CHAPTER -1

Introduction

During the recent years human-machine-interfaces have experienced a growing interest. Systems for the analysis of body motion have been developed carefully as a first step for gesture interaction between the user and the computer. **A** special case of body motion is gestures, i.e. motion of hands and **arms.** Gestures can express a certain meaning - such as pointing gestures or command gestures - and support the verbal communication. If a community of deaf people has assigned defined meanings to certain gestures, they are called signs, being part of a sign language. Many sign languages, like Sign Language of the Netherlands, ***are*** characterised by manual and non-manual parameters **[2].** Manual parameters are handshape, orientation, location and movement of hands. Non-manual parameters are line of sight, facial expression, poise, etc. If two signs only differ in one parameter they are called a minimal pair. Signs can be divided into one-handed and two-handed signs. For one-handed signs the action of only one hand is required, where a person generally takes the same hand, known as the dominant hand. For two-handed signs, the other hand, called the non-dominant hand, performs the sign together with the dominant hand. By now many systems use datagloves as input devices for the recognition of gestures or sign language, e.g. **[SI.** These approaches suffer from limiting the user's freedom of movement.

Video-based techniques are less intrusive and therefore more comfortable to utilise. The developed system should only use a single colour camera in order to minimise the necessary hardware components **and** the resulting adjustments. The following problems, resulting from the above requirements, must be taken into account: Signs vary in time and space. Even if a person tries to **perform** the same sign twice, slight changes of speed and position of the hands will occur. ***As*** the system **uses** only **a** single camera, the 3D-space is projected on a 2D-plane, resulting in **loss** of depth information. While signing some fingers can be occluded, as they are hidden behind other parts of the hands or **arms.**

The position of the test person in front of the camera can vary. **A** shift of the signer in one direction and rotation around his body axis must be considered. The property of

Hidden Markov Models (HMMs) to compensate time and amplitude variances of signals has been proven for speech and character recognition *[9]*. This property makes HMMs appear an ideal approach for sign language recognition.

Like speech, sign language can be considered as a nondeterministic time signal, not as a word or phoneme sequence but as a sequence of signs. Unlike speech recognition, where the smallest unit is the phoneme, linguists have not agreed on first proposals on sub-units for signs **[3].** Thus we model each sign with one HMM. **For** both, training and recognition, feature vectors must be extracted from each video frame and then inputted to the HMM.

The presented system for the recognition of isolated signs only regards the manual parameters, as non-manual parameters often have grammatical functions or emphasise emotions. The system is capable of recognising **262** different signs of the Sign Language **of** the Netherlands. The aim is signer dependent recognition, i.e. the same person trains and tests the system. This paper is structured as follows. Section **2** gives an overview of related work in this field. Section **3** mentions important details of the theory of HMM and section **4** explains how we adapted this theory to sign language recognition. **A** description of the experiments and their results can be found in section 5 and in the last section we summarise the main issues discussed in this paper.

1.1. Related Work

Different methods have been used for gesture recognition. Cui and Weng **[5]** recognise 262 hand gestures with a recognition rate of 93%. They analyse the shoulder-chest area and segment the handshape from the image. Therefore regions-of-interest in different sizes are compared with trained patterns. The final decision about the gesture is made using a nearest neighbor algorithm.

Yamato et al. [13] presented the first HMM approach recognizing six tennis strokes with a rate **of** about 90% using a 25x25 pixel sub-sampled video image as the feature vector. ***An*** unusual approach is reported from Schlenzig et al. [lo]. They use a single universal HMM and a finite state estimator for the determination of gestures. Hand poses are classified with a neural net. Sequences of gestures of six static hand poses are used to control a robot vehicle, yielding classification rates of 96%.

1.2. Theory Of Hidden Markov Model

Given a set of N states **si** we can describe the transitions from state to state at each time step t **as** a stochastic process. The transition probability to reach state si in the first time step is denoted **as q.** Assuming that the transition probability **aij** of state **si** to state **sj** only depends on the preceding states, we call this process a Markov chain. The further assumption, that the actual transition only depends on the very preceding state leads to a first order Markov chain.

We can now define a second stochastic process that produces, at each time step t. symbol vectors x. The emission probability **of** a vector x only depends on the actual state, but not on the way the state was reached. The emission probability density bi(x) for vector x at state **si** can either be discrete or continuous.

This doubly stochastic process is called a Hidden Markov Model (HMM) if only the vectors x are observable, but not the state sequence. A HMM ***h*** is defined by its parameters ***h*** = ***(x,* A, B).** ***x*** stands for the vector of the initial transition probabilities **q,** the **NxN** matrix A represents the transition probabilities a, from state **si** to **sj** and finally, **B** denotes the vector of the emission densities bi(x) of each state 4.

Having defined the HMM we have to cope with the problem. that given the parameters of a HMM ***h*** and an observation sequence ***0*** of vectors 0, of the signal, how to determine the state sequence, that best models the signal. In other words, how to find the state sequence, that emits, with a high probability, the same symbol vectors **as** observed from the signal. This problem can be **solved** with the Viterbi algorithm.

