Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

A Self-Organizing Map (SOM) [28] is used to preserve the topology of the high-dimensional feature space by mapping the feature vectors onto 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. A similar approach has been used by Waldherr [27].

The SOM clusters the unlabeled training feature vectors which lie near one another 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, a time-dependent learning rate, and a neighborhood function that decays 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 nodes organized in a 40×20 grid. The feature vectors are 15-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.

After the clustering process, each neuron of the network corresponds 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 [28].

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 figure 13 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. These labels were assigned to the units according to the subgesture subdivision as depicted in figure 13 by using the recorded training samples as input.

5.2. *Using DHMMs for classification*

In subsection 5.1 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 DHMM's 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 (figure 14) with as many states as

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

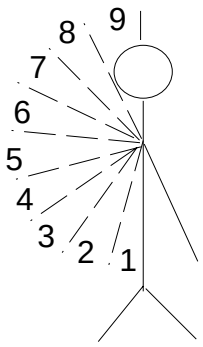

Fig. 13. Waving-right gesture hand-labeling. That movement is divided in 9 subregions each covering exactly 20 grad of the two-dimensional plane surface the gesture is projected on. Each subregion is labeled by one symbol.

the subregions which this gesture is divided in. In such a model, each DHMM state is associated with a single movement's subgesture (figure 13).
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 [19]. The 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* algorithm [29].

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

6. 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, these density functions are some parametric probability distributions or mixtures of them.

The most common parametric distribution used is the mixture of Gaussian density which can be expressed for a generic state i as

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

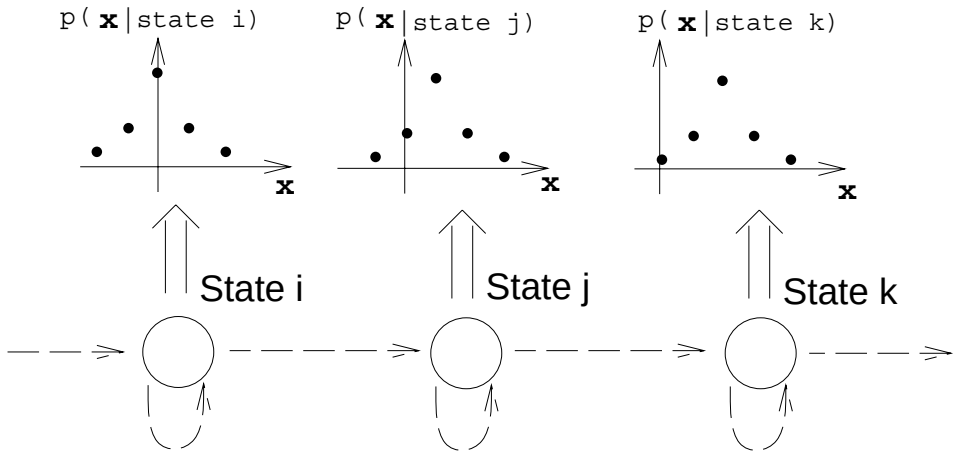

Fig. 14. Left-to-right DHMM. This model is called left-to-right or Baskis model because it has the property that as time increases the state changes proceed from left to right. The dashed arrows depict the transition probabilities among the states. Here only transitions from a state to the next one or to itself are allowed. The probability distribution functions assume discrete values.

$$p_i(\boldsymbol{X}) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\boldsymbol{X}, \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \tag{15}$$

where M is the number of mixtures ($M = 3$ in our experiments), $\boldsymbol{X}$ 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 $\boldsymbol{\Sigma}_{im}$ and mean vector $\boldsymbol{\mu}_{im}$ in state i for the m-th mixture.

