Facial Expression Recognition Using Pseudo 3-D Hidden Markov Models

Stefan Müller, Frank Wallhoff, Frank Hülsken[†]
Dep. of Computer Science, Faculty of Electrical Engineering
Gerhard-Mercator-University
47057 Duisburg, Germany
{stm, wallhoff, huelsken}@fb9-ti.uni-duisburg.de

Gerhard Rigoll
Institute for Human-Machine Communication
Munich University of Technology
80290 Munich, Germany
rigoll@ei.tum.de

Abstract

In this paper pseudo 3-D Hidden Markov Models (P3DHMMs) are applied to the task of dynamic facial expression recognition. P3DHMMs are an extension of the pseudo 2-D case, which has been successfully used for the classification of images and the recognition of faces. Although the application of P3DHMMs for image sequence recognition has been reported before, this paper provides a formal definition of the novel approach as well as a detailed explanation of a triple embedded Viterbi algorithm. Furthermore an equivalent one-dimensional structure is introduced, which allows the application of the standard Viterbi and Baum-Welch-Algorithms. The approach has been evaluated on a person independent database, which consists of 4 different facial expressions, performed by 6 individuals. The recognition accuracy achieved in the experiments is close to 90%.

1. Introduction

The modeling of entire image sequences using novel pseudo 3-D HMMs has first been described in [1]. In this publication the new approach has been evaluated on a crane signal database, which consists of 12 different predefined gestures for maneuvering cranes. It was found that the approach allows the recognition of dynamic gestures such as waving hands as well as static gestures such as standing in a certain pose. This was achieved by the use of statistically independent streams, which allow the integration of features derived from temporal as well as spatial data into a single model. Although the paper [1] introduced the P3DHMMs and showed first promising results, a formal definition of the model as well as details of the algorithms used have been omitted. The present paper provides a formal definition of P3DHMMs and a description of the triple embedded Viterbi algorithm as well as a description of the equivalent one-dimensional HMM structure. Furthermore, the P3DHMM approach is for the first time introduced to the task of dynamic facial expression recognition. Fig. 1 shows images taken from sequences that belong to our facial expression database. The sequences show individuals performing the facial expressions *anger*, *surprise*, *disgust* and *happiness*. Other publications which deal with facial expression recognition include [2, 3].

This paper is organized as follows. Section 2 gives a formal definition of the pseudo 3-D HMMs, describes the triple embedded Viterbi algorithm and introduces the equivalent one-dimensional HMM structure. Section 3 presents experimental results. A summary is given in Section 4.

2. Pseudo 3-D Hidden Markov Models

Hidden Markov Models are finite non-deterministic state machines which have been successfully applied to numerous applications. They consist of a fixed number of states with associated output density functions (pdfs) as well as transition probabilities a_{ij}. For a continuous HMM the pdf $b_j(\vec{o})$ of state S_j is usually given by a finite Gaussian mixture of the form:

$$b_j(\vec{o}) = \sum_{m=1}^{M} c_{jm}\, \mathcal{N}(\vec{o}, \vec{\mu}_{jm}, \Sigma_{jm}) \qquad (1)$$

where c_{jm} is the mixture coefficient for the mth mixture and $\mathcal{N}(\vec{o}, \vec{\mu}_{jm}, \Sigma_{jm})$ is a multivariate Gaussian density with mean vector $\vec{\mu}_{jm}$ and covariance matrix Σ_{jm}. A detailed explanation of the one-dimensional HMM-framework is given by Rabiner in [4]. Pseudo 3-D HMMs are an extension of the one-dimensional HMM paradigm, which have been developed in order to model three-dimensional data. The following section provides a formal definition of the P3DHMMs.

