometric features

The Geometric features of the whole body or the body parts like the hands or the head of the signer represent spacial information related to their position, shape or configuration. In [Rigoll & Kosmala 97, Eickeler & Kosmala+ 98, Rigoll & Kosmala+ 98], the geometric features of the whole body are extracted and used successfully to recognize 24 complex dynamic gestures like "hand waving", "clapping", "pointing", and "head moving". The geometric features of the dominant and non-dominant hand are also used successfully in [Bauer & Hienz+ 00b, Bauer & Hienz 00a] to recognize sign language words. B. Bauer et al. [Bauer & Hienz+ 00b, Bauer & Hienz 00a] extract the geometric features of the fingers, palm and back side of the dominant hand where the signer wears a colored glove with seven different colors. In this section we explain the geometric features which are extracted from the dominant hand of the signer without any glove [Zahedi & Dreuw+ 06a]. The hand is tracked by the tracking method described in [Dreuw & Deselaers+ 06a] and segmented by using a simple chain coding method [Estes & Algazi 95].

The used tracking algorithm prevents taking possibly wrong local decisions because the tracking is done at the end of a sequence by tracing back the decisions to reconstruct the best path. The geometric features extracted from the tracked hand can roughly be categorized into four groups which are going to be presented in the Sections 5.2.2, 5.2.3, 5.2.4 and 5.2.5.

5.2.1 Tracking

The tracking method can be seen as a two step procedure: in the first step, the scores are calculated for each frame starting from the first, and in the second step, the globally optimal path is traced back from the last frame of the sequence to the first.

Step 1. For each position $u = (i, j)$ in the frame x_t at the time $t = 1, ..., T$ a score $q(t, u)$ is calculated, called the local score. The global score $Q(t, u)$ is the total score for the best path until the time t which ends in the position u. For each position u in an image x_t, the best predecessor is searched among a set of possible predecessors from the scores $Q(t-1, u')$. This best predecessor is then stored in a table of backpointers $B(t, u)$ which is used for the

Figure 5.14: Two examples for linear interpolation of the image frames: the interpolated image frames are shown between the successor and the predecessor frames.

traceback in Step 2. This can be expressed in the following recursive equations:

$$Q(t,u) = \max_{u' \in M(u)} \{(Q(t-1,u') - \mathcal{T}(u',u)\} + q(t,u) \tag{5.10}$$

$$B(t,u) = \operatorname*{argmax}_{u' \in M(u)} \{(Q(t-1,u') - \mathcal{T}(u',u)\}, \tag{5.11}$$

where $M(u)$ is the set of possible predecessors of the point u and $\mathcal{T}(u',u)$ is a jump-penalty, penalizing large movements.

Step 2. The traceback process reconstructs the best path u_1^T using the score table Q and the backpointer table B. The traceback starts from the last frame of the sequence at the time T by using $c_T = \operatorname{argmax}_u Q(T,u)$. The best position at the time $t-1$ is then obtained by $c_{t-1} = B(t,c_t)$. This process is iterated up to the time $t=1$ to reconstruct the best path.

Because each possible tracking center is not likely to produce a high score, pruning can be integrated into the dynamic programming tracking algorithm for speed-up.

One possible way to track the dominant hand is to assume that this object is moving more than any other object in the sequence and to look at difference images where motion occurs to track these positions. Following this assumption, we use a motion information score function to calculate local scores using the first-order time derivative of an image. The local score can be calculated by a weighted sum over the absolute pixel values inside the tracking area. More details and further scoring functions are presented in [Dreuw & Deselaers$^+$ 06a]. Figure 5.13 shows some sample frames in which the dominant hand of the signer is tracked by the explained tracking method.

Comparing to speech recognition systems, the sample rate of the image frame sequences for sign language recognition is far less than the sequence of acoustic feature vectors. Therefore we employ a linear interpolation method to produce the intermediate frames between

the two frames of the video stream. Where x_t and x_{t+1} are the image frames recorded at the time t and $t+1$, the interpolated frame $x_{t+0.5}$ is achieved by:

$$x_{t+0.5} = \frac{x_t + x_{t+1}}{2}. \tag{5.12}$$

This interpolation not only helps the explained tracking method by providing more image frames, but also provides some missing information about the body parts of the signer during the movement between the image frames captured by the camera. Figure 5.14 shows how the linear interpolation method works.

5.2.2 Basic geometric features

The first group of the features contains the feature describing basic properties including the size of the area of the hand, the length of the border of the hand, the x and y coordinates of the center of the gravity, the most top-left and bottom-right points of the hand and the compactness. The definition of the features is based on basic methods of image processing [Sonka & Hlavac$^+$ 98]. In total, nine features are calculated, where the definition of each is very well-known, except for compactness. The compactness C of the area which ranges from 0 to 1 is calculated by:

$$C = \frac{4 \cdot \pi \cdot A}{B^2}, \tag{5.13}$$

where A is the area size and B is the border size of the tracked hand. The compactness is 0 for lines and 1 for circles.

Figure 5.15.(a,b, and e) shows some sample image frames in which the center of the gravity, the most top-left and bottom-right points and also the compactness of the dominant hand of the signer are shown by basic shapes. In Figure 5.15.(a and b) the center of the gravity and the boundary of the dominant hand is shown by points and rectangles, respectively. Also in Figure 5.15.(e) the compactness of the tracked hand is shown through the radius of a circle which is located in the center of the gravity.

5.2.3 Moments

The second group consists of features that are based on moments [Sonka & Hlavac$^+$ 98]. A total of 11 features is calculated. The two dimensional $(p+q)$th order moments of the grey-value image with the pixel intensities $X(i,j)$ are defined as:

$$m_{pq} = \sum_{i}^{I} \sum_{j}^{J} i^p j^q X(i,j). \tag{5.14}$$

If $X(i,j)$ is piecewise continuous and it only has non-zero values in the finite part of the two dimensional plane, then the moments of all orders exist and the sequence $\{m_{pq}\}$ is uniquely determined by $X(i,j)$ and vise versa. The small order moments of the $X(i,j)$ describe the shape of the region. For example m_{00} is equal to the area size, and m_{01} and m_{10} give the x and the y coordinates of the center of the gravity, and also m_{11}, m_{20} and m_{02} yield the direction of the main axis of the distribution.

Figure 5.15: Some examples of the extracted geometric features from the dominant hand of the signer.

Shifting to the center of the gravity point, the central moments μ_{pq} which are invariant to translation are calculated by:

$$\mu_{pq} = \sum_{i}^{I} \sum_{j}^{J} (i - \bar{i})^p (j - \bar{j})^q X(i, j). \tag{5.15}$$

where $\bar{i} = \frac{m_{10}}{m_{00}}$ and $\bar{j} = \frac{m_{01}}{m_{00}}$.

If $p + q > 2$, the central moments can also be invariant to the changes of scale. To make the central moment μ_{pq} invariant to scaling, the moments are divided by the properly scaled (00)th moment, by using the following formula.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+\frac{p+q}{2})}} \tag{5.16}$$

The small order of the moments is calculated in the first group of the features. The moments η_{02}, η_{03}, η_{11}, η_{12}, η_{20}, η_{21} and η_{30} which are invariant for translation and changes of scale are calculated in this group and used as features.

The inertia parallel to the main axis is named J_1 and the inertia orthogonal to the main axis is named J_2 which are invariant for translation, rotation and flipping are calculated by:

$$\begin{aligned} J_1 &= \frac{m_{00}}{2} \cdot \left(m_{20} + m_{02} + \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2} \right) \\ J_2 &= \frac{m_{00}}{2} \cdot \left(m_{20} + m_{02} - \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2} \right). \end{aligned} \tag{5.17}$$

Also the orientation of the main axis O, which is invariant for translation and scaling is calculated by:

$$O = \frac{180}{2\pi} \arctan \left(\frac{2m_{11}}{m_{20} - m_{02}} \right). \tag{5.18}$$

In Figure 5.15.(c) orientation of the tracked hands is shown by the lines located on the hands.

The eccentricity E, ranges from zero for a circle to one for a line and is calculated by:

$$E = \frac{(m_{20} - m_{02})^2 + 4{m_{11}}^2}{(m_{20} + m_{02})^2}. \tag{5.19}$$

The eccentricity is invariant for translation, rotation, scaling and flipping. Figure 5.15.(e) shows eccentricity of the dominant hand in some image examples with a circle whose center is located in the center of the gravity and its radius is equal to the eccentricity of the tracked hand.

5.2.4 Hu moments

Here, seven features are extracted by determining the first seven moment invariants as described in [Hu 62].

$$
\begin{aligned}
hu_1 &= -\log(m_{20}+m_{02}) \\
hu_2 &= -\log\left((m_{20}-m_{02})^2+4m_{11}^2\right) \\
hu_3 &= -\log\left((m_{30}-3m_{12})^2+(3m_{21}-m_{03})^2\right) \\
hu_4 &= -\log\left((m_{30}+m_{12})^2+(m_{21}+m_{03})^2\right) \\
hu_5 &= -\log\Big((m_{30}-3m_{12})(m_{30}+m_{12})\big((m_{30}+m_{12})^2-3(m_{21}+m_{03})^2\big) \\
&\quad +(3m_{21}-m_{03})(m_{21}+m_{03})\big(3(m_{30}+m_{12})^2-(m_{21}+m_{03})^2\big)\Big) \\
hu_6 &= -\log\Big((m_{20}-m_{02})\big((m_{30}+m_{12})^2-(m_{21}+m_{03})^2\big) \\
&\quad +4m_{11}(m_{30}+m_{12})(m_{21}+m_{03})\Big) \\
hu_7 &= -\log\Big((3m_{21}-m_{03})(m_{30}+m_{12})\big((m_{30}+m_{12})^2-3(m_{21}+m_{03})^2\big) \\
&\quad -(m_{30}-3m_{12})(m_{21}+m_{03})\big(3(m_{30}+m_{12})^2-(m_{21}+m_{03})^2\big)\Big)
\end{aligned}
\tag{5.20}
$$

All Hu moments are invariant for translation, rotation, scaling and flipping except hu_7 which is not invariant for flipping.

5.2.5 Combined geometric features

For this category, seven features which take the distance between the center of the gravity for the tracked object and certain positions in the images into account are calculated. Additionally, the distance between the left most point and the right most point to the main axis and the distance between the front most and the rear most point to the center of the gravity along the main axis are calculated. In Figure 5.15.(f) two lines are used to show the distance between the rear most and the front most points to the center of the gravity, respectively.

Thus, we end up with 34 geometric features that are extracted from the tracked dominant hand in the images.

6 Automatic Sign Language Recognition

Das Rätsel dieser Welt löst weder du noch ich,
Jene geheime Schrift liest weder du noch ich.
Wir wüssten beide gern, was jener Schleier birgt,
Doch wenn der Schleier fällt, bist weder du noch ich.
– Omar Khayyam *(1048–1122)*

The decision making of our system employs hidden Markov models (HMM) to recognize the sign language words and sentences. This approach is inspired by the success of the application of hidden Markov models in speech recognition [Rabiner 89, Jelinek 98, Kanthak & Molau$^+$ 00, Lööf & Bisani$^+$ 06]. Also HMMs are employed by most of the research groups to model sequential samples like gestures and human actions in [Schlenzig & Hunter$^+$ 94, Pavlovic & Sharma$^+$ 97, Bobick & Wilson 97, Brand & Oliver$^+$ 97, Rigoll & Kosmala$^+$ 98, Moore & Essa 02, Nguyen & Bui$^+$ 03].

Since the recognition of sign language words and sentences is similar to speech recognition for the modeling of sequential samples, most sign language recognition systems like [Nam & Wohn 96, Vogler & Metaxas 97, Starner & Weaver$^+$ 98, Bauer & Hienz$^+$ 00b, Bowden & Windridge$^+$ 04] employ hidden Markov models as well.

Comparing to speech recognition systems, the data sets of sign language recognition systems are rather small, and there is not always enough data available for a robust estimation of the visual models for the sign language words. When adding a new gloss to a training corpus, there is no data from the other glosses to be used in training of the new model, as the definition of phonemes or sub-word units in sign language recognition is still unclear.

In this chapter the theory of a hidden Markov models and the methods employed in our automatic sign language recognition (ASLR) system are going to be explained in detail. Furthermore, as the input of the system are video frames captured from the signers with a large visual variability of the utterances of each sign language word, we are going to explain how to make the classifier more robust against variability of signs and deformations.

The methods which are employed in speech recognition systems for feature selection and combination are used in our work to improve the accuracy of the system too. We are going to explain how these methods can be useful for a sign language recognition system.

6.1 Hidden Markov model

Given $x_1^T = x_1, ..., x_t, ... x_T$ which is a sequence of feature vectors, our decision making rule based on Bayes' decision rule chooses the best sequence of words $w_1^N = w_1, ..., w_N$ which

maximizes the a-posteriori probability:

$$x_1^T \longrightarrow r(x_1^T) = \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N|x_1^T) \right\} \tag{6.1}$$

$$= \underset{w_1^N}{\operatorname{argmax}} \left\{ \frac{Pr(w_1^N) \cdot Pr(x_1^T|w_1^N)}{Pr(x_1^T)} \right\} \tag{6.2}$$

$$= \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N) \cdot Pr(x_1^T|w_1^N) \right\}, \tag{6.3}$$

where language model $Pr(w_1^N)$ is the prior probability of the word sequence w_1^N. The $Pr(x_1^T|w_1^N)$ called visual model (cp. acoustic model in speech recognition), is the class conditional probability of observing sequence x_1^T given a word sequence w_1^N.

The architecture of automatic sign language recognition system, adopted from automatic speech recognition (ASR) system, is shown in Figure 6.1. It contains four main components which are explained in the following sections in more detail:

- The feature analysis module extracts visual features of the video input or image frame sequences and chooses the most discriminative components as a search module. This is explained in the Sections 6.1.1 and 6.5.

- The visual model $Pr(x_1^T|w_1^N)$, comparing to an acoustic model in speech recognition system, is the class conditional probability of observing sequence x_1^T given a word sequence w_1^N. In contrary to medium and large vocabulary speech recognition systems where the acoustic model is defined on a phonetic or a phoneme level, the visual model is defined on a sign language word level [Dreuw & Rybach$^+$ 07]. The visual model is explained in more detail in Section 6.1.2.

- The language model gives a statistical model from the syntax, semantics and pragmatics of a language. The language model probability is calculated on written language and it is independent of the visual model. It means even in scarceness of the video data, we can use a big dataset including only a sign language text written in glosses. In Section 6.1.3 we are going to illustrate how the language model is extracted from a written text dataset.

- The search module, minimizing the expected number of words recognized incorrectly, integrates the visual model of sign language words and the language model to determine the optimal word sequence with the highest posterior probability of $Pr(w_1^N|x_1^T)$ for a given visual sequence $x_1, ..., x_T$. The search module is broken into a training, a development, and a recognition (evaluation) part, and is going to be explained further in the Sections 6.1.4, and 6.1.5, 6.1.6.

6.1.1 Signal analysis

The signal analysis module aims at providing the sign language recognizer with a stream of visual feature vectors. The vector sequence consists of the features from the sequence of the image frames extracted from the signer. It should fulfill the following criteria:

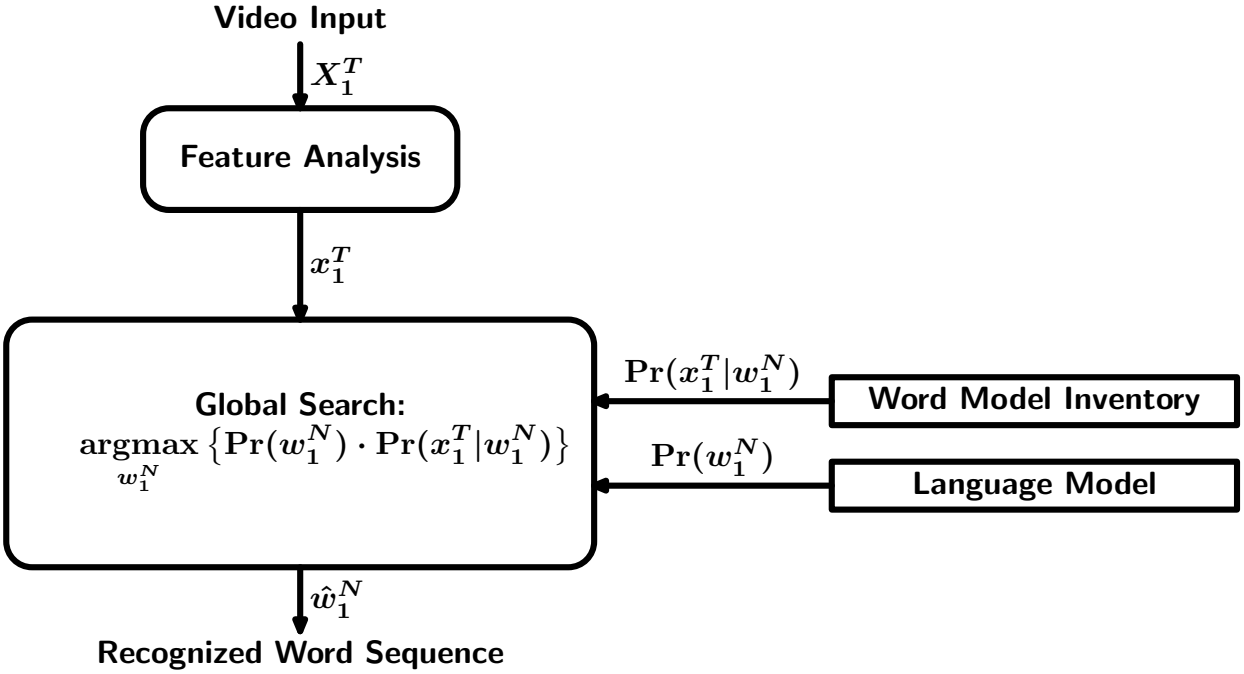


Figure 6.1: Basic architecture of the automatic sign language recognition system.

- feature vectors should be characteristic of the sign language words used in the visual modeling. In other words, the feature vectors should be similar for the same word models, but discriminative among the vocabulary set.
- they should depend on the signed concepts only; which means they have to tolerate certain recording conditions like different lightening, different signers; male or female, different pronunciations, etc.
- they have to be as small as possible to allow robust parameter estimation. Also low dimensionality causes the recognition system to run faster.

The feature extraction part is going to be explained in all details in Chapter 5. It includes appearance based features which are extracted directly from the original images as well as the geometric features of the dominant hand of the signer. In Section 6.3 we discuss about normalization schemes and how to make the classifier invariant against local and global transformations which leave the class membership of the signs unchanged.

We explain the methods used to reduce the dimensionality of the feature vectors in Section 6.5. The *linear discriminant analysis* (LDA) and *Principle component analysis* (PCA) are two transformations which are employed to extract the most discriminative coefficients of the feature vectors. Therefore we expect that the loss of information involved by the dimension reduction can be compensated by a more reliable parameter estimation in the reduced feature space.

6.1.2 Visual model

The probability $Pr(x_1^T|w_1^N)$ is defined as:

$$Pr(x_1^T|w_1^N) = \max_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1}, w_1^N) \cdot Pr(x_t|s_t, w_1^N) \right\}, \tag{6.4}$$

where s_1^T is the sequence of states, and $Pr(s_t|s_{t-1}, w_1^N)$ and $Pr(x_t|s_t, w_1^N)$ are the transition probability and the emission probability, respectively. The transition probability is estimated by simple counting. The emission probabilities can be modeled either as *discrete probabilities* [Jelinek 76], as *semi-continuous probabilities* [Huang & Jack 89], or as *continuous probability distributions* [Levinson & Rabiner$^+$ 83]. We use the latter case as Gaussian mixture densities for the emission probability distribution $Pr(x_t|s_t, w_1^N)$ in the states. The emission probability is defined as:

$$\begin{aligned} Pr(x_t|s_t, w_1^N) &= \sum_{l=1}^{L(s_t)} Pr(x_t, l|s_t, w_1^N) \\ &= \sum_{l=1}^{L(s_t)} Pr(l|s_t, w_1^N) \cdot Pr(x_t|s_t, w_1^N, l), \end{aligned} \tag{6.5}$$

where $L(s_t)$ is the number of densities in each state and

$$Pr(x_t|s_t, w_1^N, l) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{l,s_t,w_1^N,d}}} \cdot \exp\left(-\frac{(x_{t,d} - \mu_{l,s_t,w_1^N,d})^2}{\sigma^2_{l,s_t,w_1^N,d}}\right). \tag{6.6}$$

In this work, the sum is approximated by the maximum, and the emission probability is defined as:

$$\begin{aligned} Pr(x_t|s_t, w_1^N) &= \max_l \left\{ Pr(x_t, l|s_t, w_1^N) \right\} \\ &= \max_l \left\{ Pr(l|s_t, w_1^N) \cdot Pr(x_t|s_t, w_1^N, l) \right\}. \end{aligned} \tag{6.7}$$

To estimate $Pr(x_t|s_t, w_1^N)$, we use the maximum likelihood estimation method for the parameters of the Gaussian distribution, i.e. the mean $\mu_{s_t,w_1^N,d}$ and the variances $\sigma_{s_t,w_1^N,d}$. Here, the covariance matrix is modeled to be diagonal, i.e. all off-diagonal elements are fixed at zero. The number of states for the HMM of each word is determined by a fixed number, for instance the minimum sequence length of the training samples for segmented sign language recognition. Instead of a density-dependent estimation of the variances, we use pooling during the training of the HMM, which means that we do not estimate variances for each density of the HMM, but instead we estimate one set of variances for all densities in the complete model (word-dependent pooling for single word recognition and global pooling for continuous sign language recognition).

We use the Viterbi algorithm to find the maximizing state sequence s_1^T. By using the Viterbi algorithm, we calculate the score of the observation feature vector x_t in the emission probability distribution $Pr(x_t|s_t, w_1^N)$ at each state s_t. Assuming the Gaussian function with diagonal covariances for $Pr(x_t|s_t, w_1^N)$, as described above, this score is calculated as:

$$\begin{aligned} -\log Pr(x_t|s_t, w_1^N) = \min_l \Bigg\{ & \underbrace{\frac{1}{2} \sum_{d=1}^{D} \frac{(x_{t,d} - \mu_{l,s_t,w_1^N,d})^2}{\sigma^2_{l,s_t,w_1^N,d}}}_{\text{distance}} - \log Pr(l|s_t, w_1^N) + \\ & \frac{1}{2} \sum_{d=1}^{D} \log(2\pi\sigma^2_{l,s_t,w_1^N,d}) \Bigg\}. \end{aligned} \tag{6.8}$$

In this work, when the feature vector x_t is a down-scaled image at the time t, the sum $\sum_{d=1}^{D}(x_{t,d} - \mu_{l,s_t,w_1^N,d})^2/\sigma^2_{l,s_t,w_1^N,d}$ is the distance between the observation image at the time t and the mean image μ_{l,s_t,w_1^N} of the state s_t which is scaled by the variances $\sigma^2_{l,s_t,w_1^N,d}$. This scaled Euclidean distance can be replaced by other distance functions such as the tangent distance or image distortion model distance, which we are going to introduce in Section 6.3.

6.1.3 Language model

The language model models syntax, semantics and pragmatics of the sign language at the word level. The probability of $Pr(w_1^N)$ as an a-priori probability of a word sequence w_1^N is provided by a stochastic model. This probability concerns only the constraints of written glosses from the sign language and it is independent from a visual model. To estimate the language model probability we assume that a sign language word sequence follows an $(m-1)$-th order Markov chain. Therefore the probability of observing a word sequence w_1^N is calculated by

$$\begin{aligned} Pr(w_1^N) &= \prod_{n=1}^{N} Pr(w_n|w_1^{n-1}) && (6.9)\\ &= \prod_{n=1}^{N} Pr(w_n|w_{n-m+1}^{n-1}). && (6.10) \end{aligned}$$

The word sequence w_{n-m+1}^{n-1} which is denoted by h_n is named history of the word w_n with a history length of $m-1$. In other words if the history length is equal to m, a word w_n depends on its history h_n which is named *m-gram* language model [Bahl & Jelinek$^+$ 83]. In particular, The language models with a history length of one, two and three are called *unigram*, *bigram* and *trigram* respectively. This definition needs some assumption to work properly:

- If the index $n-m+1$ in w_{n-m+1}^{n-1} is smaller than one, it is equal to one.
- If the upper index $n-1$ is smaller than the lower index $n-m+1$, the w_{n-m+1}^{n-1} is an empty sequence and the language probability is equal to $Pr(w_1)$.

The maximum-likelihood principle is used for estimation and evaluation of the language model. We define an equivalent criterion of language model named *perplexity* (PP) [Bahl & Jelinek$^+$ 83]. In the following, the perplexity of a word sequence w_1^N is defined by

$$\begin{aligned} PP &= Pr(w_1^N)^{-\frac{1}{N}} && (6.11)\\ &= \left[\prod_{n=1}^{N} Pr(w_n|h_n)\right]^{-\frac{1}{N}}, && (6.12) \end{aligned}$$

which is the inverse geometric mean of the product of the conditional probability $Pr(w_n|h_n)$ of all words of the whole sequence. The perplexity is the average number of possible words at each position of the entire text. Therefore the perplexity as a criterion in the training

process has to be decreased as much as possible. The *language model score* is calculated by the negative logarithm of the language model probability:

$$\log PP = -\frac{1}{N}\sum_{n=1}^{N}\log Pr(w_n|h_n). \tag{6.13}$$

Also we have employed smoothing methods to guarantee that the probabilities of the language model are larger than zero. The methods used for the implementation of the language model are explained in very detail in [Ney & Essen$^+$ 94, Wessel & Ortmanns$^+$ 97].

