in: Proc. of the IEEE International Conference on Intelligence, Information and Systems (ICIIS'99), Washington 1999, pp. 336-341, IEEE Press

# A Hybrid Stochastic-Connectionist Architecture for Gesture Recognition

Andrea Corradini, Horst-Michael Gross
Technical University of Ilmenau
Department of Neuroinformatics
D-98684 Ilmenau, Federal Republic of Germany
andreac@informatik.tu-ilmenau.de

## Abstract

*In this paper an architecture for the recognition of dynamic gestures is described. The system implemented is designed to take a sequence of images and to assign it to one of a number of discrete classes where each of them corresponds to a gesture from a predefined vocabulary.*

*The classification task is broken down into an initial preprocessing stage following by a mapping from the preprocessed input variables to an output variable representing the class label. The preprocessing stage consists in the extraction of one translation and scale invariant feature vector from each image of the sequence. Further we utilize a hybrid combination of Kohonen Self Organizing Map (SOM) and discrete Hidden Markov Models (DHMM) for mapping an ordered sequence of feature vectors to one gesture category. We create one DHMM for each movement to be detected.*

*In the learning phase the SOM is used to cluster the feature vector space. After the self-organizing process each codebook is quantized into a symbol. Every symbol sequence underlying a given movement is finally used to train the corresponding Markov model by means of the non discriminative Baum-Welch algorithm, aiming at maximizing the probability of the samples given the model at hand.*

*In the recognition phase the SOM transforms any input image sequence into one symbol sequence which is subsequently fed into a system of DHMMs. The gesture associated with the model which best matches the observed symbol sequence is chosen as the recognized movement.*

*Preliminary experiments with our baseline system achieved a recognition accuracy of about 82%. The data was gathered from four people performing five repetitions of each of five gestures from a predefined vocabulary. The system uses input from a monocular color video camera, is user-independent but not real-time.*

## 1. Introduction

Gestures are part of everyday natural human communication. They are used as an accompaniment to spoken language and as an expressive medium in their own right. Recently, there have been strong efforts to develop intelligent, natural interfaces between users and systems based on gesture recognition. The optimal interaction has to be natural, intuitive, not require any remembrance and is similar to that we are familiar, thus the interaction with other people. The operational area of such intelligent interfaces covers a broad range of application fields in which an arbitrary system is to be controlled by an external user or in which system and user have to interact immediately [10, 13, 15].

One of the crucial problems in recognition of gestures is to deal with the varying temporal and spatial structure of dynamic gestures. The difficulty of gesture recognition stems from the high variability of each movement associated with a gesture to be detect. Gesture's segments may overlap, have varying lengths, and vary across speakers. Even the same user is not ever able to produce exactly the same movement for the same gesture. Moreover the complexity of the automatic recognition task is related to robustness to environmental conditions, vocabulary size, number and movement characteristics of users in user independent recognizers, and so on.

Throughout this paper the following definitions are considered.

**Definition 1 (Posture/Pose)** *A posture or pose is a couple determined by the only static hand locations with respect to the head position. The spatial relation of face and hands determines the behavioral meaning of each posture.*

**Definition 2 (Gesture)** *A gesture is a series of postures over a time span connected by motions.*

This paper is structured as follows. Starting from our saliency system for person localization [9], in the next section we provide an overview of the process which is to be

0-7695-0446-9/99 $10.00 © 1999 IEEE

carried out to describe the user's postures. We propose to combine skin color-based image segmentation with shape analysis by means of invariant moments. Section 3 mentions some basic ideas of the theory of SOM and DHMM and how we exploit these tools to gesture recognition while in Section 4 an alternative stochastic architecture is suggested. Finally, a description of the preliminary results and some final considerations can be found in Section 5 and Section 6, respectively.

## 2 View-based Posture Description

### 2.1. Posture Segmentation

We think that a good person localization task is essential for any further gesture recognition process. In [5, 8] we proposed a multi-cue approach consisting of three feature modules sensitive to *skin color*, *facial structure* and *structure of the head-shoulder-contour* respectively.

