Integrating Face-ID into an Interactive Person-ID Learning System

Stephan Könn, Hartwig Holzapfel, Hazım Kemal Ekenel, Alex Waibel

InterACT Research, Interactive Systems Labs,
Universität Karlsruhe, Germany
hartwig,ekenel,waibel@ira.uka.de, stephan@koenn.de

Abstract. Acquiring knowledge about persons is a key functionality for humanoid robots. By envisioning a robot that can provide personalized services the system needs to detect, recognize and memorize information about specific persons. To reach this goal we present an approach for extensible person identification based on visual processing, as one component of an interactive system able to interactively acquire information about persons.
This paper describes an approach for face-ID recognition and identification over image sequences and its integration into the interactive system. We compare the approach of sequence hypotheses against results from single image hypotheses, and a standard approach and show improvements in both cases. We furthermore explore the usage of confidence scores to allow other system components to estimate the accuracy of face-ID hypotheses.

1 INTRODUCTION

Recognizing and memorizing other people is an important part of human-human communication. A humanoid robot, if equipped with such functionality, can offer more natural ways of communication and provides a basis to facilitate personalized services. In this paper we describe a system that can recognize a person's face to identify persons during human-robot interactions. The face-ID system presented in this paper is integrated into an interactive system for human-robot interaction that can autonomously get in touch with persons [1]. A key component for this functionality is the face recognition system that reports id hypotheses on video sequences during the interaction, reports confidences how well the person's id can be predicted, and can be easily extended with new video samples and samples from new persons.

Existing work covers video or sequence based face recognition for smart environments [12] or surveillance tasks [13] - [21]. Some of these mentioned approaches use still images for the training task and video sequences only for testing, which partially evades some of the problems that occur when using video-based images for both training and testing (e.g. blurred images, differences in pose). Commonly used algorithms for face identification in terms of classification are Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). Other works also apply Decision Tree Models oder Hidden Markov

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Models (HMM) for face recognition. In this paper we focus on an instance-based learning algorithm, which is able to outperform other methods [2], and combination of single image classifications over whole image sequences together with providing a hypothesis confidence.

In the following sections we describe the system as follows. Section 2 gives an overview over the system architecture. Section 3 describes face identification on single images. Section 4 describes hypothesis computation over sequences of images during interactions and presents evaluations. Section 5 is addressed to the calculation of sequence confidence and discusses the implemented work. Section 6 concludes the paper and gives an outlook to future work.

