CONTENT-BASED INDEXING OF IMAGES AND VIDEO USING FACE DETECTION AND RECOGNITION METHODS

Stefan Eickeler [1,2], *Frank Wallhoff* [1], *Uri Iurgel* [1], *Gerhard Rigoll* [1]

(1) Gerhard-Mercator-University Duisburg
Faculty of Electrical Engineering
47057 Duisburg, Germany
{wallhoff,uri,rigoll}@fb9-ti.uni-duisburg.de

(2) Cobion AG
Miramstraße 87
34123 Kassel, Germany
stefan.eickeler@cobion.com

ABSTRACT

This paper presents an image and video indexing approach that combines face detection and face recognition methods. Images of a database or frames of a video sequence are scanned for faces by a Neural Network-based face detector. The extracted faces are then grouped into clusters by a combination of a face recognition method using pseudo two-dimensional Hidden Markov Models and a k-means clustering algorithm. Each resulting main cluster consists of the face images of one person. In a subsequent step the detected faces are labeled as one of the different people in the video sequence or the image database and the occurrence of the people can be evaluated. The results of the proposed approach on a TV broadcast news sequence are presented. It is demonstrated that the system is able to discriminate between three different newscasters and an interviewed person.

1. INTRODUCTION

The increasing amount of images and video in multimedia databases results in a demand for techniques for automatic content-based access to visual data. In recent years many different approaches to image and video indexing have been developed. Most methods for the image indexing use low-level features like texture or color. The main drawback of low-level features for image and video indexing is that the people cannot be recognized and person-based indexing is not possible. People are one of the most important objects in images and video sequences. Therefore any practical indexing approach has to combine the detection and recognition of people in images and video sequences.

This paper explores the usability of face detection and face recognition methods for image and video indexing. The face detection method is used to find the faces in images and video sequences. The face recognition method assigns the detected faces to the different characters of a movie, or groups the faces of each person for image indexing. In contrast to the system in [6], the system proposed here does not determine the name the people in the video sequence from the transcript, because the names of the people do not provide information about the structure of the video sequence.

Firstly the face detection algorithm, the face recognition method based on pseudo two-dimensional Hidden Markov Models, and the k-means clustering using HMMs are introduced. The combination of these methods to build up a video and image indexing system is explained. Finally the results on a sample video of a TV broadcast news and the conclusions are presented.

2. FACE DETECTION USING NEURAL NETWORKS

Our face detection method is very similar to the Neural Network-based method presented in [5]. A square sampling window extracts image regions from the image at all possible positions and of various sizes. These samples are then preprocessed by subtracting a best-fit brightness plane and performing a histogram equalization. The subimages are scaled to a size of 20×20 pixels. Subimages with a very low contrast or with a mean color that is not a face color ($U > 0$ or $V < 0$), are discarded. Each pixel of the 20×20 subimage is used as the input of a multi-layer perceptron that is trained on face and non-face samples. The output neuron gives a probability that the subimage contains a face. In case the probability exceeds a threshold a further postprocessing step is applied to eliminate false positives. Detections with an overlap of more than 70 % are merged, and if the overlap is less then 70 %, the less probable detection is deleted. At the end of the merging step all detections that are merged from at least three of the previous detections are extracted from the images.

3. FACE RECOGNITION USING PSEUDO 2-D HMMS

The face recognition module [2] uses pseudo 2-D Hidden Markov Models (HMM) and DCT coefficients. The image of the face is scanned with a sampling window top to bottom and left to right. The pixels in this sampling window of the size 8×8 are transformed using the DCT. The first