2.1. Definition of P3DHMM

When using a P3DHMM, the emission of observations is considered to origin from a three-staged hierarchical process. The topmost stage in this hierarchy models the dependencies of consecutive images using a first order Markov Model. The images itself are emitted by the well documented pseudo two-dimensional HMM (see e.g. [5, 6, 7]). Fig. 2 illustrates a pseudo 3-D Hidden Markov Model. The figure shows three so-called *hyperstates*, which are denoted as S_1 and S_2 and two pseudo 2-D HMMs which are assigned to those hyperstates. The P2DHMMs consist of

[†] F. Hülsken is now with the Fraunhofer Institute for Media Communication IMK, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

From: ICPR 2002, IEEE Computer Society, NY

Figure 1. Examples for the facial expressions anger, surprise, disgust and happiness (from left to right)

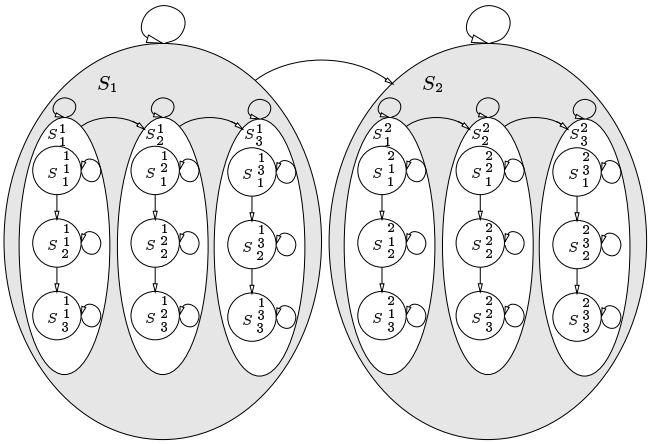

Figure 2. Pseudo 3-D Hidden Markov Model

three superstates (e.g. S_1^1 for hyperstate S_1) containing three states each (e.g. S_1^1,...,S_3^1 for superstate S_1^1).

A three-dimensional pattern O_{XYT} can be modeled by a P3DHMM in the following manner: each image ($o_{xyt}, t =$ const.) in the image sequence is assigned to a hyperstate, which results in a time warping of the pattern. Additionally, each image is modeled by a P2DHMM, which results in a non-linear warping of the image in both spacial directions. Each column ($o_{xyt}; t, x =$ const.) is assigned to a superstate which leads to a nonlinear warping in horizontal direction. The columns itself are modeled by an ordinary one-dimensional HMM and thus a warping in vertical direction is achieved.

A P3DHMM (L) is explicitely characterized by the following parameters: First the number of hyperstates (K) has to be chosen. As shown in Fig. 2 the hyperstates are denoted as S_j, where $j = 1, \ldots, K$. Transition probabilities and initial states probabilities have to be assigned to the hyperstates. Their definitions are as follows:

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i) \tag{2}$$

and

$$\pi_j = P(q_1 = S_j) \tag{3}$$

A P2DHMM is assigned to every hyperstate and thus the superscript ^{j} is added to every parameter of the P2DHMM in order to indicate the assignment to the j-th hyperstate. The P2DHMMs ($\Lambda_1, \ldots, \Lambda_K$) are characterized by the following parameters:

- L^j is the number of superstates ($S_1^j, \ldots, S_{L^j}^j$) of the P2DHMM which is assigned to hyperstate S_j
- The superstate transition probabilities a_{kl}^j

$$a_{kl}^j = P(q_{x,t} = S_l^j | q_{x-1,t} = S_k^j) \tag{4}$$

  where $q_{x,t}$ denotes the actual superstate at time t and for column x
- The initial superstate probabilities

$$\pi_i^j = P(q_{1,t} = S_i^j) \tag{5}$$

The conventional 1-D HMM which is assigned to the superstate has the following parameters:

- M_i^j is the number of states ($S_{i\,1}^{j\,i}, \ldots, S_{i\,M_i^j}^{j\,i}$) of the 1-D HMM which is assigned to hyperstate S_j and superstate S_i^j
- The transition probabilities $a_{i\,kl}^{j\,i}$

$$a_{i\,kl}^{j\,i} = P(q_{x,y,t} = S_{i\,l}^{j\,i} | q_{x,y-1,t} = S_{i\,k}^{j\,i}) \tag{6}$$

  where $q_{x,y,t}$ is a state at time t and position x, y.
- The initial state probabilities

$$\pi_{i\,k}^{j\,i} = P(q_{x,1,t} = S_{i\,k}^{j\,i}) \tag{7}$$

- The observation probabilities $b_{i\,k}^{j\,i}(\vec{o})$ for each state

$$b_{i\,k}^{j\,i}(\vec{o}) = P(\vec{o} | q_{x,y,t} = S_{i\,k}^{j\,i}) \tag{8}$$

  which can be expressed as a Gaussian mixture (see Eq. 1).