2 SYSTEM OVERVIEW

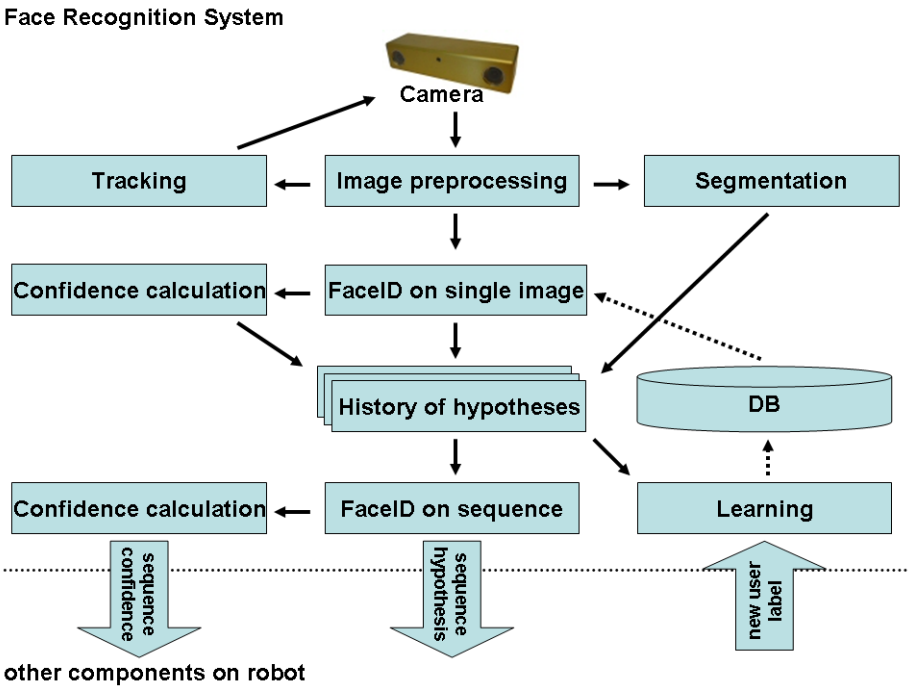

Fig. 1. System architecture of the face recognition system

In order to create an assumption about the person in front of the robot several processes have to be passed that interact with and depend on each other. Figure 1 gives an overview over the face recognition system architecture and its components. As the first component of the chain, the robot camera takes images in regular intervals. Afterwards, the image is preprocessed to detect face

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

and eyes in the image, after which discrete cosine transform (DCT) is applied to extract multidimensional features for face recognition. Concurrently, the person is tracked by a tracking unit and a segmentation unit registers start and end of the user session. Face Recognition first computes a single image hypothesis, additionally, a confidence is calculated for each image which reflects the degree of reliance on this hypothesis. Using these hypothesis-confidence-pairs a sequence hypothesis is computed over the whole sequence of images during the user session. Subsequently, a confidence is calculated which expresses the trust in the sequence hypothesis. This enables other components on the robot using the face recognition system to decide whether to accept or to reject the computed assumption.

The whole process is performed for each new image recorded by the robot camera, i.e. sequence hypotheses and confidences will be updated online for each new frame.

Additionally, the face recognition system is able to learn new persons. Other modules on the robot, e.g. a dialog system, can cause the learning unit to select recent feature vectors and add them to the database. For easily incorporating new samples into the database DCT is prefered to other methods that demand a complete revision of data, e.g. PCA [11].

3 FACE-ID ON SINGLE IMAGES

The face recognition system uses an image sequence, which is provided by a face tracker module, as input. It processes the frames to detect face and eyes of the user. If eyes can be located face recognition system then aligns the face images according to the eye center coordinates. The features that will be used for classification purposes are extracted from each face image by using a local appearance-based face recognition approach [2]. In this feature extraction approach, the input face image is divided into 8x8 pixel blocks, and on each block discrete cosine transform (DCT) is performed. The most relevant DCT features are extracted using the zig-zag scan and the obtained features are fused either at the feature level or at the decision level for face recognition [2]. The approach is extensively tested on the publicly available face databases and compared with the other well known face recognition approaches. The experimental results showed that the proposed local appearance based approach performs significantly better than the traditional face recognition approaches. Moreover, this approach is tested on face recognition grand challenge (FRGC) version 1 data set for face verification [3], and a recent version of it is tested on FRGC version 2 data set for face recognition [4]. In both tasks the system provided better and more stable results than the baseline face recognition system. For example, in the conducted experiments on the FRGC version 2 data set, 96.8% correct recognition rate is obtained under controlled conditions and 80.5% correct recognition rate is obtained under uncontrolled conditions. In these experiments, there are 120 individuals in the database and each individual has ten training and testing images. There is a time gap of approximately +six months between the capturing

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

time of the training and the test set images. The approach is also tested under video-based face recognition evaluations and again provided better results [5, 6]. For details please see [2–6].

After extracting the feature vector from a face image in the sequence, it is compared with the ones in the database using a nearest neighborhood classifier. The result is a single image hypotheses about the person in front of the robot.

4 COMPUTING SEQUENCE HYPOTHESES

The process of retrieving a single image hypothesis is repeated for each frame in the sequence. To combine the classification results a confidence value is calculated for each single image hypothesis. These confidence scores are summed up over a sliding window of frames or the whole image sequence, leaving out hypotheses with low confidence values. The first part describes the calculation process of single image confidences. In the second part hypotheses fusion is explained. An evaluation of the presented approach finalizes this section.