In the above mentioned work the three cues were assumed to be of equal importance. After a period of practical experiences we had to face that the shape-based approach supplies contribution to the localization process much more confident in contrast to the skin color and facial structure cues. The reasons are quite obvious: skin color detection is highly influenced by illumination and therefore its robust detection cannot be ensured in general. Further, solving a localization problem is particularly of interest if a person is rather distant. Necessarily, relevant features should appear even on rather coarse resolutional scales so that details, as facial structures, are less appropriate. Facial structure can be detected confidently only if the distance between person and camera is not too large. Otherwise, the region covered by the face becomes to small to be localized.

Against this background, the method for head-shoulder contour detection was improved significantly. Since the other cues can only support the person localization, but cannot ensure the localization alone, their methods were reduced to rather simple, but computationally efficient algorithms.

The utility of the different parallel processing cue modules is to make the saliency system robust and independent of the presence of one certain information source in the images. Hence, we can handle varying environmental circumstances much easier, which, for instance, make the skin color detection difficult or almost impossible. Due to its reliability and robustness against varying environmental conditions, that system represent the starting point for any further precessing step.

After detecting the location of the head and considering a subregion around it (Fig. 1,a), we characterize the distribution of the pixel values inside that window by a multidimensional normal distribution function. This function represents a parametric model for the skin color and is unequivocally described by a mean vector $\mu$ and a covariance matrix $\Sigma$ (the parameters). Using the Mahalanobis distance between the mean vector and an image pixel, this latter is classified to be or not a member of the skin class according to a threshold value.

From the resulting binary image (Fig. 1,b) we determine the centers of gravity (COG) of the hand and head regions (which we suppose to be the three greatest ones). Then to avoid problems deriving by the shape of each region due to the choice of the color threshold, we model each of these regions as a circle around their COG (Fig. 1,c).

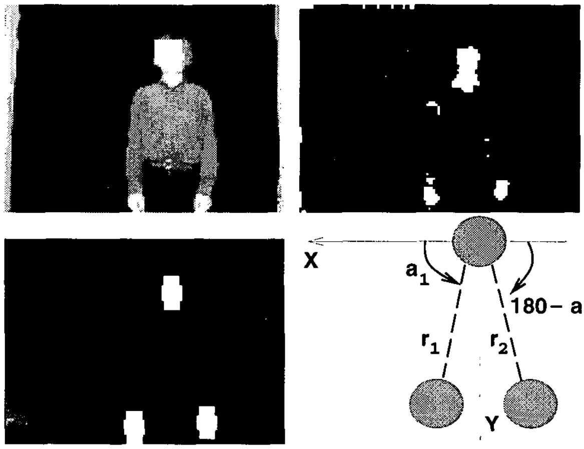

**Figure 1. From left to right and from up to down: head localization result, thresholded skin classification by means of an adapted color model derived from the pixel distribution around the head location, modeling of the hand and head regions as circle around their centers of mass, and finally the new defined coordinate system with origin centered at the center of gravity of the head.**

### 2.2. Feature Extraction

Because the decision on whether to classify an image sequence should not depend on where and how far in the image the user performing a gesture is located, our system is expected to exhibit as well translation as scale invariance. Therefore from that binary image (Fig. 1,c) we compute a feature vector $\vec{v}$ containing 13 translation and scale invariant elements characterizing the shape of the segmented scene. As first 9 feature vector elements we calculate the so-called *scale normalized moments* up to the third order. They remain unchanged under translation and size change.

Given the $(p+q)$th order central moment $\mu_{pq}$, the scale normalized moment of the same order is defined as

$$\nu_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+\frac{p+q}{2})}} \qquad (1)$$

The computation of them for binary image yields theoretically an error-free estimate of the continuous moments which is also independent of illumination as opposed to the value deriving from greyvalue images.

Further, to compensate the shift variation of the person gesticulating in front of the camera we choose for each image a suitable coordinate system by fixing its origin point at the current determined head's center of mass. It allows to calculate the remaining four feature vector elements relating to the head position and regardless to the user's position within the image. In this new coordinate system in order to ensure invariance also with respect to image size change, we use the polar coordinates of both hands's COG. If $(r_1, a_1)$ and $(r_2, a_2)$ represent the polar coordinates of the hands's COG (Fig.1,d) as last feature vector elements we take the four values $(\frac{r_1}{\max\{r_1,r_2\}}, a_1, \frac{r_2}{\max\{r_1,r_2\}}, a_2)$.