With D-dimensional data (here $D = 15$ is the dimension of the feature vectors) and using the Gaussian function as parametric probability distribution, the function $\mathcal{N}(\boldsymbol{X}, \boldsymbol{\mu}_{\wr\Updownarrow}, \boldsymbol{\Sigma}_{\wr\Updownarrow})$ in equation (15) can be expressed as

$$\mathcal{N}(\boldsymbol{X}, \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) = \frac{e^{(-1/2(\boldsymbol{X}-\boldsymbol{\mu}_{im})^{\top} \boldsymbol{\Sigma}_{im}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_{im}))}}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_{im}|^{1/2}} \tag{16}$$

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 expensive. Additionally, with insufficient training data to estimate some of these parameters will assume more or less arbitrary values.

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

To avoid a huge number of parameters and, at the same time, to have representative models, we approximate the covariance matrices by diagonal matrices, and we tie it over the whole model. Under that simplification the model parameters can be estimated faster by using the Baum-Welch learning algorithm [19] again.

7. 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 HMM's each corresponding to one class.

Table 1. Recognition results using DHMM's.

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

The sequences 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 shows the recognition performance achieved by using only CHMMs.

Table 2. Recognition results using CHMM's.

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

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 if all the outputs are under a given threshold.

Compared with continuous models, discrete distributions normally require less parameters. This means that DHMM's have less memory requirements and need

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

less training data to achieve good generalization performance. Moreover, discrete models require shorter recognition and training time since they do not have to calculate any mixture of Gaussian distribution. For discrete models only quantization of the observation vectors has to be performed while the state probability estimation is replaced with a look-up table.

From a direct comparison of the recognition rates regarding our problem, we can see how the CHMM-based system leads to slightly better results than the hybrid SOM/DHMM-based one. We think that this is mainly due to the continuous intrinsic character of the feature vectors. The conversion of them into discrete symbols via vector quantization can worsen the recognition task. In spite of our experimental results, we do not state that CHMM's outperform SOM/DHMM-based recognizers in general. Due to the limited training data it would be a shaky conclusion, strongly dependent on the implementation and the few data at hand.

Anyway, the recognition rate of both systems can be improved by using a discriminative training algorithm instead of the Baum-Welch algorithm giving arise to a poor discriminative power among different models.

8. Conclusions and Outlook to Future Work

Besides the performance concerning posture recognition, the person localization is the most crucial but absolutely necessary prerequistite for the function of the whole system. The use of multiple cues and their integration into a selection process via 3D dynamic neural fields led to a satisfying person specific saliency system. Using a CHUGAI BOYEKI CD 08 video camera with maximum wide angle mode, the multiscale representation covers a distance from 0.5 to about 2.5 meters. Within this interval, the localization is very robust against slight rotations (up to 15°), scene content, and illumination. Furthermore, the integration of the omnidirectional camera makes it easier for the system to detect a person it its surroundings, which significantly speeds up the localization process.

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 larger 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.

On the one hand HMM's 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 [23]. 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

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

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).

HMM's 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 we are currently using and testing a system with neural networks to estimate the probability for HMM states.

The overall system has to be understood as work in progress, undergoing continuous changes. Currently, the major constraints are that the gesture recognition process does not work in real-time, whereas the localization process does, and that the posture segmentation uses only skin color which causes problems when other skin-colored objects are in the scene. The latter problem is currently to be eliminated by additionally using motion information, resulting in a search for moving skin color.

In the long run, we want to develop a continuous action-perception cycle between the robot and its human user in service system domains, where the architecture described here could be one building block.

Acknowledgments

This work was supported by the TMR Marie Curie Research Training Grant # ERB FMBI CT 97 2613 to A. Corradini.

References

[1] Darrell, T., Basu, S., Wren, C., and Pentland, A. Perceptually-driven Avatars and Interfaces: active methods for direct control. In *SIGGRAPH'97*, 1997. M.I.T. Media Lab Perceptual Computation Section, TR 416.

[2] Kahn, R. *Perseus: An Extensible Vision System for Human-Machine Interaction.* PhD thesis, University of Chicago, 1996.

[3] Kortenkamp, D., Huber, E., and Bonasso, P.R. Recognizing and interpreting gestures on a mobile robot. In *Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, 1996.