4.1 Confidence scores for single images

To combine the hypotheses of single images their individual classification confidence is taken into account. This confidence is approximated by a logistic regression model [9], which is used to model the likelihood of an event as a function of predictor variables. In this case the binary, dependent variable Y models the event "classification correct" ($Y = 1$) and "classification incorrect" ($Y = 0$). In order to calculate the likelihood of a correct classification several image characteristics come into question as predictor variables. The face recognition system considers four features, namely:

- distance deviation
- mean gray value
- eye detection change
- neighbor distance

"Distance deviation" is based on the distance between the user and the robot camera. This value can be estimated by using the pixel width of the user face on the examined image. "Distance deviation" characterizes the difference between the current user's distance from the camera and the mean distance value over all frames. The variable "mean gray value" calculates the average gray value of the extracted face rectangle. "Eye detection change" is a binary variable and indicates if the face recognition system was able to detect both face and eyes of a person for the first time after at least one frame where either one of these two detectors failed. The last variable "neighbor distance" describes the distance to the next instance when using the nearest neighborhood classifier. Other features, such as hypothesis changes, angle between the robot's gaze direction and person, have been examined as well but have not shown a significant logistic correlation between the classification result and the variable value. The values of all four

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

remaining variables can be determined automatically by the face recognition system and thus be generated online for computing a confidence for a single image hypothesis.

4.2 Summation over sequences

The baseline approach to compute a sequence hypothesis is by normalizing each frame's distance to its nearest neighbor scores with Min-Max normalization method [7]. Then these scores are fused over the sequence using the sum rule [8].

In this work a slightly different approach to the sum rule is used. At first, not necessarily all hypotheses of the whole sequence are considered. A window size w is determined and only the assumptions of the last w frames are taken into account. Furthermore hypotheses with low single image confidence values are ignored and only the best p percent of hypotheses are selected to form the set of single image hypotheses which is regarded to compute a sequence hypothesis. This is realized in two more steps: for each hypothesis its confidence is used as score and summed up over the whole sequence as in equation 1

$$\mathrm{sum}(h_j) = \sum_{h_i \in H} (\mathrm{conf}(h_i) \cdot \delta(h_i, hj)) \tag{1}$$

where h_i denotes a hypothesis in the set H of remaining hypotheses, $\mathrm{conf}(h_i)$ reflects the corresponding confidence and $\delta(h_i, hj) = 1$ if h_i and h_j are equal, and 0 otherwise. Subsequently, the sequence hypothesis is computed according to equation 2

$$c = \mathrm{argmax}_{h_i \in H}\, \mathrm{sum}(h_i) \tag{2}$$

where $\mathrm{sum}(h_i)$ refers to the summation in equation 1.

4.3 Evaluation

Evaluation of these different approaches is done on a dataset consisting of 30 sequences with 11576 images overall from 16 different persons. Both approaches were trained on a distinct training set consisting of 17 sequences with 4288 images from 16 different persons. The recording of training and evaluation data was done on three consecutive days.

Figure 2 shows the successful classification rates for a single image hypothesis and the sum rule approach. For classification of a single image a 10-nearest-neighbor classifier is used. To generate scores Min-Max normalization is applied. The last bar shows the success rate using the previously introduced method to generate a sequence hypothesis. The complete classifation results of the presented method are shown in figure 3 using different configurations, which are displayed below each bar. The first number determines the window size, the second parameter corresponds to the percentage of considered hypotheses. For classification of a single image a simple nearest-neighbor classifier is used.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