6.1.4 Training

In the training process, the visual model and the language model have to be created using the training set of the database. The statistical methods are employed to model the language model and the visual model by probability distributions. The input of the training process includes Q number of sentences containing a few sign language words and a corresponding sequence of image frames. The sequences of image frames are used to train a HMM as a visual model for each sign language word and the sequences of written sign language words are used to train the language model. At this point we are going to explain how to train the language model and the visual model in detail.

Training of the visual model An *expectation maximization (EM)* algorithm is used to estimate the parameters of the visual model which includes the parameters of mixture densities of the emission probability and transition probability of the HMMs for the sign language words.

The transition probability between the states of HMMs is calculated by a simple counting method. *Mixture densities* has been made out of a weighted sum of the Gaussian distributions (Eqn. 6.14) are used to model the continuous probability distributions. Furthermore, the Viterbi approximation [Merialdo & Marchand-Maillet$^+$ 00] is applied at the density level as well (Eqn. 6.15):

$$p(x_t|s_t, w_1^N) = \sum_{l=1}^{L_(s_t)} c_{sl}\cdot\mathcal{N}(x_t|\mu_{sl},\Sigma,w_1^N) \tag{6.14}$$

$$\cong \max_l\left\{c_{sl}\cdot\mathcal{N}(x_t|\mu_{sl},\Sigma,w_1^N)\right\}, \tag{6.15}$$

where l denotes the index of the density within the mixture of the state s, and c_{sl} is the weight for the corresponding single Gaussian density. μ_{sl} is the mean vector of Gaussian density l in the state s, and Σ is the pooled diagonal covariance matrix independent of s and l.

If $\hat{\vartheta}$ is the set of the parameters of the Gaussian mixture densities for the emission probabilities, using the maximum likelihood principle, the $\hat{\vartheta}$ parameters are estimated by:

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmax}}\left\{\prod_{q=1}^{Q} p\left([x_1^{T_q}]_q \,|\, [w_1^{N_q}]_q, \vartheta\right)\right\}, \tag{6.16}$$

where $[x_1^{T_q}]_q$ is a feature vector sequence extracted from the sequence of image frames and $[w_1^{N_r}]_r$ is its corresponding sequence of the sign language words.

For training, first performing a so-called linear segmentation, each mixture is initialized by a single Gaussian density, i.e. to initialize the visual model parameters an initial alignment is done. Then collecting the feature vectors for each state, the resolution of the mixture densities increases successively by density splitting. Parameter estimation is iteratively performed according to the *maximum likelihood* principle (Eqn. 6.16) with the *expectation maximization* algorithm [Dempster & Laird$^+$ 77]. A *dynamic programming* procedure is employed to calculate the summation and the maximum approximation efficiently [Ney 84].

Training of the language model We use the text database of the training set, including the sequences of sign language glosses to estimate the language model probability $Pr(w_1^N)$. Furthermore, either the minimum perplexity or the maximum log-likelihood of the language model probabilities of the text database is the training criterion of the language model. The maximizing log-likelihood function is defined as:

$$F = \sum_{n=1}^{N} \log Pr(w_n|h_n) \qquad \text{with} \qquad \sum_{w} p(w|h) = 1 \qquad \forall h, \tag{6.17}$$

$$F = \sum_{h,w} N(h,w) \log Pr(w,h) \qquad - \qquad \sum_{h} \mu_h \left(\sum_{w} Pr(w|h) - 1 \right) \tag{6.18}$$

where $N(h,w)$ is the number of the happening word w after the word sequences h in the training set. We differentiate the criterion function F by $Pr(w|h)$ and μ_h to solve the maximization problem. The result is:

$$p(w|h) = \frac{N(h,w)}{\sum_{w'} N(h,w')} \tag{6.19}$$

$$= \frac{N(h,w)}{N(h)}. \tag{6.20}$$

However, the number of m-grams increases exponentially with an increasing length of the history h. For a large number of the history length m, the vast majority of m-grams will not be seen in the training data or occur too infrequent. Therefore the unseen m-grams would get the probability zero and could never be recognized. To ensure that the probability of all m-grams is larger than zero, a smoothing method is applied to the language models.

We have employed a smoothing method based on various *discounting* schemes [Katz 87, Ney & Essen$^+$ 94] which reduce the probability mass of the observed m-grams to distribute them among the unseen (*backing off*) or all (*interpolation*) m-grams.

The generalized language model probabilities which are calculated based on shorter histories determine the amount of the discounting mass that is assigned to each m-gram. To estimate the discounting and the generalized language model parameters, the *leaving-one-out* algorithm is run.

To train language models, we have used the SRILM toolkit [Stolcke 02] to estimate the language model parameters.

6.1.5 Recognition and evaluation

In Figure 6.1 the search module performs the recognition task of the automatic sign language recognition. It searches for the optimum word sequence $[w_1^N]_{opt}$ which maximizes the posterior probability of $Pr(w_1^N|x_1^T)$, given a sequence of feature vectors x_1^T extracted from the image sequences. The recognition is based on Bayes' decision rule and the basic decision rule for the classification of $x_1^T = x_1, ..., x_t, ... x_T$ is:

$$[w_1^N]_{opt} = \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N) \cdot Pr(x_1^T|w_1^N) \right\}, \tag{6.21}$$

where $[w_1^N]_{opt}$ is the sequence of words which are recognized, $Pr(w_1^N)$ is the language model, and $Pr(x_1^T|w_1^N)$ is the visual model (cp. acoustic model in speech recognition). For the language model $Pr(w_1^N)$ we use a trigram language model calculated by

$$Pr(w_1^N) = \prod_{n=1}^{N} Pr(w_n|w_{n-2}^{n-1}). \tag{6.22}$$

The visual model $Pr(x_1^T|w_1^N)$ is defined as:

$$Pr(x_1^T|w_1^N) = \max_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1}, w_1^N) \cdot Pr(x_t|s_t, w_1^N) \right\}, \tag{6.23}$$

where s_1^T is the sequence of the states, and $Pr(s_t|s_{t-1}, w_1^N)$ and $Pr(x_t|s_t, w_1^N)$ are the transition probability and the emission probability, respectively. In training, the model parameters are estimated from the training data using the maximum likelihood criterion and the EM algorithm with Viterbi approximation.

As the language model, transition and emission probabilities can be weighted by exponentiation with exponents α, β and γ, respectively, the probability of the knowledge sources are estimated as:

$$\begin{aligned} Pr(w_1^N) &\rightarrow p^{\alpha}(w_1^N), \\ Pr(s_t|s_{t-1}, w_1^N) &\rightarrow p^{\beta}(s_t|s_{t-1}, w_1^N), \\ Pr(x_t|s_t, w_1^N) &\rightarrow p^{\gamma}(x_t|s_t, w_1^N). \end{aligned} \tag{6.24}$$

Thus, the decision rule is reformulated as:

$$\begin{aligned} [w_1^N]_{opt} = \underset{w_1^N}{\operatorname{argmax}} \Bigg\{ & \alpha \sum_{n=1}^{N} \log p(w_n|h_n) \\ & + \max_{s_1^T} \Big\{ \sum_{t=1}^{T} \left[\beta \log p(s_t|s_{t-1}, w_1^N) + \gamma \log p(x_t|s_t, w_1^N) \right] \Big\} \Bigg\} \end{aligned} \tag{6.25}$$

$$\begin{aligned} = \underset{w_1^N}{\operatorname{argmax}} \Bigg\{ & \frac{\alpha}{\gamma} \sum_{n=1}^{N} \log p(w_n|h_n) \\ & + \max_{s_1^T} \Big\{ \sum_{t=1}^{T} \left[\frac{\beta}{\gamma} \log p(s_t|s_{t-1}, w_1^N) + \log p(x_t|s_t, w_1^N) \right] \Big\} \Bigg\}. \end{aligned} \tag{6.26}$$

The exponents used for scaling, $\frac{\alpha}{\gamma}$ and $\frac{\beta}{\gamma}$ are named language model scale and time distortion penalty, respectively.

In theory, the search has to hypothesize all possible word sequences $W = w_1, \ldots, w_N$ and find the optimum one by using the maximum likelihood according to the recognition equation. The number of possible word sequences, for a database including $|V|$ words, grows exponentially with the number of words N in the sequence:

$$V^0 + V^1 + V^2 + V^3 + \ldots + V^N = \frac{V^{N+1} - 1}{V - 1}. \tag{6.27}$$

In [Bellman 57], it is shown how to reduce the complexity of the optimization significantly by dynamic programming, which accomplishes the complex mathematical structure of the task. In other words, it decomposes the difficult global optimization problem into a number of local optimization problems which can be solved more easily. In this work, we employ a *Viterbi search* [Vintsyuk 71, Ney 84] which is applied in the RWTH automatic speech recognition system. At each time frame the likelihood of all hypotheses can be compared with each other, since the state hypotheses are expanded time-synchronously. Therefore efficient *pruning techniques*, which omit the unlikely hypotheses early from the optimization process, reduces the number of options for the state sequences significantly.

To evaluate the recognition results, we calculate the *word error rate (WER)* which is defined by:

$$WER = \frac{\#\text{substitutions} + \#\text{insertions} + \#\text{deletions}}{\#\text{reference words}} \tag{6.28}$$

It uses the Levenshtein distance [Levenshtein 66] and is called *edit distance*, between the correct word sequence w_1^N and the recognized word sequence $\hat{w}_1^{\hat{N}}$. The WER is the minimum number of *substitutions*, *insertions* and *deletions* which is necessary to transform the recognized sequence into the correct sequence. In this work, the WER is calculated with a dynamic programming algorithm as explained in [Ney 05b].

6.1.6 Development

The development process is performed to find the weights for the language model, the transition and the emission probabilities which are named α, β and γ, respectively. When a database consists of a training, a development and an evaluation part, using training set we make up the visual model and language model. Then we perform the recognition process on the development set to find the optimum weights for α, β and γ. Finally we evaluate the performance of the system on the evaluation set by using the weights obtained in the development process.

6.2 Other classification approaches

Although classification of segmented sign language words is also based on the same decision making rules which are explained for continuous sign language recognition, there are still some issues of segmented sign language recognition which are different. In this section we are going to explain other classification approaches which are particularly employed in this work for segmented sign language word recognition. First we review how Bayes' decision rule

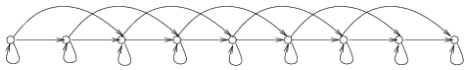


Figure 6.2: The topology of the employed HMM.

is employed for the recognition of single sign language words. Then we explain the leaving-one-out method which is employed due to the small available corpora for cross validation. Also we are going to explain how the nearest neighbor classifier is used with hidden Markov models of sign language words which is a special case of pronunciation modeling.

6.2.1 Isolated sign language word recognition

We have employed the same decision making rule to recognize the sign language words as we have used in previous section for continuous sign language recognition.

The topology of the HMM is shown in Figure 6.2. There is a transition loop at each state and the maximum allowable transition is set to two. We consider one HMM for each word $w = 1, ..., W$. The basic decision rule used for the classification of $x_1^T = x_1$, ..., x_t, ... x_T is:

$$r(x_1^T) = \underset{w}{\operatorname{argmax}} \left\{ Pr(w|x_1^T) \right\} \tag{6.29}$$

$$= \underset{w}{\operatorname{argmax}} \left\{ Pr(w) \cdot Pr(x_1^T|w) \right\} \tag{6.30}$$

where language model $Pr(w)$ is the prior probability of the class w and in the isolated sign language word recognition can only use zero-gram or unigram language models. The $Pr(x_1^T|w)$ is the class conditional probability of the x_1^T given class w. The $Pr(x_1^T|w)$ is defined as:

$$Pr(x_1^T|w) = \max_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1}, w) \cdot Pr(x_t|s_t, w) \right\} \tag{6.31}$$

where s_1^T is the sequence of states and further $Pr(s_t|s_{t-1}, w)$ and $Pr(x_t|s_t, w)$ are the transition probability and emission probability, respectively. The transition probability is calculated by simple counting. We use the Gaussian and Laplace function as emission probability distributions $Pr(x_t|s_t, w)$ in the states. To estimate $Pr(x_t|s_t, w)$ we use the maximum likelihood estimation method for the Gaussian and the Laplace functions, i.e. standard deviation and mean deviation estimation, respectively. The number of states for the HMM of each word can be determined in two ways: minimum and average sequence length of the training samples. Mixture densities with a maximum number of five densities are used in each state.

We use the Viterbi algorithm to find the sequence of the HMM. In addition to the density-dependent estimation of the variances, we use pooling during the training of the HMM which means that we do not estimate variances for each density of the HMM, but instead we estimate one set of variances for all densities in each state of the model (state-dependent pooling) or for all densities in the complete model (word-dependent pooling).

6.2.2 Leaving-one-out

The validation of the experiments performed on single word databases is an important issue. In the databases produced for isolated sign language recognition, RWTH-BOSTON-

10 and RWTH-BOSTON-50, the number of utterances for each word is not large enough to separate them into training and evaluation sets. Therefore, we employ the leaving-one-out method for training and classification [Zahedi & Keysers$^+$ 05b, Zahedi & Keysers$^+$ 05c, Zahedi & Keysers$^+$ 05a]. That is, we separate each utterance as a test sample, then train the HMM of each word with the remaining utterances, and finally classify the test utterance. We repeat this process for all utterances in the database. The percentage of the misclassified utterances is then the error rate of the system. The experiments using the leaving-one-out method on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50 are reported in the Chapter 7.

6.2.3 Nearest neighbor classifier

Nearest neighbor classification is a special case in the modeling of the different pronunciations which is going to be explained in Section 6.4. In the nearest neighbor classification the number of pronunciations is considered to be equal to the number of the training utterances for each word. Using each training utterance in the database, we create an HMM. According to the leaving-one-out method used in this work for isolated sign language recognition, we separate an utterance as a test utterance from the database. This unknown utterance is classified as belonging to the same class as the most similar or nearest utterance in the training set of the database. This process is repeated for all utterances in the database.

If y_{w1}, y_{w2},..., y_{wN_w} are the training observations of the word w, the decision rule used to classify an observation sequence $x_1^T = x_1,\ldots, x_t,\ldots x_T$ in this approach is

$$x_1^T \longrightarrow r(x_1^T) = \arg\min_w \left\{ \min_{n=1,\ldots,N_w} \left\{ d(x_1^T, y_{wn}) \right\} \right\}, \tag{6.32}$$

where y_{wn} is the n-th observation of the word w and the distance between the observation x_1^T and the utterance y_{wn} is calculated by:

$$d(x_1^T, y_{wn}) = -\log\left\{ Pr\left(x_1^T | y_{wn}\right) \right\}. \tag{6.33}$$

The probability $Pr\left(x_1^T|y_{wn}\right)$ is approximated by using the maximum approximation:

$$\begin{aligned} Pr\left(x_1^T|y_{wn}\right) &= \sum_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1}, y_{wn}) \cdot Pr(x_t|s_t, y_{wn}) \right\} && (6.34) \\ &\cong \max_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1}, y_{wn}) \cdot Pr(x_t|s_t, y_{wn}) \right\}. && (6.35) \end{aligned}$$

Here, the transition probability $Pr(s_t|s_{t-1}, y_{wn})$ is uniformly distributed, and the emission probability $Pr(x_t|s_t, y_{wn})$ is:

$$Pr(x_t|s_t, y_{wn}) = \frac{1}{\sqrt{2\pi\sigma^2}^D} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_{td} - y_{wns_td})^2}{\sigma^2} \right). \tag{6.36}$$

6.3 Invariance in classification

Some certain changes and transformations of the input video leave the classes of sign language words unchanged. For example, when signing a sign language sentence by different signers, the size of the hand, or the head or the position of the body of the signer changes. Even when a signer signs a sentence two times, there are still some small changes between the image frames of the two signings. However, all these varieties are not enough to change the concept of the sentence, and therefore the sequence of glosses of the sentence is the same. For this reason the automatic sign language recognition has to have a classifier which is invariant with respect to these certain transformations.

To reach to this goal, we have to ignore these kinds of varieties in the different levels of the classification process. First, in feature analysis we can extract features which are invariant with respect to the different transformations, or we can normalize the feature vectors in regard to the chosen transformations like rotation, scaling, and transition (RST). For example some geometric features which were introduced in Chapter 5 are invariant with respect to translation, rotation or scaling. Second, we can change the decision making part to ignore these varieties when measuring the similarity of the image frames. This deals with invariant probability density functions, which takes into the account with invariant distance functions. In the following, more details of these two methods are going to be explained.

6.3.1 Normalization and invariant features

Since some transformations do not change the class membership of the signs we can either extract the features which are invariant with respect to these transformations or make them invariant. For example, when the illumination of the room or the size of the body parts of the signer changes during the signing of a sign, this does not change the meaning of the sign. Therefore, if the illumination of the room changes due to the changes of the sunshine, location of curtains of the room or position of camera with respect to the signer's place, intensity features are not good features to be used for recognition. Because intensity features are strongly dependent on illumination, transition, scaling and rotation which vary by the change of illumination and camera movements, respectively.

To normalize the feature vectors we have to remove the factor which results in the feature vector but does not change the class membership of the signs. To make the image features invariant to illumination changes, they have to be normalized to have the same mean intensity values. Also, to achieve invariance to transition and scaling, it is sufficient to translate the center of gravity of the image frames to the origin and to normalize the image features to have an average radius, respectively. Features can be normalized for rotation of the image frames as well, but rotation of the image frames changes the meaning of the signs and its class membership.

The geometric features of the dominant hand which were introduced in Chapter 5, contain some members like size of the hand area, compactness or eccentricity which are invariant to transition, scaling or rotation. Also there are some other features like moments which are not invariant against these changes and we can make them invariant with respect to scale and transition. The experiments using these invariant features are reported in Chapter 7.

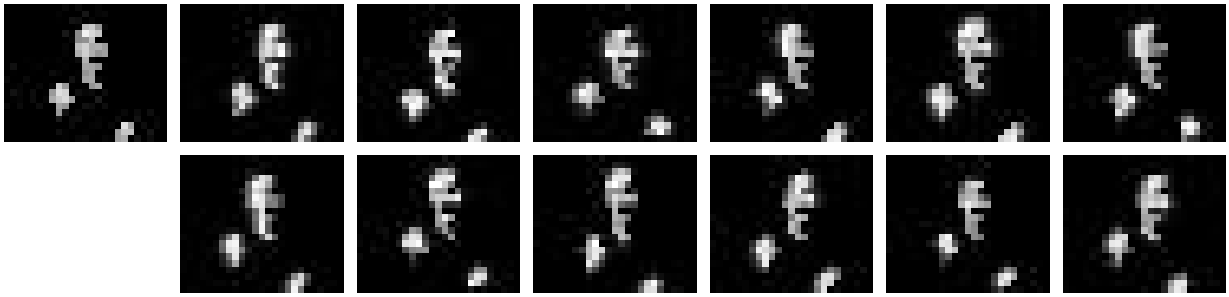


Figure 6.3: Example of the first-order approximation of the affine transformations. (Left to right: original image, ± horizontal translation, ± vertical translation, ±axis deformation, ± diagonal deformation, ± scale, ± rotation)

6.3.2 Invariant distances

Each signer may utter a sign language word differently, depending on his individual signing style or on the predecessor and on the successor of the uttered word. Therefore, a large visual variability of utterances for each word exists. Due to the visual variability of the utterances of each word, invariance is an important aspect in sign language recognition. An invariant distance measure ideally takes into account the transformations of the patterns, yielding small values for patterns which mostly differ by a transformation which does not change the class-membership. To model the variability of utterances, the tangent distance (TD) [Drucker & Schapire$^+$ 93, Keysers & Macherey$^+$ 04] and the image distortion model (IDM) [Keysers & Gollan$^+$ 04, Dreuw & Keysers$^+$ 05, Dreuw & Deselaers$^+$ 06b] can be used to account for global and local variations, respectively.

When the feature vector x_t is an original image frame or a down-scaled image at the time t, the sum $\sum_{d=1}^{D}(x_{t,d} - \mu_{l,s_t,w,d})^2/\sigma^2_{l,s_t,w,d}$ is the distance between the observation image at the time t and the mean image $\mu_{l,s_t,w}$ of the state s_t which is scaled by the variances $\sigma^2_{l,s_t,w,d}$. This scaled Euclidean distance can be replaced by other distance functions which are invariant with respect to the changes which leave the class membership unchanged.

Tangent distance

Let $x_t \in \mathbb{R}^D$ be a pattern and $f(x_t, \alpha)$ denotes a transformation of x_t that depends on a parameter L-tuple $\alpha \in \mathbb{R}^L$, where we assume that f does not affect class membership (for a small α). The set of all transformed patterns now is a manifold $\mathcal{M}_{x_t} = \left\{ f(x_t, \alpha) \ : \ \alpha \in \mathbb{R}^L \right\} \subset \mathbb{R}^D$ in the pattern space. The distance between two patterns can then be defined as the minimum distance between the manifold $\mathcal{M}_{x_t}$ of the pattern x_t and the manifold $\mathcal{M}_\mu$ of a class specific prototype pattern μ. However, the distance calculation between manifolds is a hard non-linear optimization problem in general. The manifolds can be approximated by a *tangent subspace* $\widehat{\mathcal{M}}$. The *tangent vectors* $x_{t,l}$ that span the subspace are the partial derivatives of $f(x_t, \alpha)$ with respect to α_l. Thus, a first-order approximation of $\mathcal{M}_{x_t}$ is obtained as

$$\widehat{\mathcal{M}}_{x_t} = \left\{ x_t + \sum_{l=1}^{L} \alpha_l x_{t,l} \ : \ \alpha \in \mathbb{R}^L \right\} \subset \mathbb{R}^D. \tag{6.37}$$

Using the linear approximation $\widehat{\mathcal{M}}_{x_t}$ has the advantage that distance calculations are equivalent to the solution of linear least square problems or equivalently projections into subspaces, which are computationally inexpensive operations. The approximation is valid for small values of α, which is nevertheless sufficient in many applications, as Figure 6.3 shows examples of an image frame of the RWTH-BOSTON dataset. The depicted patterns all lie in the same subspace and can therefore be represented by one prototype and the corresponding tangent vectors. The tangent distance between the original image and all transformations are therefore zero, while the Euclidean distance is significantly greater than zero. Using the squared Euclidean norm, the double-sided TD is defined as

$$d(x_t, \mu) = \min_{\alpha,\beta \in \mathbb{R}^L} \left\{ ||(x_t + \sum_{l=1}^{L} \alpha_l x_{t,l}) - (\mu + \sum_{l=1}^{L} \beta_l \mu_l)||^2 \right\} . \tag{6.38}$$

A single-sided TD can also be defined, where only one of the manifolds of the reference or the observation is approximated and the distance is minimized over all possible combinations of the respective parameters.

Image distortion model

Here, we are going to briefly review an image distortion model that is able to compensate local displacements. The efficiency of the model in handwritten character recognition is shown in [Keysers & Gollan$^+$ 04]. In this model, to calculate the distance between the image frame x_t and the mean image μ, instead of computing the square error between the pixels x_{ij} and μ_{ij}, we compute the minimum distance between x_{ij} and $\mu_{i'j'}$, where $(i', j') \in R_{ij}$, and R_{ij} is a certain neighborhood of (i, j). According to this definition, the invariant distance can be calculated by

$$d(x_t, \mu) = \sum_{ij} \min_{(i',j') \in R_{ij}} \Big\{ \underbrace{||x_{ij} - \mu_{i'j'}||}_{\text{Euclidean distance}} + C_{iji'j'} \Big\}, \tag{6.39}$$

where $C_{iji'j'} > 0$ is the displacement cost from a point (i, j) to a point (i', j') of the neighborhood region. Normally long distance displacements produce larger costs. In this work, the neighborhood R_{ij} is a region that allows one pixel displacement in eight directions and the cost function is defined by:

$$C_{iji'j'} = \begin{cases} 0 & : \quad (i', j') \in R_{ij} \\ \infty & : \quad \text{otherwise} \end{cases} . \tag{6.40}$$

The accuracy of the IDM depends on choosing useful displacements which leave the class-membership unchanged. Unwanted displacements increase the error rate of the classifier. To restrict the displacements in the IDM, instead of the brightness of each single pixel, local appearances of the image frames are used to form the feature vector. The efficiency of using the local image context such as derivatives of the image values with respect to the coordinates and local sub images is also shown in handwritten character recognition [Keysers & Gollan$^+$ 04]. Therefore, we use local sub images of the original image and its derivatives with respect to the coordinates as horizontal and vertical gradient images. Informal experiments lead us to use sub images of size 7×7 pixels instead of smaller or larger

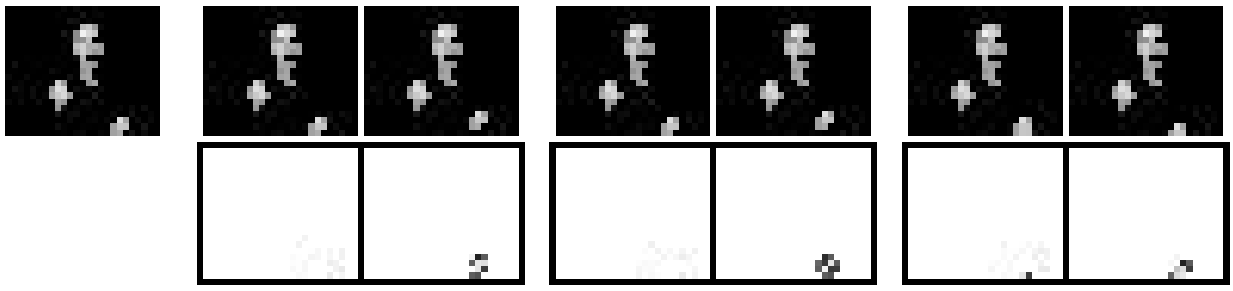


Figure 6.4: Example of the image distortion model. (first row: original image and image pairs including the transformed image which results from the IDM distance and the distorted image by displacement of the left hand of the signer, second row: the difference image between the transformed image or distorted image and the original image)

sub images to achieve better results. Therefore the Euclidean distance between pixel x_{ij} and $\mu_{i'j'}$ can be replaced by

$$\sum_m \sum_n ||x_{i+m,j+n} - \mu_{i'+m,j'+n}||^2, \tag{6.41}$$

where the m and n parameters scan the surface of the sub images. Figure 6.4 shows how the IDM works. The figure consists of an image frame and three image pairs. Each image pair includes the transformed image that results from the IDM distance calculation and the distorted image. The difference images between the transformed image or the distorted image and the original image are shown in the second row. The distorted image frames are created artificially by one pixel displacement of the left hand of the signer.