Finally, because the feature vector components have values which differ by several orders of magnitude we proceed with a rescaling of them. We perform the *whitening* linear rescaling [4] with respect to the 225 test patterns. This procedure do not treat the input variables independently but allow for correlations amongst the variables. In the transformed coordinates the data set has zero mean and a unit covariance matrix. In addition the input normalization ensures that all the feature vector elements are of order unity. In this case we can give to the network weights (Sec. 3) a suitable random initialization before training the network.

## 3 Hybrid SOM/DHMMs for Gesture Recognition

### 3.1 Self-Organizing Maps for Symbol Production

The goal of the posture analysis is the extraction of local features along the hand trajectory, yielding a sequence of time ordered multi-dimensional feature vectors. The further step is concerned with the quantization of that feature vectors into a sequence of symbols.

A Self Organizing Map (SOM) [14] is used to preserve the topology of the high-dimensional feature space by mapping the feature vectors to a two-dimensional space. Due to the sequential nature underlying each gesture such a topology-preserving map can be exploited to constitute trajectories where the SOM best-matching neurons are recorded during the process. The SOM clusters the unlabeled training feature vectors which lie near one other in the feature space. During the training phase as well the codebook vector most sensitive to the actual training vector as those in its (variable) neighborhood, are tuned maintaining a well-balanced set of weight values with respect to the input density function.

The weight adjustment is carried out using the Euclidean distance between the actual multi dimensional input vector and the connecting weight vectors, times a time-dependent learning rate, and times a neighborhood function that decay like the Gaussian probability density function when the topological distance between the best-matching unit and the actual vector increases.

We start the learning process with a large radius covering all the units in order to prevent the formation of undesired outliers in the clustering due to the limited training data set. Our SOM has 800 units organized into a $(xsize = 40 \times ysize = 20)$ square array. The feature vectors are 13-dimensional and the SOM is trained by decreasing the neighborhood radius from 6 to 1 and the learning rate from the value 0.9 to 0 in $(100 * xsize * ysize)$ iterations.

After the clustering process each neuron of the network correspond to a cluster in the input feature space. Proceeding from the self-organizing process we tune the weight vectors using the unsupervised Learning Vector Quantization (LVQ) method causing the weights to approach the decision boundaries [14].

In order to utilize the SOM for classification we divide each gesture of our vocabulary in *subgestures* or *posture classes* and we label each of them with a different symbol (see Fig.2 for the hand-waving-right movement). We divide the gestures of our vocabulary into altogether 32 subgestures (9 for each left-,right-waving; 5 for each go left/right; 4 for stop). For class discrimination purposes we hand-label each SOM cluster. That labels were assigned to the units according to the subgesture subdivision (Fig. 2) by using the recorded training samples as input.

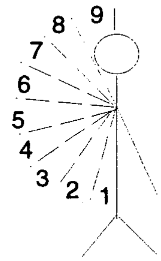

**Figure 2. Waving-right gesture hand-labeling. That movement is divided in 9 subregions each covering exactly 20 grad of the two-dimensional plane surface which the gesture is projected on. Each subregion is labeled by one symbol.**

### 3.2 Stochastic Recognition with HMMs

Hidden Markov Models [3, 17] are probabilistic finite state machines well-suited in dealing with the statistical and sequential nature of time-varying input patterns. They are the basis for a lot of applications especially in the field of speech recognition [1, 2, 6], and hand-writing recognition [7, 16].

A HMM consists of a finite number of states connected one other by directed arcs according to a predefined topology. Each arc is associated with one probability value that is called state transition probability. At regular time intervals the model undergoes a change of state according to the set of transition probabilities. Also a change back to the same state is possible. Each state compute the estimation of the likelihood for a certain input observation vector by means of a probability density distribution function which can be discrete or continuous. After defining also an initial state distribution, the HMM can be used as generator of sequence of observations or as model for how an observation sequence is generated by it.