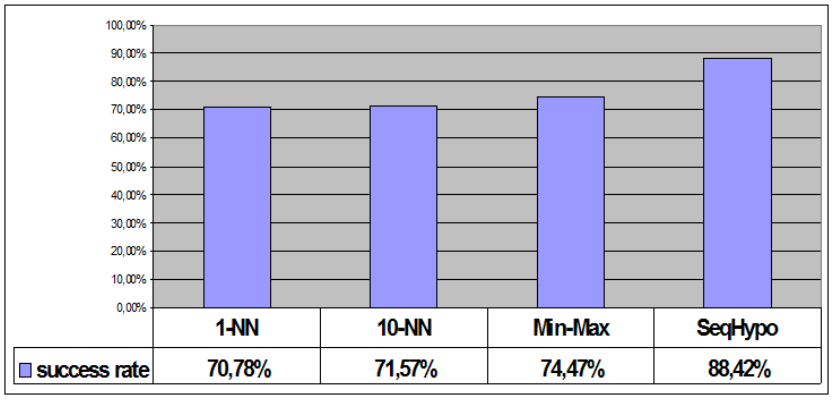

Fig. 2. Success rates of the baseline approach using Min-Max normalization method to generate scores for sum rule and the presented method to generate sequence hypotheses.

As can be seen in figure 3, any configuration of a sequence hypothesis shows improvement over single image hypotheses (70,78%). The baseline approach using Min-Max normalization method (74,47%) performs well below the values of the approach presented here. Computing sequence hypothesis with unlimited window size performs best (88,42%) on the evaluation set. Regarding the complete session a selection of the best hypotheses has only a negliable effect. Other configurations like $w = 30$ and $p = 10\%$ show slight performance differences (84,07%).

5 CONFIDENCE VALUES AND DISCUSSION

For the sake of integration the face recognition system calculates confidence values for each sequence hypothesis. This enables other components on the robot using the face recognition system to decide whether to trust in the current sequence hypothesis or reject it. Later, the discussion subsection is addressed to basic design questions and difficulties faced during development and testing of the face recognition system.

5.1 Confidence values

In order to calculate a sequence hypothesis confidence a logistic regression model is used again. The predictor variables are agreement and stability. The variable "agreement" denotes the rate of single image hypotheses that are equal with the current sequence hypothesis to the total number of frames. The feature "stability" counts the number of changes of single image hypotheses divided by the number of frames.

For these two features regression coefficients are determined. They reveal a positive correlation, which means that a higher variable value for "agreement" or

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

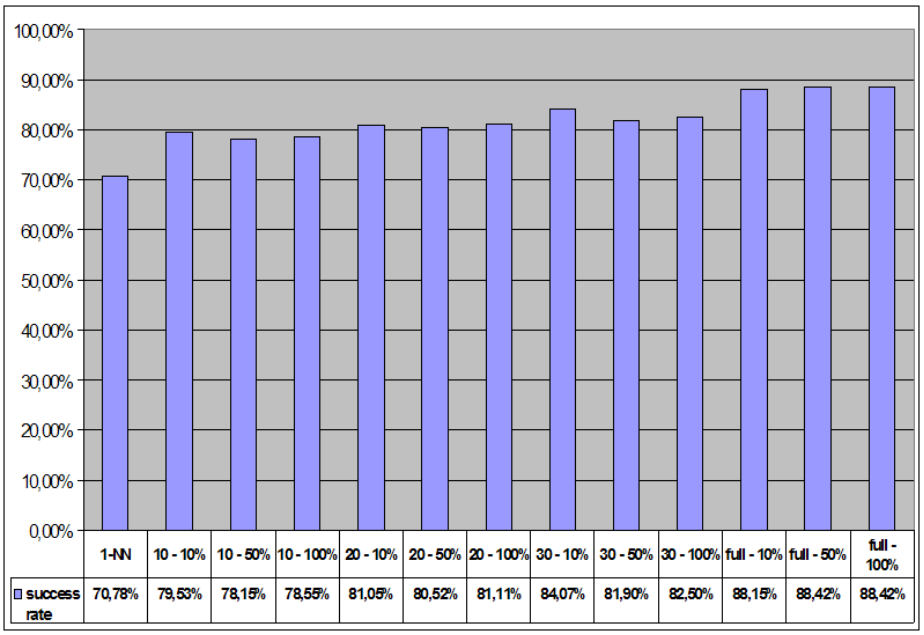

Fig. 3. Success rates for different configurations for determining the sequence hypothesis. The first number determines the window size, the second corresponds to the rate of hypotheses with highest confidence values.