Combination of TD and IDM

As explained before, each sign language word can be signed with some small visual differences. In other words, it is possible to have different image frames in two utterances of the same class where the image frames are similar in most of the parts but there exist some small local differences in size, position, and orientation of the hands and the head of the signer. To compensate for these small local variations, we combine the tangent distance and the image distortion model, which are able to compensate for global affine transformations and local displacements, respectively. The combination of these two distances makes the classifier invariant to the combination of these two kinds of distortions. Here we introduce two methods to combine the proposed distortion model and the tangent distance [Zahedi & Keysers$^+$ 05a].

Method one. In the proposed distortion model, only the displacement of the sub images is allowed. If we calculate the TD instead of the Euclidean distance between the sub images, then other transformations like axis and diagonal deformations, scaling, rotation and also sub-pixel transformations are considered. The image frames in which only the left hand of the signer is distorted by axis deformation, diagonal deformation and scaling of the left hand of the signer are shown in Figure 6.5. These distortions are tolerated by the proposed combination method and the transformed images which result from the combination method are also shown. The difference images show how the combination method tolerates the variations.

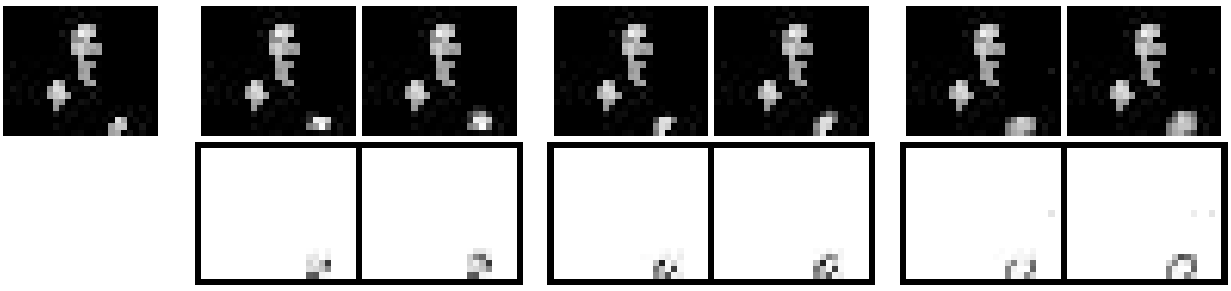


Figure 6.5: Example of the first combination method. (first row: original image and image pairs including the transformed image which results from the combination method and distorted image by axis deformation, diagonal deformation and scaling of the left hand of the signer, second row: the difference image between the transformed image or distorted image and the original image)

Method two. Another possible way to combine these two invariant distances is the use of the TD before employing the image distortion model to find the closest image frames in the manifolds. In this work, the one-sided tangent distance using the tangent vectors of the mean image μ is employed. The closest image frame in the manifold $\mathcal{M}_\mu$ to the observation image frame x_t is calculated by $\hat{\mu} = \mu + \sum_{l=1}^{L} \hat{\beta}_l \mu_l$ where

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ ||x_t - (\mu + \sum_{l=1}^{L} \beta_l \mu_l)||^2 \right\}, \qquad \beta \in \mathbb{R}^L. \tag{6.42}$$

The $\hat{\mu}$ is calculated by compensating for global affine transformations. In this combination method, we replace the μ with $\hat{\mu}$ in the proposed IDM distance. This combination method results in a distance function which is invariant to global transformations modeled by the tangent distance, and to local displacements modeled by the IDM.

There is another way to combine the tangent distance and the image distortion model by employing the closest image frame of the manifold $\mathcal{M}_\mu$ in the first combination method. Due to the compensation for affine transformations in the sub images and global transformations, this method is extensively time consuming and is ignored in the experiments.

6.4 Pronunciation clustering

Sign language, like spoken languages, is created and developed in the communities of deaf or hard-hearing people and families who have one or more deaf members. They use and shape it. Therefore different communities including different sub-communities based on regional, educational and professional backgrounds, age and sex of the members create a variety of different grammars, vocabularies and pronunciations for sign languages.

The variety of pronunciations of sign language words is visible in the databases which have been introduced in Chapter 4. Figure 6.6 shows four different pronunciations for the American sign language word “GO”. This difference between row b and c results from the direction of the signing, showing who goes into the witch direction or to whom. Although signing d is very similar to b and c signings with small differences of only one hand, the signing a is completely different to the others but expresses the same meaning. Due to the high variability of utterances for each word in the RWTH-BOSTON-50 database, we investigate

Figure 6.6: Four different pronunciation of the American sign language word "GO" from the RWTH-BOSTON-50 database; in each row the image frames of the video sequence signing the word "GO" is presented.

how to consider different pronunciations for utterances of each word which influences isolated sign language word recognition. Note that this approach involves a tradeoff; while we may be able to better model the different pronunciations when we use separate HMMs, we are left with fewer data to estimate the HMMs. We employ and compare three methods of clustering to determine the partitioning into clusters [Zahedi & Keysers$^+$ 05c].

6.4.1 Manual partitioning

As one can see in Figure 6.6, we observed that there are some large visual differences between the utterances of each word in the RWTH-BOSTON-50 database. These differences are visually distinguishable. Thus, we are able to label the utterances of different pronunciations for each word as a baseline. We separated the 483 utterances of the RWTH-BOSTON-50 database to 83 pronunciations for the 50 words. The results obtained using this method serve as a lower bound for the automatic methods described in the following because we cannot hope to obtain a better cluster structure. Obviously, for any larger task it will not be feasible to perform a manual labelling. Furthermore, manual partitioning can be done only for the utterances with a large visual variability and it is not possible to do while the difference

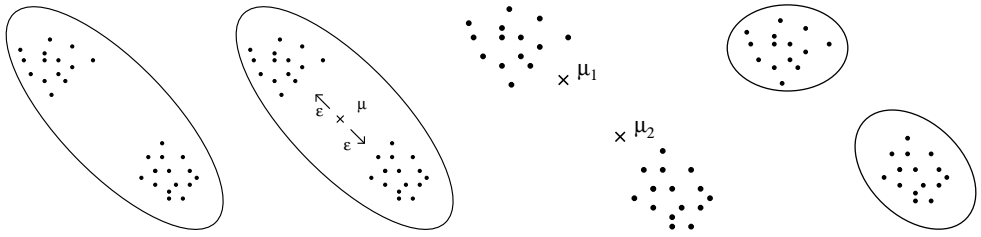


Figure 6.7: The LBG clustering.

between two utterances is not large enough to be noticed for the observer.

6.4.2 K-means clustering

One basic but very popular clustering approach is the k-means clustering method. In this method the number of clusters is assumed to be known beforehand and equal to k. We choose one utterance of each of the clusters that were labeled manually as a seed in the initialization. The algorithm continues by adding other utterances to the cluster.

In this algorithm for all words of the database: after initializing k (number of the clusters) and calculating the μ_i regarding as the mean of the Gaussian function made by the utterances of each cluster, all samples would be classified to the nearest cluster. This would be repeated until no change happens in the clusters.

K-means clustering is an intermediate level between manual clustering and full-automate pronunciation clustering methods, i.e. we need to initialize only the number of clusters and the seed of the clusters and the clustering continues automatically. This method can be employed for the databases which are too big to separate the utterances of different clusters manually.

6.4.3 LBG clustering

The k-means clustering still uses some manually extracted information, i.e. the number of clusters and the initializing seeds of the clusters. We employ the LBG clustering algorithm proposed by [Linde & Buzo$^+$ 80] to overcome this constraint and to obtain a fully automatic clustering algorithm. This method is described as follows: We perform the clustering for all words of the database as it is shown in Figure 6.7. First, we assume that all utterances belong to one cluster or to one particular pronunciation and create an HMM with all utterances existing for a word. If the criterion for dividing a cluster is met, we divide this HMM into two new cluster centers by adding or subtracting a small value to all means of the states in the model. Then we calculate the similarity between all possible pairs of cluster centers for the word and merge them if the criterion for merging is met. We continue to divide and merge the clusters until no change in the cluster assignment occurs.

If the utterances ${x_u}_1^{T_u}$ with the frames $(x_{u1}, ..., x_{ut}, ..., x_{uT_u})$ and $u = 1, ..., u, ..., U$ belong to a cluster or to two clusters that should be divided or merged, respectively, the criterion function is defined by:

$$J^* = \sqrt{\frac{\sum_{u=1}^{U} d({x_u}_1^{T_u}, (\mu, \sigma))^2}{U}} \tag{6.43}$$

where (μ, σ) is the mean/variance model for cluster center respecting ${x_1}_1^{T_1}$, ${x_2}_1^{T_2}$, ..., ${x_U}_1^{T_U}$.

The criterion function is defined to calculate the dispersion or the scattering of the utterances in a cluster. We use the mean squared distance of the utterances to the mean model as a measure of scatter and normalize that value to the range $[0, 1]$. We consider a threshold value for this criterion function to control the coarseness of the clustering.

It is necessary to mention that these three clustering methods are performed before the training of HMMs starts, i.e. the utterances of each class are clustered by pronunciation clustering methods and then the utterances of each pronunciations are used to train the HMM for the corresponding pronunciation of the word.

6.5 Dimensionality reduction

The appearance-based features, particularly the down-scaled image frames include so many feature elements creating a feature vector with high number of dimensionality. For instance, if we use an image frame which is down-scaled to 32×32 pixels, i.e. a feature vector with the size of 1024 elements, $1024 \cdot (1024/2) = 524288$ parameters need to be estimated. This process requires a huge amount of time and memory to be performed for all sentences including sign language words and nearly 100 image frames for each sentence. To cope with the problem of the high dimensionality, as it is explained before, we do a global pooling, i.e. we estimate only a single covariance matrix Σ for all densities.

Dimensionality reduction of the feature vectors is another way which is used to overcome the problem with the estimation of the covariance matrices. Feature reduction aims at selecting the most discriminative information of the feature vectors usually by means of a linear transformation of the feature space.

Although feature reduction always involves a loss of information, in practical applications, this loss introduced by feature reduction is often compensated by a more reliable parameter estimation in the reduced feature space. As scarceness of the training data in sign language recognition is a very important shortcoming for automatic sign language recognition systems, the classifier generalizes better when fewer features have to be trained.

In the following, we are going to explain the two feature reduction methods which are employed in this work to reduce the size of the appearance-based features and the geometric features of the signers' dominant hand. These are the two principle methods that are frequently used in most of the statistical pattern recognition approaches.

6.5.1 Principal component analysis

The principal component analysis (PCA), also called Karhunen-Loève is a linear transformation which aims at the reduction of the feature space and minimizing the representation error. It is very important that no class information is used for the PCA transformation. Therefore, although it is used in several pattern recognition tasks, there is nothing to say about the discriminative power of the resulting feature vectors. In other words, the PCA transforms all feature elements irrespective of their class identity.

The covariance matrix Σ can become diagonalized by using an eigenvector decomposi-

tion with the eigenvectors $v_1, \ldots, v_D$ and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_D$:

$$\begin{aligned} \Sigma &= \sum_{d=1}^{D} \lambda_d v_d v_d^T \\ &= [v_1 \cdots v_D] \operatorname{diag}(\lambda_1, \ldots, \lambda_D) [v_1 \cdots v_D]^T \end{aligned} \tag{6.44}$$

where $\lambda_d \geq \lambda_{d+1}, d = 1, \ldots, D-1$, i.e. the eigenvalues are sorted in decreasing order. The d number of the largest eigenvalues determine the corresponding eigenvectors $v_1, \ldots, v_d$ as the principal components. The PCA is a linear transformation $x \in \mathbb{R}^D \mapsto \hat{x} \in \mathbb{R}^d$ and maps each vector onto the principal components. The resulting matrix representation of the transformation is $[v_1 \cdots v_d]$. Furthermore it posses the property that the expected squared error $E\{||x - \hat{x}||^2\}$ is the smallest within all linear transformations to the d dimensions [Duda & Hart$^+$ 01].

The PCA transformation can be employed for two different purposes. First, in some pattern recognition applications, a PCA transformation is used to remove the first few eigenvectors. For example the first three eigenvectors are discarded in [Martinez & Kak 01]. Since the first large eigenvectors may capture variability in the data that is not relevant for classification, i.e. this information is common for all classes of the training set. For instance, in many image classification approaches, the first eigenvector corresponds to the brightness variability of the images which does not concern the classification of the images and it is good to reduce the impact of the brightness changes on the feature vectors by removing the first few eigenvectors.

Second, as the PCA discards the directions of small variances, it is expected that the transformation captures the most relevant part of the information contained in the vectors x. This point of view is based on magnitude of the variance and minimal reconstruction error. Therefore it may not be suitable for the classification purposes, since it does not take the class information into account. However, when the background of all image frames of a video stream does not change so much and the pixels belonging to the background have a small variance, this property makes the PCA well suited for the analysis of the video frames. If we select the first eigenvectors corresponding to the largest eigenvalues, the background pixels will be discarded by the PCA transformation. The PCA and the whitening transformation are explained in very detail in [Fukunaga 90].

6.5.2 Linear discriminant analysis

The linear discriminant analysis (LDA) which is also called Fisher's LDA aims at providing more class separability within the feature reduction process into the transformed feature space. It takes the class information into account to draw a decision region between the classes[Duda & Hart 73, pp. 118ff]. Therefore it is expected to have advantages over the PCA regarding some applications.

In LDA, we use within-class and between-class scatter to formulate the criteria of the class separability. The within-class scatter is the expected covariance of each class and the between-class scatter is the covariance of the data set including the mean vector of each class. As the class conditional distributions are implemented by Gaussian densities, the LDA tries to simultaneously maximize the distances between the class centers μ_k and to minimize the distances within each class. This can be achieved by maximizing the criterion in the LDA which is the ratio of the between-class scatter to the within-class scatter. This optimization problem leads us to a generalized eigenvalue problem.

The within-class-scatter matrix S_w and the between-class-scatter matrix S_b are calculated by:

$$S_w = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (x_{kn} - \mu_k) \cdot (x_{kn} - \mu_k)^T \tag{6.45}$$

$$S_b = \sum_{k=1}^{K} N_k \cdot (\mu_k - \mu) \cdot (\mu_k - \mu)^T \tag{6.46}$$

where there are K classes and each class k includes N_k members. To optimize the criterion we compute the eigenvectors and eigenvalues of the matrix $S_w^{-1} \cdot S_b$. Then the first d principal components of $S_w^{-1} \cdot S_b$ are computed for the projection of the data onto the new subspace. In [Duda & Hart 73], it is explained how the LDA can be performed without inversion of S_w by solving a generalized eigenvalue problem in S_w and S_b .

6.6 Combination methods

Sign language includes movements of different parts of the body which are used to convey the whole meaning of the signer. We extract two different kinds of features from the image frames where both feature groups include the different information of the signings. To use different aspect of signs, we combine the feature groups which have been defined in Section 5. As combination of the features is successfully performed in the field of automatic speech recognition [Zolnay & Schlueter$^+$ 05], we combine the features in two levels by employing three techniques. At the feature level, combination can be performed by the feature concatenation and weighting the feature groups or concatenation of the features over the time and using LDA to choose the most discriminant elements. Furthermore a log-linear model combination can be carried out at the model level. These are three combination techniques which are investigated in the following sections.

6.6.1 Feature weighting

The different features which are extracted from an image frame or from different image frames which are recorded simultaneously from the signer can be concatenated to compose a larger feature vector:

$$x_t = \begin{bmatrix} x_t^{(1)} \\ \cdots \\ x_t^{(F)} \end{bmatrix} \tag{6.47}$$

where every $x_t^{(i)}$ is a feature vector which is extracted from the front or the side camera at the time t. It can consist of geometric features of the dominant or the non-dominant hand of the signer, or the facial expressions like lip or eyebrow movements.

Also to emphasis each group of the features, the feature groups $x_t^{(i)}$ can be weighted by α_i and the visual model changes to:

$$Pr(x_t|s_t, w) = \sum_{i=1}^{F} \alpha_i \cdot Pr(x_t^{(i)}|s_t, w), \quad \sum_{i=1}^{F} \alpha_i = 1. \tag{6.48}$$

6.6.2 LDA-based feature combination

The LDA-based feature combination is used successfully to carry out an optimal linear combination of successive vectors of a single feature stream for an automatic speech recognition system [Haeb-Umbach & Ney 92]. In this approach, the feature vectors extracted by different algorithms $x_t^{(i)}$ are concatenated for all the time frames t. Then the successor and predecessor feature vectors of the current one at the time t can be concatenated to make a large feature vector which uses the context information of the visual model:

$$Y_t = \begin{bmatrix} \begin{bmatrix} x_{t-\delta}^{(1)} \\ \cdots \\ x_{t-\delta}^{(F)} \end{bmatrix} \\ \cdots \\ \begin{bmatrix} x_t^{(1)} \\ \cdots \\ x_t^{(F)} \end{bmatrix} \\ \cdots \\ \begin{bmatrix} x_{t+\delta}^{(1)} \\ \cdots \\ x_{t+\delta}^{(F)} \end{bmatrix} \end{bmatrix} \tag{6.49}$$

where Y_t is a feature vector at the time t including the features extracted from the current image frame and also from the successors and predecessors. If we consider a window with the size of $2\delta + 1$, i.e. δ successive, δ predecessive feature vectors and a current feature vector, then the resulting composite feature vector is too big. A linear discriminant analysis (LDA) based approach, selecting the most discriminative classification informations, reduces the size of the feature vector.

$$y_t = \begin{bmatrix} V^T \end{bmatrix} Y_t \tag{6.50}$$

where the LDA determines the matrix V to transfer the most discriminative classification information of Y_t to the feature vector y_t. The final feature vector is used as well in training and as in recognition.

6.6.3 Log-linear model combination

The log-linear model combination is carried out in the evaluation process, while the visual models are already trained separately by using different features extracted from the input video stream. This approach is also used successfully for speech recognition in [Beyerlein 98, Tolba & Selouani$^+$ 02] and the employ out of the log-linear combination has led to a significant improvement in WER.

As it is explained before, for the standard form of the Bayes' decision rule, while given x_1^T as a sequence of feature vectors extracted from the image frames, the best sequence of words w_1^N is chosen by maximizing the posterior probability of $Pr(w_1^N|x_1^T)$:

$$x_1^T \longrightarrow r(x_1^T) = \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N|x_1^T) \right\} \tag{6.51}$$

$$= \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N) \cdot Pr(x_1^T|w_1^N) \right\}, \tag{6.52}$$

where the posterior probability is decomposed into the language model probability $Pr(w_1^N)$ and the visual model probability $Pr(x_1^T|w_1^N)$.

In order to combine the different visual features, the visual model probabilities $Pr_i(x_1^{T^{(i)}}|w_1^N)$ are trained separately by using the sequence of the feature vectors $x_1^{T^{(i)}}$ which are extracted by the ith algorithm from the sequence of image frames. Then employing the log-linear combination of the visual probabilities, the posterior probability has the following form to recognize word sequence of $\hat{w}_1^N$.

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \left\{ Pr(w_1^N)^{\lambda_{\mathrm{LM}}} \prod_i Pr_i(x_1^{T^{(i)}}|w_1^N)^{\lambda_i} \right\}. \tag{6.53}$$

where λ_{LM} and λ_i are the language model weight and the visual model weights for the different groups of the features. The model weights have been optimized empirically in the development process. The language model $Pr(w_1^N)$ does not differ for the different features and is trained like before by using a sequence of written texts. The visual model probabilities $Pr_i(x_1^{T^{(i)}}|w_1^N)$ are trained by employing a standard maximum likelihood training to estimate the visual model parameters.

7 Experimental Results

I had seen so much of the world that I was tired of it.
– Razi (*865–925*)

In this chapter we are going to present the results of the experiments which have been performed in this work to investigate the different aspects of appearance-based sign language recognition.

In each part, regarding the goal of the specific experiment, the experiments are either performed on some or on all databases. The experiments for single word sign language recognition have been performed on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50. These two databases are the databases containing the utterances of 10 and 50 manually segmented American sign language words, respectively. For sentence sign language recognition, all experiments have been performed on the RWTH-BOSTON-104 database which is a database with a large enough number of utterances for most of the occurring sign language words in the database. Although the RWTH-ECHO databases are ready for more experiments, the preliminary results show that the utterances of each sign language word is not enough to investigate other aspects of sign language recognition on them.

7.1 Baseline results

Since the hidden Markov model is used successfully for automatic speech recognition, we employ the HMMs to recognize sign language words and sentences. In this section, we are going to present the baseline results (Table 7.1), by using simple appearance-based features on the different databases as a starting point. A sequence of the down-scaled image frames is used as feature vectors. The aim of this section is to provide an overview of the corpora used in the upcoming experiments and to give an impression of the difficulty of the individual tasks.

The choices made for the parameters of these experiments are justified in later experiments when the effect of these choices are investigated and discussed in detail. These parameters include the size of the down-scaled image frames which are used as feature vectors and the Gaussian mixture densities. In these experiments the image features are extracted from

Table 7.1: Baseline results on the databases which are introduced in Chapter 4.

Data set	WER[%]
RWTH-BOSTON-10	17
RWTH-BOSTON-50	37
RWTH-BOSTON-104	59
RWTH-ECHO-BSL	94
RWTH-ECHO-SSL	86
RWTH-ECHO-NGT	91

Table 7.2: Preliminary result on the RWTH-BOSTON-10 database using different length of HMMs and pooling of the variances [Zahedi & Keysers+ 05b].

	Pooling		
HMM size	Word-dependent	State-dependent	No pooling
Minimum seq. length	**7**	8	**7**
Average seq. length	14	15	17

the recorded image frames which are scaled down to 13×11 pixels for the RWTH-BOSTON-10 and the RWTH-BOSTON-50 and which are scaled down to 32×32 for the remaining databases which are prepared for continuous sign language recognition.

The baseline results are achieved by using very simple feature vectors without employing advanced methods of image processing or training of HMMs. Therefore it is expected to obtain better recognition rates by improving the recognition system and the extracted features. In this section we are going to show the influence of several various standard methods commonly applied in ASR to the task of ASLR for example the influence of different HMM parameters and the influence of a language model.

7.1.1 Influence of HMM parameters

To optimize the HMMs parameters, we choose the down-sampled original image after performing the skin intensity thresholding and employ the HMM classifier to classify words of the RWTH-BOSTON-10 and RWTH-BOSTON-50 databases. The results of this classification using the Gaussian distribution with different sequence lengths and pooling are shown in Table 7.2 and 7.3. By using word-dependent pooling better results are achieved than by using state-dependent pooling or density-dependent estimation of the variances. When using the Laplace distribution, the performance of the classifier is similar to these results but the Gaussian distribution performs better.

We employ an HMM of each word with the length of the minimum and the average sequence length of the training samples. As it is shown in Table 7.2 and 7.3, neglecting other parameters, the shorter HMMs give better results. This may be due to the small size of the database. There is a tradeoff for the length of the HMMs; If the HMM has fewer states, the parameters of the distribution functions will be estimated better. On the other hand, when using more states for the HMMs, there exist more parameters to model the transitions between the image frames. In informal experiments with shorter HMMs the accuracy of the classifier could not be improved. Therefore, we continue the experiments for single word recognition with the HMMs with the minimum sequence length of the training samples. Also a word-dependent pooling is concluded from the experimental results to be used for single word recognition.