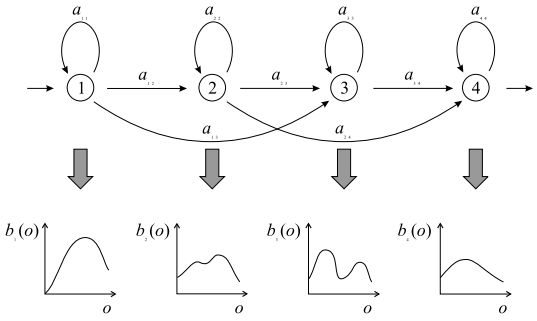

Fig. 1. 1-D Hidden Markov Model

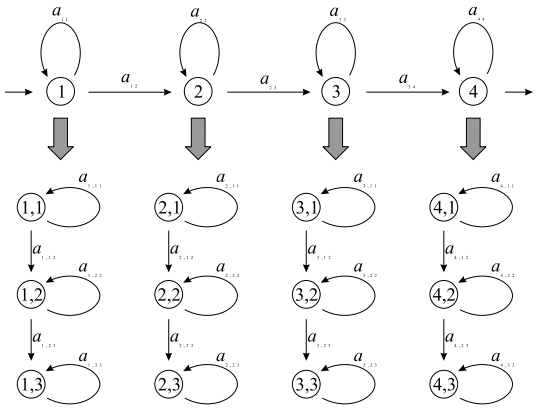

Fig. 2. Pseudo 2-D Hidden Markov Model

15 coefficients are arranged in a feature vector. The use of DCT-coefficients as features for the recognition has two important advantages: Firstly the DCT decorrelates the subimage and allows the use of diagonal covariance matrices for the probability density function of the HMMs. Secondly the face recognition can be directly applied to JPEG and MPEG compressed images. An overlap between adjacent sampling windows can be used to improve the ability of the HMM to model the neighborhood relations between the windows. The resulting array of feature vectors are classified using pseudo 2-D HMMs. A single HMM is trained for each person in the training set using the Baum-Welch algorithm. For the recognition the Viterbi algorithm is used to determine the probability of each face model for the test image.

HMMs [4] are statistical models that consist of several states. At each step a transition to another state depending on a transition probability matrix is performed and a symbol is created depending on a probability density function (pdf) that is assigned to each state. Figure 1 shows a one-dimensional Hidden Markov Model with four states and the assigned pdfs.

Pseudo 2-D HMMs are extensions of the one-dimensional case to work on two-dimensional data like images. Pseudo 2-D HMMs are nested one-dimensional HMMs: A higher level HMM models the sequence of columns in the image. Instead of a probability density function the states of the higher level model (superstates) have a one-dimensional HMM to model the cells inside the columns. Figure 2 shows a pseudo 2-D HMM with four superstates containing a three state 1-D HMM in each superstate. The probability density functions of the lower level models are omitted in this figure.

The Baum-Welch algorithm determines the parameters corresponding to a local maximum of the likelihood function depending on the parameters of the initial model [4]. Therefore it is crucial to use a good initial model for the training. We train a general initial model on all faces in the training set using the Baum-Welch Algorithm. This common model is refined on the training faces of one person to obtain the model for this person.

4. CLUSTERING USING PSEUDO 2-D HMM

The HMM-clustering is an unsupervised grouping of data into classes containing similar members. It is a k-means clustering which uses a Hidden Markov Models to represent a cluster prototype instead of a vector. Therefore it allows the clustering of 1-D vector sequences. A similar method was published in [3]. For the initialization of the clustering process, the sequences are assigned randomly to the clusters. The codebook is then iteratively refined by training the HMMs of the clusters on the assigned sequences using the Baum-Welch Algorithm and then reassigning each sequence to the cluster which HMM has the highest probability of producing that sequence by using the Viterbi Algorithm. This is repeated until the likelihoods of the clusters converge. As result we determine the clusters of data and the HMM prototypes. The advantage of the HMM-clustering versus the use of the integrated clustering capabilities of the Baum-Welch algorithm is a faster computation, because the sequences are not assigned to the prototype in every iteration of the Baum-Welch algorithm.

The use of pseudo 2-D HMMs for the representation of a cluster is possible without any other modifications of the clustering algorithm. This allows a grouping of face images into classes, which cannot be done by the classical k-means clustering, because the large variety of the facial expressions requires the incorporation of a face recognition method into the clustering method.

5. COMBINATION OF THE PRESENTED METHODS FOR IMAGE AND VIDEO INDEXING

The image and video indexing system is a combination of the methods presented above. Figure 3 shows the flow chart of the system. First the faces in the images or the video sequence are detected and extracted using the face detection method. The image regions which contain faces are enlarged such that they contain the head of the person.