"stability" corresponds to a higher outcome of the sequence confidence. Figure 4 shows the surface of the logistic regression function subject to the two predictor variables. The x-axis shows the developing of the values of "stability" whereas the y-axis characterizes the values of "agreement".

Further evaluations of the sequence confidence were performed in order to test the significance of such a value. Table 1 shows the different mean confidence and standard deviation values for correct and respectively false sequence hypotheses. These results demonstrate that the provided sequence confidence is a reasonable and reliable measure for inheriting sequence hypotheses.

sequence hypothesis	mean confidence	standard deviation
true	0,86	0,27
false	0,41	0,29

Table 1. Mean confidence and standard deviation for true respective false outcome of sequence hypothesis

Other systems on the robot accepting a sequence hypothesis if the according confidence value exceeds a threshold of 0,5 and rejecting it below this value will

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

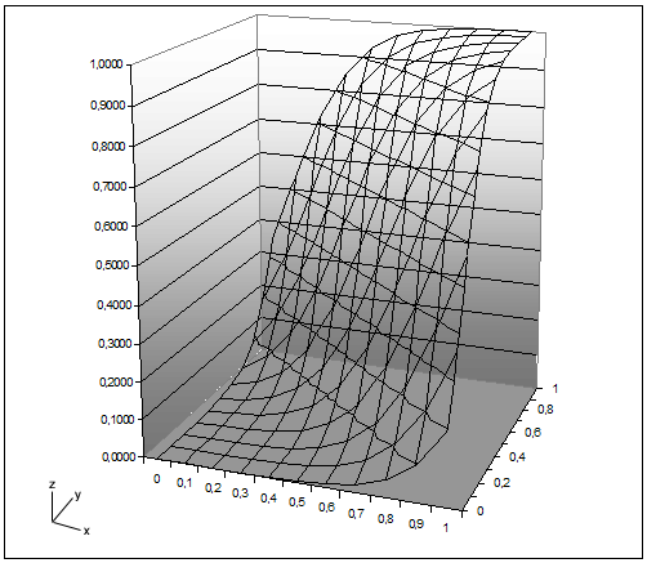

Fig. 4. The surface of the logistic regression function for calculating a sequence confidence subject to the predictor variables "agreement" and "stability".

accept correct hypotheses with a precision rate between 93% and 96% for all evaluated configurations. Increasing the threshold to 0,9 even results in precision rates of over 98% and recall rates of over 74%.

Good confidence values and high precision rates are crucial for other components on the robot, e.g. a dialogue system which needs to identify the person in front of the robot and thus can interact in a more natural way. Finally collected data from a person has to be stored in the database together with the right person ID.

5.2 Discussion

One big challenge in this work was to cope with different conditions during data recording. Recording was done inside a facility building with artificial lighting as well as natural lighting. The consequences are different lighting conditions in the image data, especially as data was taken from a Wizard-of-Oz experiment where users had to pass the robot, talk to it, and then leave again. For this reason we had to cope with a turning camera tracking the moving person and different image backgrounds. Last but not least, the persons were talking during interaction. All that leads to significantly worse results of single image classifications in comparison with the FRGC results. Nevertheless, face identification over image sequences performs much better and shows a large improvement over single image hypotheses.

In the process of computing sequence hypotheses the calculation of single image confidences is a crucial part. In particular, the choice of features as pre-

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

dictor variables is very important. One obvious feature that is not incorporated directly is the pure distance between user and robot. Single image classification requires eye detection to align the face of the user. Unfortunately, eye detection works inadequately if the person is very distant or even very close to the camera due to bad image quality. For this reason the variable value for "distance" is not strictly increasing and has to be transformed to the applied variable "distance deviation".

6 CONCLUSIONS AND FUTURE WORK

We have presented a face recognition system that is able to identify persons in front of the robot. Ids are not generated only on single images, instead whole image sequences are taken into account. The presented method of computing sequence hypotheses shows significant improvement over single image hypotheses. Additionally, the face recognition system provides a sequence confidence which is a reliable measure for other modules on the robot. Moreover, the system is able to easily incorporate new samples into its database.