The best error rate of 7% is obtained on the RWTH-BOSTON-10 database. Analyzing the errors of the classifier on the RWTH-BOSTON-10 database shows that the words which are wrongly classified are the singleton words without any similar utterance in the database. Therefore employing a leaving-one-out approach we cannot expect these words to be recognized correctly. As the RWTH-BOSTON-50 is a larger database which contains all data of the RWTH-BOSTON-10, for further study on the single word recognition, we perform the experiments only on the RWTH-BOSTON-50 database.

7.1.2 Language model

The language model which uses the information taking the syntax and the semantics of the sign language into account is expected to improve the recognition rate. Therefore we employ the different language models which have been introduced in Section 6.1.3 for the databases prepared for continuous sign language recognition. First, preliminary results on the RWTH-BOSTON-104 database are going to be shown. For the experiments, the video frames have been scaled down to the size of 32×32 pixels which has been reported to be a good size in many image recognition tasks and we will discuss it later in more detail. The performance of the system is measured by the word error rate (WER) which is equal to the number of deletions (del), substitutions (sub) and insertions (ins) of the words divided by the number of running words. The results on the development and the evaluation sets including the perplexity (PP) and WER of the system using different language models are shown in Table 7.4. The m-gram language models where the probability of a sentence is estimated from the conditional probabilities of each word while given the $m-1$ preceding words are employed in the experiments. The m-gram language models are called zerogram, unigram, bigram and trigram where m is equal to 0, 1, 2 or 3, respectively.

As expected using bigram and trigram language models with smaller perplexity helps the system to achieve a better recognition rate. When employing a zerogram language model with a high value of perplexity, due to a large number of substitutions and deletions, the error rate of the system is very high.

Also, preliminary results on the RWTH-ECHO databases with original gray valued images, cropped and down-scaled to 32×32 are shown in Table 7.5, 7.6 and 7.7.

The results on the RWTH-ECHO databases show that not only the training data is not good enough to construct a visual model, but also the sequences of sign language words are not good to construct a good language model. Therefore, it is not reasonable to do more experiments to investigate the other aspects of sign language recognition on the RWTH-ECHO

Table 7.3: Preliminary result on the RWTH-BOSTON-50 database using different length of HMMs and pooling of the variances.

	Pooling		
HMM size	Word-dependent	State-dependent	No pooling
Minimum seq. length	**32**	34	33
Average seq. length	37	36	37

Table 7.4: Preliminary results of the recognition system on the RWTH-BOSTON-104 employing different language models.

Language Model	Development set					Evaluation set				
	PP	WER[%]	del	ins	sub	PP	WER[%]	del	ins	sub
Zerogram	105	69	41	8	49	105	59	81	2	22
Unigram	36	67	55	2	38	37	58	67	4	32
Bigram	8	64	51	2	38	9	56	65	3	31
Trigram	7	64	52	2	37	6	54	61	6	30

Table 7.5: Preliminary results achieved on the RWTH-ECHO-BSL database with different language models [Zahedi & Dreuw+ 06b].

LM Type	WER[%]	del	ins	sub
Zerogram	94	111	12	102
Unigram	92	97	16	107
Bigram	91	88	17	114
Trigram	92	91	17	113

Table 7.6: Preliminary results achieved on the RWTH-ECHO-SSL database with different language models [Zahedi & Dreuw+ 06b].

LM Type	WER[%]	del	ins	sub
Zerogram	86	60	2	45
Unigram	85	52	6	47
Bigram	83	52	5	46
Trigram	82	51	5	46

Table 7.7: Preliminary results achieved on the RWTH-ECHO-NGT database with different kinds of language models [Zahedi & Dreuw+ 06b].

LM Type	WER[%]	del	ins	sub
Zerogram	91	101	4	70
Unigram	89	96	7	69
Bigram	89	83	6	83
Trigram	89	83	6	83

Table 7.8: Error rates [%] of the HMM classifier on the RWTH-BOSTON-50 database employing different clustering methods. The results are achieved by using the Euclidean distance and the tangent distance [Zahedi & Keysers+ 05c].

	Euclidean distance	Tangent distance
No clustering	28.4	27.7
Manual partitioning	22.8	**20.5**
K-means clustering	23.8	21.3
LBG clustering	23.2	21.5
Nearest neighbor	23.6	22.2

databases. Thus we are going to continue the experiments for continuous sign language recognition only on the RWTH-BOSTON-104 database containing enough training data for most of the occurring words in the data set.

7.2 Pronunciation clustering

Before starting to study the features and techniques which are used in this work to build a robust appearance-based sign language recognition system, due to the variety of pronunciations of sign language words, we perform some experiments on the RWTH-BOSTON-50 database. Although the RWTH-BOSTON-50 is a database containing 483 utterances of 50 American sign language words, one can see nearly 80 different pronunciations occurring in the database for these 50 sign language words. The experiments have been started by employing an HMM for each word of the RWTH-BOSTON-50 database resulting in an error rate of 28.4% with the Euclidean distance. We repeated the experiment using the different proposed clustering methods and the tangent distance.

The results are summarized in Table 7.8. The results show that in all experiments, the tangent distance improves the error rate of the classifiers by approximately 2 to 10 percent. Furthermore, employing clustering methods and the nearest neighbor classifier yield a lower error rate than obtained without considering the different pronunciations.

The error rate of the classifier using LBG clustering with respect to the threshold value is shown in Fig. 7.1. The threshold value used in LBG clustering is a normalized value. When the threshold value is set to one, no clustering occurs, and when it is set to zero each utterance will form a separate cluster and the classifier converges to a system which is very similar to a nearest neighbor classifier. We can observe that, with a threshold value of 1, no clustering happens and the error rate is equal to the error rate of the classifier without any pronunciation modeling. When the threshold value is decreased, the error rate is reduced and we can achieve the best error rate of 23.2% and 21.5% using the Euclidean distance and the tangent distance, respectively. The fluctuations can be observed in the diagram for the threshold values between 0 and 0.4 which lead us to the conclusion that the determination of the best threshold value is not very reliable. Nevertheless, we can observe that there is a strong trend of reducing error rates for the smaller threshold values. This leads us to consider the nearest neighbor classifier, which corresponds to the threshold value zero and achieves error rates of 23.6% and 22.2% with the Euclidean distance and the tangent distance, respectively. Since these values are only slightly worse than the best –but unstable–

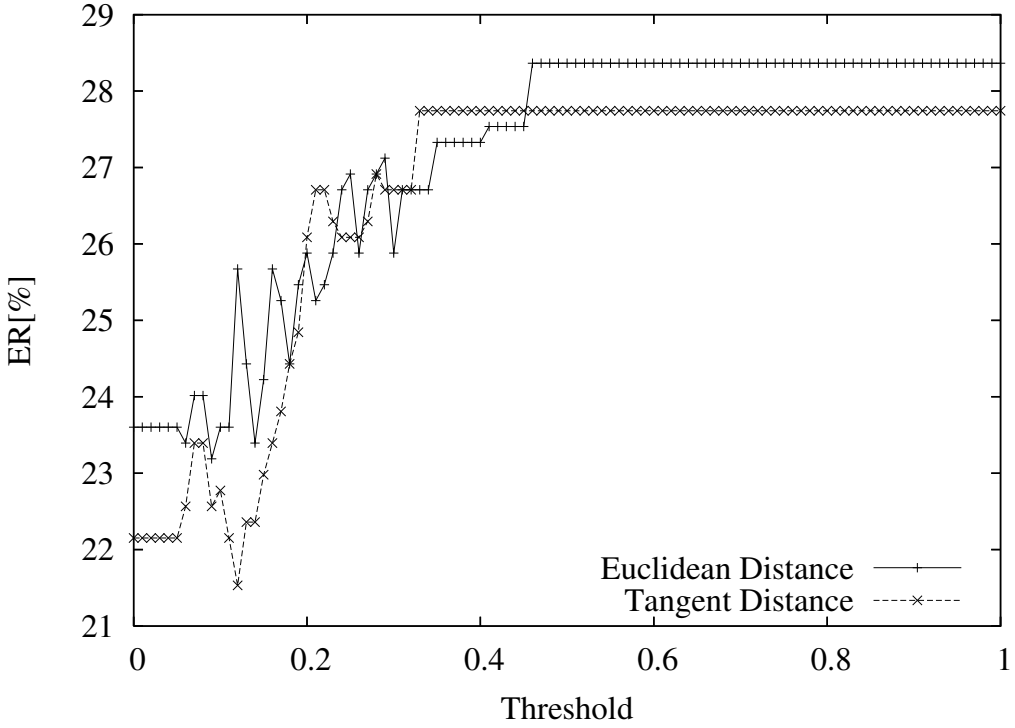


Figure 7.1: Error rate of the recognition system with respect to the threshold value of the LBG clustering. The results are obtained by employing two different distances [Zahedi & Keysers+ 05c].

results for the LBG clustering, thus this approach should be considered for tasks with a large variability of utterances.

The best error rate of 20.5% is achieved by using manual clustering and by using the tangent distance but the results which have been achieved by using other clustering methods will be preferable for large databases since they do not involve human labeling of the video sequences. The best pronunciation clustering method without human intervention is the hierarchical LBG clustering with the tangent distance which achieved an error rate of 21.5%, which is an improvement of over 22 percent relative.

In the experiments reported above, mixture densities with a maximum number of five densities are used in each state. We have repeated the experiments employing single density and mixture densities, consisting of more densities, in the states of the HMMs. Table 7.9 shows the results of the experiments employing the tangent distance and the different clustering methods. The results show that using a higher number of densities within a mixture density improves the accuracy of the system. In other words, the mixture densities can model the variability of the utterances even without employing the clustering methods. The error rate of the system without any clustering methods is 22.8%. In most experiments, the better results have been achieved when the mixture densities have been used in the states. When mixture densities are used, the influence of the different clustering methods on the error rate of the system is much less than single density experiments.

According to the results performed for the modelling of pronunciations employing three kinds of the clustering methods which are: (a) manual clustering, (b) k-means clustering and (c) hierarchical LBG clustering. One of these methods can be chosen according to the size

Table 7.9: Error rates [%] of the HMM classifier on the RWTH-BOSTON-50 database, employing different clustering methods. The presented results are obtained by using single and mixture densities.

	Single density	Mixture density
No clustering	47.4	22.8
Manual partitioning	35.4	21.9
K-means clustering	33.1	21.1
LBG clustering	21.7	22.1

of the database in different applications. Although manual clustering gives more accuracy in most of the experiments, it needs manually extracted information and can therefore only be employed for small sets of data. The k-means clustering needs less initial information and only needs to be initialized with the number of clusters and manually selected seed utterances, so this method is also suitable for medium size databases. In contrast, the LBG clustering method partitions the data automatically and is preferable for large databases where extracting labels manually is unfeasible. According to the results of the experiments on the RWTH-BOSTON-50 database, LBG clustering leads us to use the nearest neighbor classifier which performs surprisingly well. In all experiments, the tangent distance was compared to the Euclidean distance within the Gaussian emission densities. By using the tangent distance which models small global affine transformations of the images improves the accuracy of the classifier significantly. In the Section 7.4, we are going to do more experiments to model the visual variability of the utterances by using invariant distances.

7.3 Appearance-based features

The appearance-based features which are extracted directly from the image frames are the most significant advantage of this work comparing to the other approaches. In this section, we are going to investigate the different aspects of these features like image resolution, temporal derivative features and gradient images.

7.3.1 Image resolution

The original image frames recorded by the cameras are the simplest appearance-based features which can be used for the recognition of sign language. However, the original image frame includes all image pixels which is a large amount of data. A large feature vector causes the dimensionality problem which has been discussed before in detail. Furthermore, there is not enough training data to train the feature vectors which contain so many feature elements. Although the down-scaling of the image frames causes a loss of some information of the whole image, it helps by decreasing the number of elements of the feature vectors to have more information to construct the visual model for each pixel of feature vector. Therefore image resolution is a critical issue for using appearance-based features and finding the optimum size of the down-scaled image frames which makes the feature vector small enough without increasing the error rate is important as well. We are going to perform some experiments on the RWTH-BOSTON-10 and the RWTH-BOSTON-50 databases which are the databases prepared for segmented sign language recognition and also on the RWTH-BOSTON-104 for

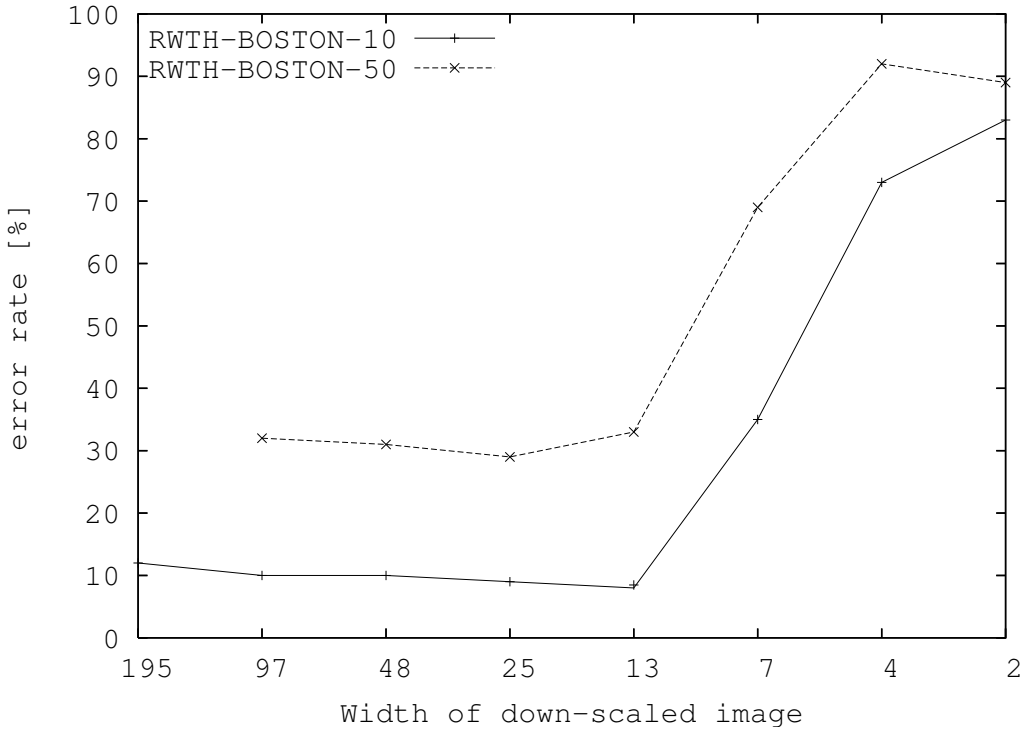


Figure 7.2: Error rate of the system on the databases of segmented words by using the leaving-one-out method and varying the image scales.

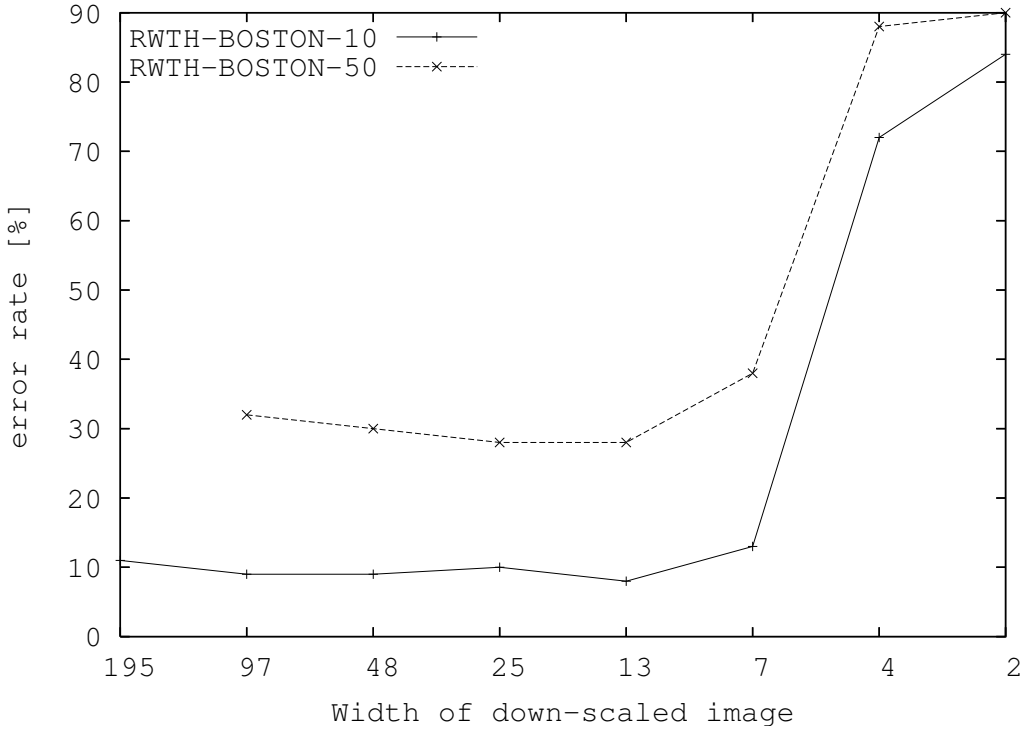


Figure 7.3: Error rate of the system on the databases of segmented words by using the nearest neighbor classifier and the down-scaled image frames.

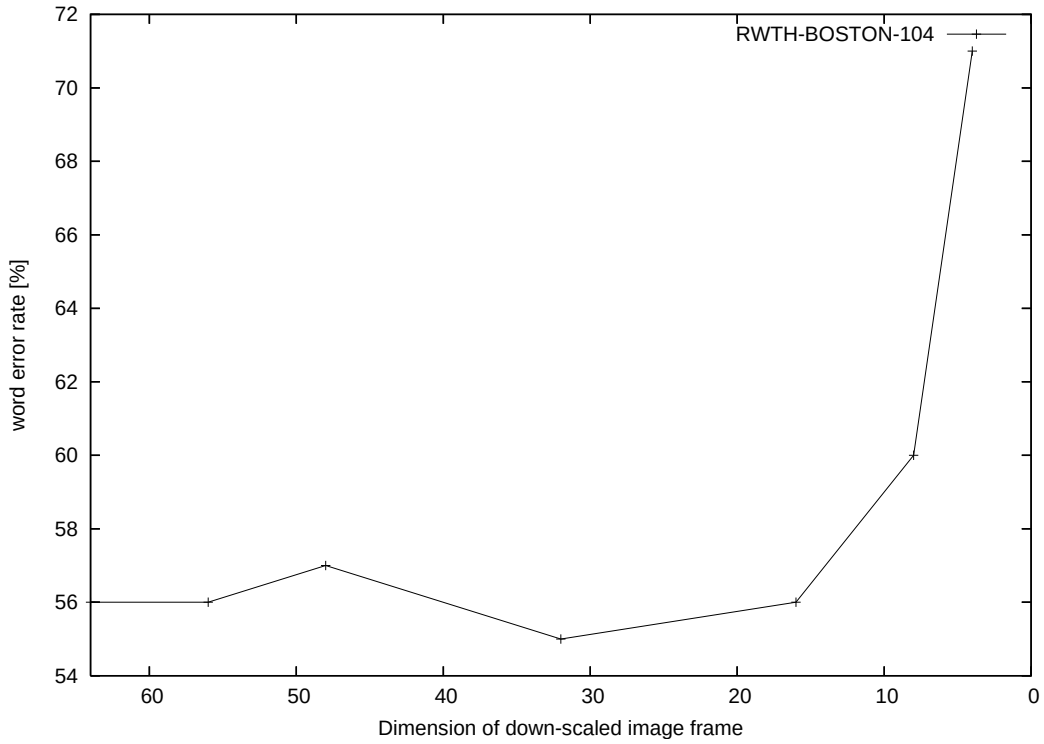


Figure 7.4: WER [%] of the system on the RWTH-BOSTON-104 using the down-scaled original image.

continuous sign language recognition to observe the influence of down-scaling of the image frames.

The error rate of the recognition system employing the leaving-one-out method and the nearest neighbor classifier on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50 databases is shown in Figures 7.2 and 7.3. The down-scaled image frames of the front camera are used as a feature vector and the error rates are presented with respect to the width of the down-scaled images.

Also the word error rate of the recognition system with respect to the width of the down-scaled image frames on the RWTH-BOSTON-104 database is shown in Figure 7.4.

One can see by using small feature vectors with the width of seven, four or two which causes a loss of so much information of the original images which therefore yields higher error rates. On the other hand, by using very large feature vectors like the whole image frame with the width of 195, 97, or 48 pixels does not help the classifier to achieve better results. According to the experimental results which are shown in the Figures 7.2 and 7.3, the down-scaling of the image frames to the size of 13×11 pixels for the RWTH-BOSTON-10 and the RWTH-BOSTON-50 database and 32×32 pixels for the RWTH-BOSTON-104 is reasonable. Since the RWTH-BOSTON-104 database contains more training data and consequently includes more image frames in the training set it may help the classifier to estimate the parameters of the Gaussian distributions for larger feature vectors.

7.3.2 Lateral camera

The RWTH-BOSTON databases contain image frames from a camera installed in front of a signer and also another camera positioned aside of the signer 4.3. To investigate the affect of using image frames from these two cameras, we have performed some experiments on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50 database. The experiments use original image frames which are scaled down to 13×11 pixels as feature vectors and the image frames from the front and from the side camera are concatenated to make the feature vector. In the Figures 7.5 [Zahedi & Keysers+ 05b] and 7.6 the error rate of the system on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50 database using the image frames from a camera fixed in front of the signer and another camera aside is shown. The error rate is shown with respect to the weight of the cameras. On the left hand side where the weight is equal to zero only the image frames of the front camera are used and the features of the side camera are weighted by zero. In contrary on the right hand side the weights of the front camera is set to zero.

The experiments are performed by using different topology for HMMs with a minimum and an average sequence length of training utterances for each sign language word and the error rates are shown in different curves.

Although the influence of using the lateral camera for the RWTH-BOSTON-10 database which is a very small database is not so clear, nevertheless it is clearly useful on the RWTH-BOSTON-50 database. In the experiments on both databases the minimum error rate occurs when the feature weights of the lateral camera and the front camera are set to 0.26 and 0.74, respectively. The error rate grows with increasing weight of the lateral camera. This result is probably caused by the occlusion of the hands. It means the side camera is not a good choice to be used alone to record the signings, but it can improve the recognition rate when we concatenate the image frames recorded by a side camera to the image frames of the front camera with a proper value of weighting.

7.3.3 Temporal derivative features

To investigate the results of the classifier using different appearance-based features we are going to perform the experiments using the original image, the different appearance-based features and the concatenation of the original image and the other features on the RWTH-BOSTON databases.

Since temporal derivative features lead to an improvement of automatic speech recognition systems, it inspires us using these kind of features including the information of the changes over the time for sign language recognition system as well. The error rate of a classifier employing the leaving-one-out method and using the different derivative features on the RWTH-BOSTON-10 and on the RWTH-BOSTON-50 is shown in Table 7.10. Also we have performed the experiments on the RWTH-BOSTON-104 database as well. According to the experimental results of the Section 7.3.1, the feature vectors are extracted from the image frames of the front camera and are scaled down to 13×11 pixels for the single word recognition and 32×32 for the continuous sign language recognition. The feature vectors are thresholded with the skin intensity. The error rates of the recognition system which use the temporal derivative features and the concatenation of the original image and the derivative features are shown in the table. As one can see, although using temporal derivative features with the image features yields better results than using them alone, the improvement

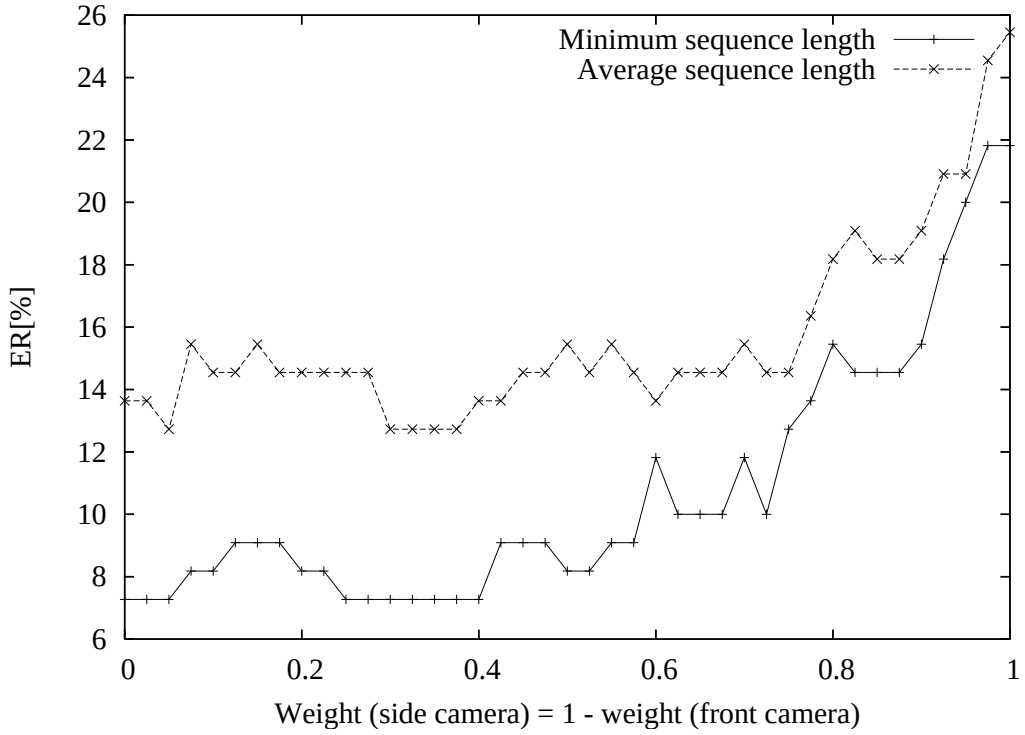


Figure 7.5: Error rate of the system with respect to the weight of the cameras (RWTH-BOSTON-10).