1.3 Hidden Markov Model For Sign Language Recognition

Having discussed the theory of HMMs the question arises. How one *can* find the observation and state sequence in sign language. Fig. 1 illustrates the modelling of a sign with a HMM.The upper line shows four images **of** the sign "WOFUR" as observed by the video camera. For each image an observation vector **is** extracted, which is - for simplicity - displayed only with two features, representing the position **of** the right hand. Assuming that always four different images are recorded for this sign it would **be** suitable to choose a linear four-state HMM, with transitions only from one state to the next. For the training procedure, several four-state image sequences would be recorded and **as** the assignment of the feature vectors to the states would be obvious, emission distributions for every feature of every state could be calculated. Dropping the assumption of a constant image number, the signer is allowed to perform the sign slower or faster. Choosing the Bakis topology for the HMM with additional transitions as displayed in Fig. 1 the system can compensate different speed of signing. This model allows transitions to the same state, the next and the one after next, and has been frequently used for speech recognition ***[9].***

Now the assignment of the vectors to the states is no longer trivial. i.e. it **is** hidden.

1.4. Feature Extraction

As the HMMs require feature vectors, an important step is the determination and extraction **of** features. In order to allow real time data acquisition and to easily retrieve information about the performed handshape, the signer wears simple coloured cotton gloves. Taking into account the different amount of information represented by the handshape **of** the dominant and non-dominant hand and the fact that many signs can be discriminated only by looking at the dominant hand, different gloves have been chosen: one with seven colours – marking each finger, the palm and the back of the dominant hand – and a plain glove in an eighth colour for the non-dominant hand.

A threshold algorithm generates idout-code for the colours of the gloves, skin, body and background. In the next processing step the size and the centre of gravity (COG) of the colour areas are calculated and a rule-based classifier estimates the position of the shoulders and the central vertical axis **of** the body silhouette **[7].** Using this information we build a feature vector that reflects the manual parameters of sign language, without explicitly modelling them. In other words at the end of the recognition the system outputs the whole sign, but not the characterising parameters.

Tab. 1 shows how the parameters of sign language are represented by the feature vector. Extracting the position of the hands during signing, we obtain the location of a sign. Thus the COGS of each hand are taken into account, where the COG for the domnant hand is determined **as** the mean position of all **COGS** of its colour areas. To compensate the shift variance of the signer in front of the camera the co-ordinates are related to a fixed point on the body. The x-co-ordinates are calculated relative to the central vertical axis of the body silhoutte, the yco- ordinates to the height of the right shoulder. **As** we use only one camera, the z-co-ordinate is neglected.

An optimal representation of the handshape of the dominant hand must be shift and rotation invariant. .

CHAPTER -2

Training

Having defined the feature vector HMMs can be trained. The Viterbi training of HMMs for isolated signs is shown in Fig. 2. The first step is the determination of the number of states of the HMM. A fixed number of states for all signs is not suitable, **as** the database contains very short signs with around four image frames and different, longer signs with about 30 frames. Even the length of one sign can vary considerably. Therefore the number of vectors in the shortest training sequence is chosen **as** the initial number of states for the HMM of the corresponding sign. Next, the system assigns the vectors of each sequence evenly to the states and initialises the matrix A, i.e. all transitions are set equally probable. Using the initial assignment the mean and deviation values of all components of the emission distributions of each state can be calculated. In later iterations, when the algorithm created more than one mixture component per state, in this step the deviation is pooled over all components

The initial HMM **A(')** is now complete. **as** xl=l for a Bakis-HMM. Given ***I.")*** the Viterbi algorithm determines a new assignment **q(")*** for each training sequence ***O(")***W**.** ith **q(")*,** the transition probabilities A and, in the next iteration, the mean and deviation values are updated. In the next step the algorithm checks each mixture component of the emission distributions, if it is preferable to split it in two.

The criterion is denoted by where **MU)** is the number of mixture components of state **j** and U is the number of training sequences. The negative logarithm of a probability, **as** used in the above equation, is also called score. Thus 8, stands for the sum of all scores, produced by component m of state **sj** for all paths **q(")*.** In other words e, is a measure for the probabilities of component m and the right side of the inequality is the mean of this measure over all components. Next the procedure decides, if the **HMM** produces the observation vectors ***O(")*** with a sufficient probability. We use the criterion where n denotes the iteration number. The algorithm **recurses** until the difference between the mean score of all U training sequences of iteration n and the mean score of all U training sequences of the previous iteration n-1 falls below a limit. Finally the algorithm verifies the initial hypothesis **for** the number of states. In

particular, at the beginning and also at the end of a sign there are similar feature vectors that **need** not be represented by separate states. Thus, states are deleted if the number of vectors that are assigned to a state are not sufficient or if the condition is met. All components n of state **si** are compared with all components m of state **sj.** If the smallest score sum of the other component falls below a limit, a state is deleted. In this case the parameters must be updated, otherwise the procedure has come to an end and the **HMM** can be stored for recognition.