[4] Corradini, A., Braumann, U.-D., Boehme, H.-J., and Gross, H.-M. Contour-based person localization by 3dneural fields and steerable filters. *Proceedings of the IAPR*

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

Workshop on Machine Vision Applications (MVA '98), pages 93–96, 1998.

[5] Wren, C., Azarbayejani, A.and Darrell, T., and Pentland, A. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997. M.I.T. Media Lab Techreport TR 353.

[6] Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. User Localisation for Visually-based Human-Machine-Interaction. In *International Conference on Automatic Face- and Gesture Recognition*, pages 486–491. IEEE Computer Society Press, 1998.

[7] Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. Neural Networks for Gesture-based Remote Control of a Mobile Robot. In *International Joint Conference on Neural Networks*, volume 1, pages 372–377. IEEE Computer Society Press, 1998.

[8] Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[9] K. Kopecz. Neural field dynamics provide robust control for attentional resources. In *Aktives Sehen in technischen und biologischen Systemen*, pages 137–144. Infix-Verlag, 1996.

[10] Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture and Sign-Language in Human-Computer Interaction*, Lecture Notes in Artificial Intelligence, pages 219–232. Springer, 1998.

[11] Pomierski, T. and Gross, H.-M. Biological Neural Architectures for Chromatic Adaptation resulting in Constant Color Sensations. In *ICNN'96, IEEE International Conference on Neural Networks*, pages 734–739. IEEE Press, 1996.

[12] Jones, J.P. and Palmer, L.A. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology*, 56(8):1233–1258, 1987.

[13] Koenderink, J.J. and van Doorn, A.J. Receptive Field Families. *Biological Cybernetics*, 63:291–297, 1990.

[14] Freeman, W.T. and Adelson, E.H. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(9):891–906, 1991.

[15] Hunke, M.H. Locating and Tracking of Human Faces with Neural Networks. Technical report, Carnegie Mellon University Pittsburgh, 1994. CMU-CS-94-155.

[16] Sim, T., Sukthankar, R., Mullin, M., and Baluja, S. High"=Performance Memory"=based Face Recognition for Visitor Identification. Technical report, Carneghie Mellon University, Institute of Computer Science, 1999.

[17] Duda, R.O. and Hart P.E. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

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

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

[20] R. Rabiner L.' A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[21] Bahl L.R., Brown P.F., de Souza P.V., and Mercer R.L. Maximum mutual information estimation of hidden markov model parameters for speech recognition. *Proceedings of the IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP), Tokyo*, pages 49–52, 1986.

[22] J.K. Baker. The dragon system - an overview. *IEEE Transaction on Acoustics, Speech,*

Int. Journal on Artificial Intelligence Tools (IJAIT), vol. 9 (2000), no. 2, pp. 177-204

and Signal Processing, 23(1):24–29, 1975.

[23] H. Bourlard and N. Morgan. Connectionist speech recognition. *Kluwert Academic Publishers, Dordrecht, The Nederlands*, 1994.

[24] Cho S.-B. A hybrid method of hidden markov model and neural network classifier for on-line handwritten character recognition. *Kohonen T., M"akisara K., Simula O. and Kangas J. ed., Proceedings of the 1991 International Conference on Artificial Neural Networks*, pages 741–744, 1991.

[25] Amlan Kundu. *Handbook of Character recognition and Document Image Analysis*, chapter Handwritten word recognition using Hidden Markov Model, pages 157–182. World Scientific Publishing Company, 1997.

[26] Hienz, H., Bauer, B., and Kraiss, K.-F. HMM-Based Continuous Sign Language Recognition Using Stochastic Grammars. In *GW'99 - The 3rd Gesture Workshop, Gif"=sur"=Yvette, France*, pages 185–196. Springer, 1999.

[27] Waldherr, S., Thrun, S., and Romero, R. A gesture-based interface for human-robot interaction. *Autonomous Robots*, page to appear, 2000.

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

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