After specifying the parameters defined above, a P3DHMM is fully characterized.

If we assume that there exists a special version of the Viterbi algorithm for P3DHMMs, it is possible to classify image sequences with these models and it is also possible to adapt the models to training data. The latter is possible due to the segmental k-means algorithm, which is also known as Viterbi training ([4]). Such a special version of the Viterbi algorithm exists as triple embedded Viterbi algorithm. This algorithm is explained in the following section.

2.2. Triple Embedded Viterbi Algorithm

The triple embedded Viterbi algorithm is an extension of the double embedded Viterbi algorithm which has been introduced by Kuo and Agazzi for P2DHMMs in [5]. The triple embedded version is build by performing an additional execution of the Viterbi algorithm for the time dimension.

Like in the one-dimensional case the most probable state sequence $Q_{XYT}^\star$ is calculated and based on this sequence an estimate for the probability that the observation sequence has been generated by the model is determined ($P(O_{XYT}, Q_{XYT}^\star | \lambda) = P^\star(O_{XYT} | \lambda)$). Because the observation O_{XYT} is a three-dimensional matrix, also the

state sequence $Q^\star$ can be arranged on a three-dimensional grid.

Due to the hierarchical structure of the P3DHMM, the most likely state sequence is calculated in three stages. The first stage is to calculate the probability that the columns of the individual images have been generated by the one-dimensional HMMs, that are assigned to the superstates of the P2DHMMs. In the next step these probabilities are used as observation probabilities of the superstates of the P2DHMMs. A second Viterbi algorithm is executed for the superstates. Finally, on the top hierarchy level a third Viterbi algorithm is executed. The complete triple embedded Viterbi algorithm is stated explicitly as follows:

1.1 Initialization

$$\begin{aligned} \vartheta^{\ j}_{i\ x1t}(l) &= \pi^{\ j}_{i} b^{\ j}_{i}(\vec{o}_{x1t}) \\ \psi^{\ j}_{i\ x1t}(l) &= 0 \end{aligned} \tag{9}$$

1.2 Recursion

$$\begin{aligned} \vartheta^{\ j}_{i\ xyt}(l) &= \max_{l-2\le k\le l}(\vartheta^{\ j}_{i\ x,y-1,t}(k)\cdot a^{\ j}_{i\ kl}) b^{\ j}_{i}(\vec{o}_{xyt}) \\ \psi^{\ j}_{i\ xyt}(l) &= \operatorname*{argmax}_{l-2\le k\le l}(\vartheta^{\ j}_{i\ x,y-1,t}(k)\cdot a^{\ j}_{i\ kl}) \end{aligned} \tag{10}$$

1.3 Termination

$$\begin{aligned} p^j_i(x,t) &= \max_{1\le l\le M^j_i}(\vartheta^{\ j}_{i\ xYt}(l)) \\ n^j_i(x,t) &= \operatorname*{argmax}_{1\le l\le M^j_i}(\vartheta^{\ j}_{i\ xYt}(l)) \end{aligned} \tag{11}$$

The quantity $\psi^{\ j}_{i\ xyt}(l)$ is used to keep track of the optimum states that maximize $\vartheta^{\ j}_{i\ xyt}(l)$ which is the highest probability along a single path which ends in state $S^{\ j}_{i}$ and accounting for the first x observations in column y for the image at time t. The second application of the Viterbi algorithm is as follows:

2.1 Initialization

$$\begin{aligned} D^j_{1,t}(i) &= \pi^j_i p^j_i(1,t) \\ \gamma^j_{1,t}(i) &= 0 \end{aligned} \tag{12}$$

2.2 Recursion

$$\begin{aligned} D^j_{x,t}(i) &= \max_{i-2\le k\le i}(D^j_{x-1,t}(k)\cdot a^j_{ki}) p^j_i(x,t) \\ \gamma^j_{x,t}(i) &= \operatorname*{argmax}_{i-2\le k\le i}(D^j_{x-1,t}(k)\cdot a^j_{ki}) \end{aligned} \tag{13}$$