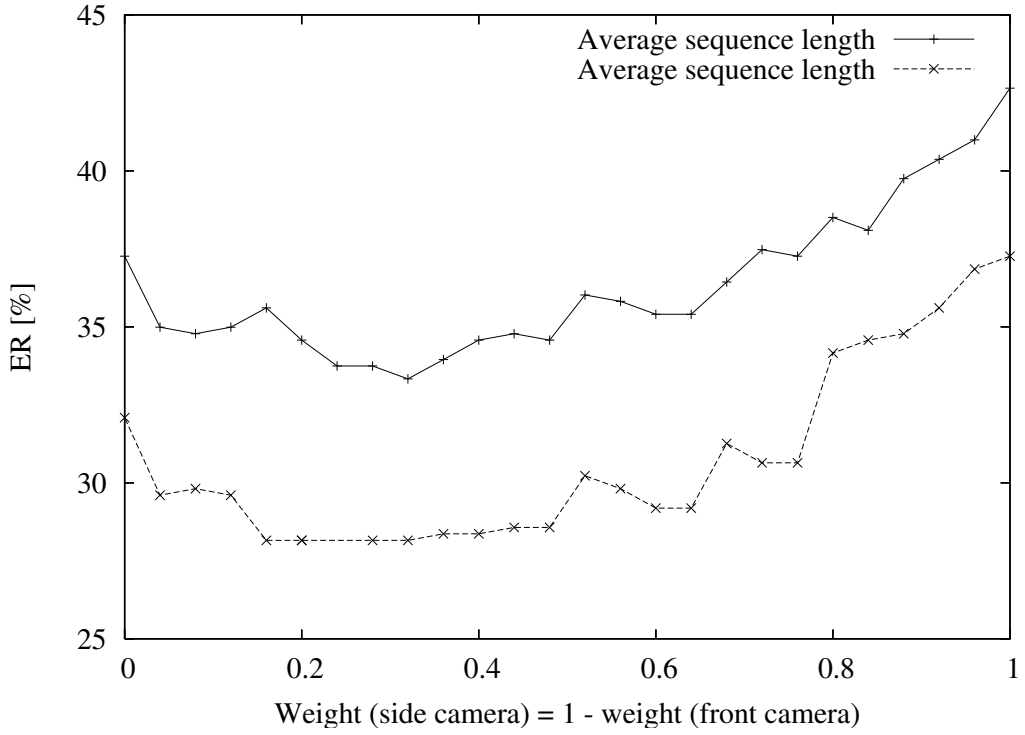


Figure 7.6: Error rate of the system with respect to the weight of the cameras (RWTH-BOSTON-50).

Table 7.10: Word error rate [%] of the recognition system on the RWTH-BOSTON databases using the skin intensity thresholded (SIT) image frame and the different temporal derivative features.

	RWTH-BOSTON-10		RWTH-BOSTON-50		RWTH-BOSTON-104	
Features	Features	+ SIT	Features	+ SIT	Features	+ SIT
Skin intensity threshold	7		32		54	
First derivative	18	10	47	32	58	55
Positive first derivative	27	9	53	32	52	55
Negative first derivative	31	10	56	32	60	53
Absolute first derivative	21	10	47	33	53	52
Second derivative	32	10	70	34	74	58

achieved by using the derivative features is not large enough to conclude that using temporal derivative features are more useful.

7.3.4 Gradient features

The gradient images including the information of the edges in the image frames are going to be investigated in this section. We perform the experiments by using the image features and the gradient images as the feature vector for the classifier. The error rate of the leaving-one-out method for single word recognition on the RWTH-BOSTON-10 and the RWTH-BOSTON-50 and for the continuous sign language recognition on the RWTH-BOSTON-104 database performed under the same conditions like the experiments in the previous section is shown in Table 7.11.

All gradient images used for the above experiments are obtained by employing standard gradient filters with the size of 3×3, 3×5, 5×3 or 5×5 which are introduced in the Section 5.1.5. Furthermore we have performed more experiments using larger gradient filters created by Gaussian functions and the word error rate of system is presented in the Tables 7.12, 7.13 and 7.14.

Table 7.11: Word error rate [%] of the recognition system on the RWTH-BOSTON databases using the skin intensity thresholded (SIT) image frame and the gradient features.

	RWTH-BOSTON-10		RWTH-BOSTON-50		RWTH-BOSTON-104	
Features	Features	+ SIT	Features	+ SIT	Features	+ SIT
Skin intensity threshold	7		32		54	
Diagonal Sobel filter	10	9	36	34	54	55
Horizontal Sobel filter	9	9	38	32	52	53
Vertical Sobel filter	12	9	36	35	58	55
Diagonal Jähne filter	10	10	33	34	52	55
Horizontal Jähne filter	8	9	36	32	50	52
Vertical Jähne filter	11	9	35	33	56	57
Diagonal PLUS filter	10	8	32	31	54	55
Horizontal PLUS filter	8	9	36	31	51	53
Vertical PLUS filter	10	8	38	34	56	56
Laplace filter	12	8	37	34	54	55

Table 7.12: Word error rate [%] of the classifier on the RWTH-BOSTON-104 database using gradient image features employing different size of Gaussian filters.

	3×3	5×5	9×9	15×15	21×21	27×27
Basic features	57	65	58	56	51	55
Basic features+SIT	54	52	53	55	57	54

Table 7.13: Word error rate [%] of the classifier on the RWTH-BOSTON-104 database using gradient image features employing larger size of horizontal gradient filters.

	5×9	7×13	11×21	15×29
Basic features	54	51	52	53
Basic features+SIT	55	53	53	53

Table 7.14: Word error rate [%] of the classifier on the RWTH-BOSTON-104 database using gradient image features employing a larger size of vertical gradient filters.

	9×5	13×7	21×11	29×15
Basic features	52	52	52	52
Basic features+SIT	55	52	54	57

In most of the experimental results which are presented above, the classifier which uses the image features, the temporal derivatives and the gradient images and also the concatenation of the original image and the other appearance-based features, have not been achieved better results than by using the original images alone as a feature vector. The small improvement which is obtained in some cases can happen due to a noise and is not reliable. Therefore we continue the experiments using the original image frames and we are going to look back once again to investigate the influence of using the gradient images and the temporal derivatives at the end when a well tuned system which uses the proposed methods is built.

7.4 Invariant distances

When using image features, we can model the visual variability of utterances of each word by employing invariant distances instead of the Euclidean distance. An invariant distance measure ideally takes the transformations of the patterns into account, yielding small values for the patterns which mostly differ by a transformation that does not change the class-membership. To model the variability of utterances, the tangent distance (TD) [Drucker & Schapire$^+$ 93, Keysers & Macherey$^+$ 04] and the image distortion model (IDM) [Keysers & Gollan$^+$ 04, Dreuw & Keysers$^+$ 05] can be used to account the global and the local variations, respectively.

In this section, we are going to perform experiments by employing the nearest neighbor classifier on the RWTH-BOSTON-50 database to show how invariant distances tolerate the visual variability of utterances to improve the recognition rate of the sign language recognition system [Zahedi & Keysers$^+$ 05a]. In these experiments we make use of the different proposed distances and the different local image context features. The results are summarized

Table 7.15: Error rates [%] of the classifier employing the TD, the IDM distance and the two combination methods. These invariant distances are used to calculate the distance between the image features and the mean images [Zahedi & Keysers+ 05a].

Distance	Original image	Horizontal gradient	Vertical gradient	Horizontal & vertical gradients
Euclidean distance	23.6	24.0	25.2	24.8
TD	22.2	22.8	23.4	21.3
IDM	21.9	21.5	24.6	23.4
IDM,TD (method 1)	**17.2**	18.8	18.2	18.4
IDM,TD (method 2)	20.3	21.1	21.5	20.9

in Table 7.15.

Not only TD and IDM decrease the error rate of the system comparing to the Euclidean distance, but also both of the combination methods of TD and IDM improve the accuracy of the classifier. The best error rates are achieved by using the first combination method of the TD and IDM in each column. This combination enables the classifier to model local distortions in the sub images of the image frames by calculating the tangent distance instead of the Euclidean distance between the sub-images of image distortion model.

More experiments are going to be performed to investigate how to weight the importance of the local sub images of the original image with respect to the gradient images. Figure 7.7 shows the error rate of the classifier using IDM and the two combination methods of the TD and IDM with respect to the relative weight of the original images. A relative weight of zero means that only gradient images are used. The graphs show that the best results are achieved by only using the local sub images of the original images. Using the first combination method gives the best error rate of the 17.2% on the RWTH-BOSTON-50 database which is an improvement of 27.1%. About 65% of the remaining misclassified utterances of the data are due to a strong visual difference from the other utterances in the database.

Each sign language word is signed with small visual differences regarding position, orientation or size of the hands and the head of the signer. We have presented how two different distances which are invariant with respect to the global affine transformations and the local displacements compensate for these variations. The combination of these two kinds of distances enables the nearest neighbor classifier based on HMMs to compensate for a combination of these two kinds of distortions. Two methods for combination of the invariant distances improve the accuracy of the nearest neighbor classifier significantly.

The tangent distance and the IDM distance are also employed on the RWTH-BOSTON-104 database. The word error rate of the recognition system using these two invariant distances is presented in Table 7.16. The best word error rate is achieved by employing the tangent distance using the gradient image frames obtained by employing a horizontal Sobel filter. Further experiments employing a larger size of gradient filters are employed to extract the gradient images but they have not improved the recognition rate.

7.5 Geometric features

Here, we are going to discuss the results from the experiments which are performed on the RWTH-BOSTON-50 and the RWTH-BOSTON-104 database using the described recognition

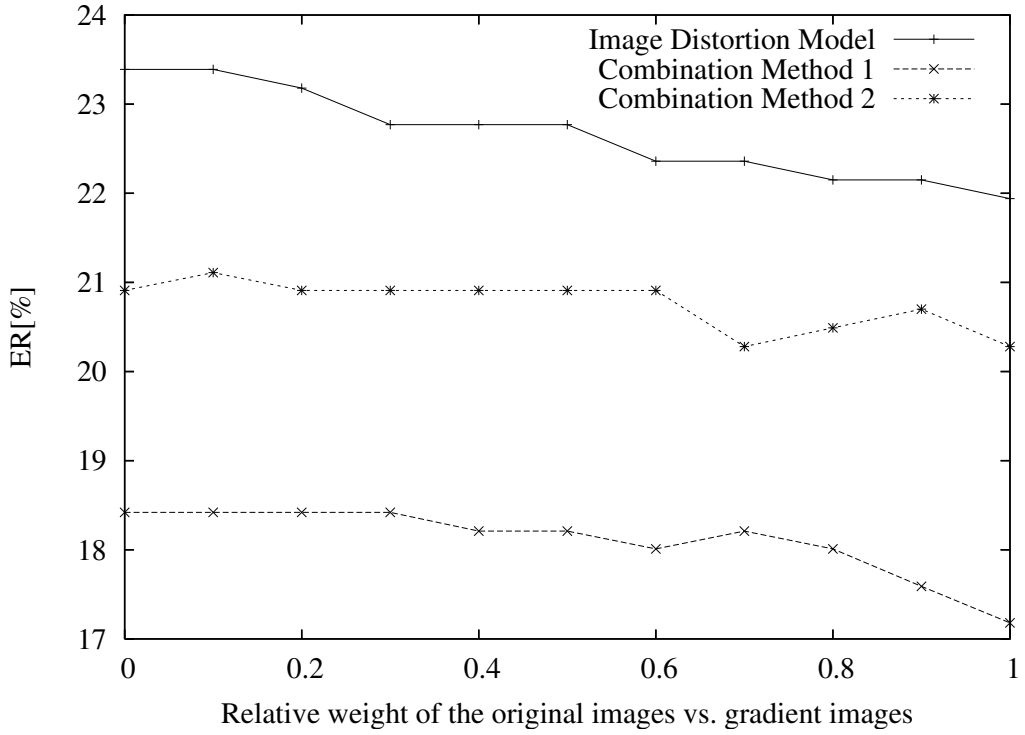


Figure 7.7: Error rate of the recognition system on the RWTH-BOSTON-50 database employing the IDM distance and the two combination methods by using the concatenation of the image features and the gradient features. The results are presented with respect to the gradient features weight [Zahedi & Keysers+ 05a].

Table 7.16: WER [%] of the classifier on the RWTH-BOSTON-104 with the TD and the IDM distance using image features and gradient features.

	TD		IDM	
Features	Development	Evaluation	Development	Evaluation
Original image	66	55	67	55
Horizontal gradient	59	46	60	42
Vertical gradient	59	48	64	56
Horizontal & vertical gradients	59	47	59	44

Table 7.17: Word error rate [%] of the recognition system on the RWTH-BOSTON-104 database using image features and geometric features of the signers' dominant hand [Zahedi & Dreuw+ 06a].

Features	Development set	Evaluation set
Image down-scaled to 32 × 32	64	54
Geometric features	61	**50**

Table 7.18: Word error rate [%] of the system on the RWTH-BOSTON-104 database with image features and geometric features extracted from the sequence of image frames plus interpolated image frames.

Features	Development set	Evaluation set
Image down-scaled to 32 × 32	69	60
Geometric features	59	**37**

framework with the geometric features (Section 5.2) extracted from the dominant hand of signer. The geometric features describing the hand shape and configuration of the dominant hand contacting the most meaningful information of the signings is expected to be useful as a feature vector for sign language recognition.

In Table 7.17 the word error rate (WER) of the recognition system using geometric features and image features which have been down-scaled to 32×32 pixels as baseline features on the development and the evaluation set are reported. It can be seen that geometric features alone slightly outperform better than appearance-based features for the development and for the evaluation set.

Comparing the sample rate of the video streams to the input data of the speech recognition systems inspires us to use more image frames by employing a linear interpolation method between the image frames (Section 5.2.1). In other words, We have performed the experiments which use more image frames to investigate the influence of scaling the time axis. The linear interpolation method creates the new interpolated image frames between the two sequential frames. Since the tracking method uses the motion information to track the dominant hand of the signer, we expect the tracking method works more precisely by using the new sequence of image frames including the sequence of image frames recorded by the camera and the interpolated frames. Consequently it leads us to the extraction of the geometric features which yield better results.

The word error rate of the recognition system on the RWTH-BOSTON-104 database in Table 7.18 shows although adding interpolated image frames does not decrease the error rate of the system when down-scaled image features are used, however it improves the system error rate when geometric features are used from an error rate of 50% to 37%.

The described experiments are performed on the RWTH-BOSTON-50 database which includes segmented sign language words as well. The experiments on this database yield high error rates of the recognition system. The visualization of the tracked dominant hand in the utterances of the database shows that the tracking method does not work properly for two-handed signs. Since in the RWTH-BOSTON-50 database the signs are segmented, concerning the two-handed signs there is no difference between the dominant and the non-dominant hand. For instance for some signs like "BOOK" both hands move completely symmetric and thus the tracking method is unable to detect the *correct* hand properly.

Since the geometric features rely on the tracking of the dominant hand, the wrong tracked hand causes meaningless geometric features which result in a high error rate.

On the contrary, in the RWTH-BOSTON-104 database the tracking method can cope with this problem because it includes complete sentences of sign language. The sentences commonly include at least one one-handed sign, and the trace-back of the tracking sequence is done at the end of the sequence. Therefore the tracker is able to capture the correct hand and to extract a meaningful path which in turn allows to extract meaningful geometric features.

7.6 Reduction of feature dimensionality

To cope with the problem of dimensionality which is explained in the Section 6.5 and also to select the most discriminative features of the image features and the geometric features we study how to employ feature reduction techniques for sign language recognition. Two kinds of feature reduction methods are employed to select more discriminative or relevant features from the appearance-based features and from the geometric features of signers' dominant hand. First, linear discriminant analysis (LDA) which takes the class membership information into account is employed for the feature reduction of both groups of the features. Then we are going to investigate how a principle component analysis can be employed to find the most relevant feature components of the image features by discarding the pixels with low variances from the image frames.

7.6.1 Linear discriminant analysis

In the following, we are going to illustrate the experiments which are performed to find the best setup for the dimensionality reduction using LDA for the image features and the geometric features individually. The results are given in Figure 7.8 and 7.9 for image features and geometric features, respectively. When using the original image features with a very small number of components, the smaller number of components yields larger error rates because the system looses the information of the image frames. However, when the feature vectors are too large, the error rate increases because too much data that is not relevant for the classification disturbs the recognition process.

Using the image features with the dimensionality of 90 the best error rate of 60% is obtained on the development set which leads the evaluation process to yield an error rate of 36%. Also we have employed the LDA using the image features of the recorded image frames plus the interpolated frames which has resulted in a high word error rate. For the geometric features extracted from the image frames plus the interpolated frames the best dimensionality is 20, giving the error rate of 57% and 29% on the development and the evaluation set, respectively. The experimental results of the feature reduction employing the LDA is summarized in Table 7.19. It shows that the LDA which takes the class membership into account is a powerful mean to select the most discriminative features.

7.6.2 Principle component analysis

Since principle component analysis discards the low variance pixels of the image frames, we employ it to transform the image features to a smaller feature vector expecting it to remove the background pixels. The experimental results of employing PCA on the image frames

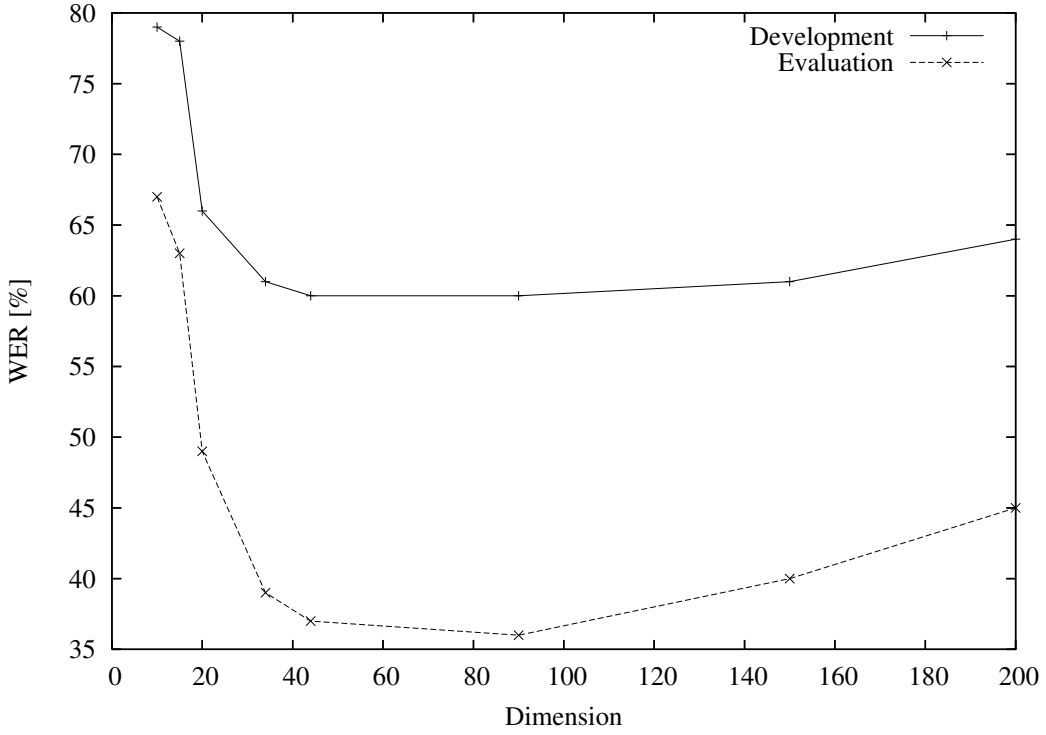


Figure 7.8: Word error rate of the recognition system on the RWTH-BOSTON-104 using the recorded image frames and employing the LDA [Zahedi & Dreuw$^+$ 06a].

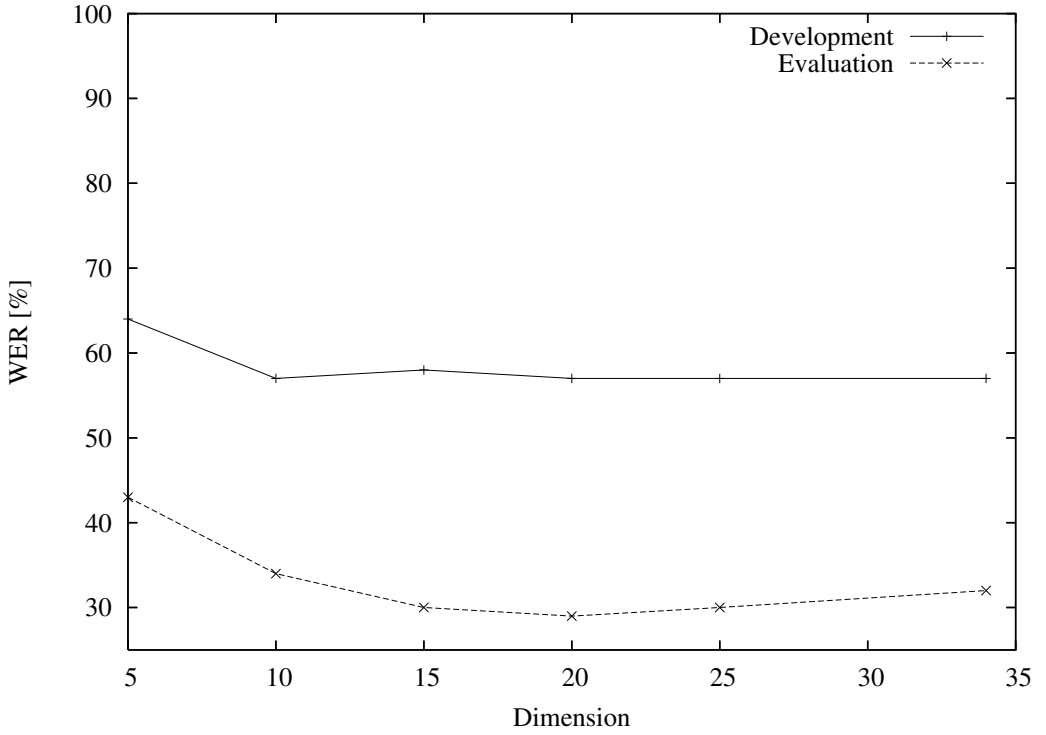


Figure 7.9: Word error rate of the system using the geometric features of the dominant hand extracted from the interpolated image frames and employing the LDA.

which have been recorded directly by the camera and on the image sequences including the interpolated image frames are shown in Figure 7.10 and 7.11.

The experimental results are summarized in Table 7.20 which presents the best error rate of 45% on the development set and the corresponding error rate of 25% for the evaluation set. They are obtained by using 225 components of the feature vectors including the image frames and the interpolated ones. The result shows that PCA is a very powerful transformation to select more relevant features when using image frames directly for video processing. It removes consistent background pixels which do not change the class membership of the sign language words.

Comparing the results of the recognition system which have been achieved by employing the LDA and the PCA when using image features shows that the PCA is significantly more useful to improve the error rate. As the relationship between linear discriminant analysis and maximum entropy framework for log-linear models is studied in [Keysers & Ney 04] in detail, it is expected that the LDA leads to better results for a lower dimensionality of the feature space. When the feature reduction factor is large, i.e. the large feature space with respect to the number of classes, it is shown in [Keysers & Ney 04] that the model distributions are left unchanged by a non-singular linear transformation of the feature space when a log-linear model for the class posterior probability is employed. Furthermore, an explanation for this could be that the LDA expects the samples of one class to be relatively similar. This may not happen for the sign language words where the average distance of the utterances from their class is larger than the mean distance between the classes. It is also commented in [Duda & Hart$^+$ 01] that the LDA is problematic, if the classes are not compact.

As principle component analysis is used successfully to improve the accuracy of the recognition system by reducing the size of the image features, we are encouraged to use it with a larger size of image frames which include more information. Although down-scaling of the image frames reduces the size of the feature vectors and consequently the size of the covariance matrix, it causes a slight distortion in the image frames. We expect to obtain better results by using larger image frames with more components transformed by the PCA. The experimental results summarized in Table 7.21 shows using larger image frames with

Table 7.19: Word error rate of the system on the RWTH-BOSTON-104 employing the LDA and using the image features [Zahedi & Dreuw$^+$ 06a] and the geometric features.

Features	Feature dimension	Development set	Evaluation set
Image down-scaled to 32 × 32	–	64	54
+ LDA	90	60	36
Geometric features	–	59	37
+ LDA	20	57	29

Table 7.20: The word error rate of the system on the RWTH-BOSTON-104 using the image features and employing the PCA.