There are two concurrent stochastic processes associated with each HMM: a set of state output processes that model the local stationary character of the observation at each time step, and the state sequence that models the temporal structure of the signal being modeled. Because this latter state sequence is not directly observable the Markov model is called 'hidden'.

In our work we used as many Hidden Markov Models as the number of gesture to be detected. The training and decoding of the models are based on the posterior probability $P(M|\mathbf{X}_0^t)$ that the feature vector sequence $\mathbf{X}_0^t$ has been produced by the model $M$. In the learning phase the set of parameters maximizing that probability are sought for every sequence $\mathbf{X}_0^t$ associated with the model $M$. This strategy is referred to as the Maximum a Posteriori (MAP) criterion [4]. During the recognition stage, given an observation sequence $\mathbf{X}_0^t$ and a fixed set of parameters the goal is to find out among many models the one model $M$ that maximizes $P(M|\mathbf{X}_0^t)$..

Unfortunately, the learning process generally does not consent to expressly characterize $P(M|\mathbf{X}_0^t)$ but permits the characterization of the probability $P(\mathbf{X}_0^t|M)$ that a given model generates certain feature sequences. Using the Bayes' rule one can express $P(M|\mathbf{X}_0^t)$ in terms of $P(\mathbf{X}_0^t|M)$ as

$$P(M|X_0^t) = \frac{P(\mathbf{X}_0^t|M)P(M)}{P(\mathbf{X})} \tag{2}$$

where $P(M)$ is the prior probability of the model, $P(\mathbf{X})$ is the prior probability of the vector sequence, and $P(\mathbf{X}_0^t|M)$ is referred to as the likelihood of the data given the model. Because $P(M)$ can be calculated without using the feature vector sequences, and $P(\mathbf{X})$ can be assumed constant since it does not depend on the models, the estimation of equation 2 amounts to calculating the only likelihood $P(\mathbf{X}_0^t|M)$. In that case, when the training criterion aims at the maximization of the quantity $P(\mathbf{X}_0^t|M)$, it is referred to as Maximum Likelihood (ML) criterion [4]. This is exactly the learning criterion we adopted.

### 3.3 Using Discrete HMMs for Classification

In subsection 3.1 we assigned each feature vector to a symbol which corresponds to a codeword in the codebook created by LVQ. The feature vectors of the data set for training were vector quantized. The need of a vector quantizer to map the continuous observation vectors into discrete symbols arises from the choice to use discrete Hidden Markov Models (DHMM) as recognizer.

For the choice of the model topology there is no theoretically way to rely on. The choices we made depend on the gesture being modeled. For each movement to be detected, we create one left-to-right DHMM (Fig. 3) with as many states as the subregions which this gesture is divided in. In such a model each DHMM state is associated with a single movement's subgesture (Fig. 2).

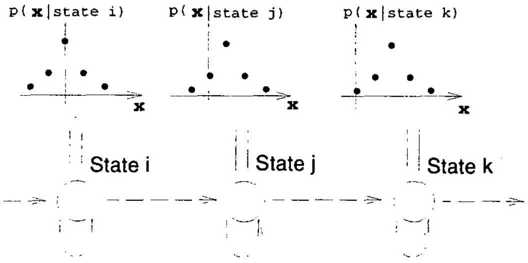

**Figure 3. Left-to-right discrete Hidden Markov model. This model is called left-to-right or Baskis model because it has the property that as time increases the state changes proceed from left to right. The dashed arrows depict the transition probabilities among the states. Here only transition from a state to the next one or to itself are allowed. The probability distribution functions assume discrete values.**

In the learning phase the parameters of each DHMM are optimized so as to model the training symbol sequences from the corresponding gesture. More precisely, the parameter of each model are estimated with symbol sequences of the according gesture samples applying the Baum-Welch training algorithm [3]. This latter is an iterative procedure based on the Maximum Likelihood criterion aiming at maximizing the probability of the samples given the model at hand and can be considered as a form of the Expectation-Maximization (EM) algorithm [11].

Because we consider a gesture as a sequence of subgesture the recognition process consists in comparing a given sequence of symbols with each DHMM. The gesture associated with the model which best matches the observed symbol sequence is chosen as the recognized movement.