The clustering of the faces is done by the face recognition method embedded into the Pseudo 2-D HMM based

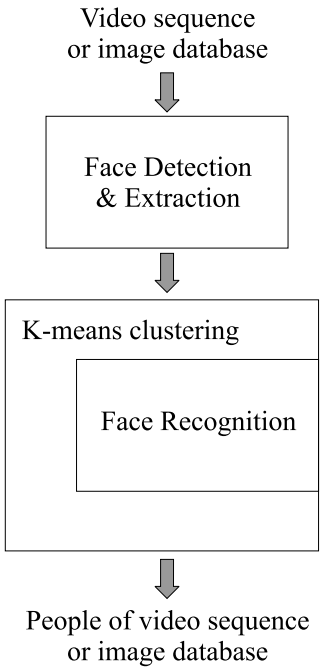

Fig. 3. Combination of the presented methods for image and video indexing

k-means clustering. The features for the face recognition do not change during the iterations of the HMM-clustering, therefore they are extracted only once before the start of the loop, to increase the speed. This is not illustrated in Figure 3, because it does not affect the functionality. The feature extraction uses for each extracted face a blocksize for the DCT that gives approximately the same amount of features vectors (25×30) for all face images, to get a rough size normalization of the faces. A common initial model is trained on all faces and this model is used as prototype for each class. The clustering works as described in Section 4, but a smoothing of the variances of the HMMs with the variances of the common initial model prevents clusters with only a few members from overfitting and gives a better similarity inside the resulting classes. The result of the clustering algorithm are clusters of people. The biggest clusters contain the main people of the video sequence or the image database. Small clusters contain people with a occurrence that is too low to build individual clusters. Our clustering method uses the same techniques for image database indexing and video indexing. For the case of video indexing the method can be improved by evaluating the temporal order of the frames.

6. EXPERIMENTS AND RESULTS

To demonstrate the capabilities of the proposed approach we applied it to the indexing of TV news. The TV broadcast was captured in a resolution of 384×288 with a frame rate of 0.2 fps. The news are presented by three different people, therefore the approach presented in [1] cannot be used in this case, because it can cope with only one newscaster. The detection method detected 706 faces in the sequence. The clustering approach was able to assign the faces of the three newscasters and an interviewed person correctly. Figures 4, 5, and 6 show images of the three clusters of people representing the three newscasters. Figure 7 shows images of the cluster of the interviewed person. The faces in these images that are assigned to the same cluster are marked by the white rectangle.

7. CONCLUSIONS

This paper presented an image and video indexing approach based on the detection and recognition of faces. It was shown that in the case of video indexing this method has advantages compared to our previous video indexing method. The proposed approach is capable of indexing a video sequence without any prior knowledge of the sequence, because in contrast to the approach in [1] no video model has to be trained on the training samples. The method presented here can be further improved by using the temporal information of the video sequences, like detection of cuts and other edit effects and a tracking of faces. This simplifies the detection of the face of the same person in consecutive frames of the video sequence. In the future we will apply this method to detect the main characters in movies.

8. REFERENCES

[1] S. Eickeler and S. Müller. Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2997–3000, Phoenix, Mar. 1999.

[2] S. Eickeler, S. Müller, and G. Rigoll. Recognition of JPEG Compressed Face Images Based on Statistical Methods. *Image and Vision Computing Journal, Special Issue on Facial Image Analysis*, 18(4):279–287, Mar. 2000.

[3] L. M. D. Owsley, L. E. Atlas, and G. D. Bernard. Automatic Clustering of Vectortime-Series for Manufacturing Machine Monitoring. In *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3393–3396, Munich, Apr. 1997.

[4] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–285, Feb. 1989.

[5] H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, Jan. 1998.

[6] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and Detecting Faces in News Videos. *IEEE Multimedia*, 6(1):22–35, 1999.

Fig. 4. First newscaster

Fig. 5. Second newscaster

Fig. 6. Third newscaster

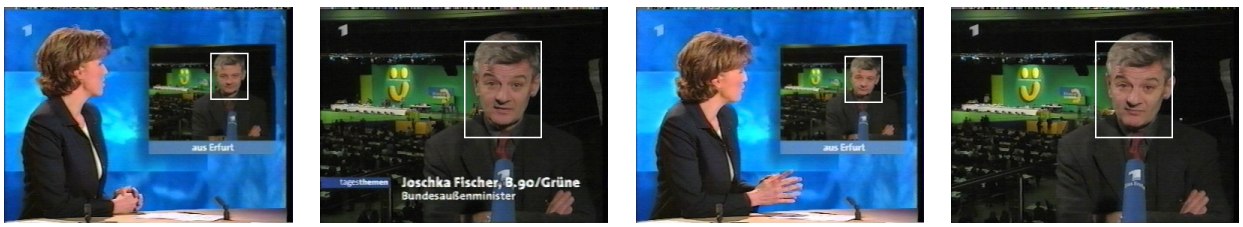

Fig. 7. Interviewed person