Features	Feature dimension	Development set	Evaluation set
Image frames	–	64	54
+ PCA	250	47	37
Image frames + interpolated frames	–	69	60
+ PCA	225	45	25

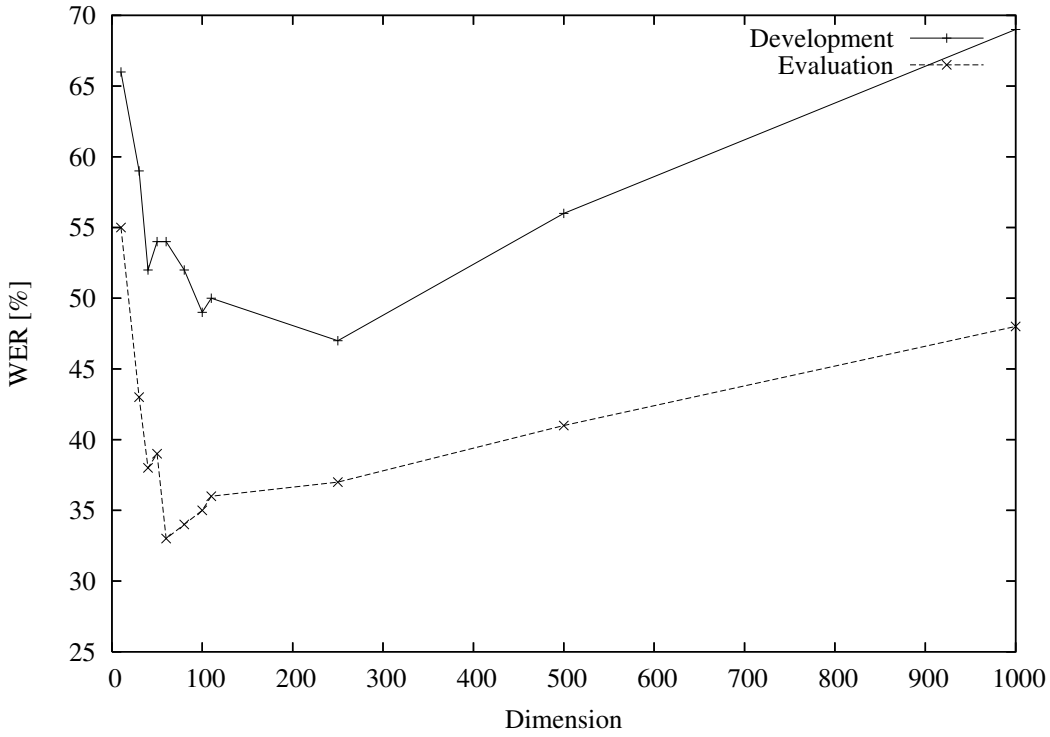


Figure 7.10: The word error rate of the recognition system on the RWTH-BOSTON-104 database using the image frames and the employing PCA.

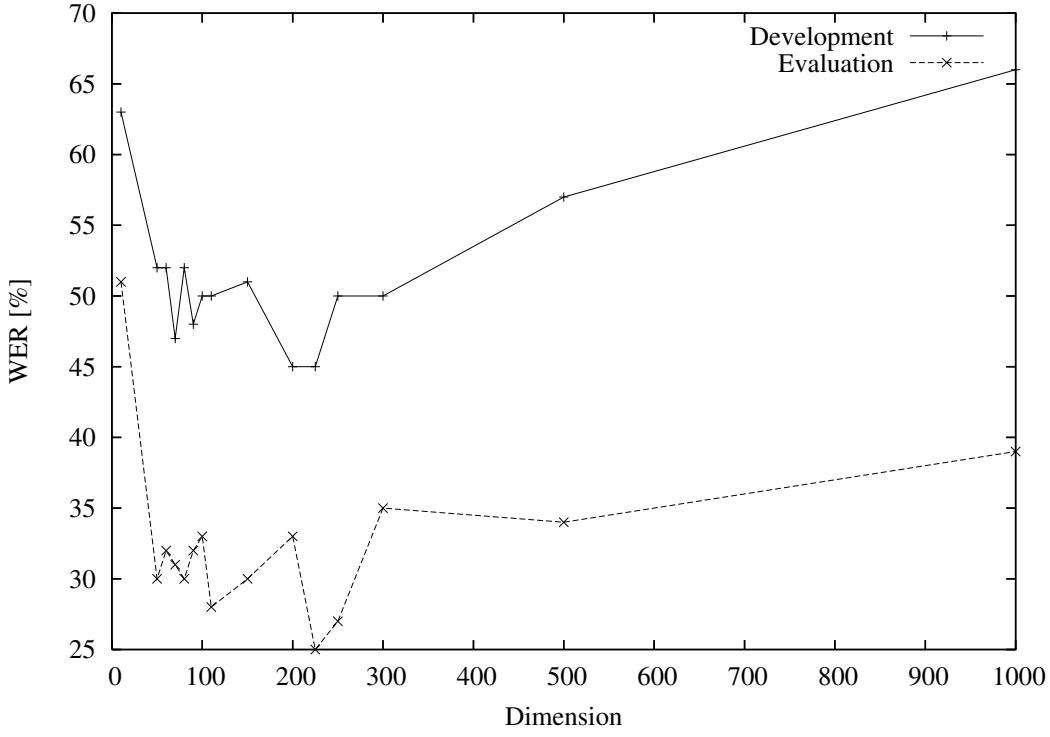


Figure 7.11: The word error rate of the system on the RWTH-BOSTON-104 database using the interpolated image frames and employing the PCA.

more information does not help the classifier to improve the results. It may happen due to a loss of some pixels including signings movements, when larger image frames are used and the size of feature vector remains the same by employing the PCA. However, we could not obtain better results by using larger image frames with more principle components. It may happen due to the scarceness of training data where there is not enough training data to train larger feature vectors.

7.7 Combination methods

In the previous sections, two groups of features extracted from the image frames have been investigated in detail. The geometric features representing the position and the configuration of the signers' dominant hand which convey most of the information about of the meaning of the sign yields a word error rate of 37%. Selecting 20 of the most discriminative features of the geometric features helps the recognition system to obtain an error rate of 29%. On the other hand, The first 225 principle components of the image features which include all information of the signing without emphasizing any part of the signers' body results in a very good error rate of 25%. It is expected that a proper combination of these two feature groups which represent different aspects of the signing can improve the accuracy of the recognition system. The combination can be done in two levels consisting out of the feature level and the model level.

7.7.1 Feature combination

Feature level combination can be performed by (a) feature weighting, i.e. concatenation of the features and weighting them to emphasize the influence of each group of features and (b) LDA-based feature combination, i.e. concatenation of feature vectors over the time to incorporate context information.

Table 7.21: The word error rate of the system on the RWTH-BOSTON-104 database using the image frames with a different size of scaling and employing the PCA.

Image frame size	Feature dimension	Development	Evaluation
16×16	50	50	27
	150	49	26
	225	59	34
32×32	225	45	25
48×48	225	53	31
	250	51	29
	300	51	34
56×56	225	52	30
	300	54	37
	500	56	35
64×64	225	50	37
	300	54	44
	500	57	43
	1000	66	48

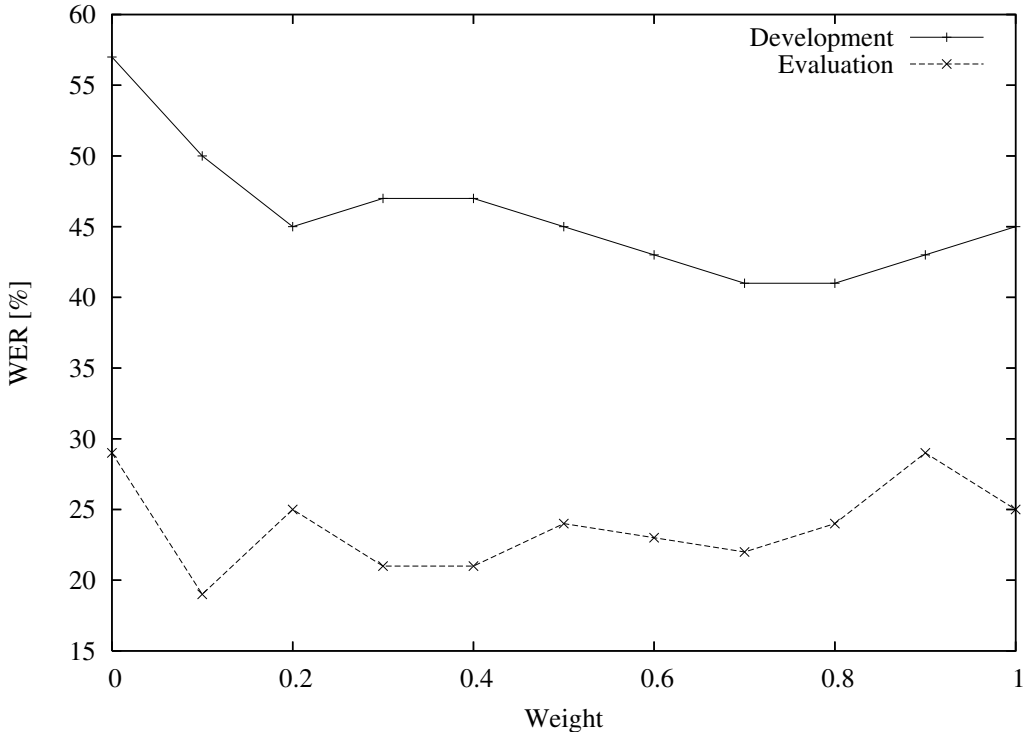


Figure 7.12: The word error rate of the recognition system on the RWTH-BOSTON-104 with weighting of the geometric and the intensity features.

First, we concatenate and weight the feature vectors which are selected by the LDA or the PCA from the image intensity features and from the geometric features. As mentioned before, the image features after the PCA has been employed including 225 components and the geometric features after the LDA has been employed with 20 elements yield the best error rate of 29% and 25%. The results obtained by concatenation and weighting of the feature groups are shown in Figure 7.12. The graphs show the error rate with respect to the weight of the intensity features on the development and the evaluation set. The weight of the intensity image features and the geometric features are chosen to add up to 1.0. The best error rate of 41% is achieved on the development set which corresponds to an error rate of 22% on the evaluation set when weighting the image features and the geometric features with 0.7 and 0.3 respectively. Although an error rate of 19% is obtained on the evaluation set, it is not approved by the development set. It may occur due to the small training data of the development set.

To use the context information we have used temporal derivative feature vectors concatenated to the current image features in the Section 7.3.3 which was not able to decrease the error rate of the system using simple appearance-based features. Since the context information is used for automatic speech recognition successfully by using a window of the feature vectors over the time [Zolnay & Schlueter$^+$ 05], we are going to perform some experiments with different sizes of the windows for the feature vectors and employing the LDA to select the most discriminative feature components. Furthermore, it is expected if the alignment is not good, the context information might partly recover from that. To ensure that the results are comparable, we use the LDA to reduce the size of the feature vectors to 245 elements like when using feature weighting. The results are presented in Table 7.22 which shows that

Table 7.22: The word error rate [%] of the system on the RWTH-BOSTON-104 database using the LDA-based feature combination of the image features and the geometric features with a different size of the window.

Window size	Development set	Evaluation set
1	52	26
3	51	23
5	52	26
7	52	25
11	54	27

the best error rate is obtained with a window size of three. According to these results using context information with a proper size of the window and employing an LDA to select the most relevant features can improve the recognition rate.

Further experiments are performed using a window of the features over the time with the size of three when the other size of the feature vectors are transformed by the LDA. Figure 7.13 shows the best error rate on the development set is 46% which leads us to a word error rate of 24% on evaluation set. The fact that for a different setting, where the error rate on the development set is not optimal, but a better error rate than 24% is obtained for the test set can be explained by different effects:

- the corpus is very small and thus this may be a non-significant change;
- overfitting of the development set.

However, we cannot circumvent these problems as there is no bigger corpus available and we cannot afford to have a larger development corpus since that would reduce the size of our training data too far.

Since adequate training data is a very important issue in the statistical pattern recognition, it seems the results obtained on the evaluation set which contains more data is more reliable than the results on the development set.

7.7.2 Model combination

In contrary to the combination methods in the feature level, for model combination we train two separated models using the feature groups which give the best results. Then we weight the scores using the separated models in the recognition stage. The experiments are going to be performed for both the evaluation set and the development set. The results presented in Figure 7.14 show that the best error rate of 40% is obtained on the development set with a corresponding error rate of 22% when weighting the scores of the model which has been trained by the image features and the geometric features with 0.6 and 0.4 respectively. It can be seen that optimizing the settings on the development set leads to good results on the evaluation data as well. This may occur because two separated visual models are trained by the training data. The visual models using a smaller size of feature vectors comparing to the feature combination methods which use concatenation of the feature groups need fewer parameters to be estimated in training process. Therefore we do not overfitt the development set which contains less training data comparing to the evaluation set.

The experimental results of the three combination methods are summarized in Table 7.23. The feature weighting method and the model combination method in which the

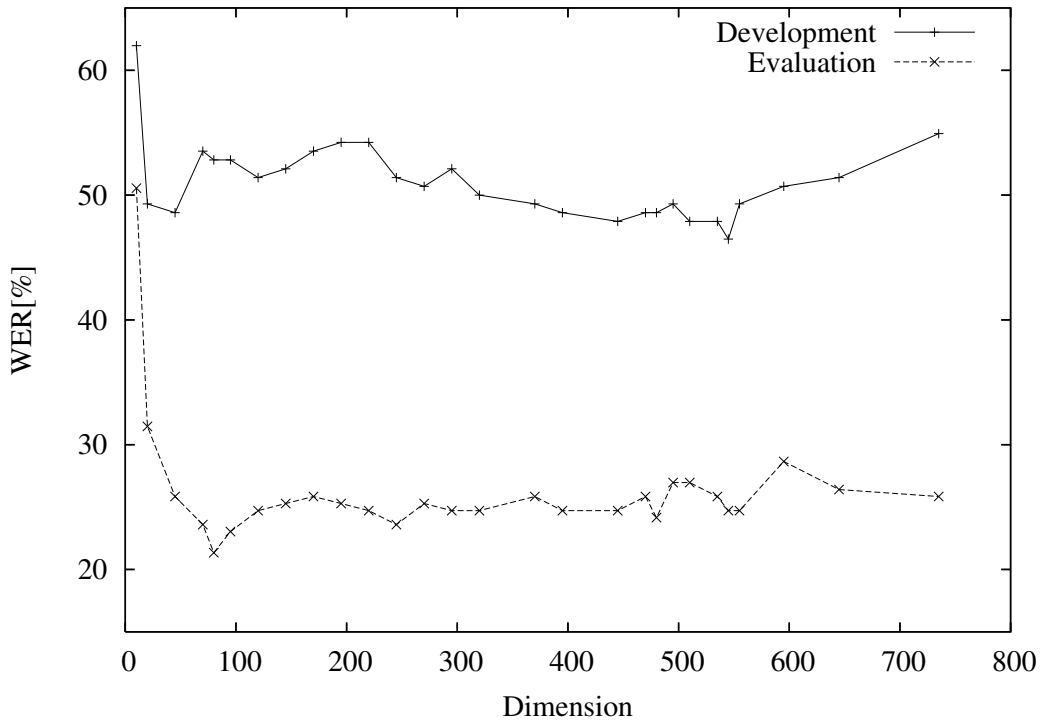


Figure 7.13: The word error rate of the system using the LDA-based combination of the geometric and the intensity features.

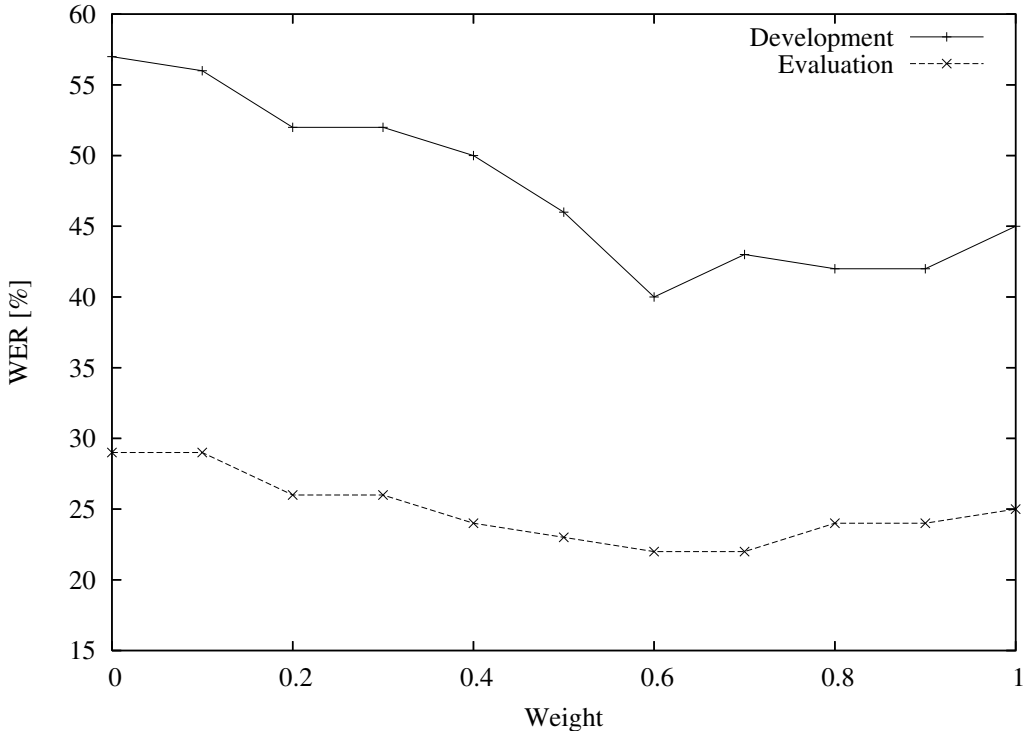


Figure 7.14: The word error rate of the system employing model combination of the HMM models which use geometric and intensity features.

Table 7.23: The word error rate [%] of the system on the RWTH-BOSTON-104 database with different combination methods for the image features and the geometric features. Using the image features and the geometric features yields an error rate of 25% and 29% on the evaluation set, respectively.

	Development set				Evaluation set			
	WER[%]	del	ins	sub	WER[%]	del	ins	sub
Feature weighting	41	28	6	25	22	16	3	21
LDA-based feature combination	46	34	8	24	24	21	3	20
Model combination	40	25	7	26	22	15	6	19

models have been trained by the features separately give better results. This may occur because the dimension of the feature groups are not the same and weighting them helps the system to emphasize their influence in the training and the recognition process.

7.8 Discussion

The best error rate on the RWTH-BOSTON-104 is obtained when combining the models which have been trained separately by the image frames and the geometric features, i.e. an error rate of 40% on development set which leads to an error rate of 22% on evaluation set. We are going to observe the results more carefully to analyze the errors occurring during recognition. A list of the sentences recognized wrongly is presented in Table 7.24. The deleted words are shown with a dashed line, the substituted words in a *red* color and the inserted words in a *green* color. One can see that the word "IX" is deleted four times occurring in the sentence number one and two. Observing the corresponding video streams shows that the sign language word "IX" is signed in these two sentences very shortly and it can be the reason that the "IX" words are deleted during recognition. Furthermore, the word "JOHN" is replaced at the beginning of the four sentences as well. Observing the training set one can see that the word "JOHN" occurs mostly at the beginning of the sentences and the high probability of the language model for this coincidence causes the mistake in the recognition process. Using larger databases can help the system to cope with the errors like these by training the visual model and language model with more training data. The error analysis shows there is no word which is recognized frequently as another word. The confusion matrix of these results presented in Figure 7.15 shows that only the word "IX" is deleted four times and the other errors do not happen more than once for a word. The reference words and the recognized words are shown in vertical and horizontal axis, respectively.

7.8.1 Language model

In the Section 7.1.2, the preliminary results on the databases using the different language models are presented. When using simple features, the impact of the language model is not observable clearly and the small improvements are not enough to see the language model impact in the recognition system. To observe the influence of the language model, we have repeated the experiments of the combination methods which have resulted in the best error rates using different kind of language models with different LM scales. The results of the classifier on the database RWTH-BOSTON-104 employing the feature weighting, the LDA-

Table 7.24: A list of the sentences recognized wrongly in a system employing model combination, i.e. obtaining the best results.

Number	Reference and recognized sentences	del	ins	sub
1	JOHN LIKE IX IX IX JOHN LIKE IX __ __	2	0	0
2	JOHN LIKE IX IX IX JOHN LIKE IX IX __	1	0	0
3	JOHN LIKE IX IX IX JOHN LIKE IX IX __	1	0	0
4	MARY VEGTABLE KNOW IX LIKE CORN MARY VEGTABLE KNOW IX LIKE MARY	0	0	1
5	JOHN [UNKNOWN] BUY HOUSE JOHN _____ BUY HOUSE	1	0	0
6	FUTURE JOHN BUY CAR SHOULD _____ JOHN BUY CAR SHOULD	1	0	0
7	JOHN SHOULD NOT BUY HOUSE JOHN _____ ____ BUY HOUSE	2	0	0
8	JOHN DECIDE VISIT MARY JOHN _____ VISIT MARY	1	0	0
9	JOHN NOT VISIT MARY JOHN FRANK _____ _____	2	0	1
10	ANN BLAME MARY JOHN CAN BLAME MARY	0	1	1
11	IX-1P FIND SOMETHING-ONE BOOK JOHN POSS _____ BOOK	1	0	2
12	JOHN IX GIVE MAN IX NEW COAT JOHN LIKE IX IX IX IX WOMAN COAT	0	1	3
13	POSS NEW CAR BREAK-DOWN JOHN POSS NEW CAR BREAK-DOWN	0	1	0
14	JOHN LEG JOHN POSS LEG	0	1	0
15	WOMAN ARRIVE JOHN ARRIVE HERE	0	1	1
16	IX CAR BLUE SUE BUY IX CAR BLUE SUE ____	1	0	0
17	SUE BUY IX CAR BLUE SUE BUY IX CAR JOHN	0	0	1
18	JOHN MARY BLAME JOHN IX GIVE	0	0	2
19	JOHN ARRIVE JOHN GIVE IX	0	1	1
20	JOHN GIVE GIRL BOX JOHN GIVE IX ____	1	0	1
21	JOHN GIVE GIRL BOX JOHN GIVE JOHN BOX	0	0	1
22	LIKE CHOCOLATE WHO JOHN ARRIVE WHO	0	0	2
23	JOHN TELL MARY IX-1P BUY HOUSE JOHN _____ MARY IX-1P GO IX	1	0	2

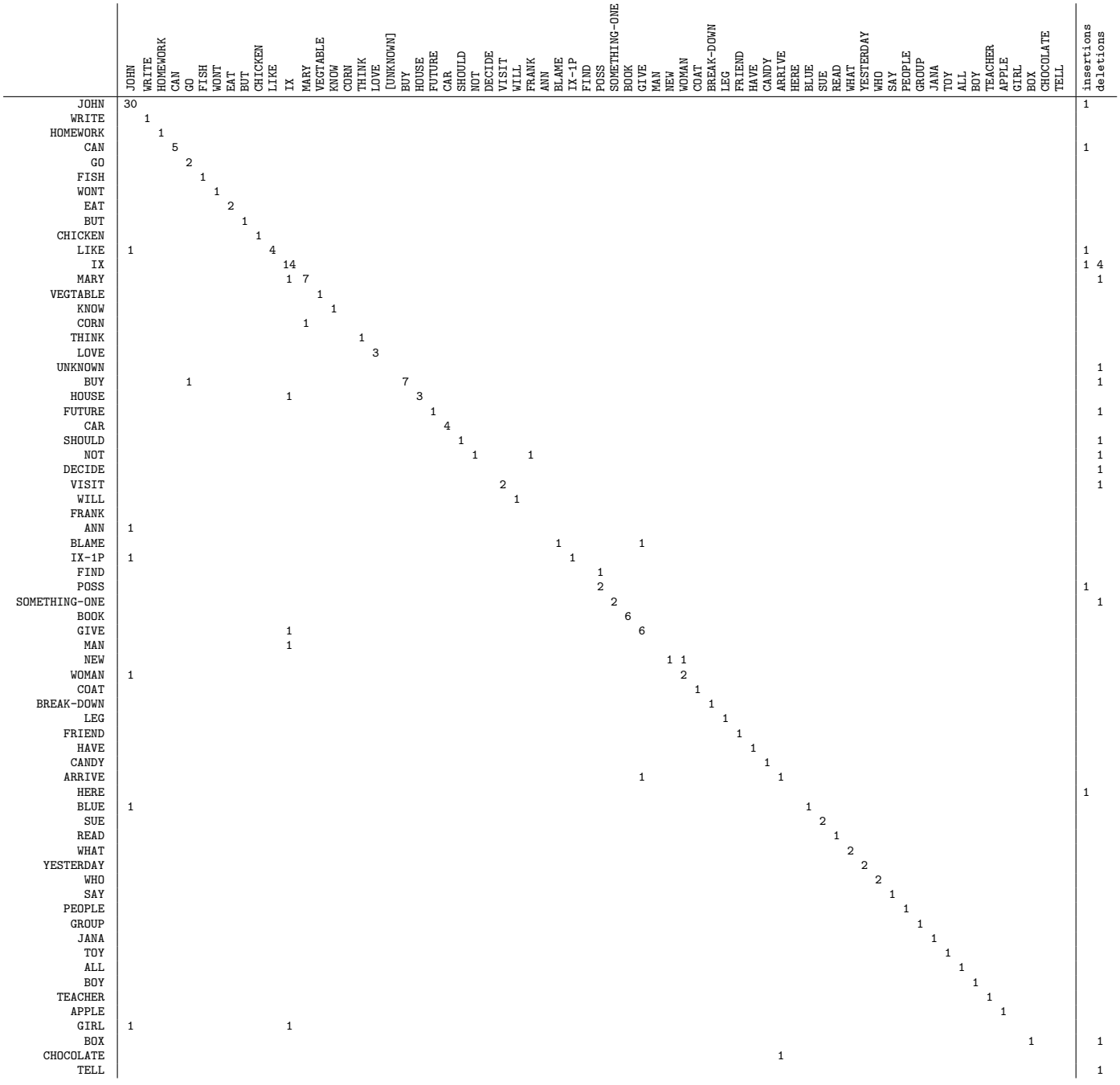


Figure 7.15: Confusion matrix for the results employing model combination of the HMM models which use geometric and intensity features.

based feature combination and the model combination is shown in the Figures 7.16, 7.17 and 7.18, respectively.