In the future we plan to conduct further experiments in different environments and in different scenarios to verify the accuracy of our presented results.

Further work is necessary to solve the problem of large datasets. On the one hand DCT enables easy incorporation of new samples into the database, on the other hand without some kind of clustering the amount of data grows exceedingly large and downgrades performance of the system in terms of processing time.

As said before the selection of features for calculating a confidence value is a very essential part in the presented approach. This choice was more or less done with respect of the suitability in a logistic regression model. Although the selected features perform well, extensive studies still have to be done to detect those features that are best as predictor variables for our task.

In addition, we plan to establish a face verification component to enable the face recognition system to decide itself whether a person is unknown to the system or not.

References

1. Holzapfel, H., Schaaf, T., Ekenel, H.K., Schaa, C., Waibel, A. A Robot learns to know people - First Contacts of a Robot. KI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science, volume 4314
2. Ekenel, H.K., Stiefelhagen, R.: Local Appearance based Face Recognition Using Discrete Cosine Transform. Proceedings of the 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey, (2005)
3. Ekenel, H.K., Stiefelhagen, R.: A Generic Face Representation Approach for Local Appearance based Face Verification. Proceedings of the CVPR IEEE Workshop on FRGC Experiments, San Diego, CA, USA (2005)
4. Ekenel, H.K., Stiefelhagen, R.: Analysis of Local Appearance-based Face Recognition on FRGC 2.0 Database. Face Recognition Grand Challenge Workshop (FRGC), Arlington, VA, USA (2006)

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

5. Ekenel, H.K., Pnevmatikakis, A.: Video-Based Face Recognition Evaluation in the CHIL Project - Run 1. Proceedings of the 7th Intl. Conf. on Automatic Face and Gesture Recognition (FG 2006), Southampton, UK (2006)
6. Ekenel, H.K., Jin, Q.: ISL Person Identification System in the CLEAR Evaluations. Proceedings of the CLEAR Evaluation Workshop, Southampton, UK (2006)
7. R. Snelick, M. Indovina: Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Trans. Pattern Anal. Mach. Intell., 27(3) (2005) 450–455
8. J. Kittler, M. Hatef, R. Duin, J. Matas: On Combining Classifiers. IEEE Trans. Pattern Anal. Mach. Intell., 20(3) (1998)
9. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. 1989, Wiley.
10. Mitchell, T.: Machine Learning. 1997, McGraw Hill.
11. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Science, pages 7186, 1991.
12. Waibel, A., et al.: CHIL: Computers in the Human Interaction Loop. WIAMIS 2004.
13. J. Weng, C.H. Evans, W.S. Hwang,: An Incremental Learning Method for Face Recognition under Continuous Video Stream. AFGR 2000.
14. V. Krger, S. Zhou, Exemplar-Based Face Recognition from Video. AFGR 2002.
15. X. Liu, T. Chen, Video-Based Face Recognition Using Adaptive Hidden Markov Models. CVPR 2003.
16. S. Zhou, V. Krueger and R. Chellappa, Probabilistic Recognition of Human Face from Video. CVIU, Vol. 91, pp. 214-245, July 2003.
17. B. Raytchev, H. Murase, Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion. CVIU(91), No. 1-2, pp. 22-52, 2003.
18. S. Zhou, R. Chellappa, B. Moghaddam, Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Trans. on Image Processing, Vol. 13, No. 11, pp. 1491-1506, 2004.
19. G. Aggarwal, A.K.R Chowdhury, R. Chellappa, A system identification approach for video-based face recognition. ICPR 2004.
20. C. Xie et al, A Still-to-Video Face Verification System Using Advanced Correlation Filters. ICBA 2004.
21. K.C. Lee, D. Kriegman, Online Learning of Probabilistc Appearance Manifolds for Video-based Recognition and Tracking. CVPR 2005.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de