2.3 Termination

$$\begin{aligned} P_j(t) &= \max_{1\le i\le L^j}(D^j_{X,t}(i)) \\ N_j(t) &= \operatorname*{argmax}_{1\le i\le L^j}(D^j_{X,t}(i)) \end{aligned} \tag{14}$$

The quantities $\vartheta^{\ j}_{i\ xyt}(l)$ and $D^j_{x,t}(i)$ as well as $\psi^{\ j}_{i\ xyt}(l)$ and $\gamma^j_{x,t}(i)$ correspond to each other. The third application of the Viterbi algorithm is as follows:

3.1 Initialization

$$\begin{aligned} E_1(j) &= \pi_j P_j(1) \\ \delta_1(j) &= 0 \end{aligned} \tag{15}$$

3.2 Recursion

$$\begin{aligned} E_t(j) &= \max_{j-2\le h\le j}(E_{t-1}(h)\cdot a_{hj}) P_j(t) \\ \delta_t(j) &= \operatorname*{arg\,max}_{j-2\le h\le j}(E_{t-1}(h)\cdot a_{hj}) \end{aligned} \tag{16}$$

3.2 Termination

$$\begin{aligned} P^\star &= \max_{1\le j\le K}(E_T(j)) \\ q_T &= \operatorname*{argmax}_{1\le j\le K}(E_T(j)) \end{aligned} \tag{17}$$

The value $P^\star$ obtained from Eq. 17 is a measure of how well the P3DHMM models the data O. Finally, by backtracking through $\delta_t(j)$,$\gamma^j_{x,t}(i)$ and $\psi^{\ j}_{i\ xyt}(l)$ it is possible to find the maximum likelihood state sequence.

2.3. Equivalent 1-D Topology

Samaria shows in [6] that a P2DHMM can be transformed into an equivalent one-dimensional HMM by inserting special *start-of-column* states and features. This has been verified by experiments in the domain of speech recognition by Weber et al. [8]. Consequently, it is possible to obtain a one-dimensional HMM which is equivalent to a P3DHMM by applying the technique suggested by Samaria twice, i.e. by additionally inserting special *start-of-image* states and features. Figs. 3 and 4 illustrate the equivalent one-dimensional modeling technique. Fig. 3 shows the

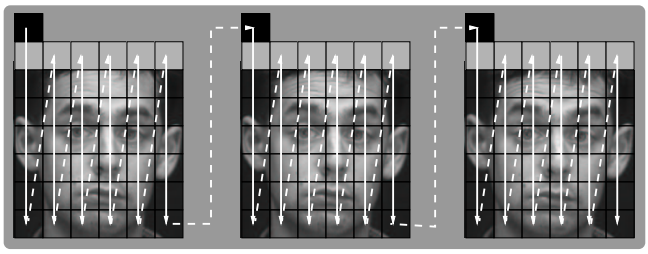

Figure 3. Generation of a one-dimensional observation sequence from an image sequence

three-dimensional pattern, which is a part of an image sequence representing the expression *surprise* which shall be modeled by the HMM in Fig. 4. In order to be able to represent the pattern by a one-dimensional structure, it is necessary to transform the pattern into an observation sequence. As illustrated in Fig. 3 the sequence is obtained in the following manner: Each image of the sequence is scanned with a sampling window from top to bottom and from left to right. The beginning of each column is indicated by a special feature set (start-of-column feature) which is shown in light gray in Fig. 3. The feature sequences of adjacent images are concatenated via start-of-image features (marked in black in Fig. 3).

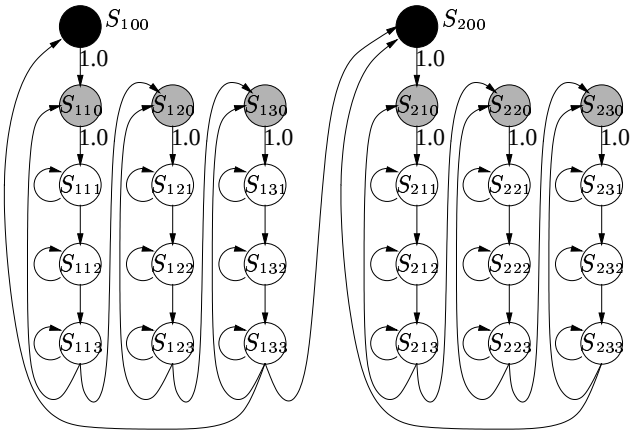

Figure 4. One-dimensional HMM structure which is capable of modeling whole image sequences