The results show that the zerogram which does not use any information of the language model and the unigram which uses only the frequency of the words occurring in the training set and neglect the coincidence of the words do not improve the results. The zerogram distributes language model probability uniformly for all words existing in the vocabulary list and the unigram does not regard the coincidence of the words in the training data. Therefore when we increase the scale value of the language model for these two kinds of language models, it means the scale of the language model increases and the influence of the visual model is decreased. The recognition system with the high scale value of the language models which use poor information and a low scale value of the visual model yields high error rates. In contrary, using the bigram and the trigram language models which take the coincidence of sign language words into account helps the system to obtain better results. The best error rates are obtained by using a middle range of LM scales where the impact of the visual model is taken fairly into account.

Based on theoretical assumptions a good correlation is expected between the WER and the PP of the language model, however it is shown in [Klakow & Peters 02] on different data sets that there are uncertainties in the WER and the PP since they are measured on finite test sets. Although the impact of a good language model in the preliminary results was not as good as expected, in the well tuned system, which uses a very well trained visual model, the trigram and the bigram language models with a small PP and a proper scale with respect to the visual model have improved the word error rate of the recognition system strongly to nearly the same degree as expected according to [Klakow & Peters 02]. A better consistency to fulfill the expectations might be obtained if a larger corpus was available.

7.8.2 Temporal derivative features

In the Section 7.7.1 the results of using a window of feature vectors over the time which improves the accuracy of the recognition system is reported. Another way to use the context information is by using temporal derivative features which was inspired by a successful employment in automatic speech recognition systems. The image frames extracted from the temporal derivatives in video processing contain the information about the movement of the objects in the video stream.

Using these kinds of features did not show an influence in the preliminary experiments in which very simple features have been used. In this section we are going to investigate the influence of the temporal derivative features by performing the experiments which employ the model resulting in the best error rate by using the image frames. Since the 225 principle components of interpolated image frames give the best error rate of 45% and 25% on the development and the evaluation set, we add the temporal derivative features to the image features which yield the best error rate. The results of the system using image frames plus first and second derivatives are shown in Table 7.25. To ensure the results are comparable for all experiments 225 principle components are used as feature vector.

The results show that although by adding derivative features an error rate of 24% can be gained on the evaluation set, however it results in a higher error rate on the development set. This may happen due to the scarceness of the data, i.e. the training data is not large enough to estimate the parameters in the development set which uses less training data.

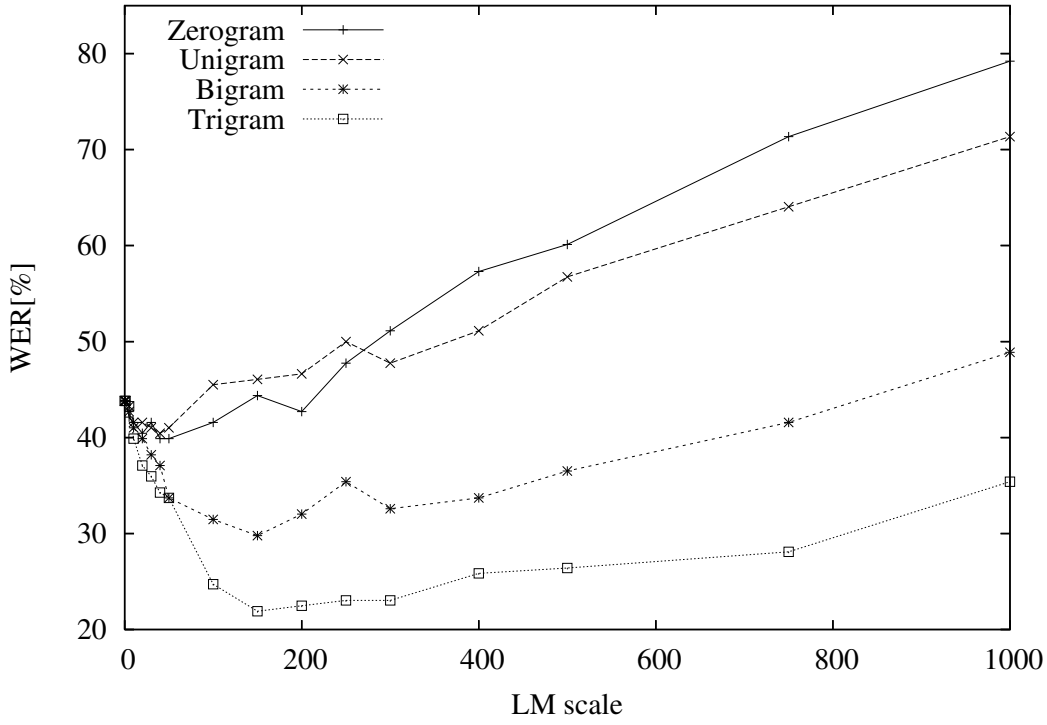


Figure 7.16: WER[%] of the classifier using the different LM scales employing feature weighting.

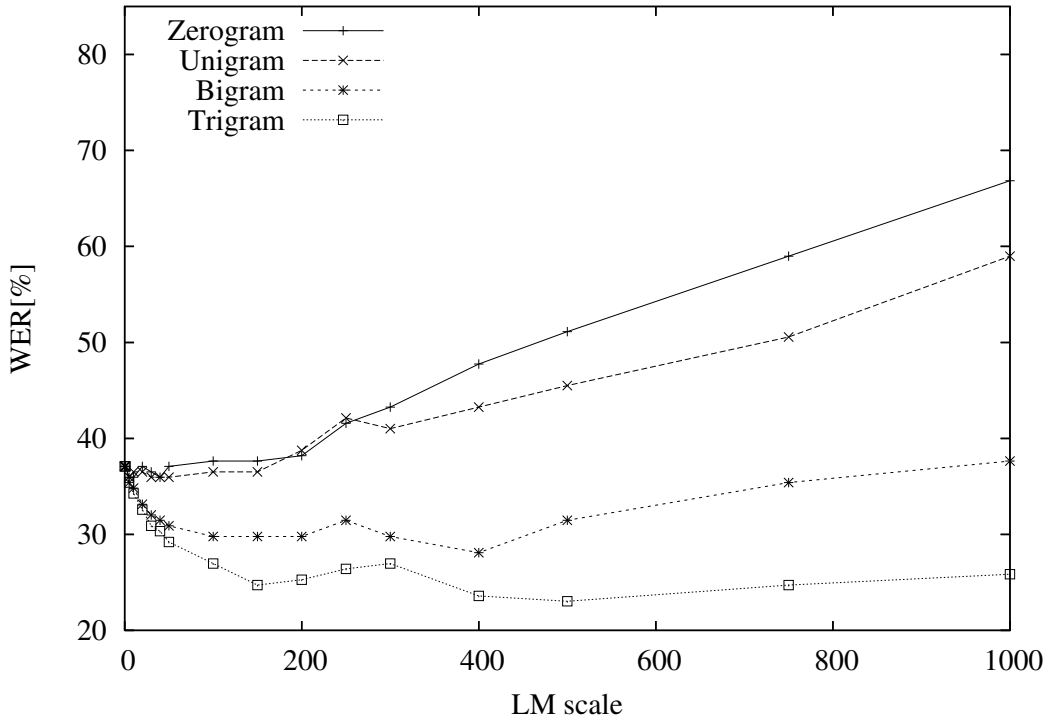


Figure 7.17: WER[%] of the classifier using the different LM scales employing the LDA-based feature combination.

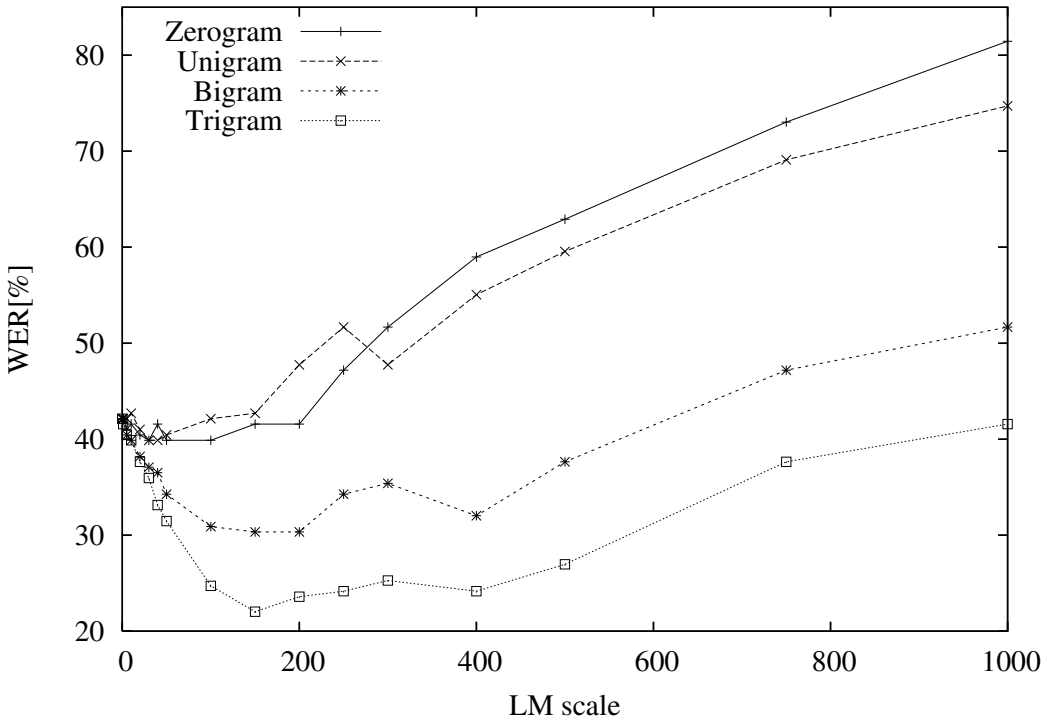


Figure 7.18: WER[%] of the classifier using the different LM scales employing the model combination.

Table 7.25: The word error rate [%] of the classifier on the RWTH-BOSTON-104 using the temporal derivatives and the 225 principle components.

Features	Development	Evaluation
Original image	45	25
+FD	47	25
+SD	47	24
+FD +SD	49	24

Table 7.26: The word error rate [%] of the classifier on the RWTH-BOSTON-104 using the gradients and the 225 principle components.

Features	Development	Evaluation
Original image	45	25
+H gradient	48	26
+H gradient	47	29
+H & V gradients	50	24

7.8.3 Gradient features

As we did for the temporal derivative features, we have repeated the experiments which have given the best error rates using image features with the gradient features as well. The gradient features enhancing the edges in the image frames can add some information about the appearance of the signers' body parts to the feature vector. Table 7.26 presents the word error rate of the system using a feature vector with the size of 225 from the original images and gradients.

Using the gradient images helps to obtain an error rate of 24% using the original image frame plus the corresponding horizontal and vertical gradients. The results on the development does not look like it is obeying a consistent rule just like the previous results where the best error rate on the evaluation set has not been approved by the development process which seems to be due to the overfitting on the development set.

7.8.4 Summary

We are now going to summarize the experimental results of this work for segmented and continuous sign language recognition. For segmented sign language recognition the results on the RWTH-BOSTON-50 database are presented in Table 7.27.

The baseline result gives an error rate of 37% using down-scaled image frames recorded by a camera fixed in front of the signer. This is a good starting point showing the appearance-based features can work well for sign language recognition. Optimizing HMM parameters and using additional image frames recorded by a lateral camera, we have obtained the error rate of 28%. To model different pronunciations, three kinds of clustering methods: manual partitioning, K-means clustering and LBG clustering have been investigated and finally the results lead us to use the nearest neighbor classifier for segmented sign language recognition giving an error rate of 23%.

The analysis of the training data set shows that a large visual variability of the utterances for the sign language word occurring in the database exists. To model the image variability the TD, accounting for global affine transformations, and the IDM, accounting for local deformations are employed therefore improving the accuracy. The TD and the IDM, compensate each other and thus allow a combination of global and local transformations, i.e. the best error rate of 17% has been achieved.

Analyzing the sign language words which have been recognized wrongly shows that most of the remaining errors are due to the visual singletons in the dataset, which cannot be classified correctly using the leaving-one-out approach. This means that one word has been uttered in a way which is visually not similar to any of the remaining utterances of

Table 7.27: The results summary on the RWTH-BOSTON-50 database.

	Error rate [%]
Baseline	37
HMM parameters	32
Lateral camera	28
Nearest neighbor	23
TD	22
IDM	21
TD + IDM	17

Table 7.28: The results summary on the RWTH-BOSTON-104 database.

	Word error rate [%]
Baseline (image features)	54
+ PCA	37
+ interpolation	25
Geometric features	50
+ interpolation	37
+ LDA	29
Feature combination	22
Model combination	22

this particular word. For example, all but one of the signs for "POSS" show a movement of the right hand from the shoulder towards the right side of the signer, while the remaining one shows a movement which is directed towards the center of the body of the signer. Thus this utterance cannot be classified correctly without further training material that shows the same movement. This is the major problem caused by the small available amount of training data.

The results for continuous ASLR on the RWTH-BOSTON-104 are summarized in Table 7.28. After optimizing the preliminary setting of the visual model and the language model parameters, the baseline results have been achieved by two groups of features named image frames and geometric features of the dominant hand of the signer. We have investigated the conventional approaches which are applied in automatic speech recognition and image processing, for our particular case of sign language recognition.

The preliminary results show that appearance-based features work well for sign language recognition. We could obtain an error rate of 37% by using 225 principle components of the image features extracted from the recorded image frames of the camera. This error rate is improved significantly to 25% by employing a linear interpolation method to insert interpolated frames between the recorded image frames. Also using geometric features of the signer's dominant hand improves the error rate and the best error rate of 29% is obtained by employing the LDA to choose the most discriminant feature elements. For both feature groups adding interpolated frames to the sequence of the recorded frames of the camera has helped the system to obtain better results. Since these two feature groups convey different aspects of the signings, the combination methods at the feature and the model level are used for finding a proper way to make use of the information of both feature groups simultaneously. The combination methods which are employed in the feature level and the model level help to improve the accuracy of the system resulting in the best error rate of 22%. As explained before, the development set contains less training data than the evaluation set; if we ignore the development process, further improvement is possible by tuning the knowledge source scales to obtain an error rate of 19% on the evaluation set.

8 Conclusion

Hafez! hypocrisy and dissimulation give not purity of heart;
Choice of the path of profligacy and of love, I will make.
– Hafez (*1320– 1390*)

In this work, an automatic sign language recognition system based on a large-vocabulary speech recognition system has been presented. In the course of this work a new approach to sign language recognition was presented and opened a new view into this field of research which hopefully may inspire other research groups to investigate the remaining issues in more detail.

This approach is unique due to some novel features and properties which are enlisted here:(a) using a standard stationary camera, (b) the feature vectors which include whole image frames containing all the aspects of the signing (c) using geometric features describing the shape of the signers' dominant hand in detail. We have investigated the different issues of this new approach to sign language recognition in a system which is able to recognize on the one hand segmented sign language words and on the other hand continuous sign language sentences and transcribe them properly to glosses.

At this point, we conclude this work summarizing the major findings:

- Appearance-based features such as the original image frames and their transformations like gradients and temporal derivative features work well for sign language recognition. Using appearance based features which are extracted directly from a video stream recorded with a conventional camera makes recognition system more practical. The system is made more robust with respect to various sources of variability by invariant distances such as the tangent distance and the image distortion model, which allow to model global and local changes in appearance respectively. Furthermore, TD and IDM, complement each other and additional performance gain is obtained by combining both. To reduce the size of the image feature vectors and thus to make estimation of a model easier, a principle component analysis has been used very successfully, discarding those pixels with low variances (in fact, variances close or equal to zero from pixel areas mainly in the background) which can therefore not help the classification.

- Although signing contains many different aspects from manual and non-manual cues, the position, the orientation and the configuration or shape of the dominant hand of the signer conveys a large portion of the information of the signings. Therefore, the geometric features which are extracted from the signers' dominant hand, improve the accuracy of the system to a great degree. We have employed a dynamic programming method to track the dominant hand of the signer for succeeding extraction of the geometric features. The accuracy of the tracker is improved by adding interpolated image frames between each pair of frames from the original video, in turn, leading to a better recognition result. Another improvement of the recognition was obtained by linear discriminant analysis reducing the 34 geometric features to 20 coefficients.

- To capture the different aspects of sign language the appearance cue and the geometric cue are fused together by three different combination methods. The experimental results show that all three combination methods help to improve the recognition rate but the feature weighting and the weighted model combination lead to a higher accuracy. It has been shown that a suitable combination of the different features yields an improved error rate over the two different baseline systems.

The experiences gained so far lead us to name remaining issues that are worth further investigations. These issues are directly motivated by some findings of this work:

- In this work, we have used a publicly available database which allows other research groups to compare their system to ours. From our view, comparing systems quantitatively is vital to foster advances. However, the results in this work are biased by the size of the database: on the one hand, the lack of training data and the large amount of singletons leads to a very difficult task and on the other hand the small database makes it difficult to interpret the results as slight improvements can easily be due to random effects. The first problem was counteracted by training mixture densities with relatively small amounts of densities and a novel nearest-neighbor approach to gesture classification. The latter problem was counteracted by motivating the experiments and carefully selecting the experiments performed and analyzing the outcomes of the experiments.

  For future experiments it is advisable to create a larger database.

- In this work, cues describing the dominant hand and cues describing the overall appearance of the gestures have been used jointly. We did not focus on facial expressions although it is well known that facial expressions convey important part of sign-languages but relied on the overall appearance to capture this information. The facial expressions can e.g. be extracted by tracking the signers' face. Then, the most discriminative features can be selected by employing a dimensionality reduction method and this cue could also be fused into the recognition system.

- Epenthesis movements are transition movements between two signs which begin at end of a current sign and finish at starting point of the succeeding sign [Gao & Fang$^+$ 04]. When starting and ending a sign language sentence, signers move one hand or both hands from and to a base point respectively which is often located close to the bottom of the signing space, e.g. on a table. In our work, these movements are modelled as "Silence", but the silence model which is employed in this work, is probably not able to capture enough variability and very likely a better model for epenthesis would also lead to further improvements.

  For the future, investigations on a proper epenthesis model are very promising.

- In this work, words are modelled by whole-word models which allow for a good recognition but make it impossible to recognize unseen words, which is possible in large-vocabulary speech recognition systems using phonemes and a pronunciation lexicon. The phoneme concept from spoken language recognition cannot directly be transfered to sign language recognition due to the various special issues in the grammar of sign-language: for example sub-lexical units can be signed simultaneously by different parts of the body. Indexing is a unique property of sign language grammar where the signer

defines persons and objects by classifiers or by spelling their name by using the sign language alphabet and locates them in the signing space around himself.

Therefore, we expect some major improvements in ASLR by a suitable model of subword units and a proper way of transcribing words into a phoneme-like way.

- The methods in this work are focussed towards good recognition accuracy and not toward real-time performance and thus many of the methods are computationally expensive and not applicable in real-life situations. However, we expect that due to upcoming developments in computing, even handheld devices will have considerable computing power and that therefore many of the presented techniques might be applicable in near future.

List of Acronyms

2D	2-Dimensional
3D	3-Dimensional
AFD	Absolute First Temporal Derivative
ANN	Artificial Neural Networks
ASL	American Sign Language
ASLR	Automatic Sign Language Recognition
ASR	Automatic Speech Recognition
BSL	British Sign Language
D	Diagonal Sobel Filter
DEL	Deletion
DGS	German Sign Language (Deutsche Gebärdensprache)
DJ	Diagonal Jähne Filter
DL	Diagonal Right Sobel Filter
DLJ	Diagonal Left Jähne Filter
DLP	Diagonal Right PLUS Filter
DP	Diagonal PLUS Filter
DR	Diagonal Left Sobel Filter
DRJ	Diagonal Right Jähne Filter
DRP	Diagonal Left PLUS Filter
EGM	Elastic Graph Matching
ER	Error Rate
EM	Expectation-Maximization
FD	First Temporal Derivative
G	Gaussian Filter
GMD	Gaussian Mixture Density
H	Horizontal Sobel Filter
HamNoSys	The Hamburg Sign Language Notation System
HJ	Horizontal Jähne Filter
HP	Horizontal Sobel Plus Filter
HPC	High Performance Computing
HPSG	Head-driven Phrase Structure Grammars
HMM	Hidden Markov Model
i6	Chair of Computer Science VI (Lehrstuhl für Informatik VI) of RWTH Aachen University
ICA	Independent Component Analysis
IDM	Image Distortion Model
INS	Insertion
KD	Kernel Density
L	Laplace Filter
LBG	Linde-Buzo-Gray Clustering

LDA	Linear Discriminant Analysis
LM	Language Model
LTI	Lehrstuhl für Technische Informatik of RWTH Aachen University
ME	Maximum Entropy
ML	Maximum Likelihood
NFD	Negative First Temporal Derivative
NGT	Sign Language of the Netherlands
NN	Nearest Neighbor
OCR	Optical Character Recognition
OOV	Out-Of-Vocabulary
PCA	Principal Component Analysis
PFD	Positive First Temporal Derivative
PP	Perplexity
RST	Rotation, Scaling, Translation (Invariance)
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University)
SD	Second Temporal Derivative
SiGML	The Signing Gesture Markup Language
SOFM/HMM	Self-Organizing Feature Maps/Hidden Markov Model
SSL	Swedish Sign Language
SUB	Substitution
TD	Tangent Distance
V	Vertical Sobel Filter
VJ	Vertical Jähne Filter
VP	Vertical PLUS Filter
WER	Word Error Rate
XML	Extensible Markup Language

List of Tables

List of Figures

References

[Abe & Saito+ 02] K. Abe, H. Saito, S. Ozawa: Virtual 3D Interface System via Hand Motion Recognition From Two Cameras. *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 32, No. 4, pp. 536–540, July 2002.

[Akyol & Canzler+ 00] S. Akyol, U. Canzler, K. Bengler, W. Hahn: Gesture Control for Use in Automobiles. In *Proc. IAPR Workshop Machine Vision Applications*, pp. 349–352, Tokyo, Japan, Nov. 2000.

[Akyol & Zieren 02] S. Akyol, J. Zieren: Evaluation of ASM Head Tracker for Robustness against Occlusion. In *Proceedings of the International Conference on Imaging Science, Systems, and Technology (CISST 02)*, Vol. 1, June 2002.

[A.Pelkmann 99] A.Pelkmann. Entwicklung eines Klassifikators zur videobasierten Erkennung von Gesten. Studienarbeit, Diploma thesis, RWTH Aachen University, Aachen, Germany, feb 1999.

[Bahl & Jelinek+ 83] L.R. Bahl, F. Jelinek, R.L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 179–190, March 1983.

[Bauer & Hienz 00a] B. Bauer, H. Hienz: Relevant Features for Video-based Continuous Sign Language Recognition. In *Proceedings of the 4th International Conference Automatic Face and Gesture Recognition*, pp. 440–445, Grenoble, France, march 2000.

[Bauer & Hienz+ 00b] B. Bauer, H. Hienz, K. Kraiss: Video-Based Continuous Sign Language Recognition Using Statistical Methods. In *Proceedings of the International Conference on Pattern Recognition*, pp. 463–466, Barcelona, Spain, 2000.

[Bauer & Nießen+ 99] B. Bauer, S. Nießen, H. Hienz: Towards an automatic sign language translation system. In *Proc. of the International Workshop on Physicality and Tangibility in Interaction*, Siena, Italy, Oct. 1999.

[Bellman 57] R.E. Bellman. Dynamic Programming. Princeton University Press, Princeton, NJ, 1957, 1957.

[Beyerlein 98] P. Beyerlein: Discriminative model combination. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 481–484, Seattle, WA, May 1998.