## 4 Continuous HMMs for Automatic Gesture Recognition

Up to this point, we have considered the case when the observations were characterized as discrete symbols from a finite alphabet. In this situation we could use only discrete probability density functions within each model state. The main problem with this approach is the need to quantize the continuous feature vectors via codebooks. Because that quantization process might be accompanied by distortion or loss of information, it could be advantageous to utilize the HMMs with continuous observation density functions. In this case the model probability density functions are some parametric pdfs or mixture of them. The most common parametric pdf used is the mixture of Gaussian density which can be expressed for a generic state $i$ as

$$p_i(\mathbf{X}) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\mathbf{X}, \mu_{im}, \mathbf{\Sigma}_{im}) \qquad (3)$$

where $M$ is the number of mixtures ($M = 3$ in our experiments), $X$ is the vector being modeled, $c_{im}$ is the mixture coefficient for the $m$th mixture in state $i$ and $\mathcal{N}$ is any strictly log-concave or elliptically symmetric density function with covariance matrix $\mathbf{\Sigma}_{im}$ and mean vector $\mu_{im}$ in state $i$ for the $m$th mixture.

With $D$-dimensional data (here $D = 13$ is the dimension of the observation vectors) and using the Gaussian function as parametric pdf, the function $\mathcal{N}(\mathbf{X}, \mu_{im}, \mathbf{\Sigma}_{im})$ in equation 3 can be expressed as

$$\mathcal{N}(\mathbf{X}, \mu_{im}, \mathbf{\Sigma}_{im}) = \frac{e^{(-1/2^t(\mathbf{X}-\mu_{im})\mathbf{\Sigma}_{im}^{-1}(\mathbf{X}-\mu_{im}))}}{(2\pi)^{D/2}|\mathbf{\Sigma}_{im}|^{1/2}} \qquad (4)$$

As the dimension of the feature vectors increases as well the length of the mean vectors as the size of the covariance matrices becomes greater. But while the increase in size of the mean vectors is proportional to the one of the observation vector, the enlargement in size of the covariance matrices is even square proportional to the vector dimension. Hence, with multi-dimensional observation vectors the number of parameters of the mixture of Gaussian is very large and its estimation becomes computationally excessive. In addition with insufficient training data some of these parameters to estimate will assume more or less arbitrary values. A good way to avoid a huge number of parameters and, at the same time to have representative models, consists in approximating the covariance matrices by diagonal matrices. Under that simplification the model parameters can be estimated faster maximizing the likelihood of the data using another time the Baum-Welch learning algorithm [3].

## 5 Preliminary Results

To train and test each HMM in both discrete and continuous case, we gathered the data from four people performing five repetitions of the gesture to be described. The categories to be recognized are five. Therefore we take the same number of left-to-right Markov models each corresponding exactly to one class.

| Gesture | % of not classified patterns | % of false classified patterns | Recognition rate in % |
|---|---|---|---|
| stop | 9.2 | 13.2 | 77.6 |
| waving right | 8.4 | 11.1 | 80.5 |
| waving left | 8.7 | 10.0 | 81.3 |
| go right | 9.6 | 8.6 | 81.8 |
| go left | 10.2 | 9.6 | 80.2 |

**Table 1. Recognition results using discrete Hidden Markov Models.**

The performances were captured by a color camera at a frequency of 25 frames per second and digitized into 120 × 90 pixel RGB images. Table 1 summarizes the achieved performance concerning the recognition task by utilizing a recognizer based on the SOM/DHMM hybrid architecture; Table 2 the recognition performance achieved by using only CHMMs.

| Gesture | % of not classified patterns | % of false classified patterns | Recognition rate in % |
|---|---|---|---|
| stop | 10.4 | 10.0 | 79.6 |
| waving right | 7.3 | 10.3 | 82.4 |
| waving left | 8.8 | 8.5 | 82.7 |
| go right | 7.4 | 7.8 | 84.8 |
| go left | 8.1 | 8.0 | 83.9 |

**Table 2. Recognition results using continuous Hidden Markov Models.**

We consider an input as not classified if after feeding it into each HMM either the difference between the highest and the second highest output is not over an heuristically determined threshold or all the outputs are under a given threshold.