Fig. 4 illustrates the 1-D HMM topology which can be used in order to represent whole image sequences. The states which are shaded light gray generate a high probability for the emission of start-of-column features whereas the states which are black generate a high probability for the emission of start-of-image features. When using the structure in Fig. 4 one has to take care of the fact that the values for the start-of-column and start-of-image features are different from all possible ordinary features. These equivalent HMMs can be trained by the standard Baum-Welch algorithm and the recognition step can be carried out using the standard Viterbi algorithm.

3. Experiments and Results

In order to obtain a detailed evaluation of the P3DHMM approach, experiments on a facial expression database consisting of 4 classes (see also Fig. 1) have been performed. In the experiments a one-dimensional structure was used, which is equivalent to a P3DHMM with four hyperstates and encapsulated P2DHMMs sized (4×4).

As feature extraction a discrete cosine transform (DCT) was applied to blocks of size (16×16) pixels. The blocks did not contain the gray values itself, but the difference of neighboring frames (difference image). A block overlap of 75% was utilized. A single P3DHMM was trained for each of the facial expressions using 3 sequences for each of the 6 different persons.

The training of the equivalent 1-D HMMs is a very time consuming process. This is due to the complexity of the Forward-Backward algorithm which is proportional to $N^2 \cdot T$, where N is the number of states in an HMM and T is the length of the observation sequence. Therefore the training has been divided into two steps: In a first step, sections of the sequence have been used to train 1-D HMMs which are equivalent to P2DHMMs. This can be considered as an initialization step. Afterwards the initialized HMMs are concatenated to form an HMM with a topology similar to Fig. 4 and a training of the model with complete image sequences is performed. Due to the initialization of parts of the model, the required number of training iterations for the whole model is reduced drastically. This is particularly important, because the whole model consists of the large number of 64 states in total. Table 1 shows the recognition accuracies achieved in the experiments for a varying number of Gaussian mixtures (see Eq. 1). As can be seen the recognition accuracy increases with the number of mixtures and does not improve further after using 3 mixtures.

Table 1. Recognition accuracies achieved in the experiments

rec.rate	1 mixture	2 mixtures	3 mixtures	4 mixtures
on rank 1	75.00%	79.17%	87.50%	87.50%
on rank 2	79.17%	87.50%	95.83%	95.83%

4. Summary

This paper presented dynamic facial expression recognition based on pseudo 3-D Hidden Markov Models. A formal definition of P3DHMMs as well as a triple embedded Viterbi algorithm have been described. Experiments on a person-independent facial expression database have been carried out and recognition accuracies of up to 87.5% have been achieved. In the experiments a one-dimensional HMM structure has been utilized which has similar modeling capabilities with respect to the P3DHMM approach.

References

[1] S. Müller, S. Eickeler, and G. Rigoll, "Pseudo 3-D HMMs for Image Sequence Recognition", In Proc. IEEE-ICIP, Kobe, Japan, Oct. 1999, pp. 237–241.

[2] Y.-I. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis", IEEE Trans. on PAMI, Vol 23, No. 2, Feb. 2001, pp. 97–115.

[3] M. Rosenblum, Y. Yacoob, and L. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture", Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, Texas, Nov. 1994.

[4] 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.

[5] S.-S. Kuo and O. E. Agazzi, "Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models", IEEE Trans. on PAMI, Vol 16, No. 8, August 1994, pp. 842–848.

[6] F.S. Samaria, "Face Recognition Using Hidden Markov Models", PhD Thesis, Engineering Department, Cambridge University, October 1994.

[7] S. Eickeler, S. Müller, G. Rigoll, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", Image and Vision Computing Journal, Vol 18, No 4, pp. 279–287, March 2000.

[8] K. Weber, S. Bengio, and H. Bourlard, "A Pragmatic View of the Application of HMM2 for ASR", IDIAP Research Report, IDIAP-RR 01-23, Martigny, Suisse, July 2001.