[Birk & Moeslund+ 97] H. Birk, T. Moeslund, C. Madsen: Real-Time Recognition of Hand Alphabet Gestures Using Principal Component Analysis. In *Proc. of the 10th Scandinavian Conference on Image Analysis*, Laeenranta, Finland, June 1997.

[Bobick & Wilson 97] A.F. Bobick, A.D. Wilson: A State-Based Approach to the Representation and Recognition of Gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 12, pp. 1325–1337, 1997.

[Bowden & Windridge+ 04] R. Bowden, D. Windridge, T. Kabir, A. Zisserman, M. Bardy: A Linguaistic Feature Vector for the Visual Interpretation of Sign Language. In *Proceedings of ECCV 2004, the 8th European Conference on Computer Vision*, Vol. 1, pp. 391–401, Prague, Czech Republic, 2004.

[Brand & Oliver+ 97] M. Brand, N. Oliver, A. Pentland: Coupled hidden Markov models for complex action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–, Washington, DC, USA, June 1997. IEEE Computer Society.

[Brown & Pietra+ 93] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, R.L. Mercer: The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, Vol. 19, No. 2, pp. 263–311, 1993.

[Bungeroth & Ney 04] J. Bungeroth, H. Ney: Statistical Sign Language Translation. In *Proc. of the Workshop on Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 105–108, Lisbon, Portugal, May 2004.

[Bungeroth & Stein+ 06] J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, H. Ney: A German Sign Language Groups of the Domain Weather Report. In *Proceedings of the 5th Iternational Conference on Language Resources and Evaluation*, Genoa, Itely, 2006.

[Canzler & Dziurzyk 02] U. Canzler, T. Dziurzyk: Extraction of Non Manual Features for Videobased Sign Language Recognition. In *Proc. IAPR Workshop Machine Vision Applications*, pp. 318–321, 2002.

[Cascia & Sclaroff+ 00] M.L. Cascia, S. Sclaroff, V. Athitsos: Fast, reliable head tracking under varying illumination: an approach based on registration of texture– mapped 3D models. *IEEE Transactions PAMI*, Vol. 22, No. 4, pp. 322–336, 2000.

[Chiu & Wu+ 07] Y.H. Chiu, C.H. Wu, H.Y. Su, C.J. Cheng: Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 1, No. 28, pp. 208–309, Jan. 2007.

[Crasborn & van der Kooij+ 04] O. Crasborn, E. van der Kooij, A. Nonhebel, W. Emmerik. *ECHO Data Set for Sign Language of the Netherlands (NGT).* Department of Linguistics, Radboud University Nijmegen, http://www.let.ru.nl/sign-lang/echo, 2004.

[Cui & Swets+ 95] Y. Cui, D. Swets, J. Weng: Learning-Based Hand Sign Recognition Using SHOSLIF-M. In *In Proceeding of Int. Workshop on Automatic Face and Gesture Recognition*, pp. 201–206, Zurich, 1995.

[Dempster & Laird+ 77] A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38, 1977.

[Deselaers & Criminisi+ 07] T. Deselaers, A. Criminisi, J. Winn, A. Agarwal: Incorporating On-demand Stereo for Real Time Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, in press, Minneapolis, MN, USA, June 2007.

[Deselaers & Hegerath+ 06] T. Deselaers, A. Hegerath, D. Keysers, H. Ney: Sparse Patch-Histograms for Object Classification in Cluttered Images. In *DAGM 2006, Pattern Recognition, 26th DAGM Symposium*, Vol. 4174 of *Lecture Notes in Computer Science*, pp. 202–211, Berlin, Germany, Sept. 2006.

[Deselaers & Keysers+ 04] T. Deselaers, D. Keysers, H. Ney: FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In *CLEF 2004*, Vol. 3491 of *Lecture Notes in Computer Science*, pp. 688–698, Bath, UK, Sept. 2004.

[Deselaers & Keysers+ 05a] T. Deselaers, D. Keysers, H. Ney: Discriminative Training for Object Recognition using Image Patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 157–162, San Diego, CA, USA, June 2005. IEEE.

[Deselaers & Keysers+ 05b] T. Deselaers, D. Keysers, H. Ney: Improving a Discriminative Approach to Object Recognition using Image Patches. In *DAGM 2005, Pattern Recognition, 26th DAGM Symposium*, Lecture Notes in Computer Science, pp. 326–333, Vienna, Austria, Aug. 2005.

[Deselaers & Müller+ 07] T. Deselaers, H. Müller, P. Clogh, H. Ney, T.M. Lehmann: The CLEF 2005 Automatic Medical Image Annotation Task. *International Journal of Computer Vision*, Vol. 74, pp. 51–58, 2007.

[DESIRE 04] DESIRE. Aachener Glossenumschrift. Technical report, RWTH Aachen, 2004. Übersicht über die Aachener Glossennotation.

[Dreuw & Deselaers+ 06a] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, H. Ney: Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *Proceedings of the 7th International Conference of Automatic Face and Gesture Recognition*, Southampton, UK, 2006.

[Dreuw & Deselaers+ 06b] P. Dreuw, T. Deselaers, D. Keysers, H. Ney: Modeling Image Variability in Appearance-Based Gesture Recognition. In *ECCV 2006 3rd Workshop on Statistical Methods in Multi-Image and Video Processing*, pp. 7–18, Graz, Austria, May 2006.

[Dreuw & Keysers+ 05] P. Dreuw, D. Keysers, T. Deselaers, H. Ney: Gesture Recognition Using Image Comparison Methods. In *GW 2005, 6th Int. Workshop on Gesture in Human-Computer Interaction and Simulation*, Vannes, France, May 2005.

[Dreuw & Rybach+ 07] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, H. Ney: Speech Recognition Techniques for a Sign Language Recognition System. In *Interspeech/ICSLP 2007*, accepted for publication, Antwerp, Belgium, August 2007.

[Dreuw 05] P. Dreuw. Appearance-Based Gesture Recognition. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Jan. 2005.

[Drucker & Schapire+ 93] H. Drucker, R. Schapire, P. Simard: Boosting Performance in Neural Networks. *Int. J. Pattern Recognition Artificial Intelligence*, Vol. 7, No. 4, pp. 705–719, 1993.

[Duda & Hart 73] R. Duda, P. Hart: *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[Duda & Hart+ 01] R.O. Duda, P.E. Hart, D.G. Stork: *Pattern Classification*. New York: John Wiley & Sons, 2nd edition, 2001.

[Eickeler & Kosmala+ 98] S. Eickeler, A. Kosmala, G. Rigoll: Hidden Markov Model Based Continuous Online Gesture Recognition. In *Int. Conference on Pattern Recognition (ICPR)*, pp. 1206–1208, Brisbane, 1998.

[Elliott & Glauert+ 01] R. Elliott, J. Glauert, J. Kennaway, K. Parsons. D5-2: SiGML Definition. Technical Report working document, ViSiCAST Project, Nov. 2001.

[Erdem & Sclaroff 02] U.M. Erdem, S. Sclaroff: Automatic Detection of Relevant Head Gestures in American Sign Language Communication. In *Proceedings of the International Conference on Pattern Recognition, ICPR*, Qubec, Canada, Aug. 2002.

[Estes & Algazi 95] J.R.R. Estes, V.R. Algazi: Efficient error free chain coding of binary documents. In *Data Compression Conference 95*, pp. 122–131, Snowbird, Utah, March 1995.

[F. Quek 96] M.Z. F. Quek: Inductive Learning in Hand Pose Recognition. In *Proc. of Second IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, 1996. IEEE.

[Fang & Gao+ 04] G. Fang, W. Gao, D. Zhao: Large Vocabulary Sign Language Recognition Base on Fuzzy Decision Trees. *IEEE Transaction on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 34(3), pp. 305–314, 2004.

[Fukunaga 90] K. Fukunaga: *Introduction to Statistical Pattern Recognition.* Computer Science and Scientific Computing Academic Press Inc., San Diego, CA, 2nd edition, 1990.

[Gao & Fang$^+$ 04] W. Gao, G. Fang, D. Zhao, Y. Chen: Transition Movement Models for Large Vocabulary Continuous Sign Lamguage Recognition. In *Proceedings of the sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 553–558, Seoul, Korea, 2004.

[Gauvain & Lamel$^+$ 05] J. Gauvain, L. Lamel, H. Schwenk, F. Brugnara, R. Schlüter, M. Bisani, S. Stüker, T. Schaaf, S. Mohammed, M. Bacchiani, M. Westphal, S.Sivadas, I. Kiss, F. Giron. Technology and Corpora for Speech to Speech Translation (TC-STAR). Technical Report ASR Progress Report, May 2005.

[Gavrila 99] D.M. Gavrila: The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82–98, February 1999.

[Haeb-Umbach & Ney 92] H. Haeb-Umbach, H. Ney: Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. ICASSP 1992*, pp. 13–16, 1992.

[Hegerath & Deselaers$^+$ 06] A. Hegerath, T. Deselaers, H. Ney: Patch-based Object Recognition Using Discriminatively Trained Gaussian Mixtures. In *BMVC 2006, 17th British Machine Vision Conference*, Vol. 2, pp. 519–528, Edinburgh, UK, Sept. 2006.

[Hernandez-Rebollar & Lindeman$^+$ 02] J. Hernandez-Rebollar, R. Lindeman, N. Kyriakopoulos: A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation. In *Proc. of the 4th IEEE International Conference on Multimodal Interfaces*, pp. 185–190, Pittsburgh, PA, Oct. 2002.

[Hu 62] M.K. Hu: Visual pattern recognition by moment invariants. *IRE transactions on Information Theory*, Vol. 8, pp. 179–187, 1962.

[Huang & Jack 89] X.D. Huang, M.A. Jack: Semi-Continuous Hidden Markov Models for Speech Signals. *Computer Speech and Language*, Vol. 3, pp. 239–252, 1989.

[Huenerfauth 04] M. Huenerfauth: A Multi-path Architechture for Machine Translation of English text into American Sign Language Animation. In *Proc. of Student Workshop at Human Language Technologies Conference HLT-NAACL*, Boston, MA, USA, May 2004.

[Jähne & Scharr$^+$ 99] B. Jähne, H. Scharr, S. Körkel: *Handbook of Computer Vision and Applications*, Vol. 2. Academic Press, 1999.

[Jelinek 76] F. Jelinek: Continuous Speech Recognition by Statistical Methods. In *Proc. of the IEEE*, Vol. 64, pp. 532–556, April 1976.

[Jelinek 98] F. Jelinek: *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, Massachusetts, January 1998.

[Jones & Rehg 98] M. Jones, J. Rehg. Statistical Color Models with Application to Skin Color Detection. Technical Report CRL 98/11, Compaq Cambridge Research Lab, pp. 274–280, 1998.

[Jones & Rehg 02] M.J. Jones, J.M. Rehg: Statistical color models with application to skin detection. *International Journal of Computer Vision*, Vol. 46, No. 1, pp. 81–96, January 2002.

[Kadous 96] M.W. Kadous: Machine Recognition of Auslan Signs using Powergloves: Toward Large-lexicon Recognition of Sign Language. In *Proc. Workshop Integration Gesture Language Speech*, pp. 165–174, Newark, Delaware, Oct. 1996.

[Kanthak & Molau$^+$ 00] S. Kanthak, S. Molau, A. Sixtus, R. Schlüter, H.Ney: The RWTH Large Vocabulary Speech Recognition System for Spontaneous Speech. In *Proceedings of the Konvens 2000*, pp. 249–254, Ilmenau, Germany, 2000.

[Kashima & Hongo+ 01] H. Kashima, H. Hongo, K. Kato, K. Yamamoto: A Robust Iris Detection Method of Facial and Eye Movement. In *VI2001, Vision Interface Annual Conference*, Ottawa, Canada, June 2001.

[Katz 87] S.M. Katz: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech and signal Processing*, Vol. 35, pp. 400–401, March 1987.

[Keysers & Deselaers+ 04] D. Keysers, T. Deselaers, H. Ney: Pixel-to-Pixel Matching for Image Recognition using Hungarian Graph Matching. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, Vol. 3175 of *Lecture Notes in Computer Science*, pp. 154–162, Tübingen, Germany, Aug. 2004. Received DAGM prize.

[Keysers & Deselaers+ 07] D. Keysers, T. Deselaers, C. Gollan, H. Ney: Deformation Models for Image Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. to appear, Dec. 2007.

[Keysers & Gollan+ 04] D. Keysers, C. Gollan, H. Ney: Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *ICPR 2004, 17th Int. Conf. on Pattern Recognition*, Vol. IV, pp. 511–514, Cambridge, UK, Aug. 2004.

[Keysers & Macherey+ 04] D. Keysers, W. Macherey, H. Ney, J. Dahmen: Adaptation in Statistical Pattern Recognition using Tangent Vectors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 26, No. 2, pp. 269–274, Feb. 2004.

[Keysers & Ney 04] D. Keysers, H. Ney: Linear Discriminant Analysis and Discriminative Log-linear Modeling. In *ICPR 2004, 17th Int. Conf. on Pattern Recognition*, Vol. I, pp. 156–159, Cambridge, UK, Aug. 2004.

[Klakow & Peters 02] D. Klakow, J. Peters: Testing the correlation of word error rate and perplexity. *Speech Commun.*, Vol. 38, No. 1, pp. 19–28, 2002.

[Levenshtein 66] V.I. Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, Vol. 10, pp. 707–710, 1966.

[Levinson & Rabiner+ 83] S.E. Levinson, L.R. Rabiner, M.M. Sondhi: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technology Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[Liang & Ouhyoung 98] R.H. Liang, M. Ouhyoung: A Real-time Continuous Gesture Recognition System for Sign Language. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 558–565, Nara, Japan, April 1998. IEEE.

[Liddell 93] S. Liddell: Holds and Positions: Comparing Two Models of Segmentation in ASL. In *In Phonetics and Phonology: Current Issues in ASL Phonology*, pp. 189–211. Academic Press Inc., 1993.

[Linde & Buzo+ 80] Y. Linde, A. Buzo, R. Gray: An Algorithm for Vector Quantizer Design. *IEEE Trans. Communications*, Vol. 28, No. 1, pp. 84–95, 1980.

[Lööf & Bisani+ 06] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schlüter, H. Ney: The 2006 RWTH Parliamentary Speeches Transcription System. In *In Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP 2006)*, pp. 105–108, Pittsburgh, PA, sep. 2006.

[Malassiottis & Aifanti+ 02] S. Malassiottis, N. Aifanti, M. Strintzis: A Gesture Recognition System Using 3D Data. In *Proc. IEEE 1st International Symposium on 3D Data Processing, Visualization, and Transmission*, Vol. 1, pp. 190–193, Padova, Italy, June 2002.

[Martinez & Kak 01] A. Martinez, A. Kak: PCA versus LDA. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 228–233, Feb. 2001.

[Matsuo & Igi$^+$ 97] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, T. Teshima: The Recognition Algorithm with Noncantact for Japanese Sign Language Using Morphological Analysis. In *Proc. Int. Gesture Workshop*, pp. 273–284, Bielefeld, Germany, Sept. 1997.

[Mehdi & Khan 02] S. Mehdi, Y. Khan: Sign Language Recognition Using Sensor Gloves. In *Proc. of the 9th International Conference on Neural Information Processing*, Vol. 5, pp. 2204–2206, Singapore, 2002.

[Merialdo & Marchand-Maillet$^+$ 00] B. Merialdo, S. Marchand-Maillet, B. Huet: Approximate Viterbi Decoding for 2D-Hidden Markov Models. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. IV, pp. 2147–2150, Istanbul, Turkey, June 2000.

[Moore & Essa 02] D. Moore, I. Essa: Recognizing multitasked activities from video using stochastic context-free grammar. In *18th national conference on Artificial Intelligence*, pp. 770–776. American Association for Artificial Intelligence, July 2002.

[Morrissey & Way 05] S. Morrissey, A. Way: An Example-Based Approach to Translating Sign Language. In *Workshop Example-Based Machine Translation (MT X–05)*, pp. 109–116, Phuket, Thailand, September 2005.

[Nam & Wohn 96] Y. Nam, K. Wohn: Recognition of Space-Time Hand-Gestures Using Hidden Markov Model. In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*, pp. 51–58, Hong Kong, July 1996.

[Neidle & Kegl$^+$ 00] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, R. Lee: *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press, 2000.

[Ney & Essen$^+$ 94] H. Ney, U. Essen, R. Kneser: On Structuring Probabilistic Dependencies in Language Modeling. *Computer Speech and Language*, Vol. 2, pp. 1–38, 1994.

[Ney 84] H. Ney: The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Transactions on Acoustics, Speech and signal Processing*, Vol. 32, pp. 263–271, April 1984.

[Ney 05a] H. Ney: One Decade of Statistical Machine Translation: 1996-2005. In *the MT Summit X*, Vol. I, pp. 12–17, Phuket, Thailand, Sept. 2005.

[Ney 05b] H. Ney. Speech Recognition. Script to the Lecture on Speech Recognition Held at RWTH Aachen University, 2005.

[Nguyen & Bui$^+$ 03] N. Nguyen, H. Bui, S. Venkatesh, G. West: Recognising and monitoring highlevel behaviours in complex spatial environments. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 620–625, Madison, Wisconsin, June 2003.

[Ong & Ranganth 05] S.C. Ong, S. Ranganth: Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 27, No. 6, pp. 873–891, 2005.

[Pavlovic & Sharma$^+$ 97] V. Pavlovic, R. Sharma, T.S. Huang: Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 677–695, July 1997.

[Pollard & Sag 87] C. Pollard, I.A. Sag. Information-based Syntax and Semantics. Technical report, Center for the Study of Language and Information, 1987.

[Prillwitz 89] S. Prillwitz: HamNoSys. Version 2.0; Hamburg Notation System for Sign Language. An Introduction Guide. In *Signum Verlag*, 1989.

[Rabiner 89] L.R. Rabiner: A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, February 1989.

[Raja & McKenna+ 98] Y. Raja, S.J. McKenna, S. Gong: Tracking and Segmenting People in Varying Lighting Conditions using Colour. In *3rd IEEE International Conference on Face and Gesture Recognition*, pp. 228–233, Nara, Japan, April 1998.

[Rigoll & Kosmala 97] G. Rigoll, A. Kosmala: New Improved Feature Extraction Methods for Real-Time High Performance Image Sequence Recognition. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2901–2904, Munich, 1997.

[Rigoll & Kosmala+ 98] G. Rigoll, A. Kosmala, S. Eickeler: High Performance Real-time Gesture Recognition Using Hidden Markov Models. In *Proceedings of Iternational Gesture Workshop*, Vol. 1371 of *LNCS*, pp. 69–80, Bielefeld, Germany, 1998. Springer Verlag.

[Sáfár & Marshall 02] E. Sáfár, I. Marshall: Sign Language Generation Using HPSG. In *Proc. of 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 105–114, TMI, Japan, March 2002.

[Schlenzig & Hunter+ 94] J. Schlenzig, E. Hunter, R. Jain: Recursive identification of gesture inputs using hidden markov models. In *Second Annual Conference on Applications of Computer Vision*, pp. 187–194, Sarasota, FL, USA, December 1994.

[Sigal & Sclaroff+ 00] L. Sigal, S. Sclaroff, V. Athitsos: Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *Computer Vision and Pattern Recognition*, Vol. 2, pp. 152–159, Hilton Head Island, SC, USA, June 2000.

[Sixtus & Molau+ 00] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney: Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1671–1674, Istanbul, Turkey, June 2000.

[Sonka & Hlavac+ 98] M. Sonka, V. Hlavac, R. Boyle: *Image Processing, analysis and Machine Vision.* Books Cole, 1998.

[Starner & Weaver+ 98] T. Starner, J. Weaver, A. Pentland: Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *Transaction of Pattern Analysis and Machine Intelligence*, Vol. 20(2), pp. 1371–1375, 1998.

[Stein & Bungeroth+ 06] D. Stein, J. Bungeroth, H. Ney: Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *Proc. of the 11th Annual conference of the European Association for Machine Translation*, Oslo, Norway, 2006.

[Stolcke 02] A. Stolcke: SRILM – An Extensible Language Modelling Toolkit. In *Proc. of Int. Conf. on Spoken Language Processing (ICSLP 2002)*, Vol. 2, pp. 901–904, Denver, CO, sep. 2002.

[Tolba & Selouani+ 02] H. Tolba, A. Selouani, D.O. Shaughnessy: Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 837–840, Orlando, FL, May 2002.

[Traxler 00] C.B. Traxler: The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education*, Vol. 5, No. 4, pp. 337–348, 2000.

[Triesch & von der Malsburg 01] J. Triesch, C. von der Malsburg: A System for Person-Independent Hand Posture Recognition against Complex Backgrounds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 12, pp. 1449–1453, Dec. 2001.

[Verbeek & van Vliet 94] P.W. Verbeek, L.J. van Vliet: On the Location Error of Curved Edges in Low-Pass Filtered 2D and 3D Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 16, No. 7, pp. 726–733, 1994.

[Vintsyuk 71] T.K. Vintsyuk: The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 7, pp. 133–143, March 1971.

[Vogler & Goldenstein 05] C. Vogler, S. Goldenstein: Analysis of Facial Expressions in American Sign Language. In *Proceedings of the 3rd Intl. Conf. on Universal Access in Human-Computer Interaction (UAHCI)*, Las Vegas, Nevada, USA, July 2005.

[Vogler & Metaxas 97] C. Vogler, D. Metaxas: Adapting Hidden Markov Models for ASL Recognition by Using Three-dimentional Computer Vision Methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 156–161, Orlando, FL, 1997.

[Vogler & Metaxas 99] C. Vogler, D. Metaxas: Parallel Hidden Markov Models for American Sign Language Recognition. In *Proceedings of the International Conference on Computer Vision*, Kerkyra, Greece, sep 1999.

[Wessel & Ortmanns$^+$ 97] F. Wessel, S. Ortmanns, H. Ney: Implementation of Word Based Statistical Language Models. In *SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pp. 55–59, Pilsen, Czech Republic, April 1997.

[Zahedi & Dreuw$^+$ 06a] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, H. Ney: Using Geometric Features to Improve Continuous Appearance-based Sign Language Recognition. In *Proceedings of BMVC 06, 17th British Maschine Vision Conference*, Vol. 3, pp. 1019–1028, Edinburgh, UK, 2006.

[Zahedi & Dreuw$^+$ 06b] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaesrs, J. Bungeroth, H. Ney: Continuous Sign Language Recognition – Approaches from Speech Recognition and Available Data Resources. In *Proceedings of the 5th Iternational Conference on Language Resources and Evaluation*, Genoa, Itely, 2006. In press.

[Zahedi & Keysers$^+$ 05a] M. Zahedi, D. Keysers, T. Deselaers, H. Ney: Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition. In *Proceedings of DAGM 2005, 27th Annual meeting of the German Association for Pattern Recognition*, Vol. 3663 of *LNCS*, pp. 401–408, Vienna, Austria, 2005. Springer Verlag.

[Zahedi & Keysers$^+$ 05b] M. Zahedi, D. Keysers, H. Ney: Appearance-Based Recognition of Words in American Sign Language. In *Proceedings of IbPRIA 2005, 2nd Iberian Conference on Pattern Recognition and Image Analysis*, Vol. 3522 of *LNCS*, pp. 511–519, Estoril, Portugal, June 2005. Springer Verlag.

[Zahedi & Keysers$^+$ 05c] M. Zahedi, D. Keysers, H. Ney: Pronunciation Clustering and Modelling of Variability for Appearance-Based Sign Language Recognition. In *Proceedings of GW 2005, 6th Interactional Gesture Workshop*, Vol. 3881 of *LNAI*, pp. 68–79, Berder Island, France, May 2005. Springer Verlag.

[Zhu & Yang$^+$ 00] X. Zhu, J. Yang, A. Waibel: Segmenting Hands of Arbitrary Color. In *Automatic Face and Gesture Recognition*, pp. 446–453, Grenoble, France, March 2000.

[Zolnay & Schlueter$^+$ 05] A. Zolnay, R. Schlueter, H.Ney: Acoustic Feature Combination for Robust Speech Recognition. In *ICASSP 2005, Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 457–460, Philadelphia, PA, USA, 2005.

Lebenslauf — Curriculum Vitae

Angaben zur Person

Name	Morteza Zahedi
Anschrift	Shahrood University of Technology, Shahrood, Iran
E-Mail	zahedi@ganjineh.co.ir
Geburtsdatum	25. June 1974
Geburtsort	Gorgan, Iran
Staatsangehörigkeit	Iraner
Familienstand	verheiratet, eine Tochter

Studium

Sep. 1992 – Sep. 1996	Computer-Engineering-Studium mit Schwerpunkt Hardware an der Amirkabir University of Technology Abschluss: Bachelor of Science
Sep. 1996 – Juni 1998	Computer-Engineering-Studium mit Schwerpunkt künstliche Intelligenz und Robotik an der Teheran University Abschluss: Master of Science mit Auszeichnung Titel der Arbeit: "The Design and Implementation of an Intelligent Program for Punctuation in Persian Texts"
seit November 2003	Promotionsstudent an der RWTH Aachen

Arbeitstätigkeiten

Sep. 1998 – Sep. 2003	Dozent für Informatik an der Shahrood University of Technology