From a direct comparison of the recognition rates we can see how the CHMM-based system leads better results than the hybrid SOM/DHMM-based one. We think that this is mainly due to the continuous inherent character of the feature vectors. The conversion of them into discrete symbols via vector quantization can namely worsen the recognition task.

## 6 Conclusions

So far, both methods proposed for gesture recognition were tested on a small set of simple gestures and thus have very limited scope. We are currently extending both systems in order to overcome this limitation. The aim is to design a system that can work with a large "vocabulary" of gestures, and remain user independent. The performances of the two architectures depend strongly on the number of training pattern and also how well that patterns are representative for each class. It means that the training patterns have to cover the maximum test pattern range as possible.

If one the one hand HMMs provide a good representation of the sequential nature of the human movements, on the other they suffer from several limitations and drawbacks because of the assumptions exploited for the implementation of their learning and decoding algorithms [6]. We refer, for example, to the strong statistical assumption that the probability density functions associated with the states can be described by a fixed parametric function. Again, it is supposed every state change to depend only on the current and previous state and not on all the predecessor ones (*first-order HMM*). Also the likelihood of an observation vector is assumed not to depend on the previous observations but only on the current state (*context-independent assumption*).

In addition, HMMs consider the sequence of feature vectors as a piecewise stationary process. Hence, even though gesticulating is a non-stationary process we have to assume that over a short period of time the statistics of the movement underlying the gesture do not differ from sample to sample neglecting the correlations between successive feature vectors (*statistical time-independence of the observation vectors*). HMMs trained with the non-discriminative Baum-Welch algorithm show also poor discriminative capability among different models. Namely, by maximizing the Maximum Likelihood instead of the Maximum a Posteriori the HMMs are trained only to generate high probabilities for its own class and not to discriminate against models.

Due to their inherently discriminant nature and lack of distributional assumptions it is our intention to further use Neural Networks as emission probability for HMM states.

## References

[1] L. Bahl, F. Jelinek, and R. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transaction on PAMI*, 5(2):179–190, 1983.

[2] J. Baker. The dragon system - an overview. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975.

[3] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, (37):1554–1563, January 1966.

[4] C. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[5] H.-J. Boehme, U.-D. Braumann, A. Brakensiek, A. Corradini, M. Krabbes, and H.-M. Gross. User localisation for visually-based human-machine-interaction. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 486–491, 1998.

[6] H. Bourlard and N. Morgan. *Connectionist Speech Recognition.* Kluwert Academic Publishers, Dordrecht, The Nederlands, 1994.

[7] S.-B. Cho. A hybrid method of hidden markov model and neural network classifier for on-line handwritten character recognition. *Proceedings of the 1991 International Conference on Artificial Neural Networks (ICANN)*, pages 741–744, 1991.

[8] A. Corradini, H.-J. Boehme, and H.-M. Gross. Visual-based posture recognition using hybrid neural networks. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN '99)*, pages 81–86, 1999.

[9] A. Corradini, U.-D. Braumann, H.-J. Boehme, and H.-M. Gross. Contour-based person localization by 3d neural fields and steerable filters. *Proceedings of the IAPR Workshop on Machine Vision Applications (MVA '98)*, pages 93–96, 1998.

[10] T. Darrell, S. Basu, C. Wren, and A. Pentland. Perceptually-driven avatars and interfaces: Active methods for direct control. *M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 416*, 1997.

[11] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1(B 39):1–38, 1977.

[12] K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, pages 179–187, 1962.

[13] R. Kahn. Perseus: An extensible vision system for human-machine interaction. *PhD-Thesis, University of Chicago, Departement. of Computer Science*, 1996.

[14] T. Kohonen. *Self-Organizing Maps.* Springer Verlag, 2nd Edition, 1997.

[15] D. Kortenkamp, E. Huber, and P. Bonasso. Recognizing and interpreting gestures on a mobile robot. *Proceedings of the 13th National Conference on Artificial Intelligence (AAI '96)*, pages 179–187, 1996.

[16] A. Kundu. Handbook of character recognition and document image analysis. *Handwritten Word Recognition using Hidden Markov Model*, pages 157–182, World Scientific Publishing Company, 1997.

[17] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, 1986.