CHAPTER -3

Recognition

If all **HMMs** & have been trained, signs can be recognised.Using the extracted observation sequence 0 of the sign to recognise the Viterbi procedure sequentially calculates the probabilities P*(Olu for all **HMMs.** Choosing the highest P*(01&) the most probable **HMM,** i.e. sign, can be identified. **As** with the training, the path probabilities P* must be scaled in order to avoid exceeding the precision range of the computer. Therefore we use the natural logarithm of P* for calculation.

CHAPTER -4

Experiments

The experimentation system consists of a CCD video camera and a Pentium PC with an integrated modular image processing system allowing the calculation of the feature vectors at a processing rate of 13 frames per second. In order to ensure correct segmentation, there ***are*** few restrictions for the clothing of the signer and the background must have one colour only **@I.**

The vocabulary of the database consists of **262** signs representing words from ten word types such as nouns, verbs, adjectives etc. The choice of signs used was aimed at simple stories to **be** told without avoiding minimal pairs. The tests were carried out by two persons, who learned the signs for the experiments. Each person performed signs firstly for the training database and then repeated them for the testing database.

Tab. **2** shows how the different training sets were built. For example Person **1** performed ten samples for each of the **262** signs, i.e. **2620** samples altogether. Set Training 3 in Tab. **2** contains the training data of both persons yielding a total num**ber** of 3930 samples. The test sets displayed in Tab. 3 were built in the same way **as** the training sets, only that Person 1 performed fewer repetitions per sign. For the recognition task, two additional sub-sets were built of each training set and each test set with a vocabulary size of 43 and 150 of the **262** signs. Tab. **4** shows the results. If Person 1 trains and tests the system, the recognition rate amounts to 98.8%. choosing from **43** different signs. If the system has to decide between **262** signs it still recognises 91.1% of the signs.

Chapter -5

Conclusion

We present a video-based system for the recognition of isolated signs with little intrusion **on** the signer. The system is equipped with a single camera in order to minimise the hardware requirements. We described how the HMM theory adapts to sign language recognition and presented details for the training and recognition procedures. With a feature vector of relatively simple features the results prove that even 262 different signs from two signers can be discriminated with a high probability.

Reference

[1] M. Assan, "Videobasierte **Gebiirdenspracherkennung** mit **Hidden-Markov-Modellen", *Diplomarbeit,*** Lehrstuhl **fiir** Technische Informatik. Aachen, 1997.

[2] P. Boyes Braem ***Einjiihrung in die Gebdrdensprache*** *und* ***ihre Erforschung,*** Signum Verlag, Hamburg, 1995.

[3] D. Brentari. "Sign Language Phonology: ASL. Handbook ***of Phonological Theory.*** Basil Blackwell. New York,

[4] LW. Campbell, D.A. Becker, A. Azarbayejani, A.F. **Bo**bick and A. Pentiand. "Invariant features for 3-D gesture recogniton", Proc. of ***the* 2. *Int'l. Con$* on *Autom. Face*** & ***Gesture Rec.,*** Killington, 1996, pp. 617-621.

[5] Y. Cui and J.J. Wenig, "View-based Hand Segmentation and Hand-Sequence Recognition with Complex Backgrounds",***Proc. oflCPR.*** 1996, pp. 617-621.

[6] K. Grobel and H. Hienz, "Video-based recognition of fingerspelling in real-time", ***Proc. des Workshops Bildverarbeitungfiir die Medizin,*** Aachen, 1996, pp. 197-202.

[7] K. Grobel und H. Hienz, "Videobasierte Erkennung von Korperregionen zur Bestimmung der Ausfihrungsstelle einer Gebiirde", ***Proc des*** 9. ***Aachener Kolloquiwns Signaltheorie*** M&z 18-20, 1997.

18) H. Liang and M. Ouhyoung, "A sign language recognition system using Hidden Markov Model and Context Sensitive Search,ACM ***VRST,*** 1996.

[9] R. Rabiner, "A Tutorial on Hidden Markov Models and selected applications in speech recognition", ***Proc.*** of ***the IEEE,*** vol. 77, **no.** 2, 1989. pp. 257-285.

[lo] J. Schlenzig, E. Hunter and R. Jain, "Video Based Hand Gesture interpretation Using Recursive Estimation". Proc ***ofACSSC,*** Pacific Grove, Nov. 1995, pp.1267-1271.

[111 E.G. Schukakat-Talamanzini, ***Auromarische Spracherkennung,*** Vieweg Verlag, Braunschweig. 1995, pp. 121-164.

[12]T.E. Stamer and A. Pentland. "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models", ***Technical Report no. 375.*** MIT Cambridge, Media Laboratory. 1995.

[13] J. Yamato, J. Ohya and K. Ishii. "Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model".