Comparison of Confidence Measures for Face Recognition

Stefan Eickeler, Mirco Jabs, Gerhard Rigoll
Gerhard-Mercator-University Duisburg
Department of Computer Science
Faculty of Electrical Engineering
47057 Duisburg, Germany
{eickeler,mirco,rigoll}@fb9-ti.uni-duisburg.de

Abstract

This paper compares different confidence measures for the results of statistical face recognition systems. The main applications of a confidence measure are rejection of unknown people and the detection of recognition errors. Some of the confidence measures are based on the posterior probability and some on the ranking of the recognition results. The posterior probability is calculated by applying Bayes' rule with different ways to approximate the unconditional likelihood. The confidence measure based on the ranking is a new method, that is presented in this paper. Experiments to evaluate the confidence measures are carried out on a pseudo 2-D Hidden Markov Model based face recognition system and the Bochum face database.

1 Introduction

In recent years, many different approaches to face recognition have been developed. Most approaches are used for face identification with a test set containing only images of a group of people that is identical to or a subset of the group that is used for the training of the algorithm. In this case the rejection of images of unknown people (people not in the training database) is not necessary. The rejection of unknown people is very important for many applications of face recognition like access control and image and video indexing. Without rejection an image of any object that is presented to the face recognition system is recognized as one of the trained people. Rejection can be done based on confidence measures. If the confidence measure is below a threshold, the person is rejected. Confidence measures are commonly used in speech recognition [1, 2] to detect out-of-vocabulary words, which correspond to the unknown people in face recognition, and to detect recognition errors. The confidence measures in speech recognition are based on the posterior probability. This paper compares different methods to calculate the posterior probability for face recognition and presents a new confidence measure based on the ranking of the recognition results.

For the rejection of images three cases can be distinguished:

- Face of a person or image of an object that is not member of the training set and should be rejected.
- The algorithm assigned the face to the wrong person (recognition error). In this case a rejection of the face is preferable to a misclassification.
- The person was correctly identified. The effect of the rejection method on such a recognition case should be as minimal as possible.

This paper first presents our pseudo 2-D Hidden Markov Model based face recognition algorithm to motivate the different confidence measures. Then the different methods to calculate the posterior probability and algorithms that evaluate the ranking of the results are described. The results of confidence measures are compared based on experiments.

2 Face recognition using Pseudo 2-D HMMs

The face recognition module uses pseudo 2-D Hidden Markov Models (P2DHMM) and DCT coefficients for the recognition [3]. The image of the face is scanned with a sampling window top to bottom and left to right. The size of the window is 16 × 16 pixels and the overlap between neighboring windows is 75 %. The pixels inside the sampling window are transformed using the DCT and the first fifteen coefficients are extracted. The result of the feature extraction is an array of feature vectors, that is classified using pseudo 2-D HMMs. A single P2DHMM is trained for each person on the training images. For the recognition the class-conditional probability of the test image for each face

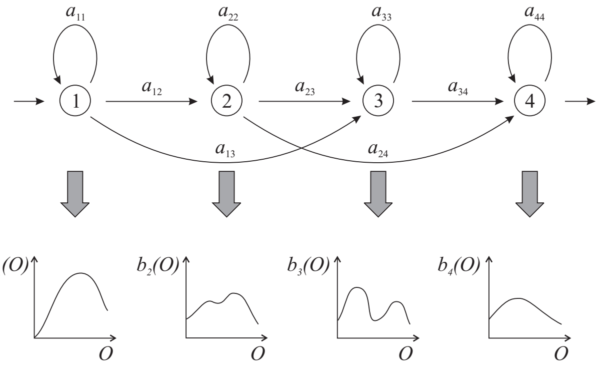

Figure 1. 1-D Hidden Markov Model

model is determined. A maximum selection chooses the person with the model for which the image has the highest probability as recognized person.

One-dimensional Hidden Markov Models (HMM) [4] are commonly used in speech recognition. A HMM λ_i consists of N states $\mathbf{Q} = s_1, s_2, \ldots, s_N$. At each step a transition to another state depending on a transition probability matrix $\mathbf{A} = [a_{ij}]$ is performed and a symbol is created depending on a probability density function (pdf):

$$b_j(O_t) = \sum_{k=1}^{K} c_{k\,j} \cdot \mathcal{N}(O_t, \boldsymbol{\mu}_{k\,j}, \boldsymbol{\Sigma}_{k\,j}) \tag{1}$$

The pdf is a weighted sum of Gaussian mixtures with the mean vector $\boldsymbol{\mu}_{k\,j}$ and the covariance matrix $\boldsymbol{\Sigma}_{k\,j}$.

Figure 1 shows a one-dimensional Hidden Markov Model with four states and assigned pdfs. Possible transitions are indicated by arcs. The main problem dealing with Hidden Markov Models is to compute the probability of an observation sequence $\mathbf{O}$ for the given model λ:

$$P(\mathbf{O}|\lambda_i) = \sum_{\mathbf{q}\in\mathbf{Q}^T} \prod_{t=1}^{T} b_{q_t}(O_t)\pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}q_t} \tag{2}$$

T is the length of the observation sequence and $\boldsymbol{\pi}$ is the initial state distribution.

Pseudo 2-D HMMs are extensions of the one-dimensional case for modeling of two-dimensional data $\mathbf{O}$ with the size $X \times Y$, like images. Pseudo 2-D HMMs are nested one-dimensional HMMs: A superior HMM models the sequence of columns in the image. Instead of a probability density function the states of the superior 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 inferior models are omitted in this figure.

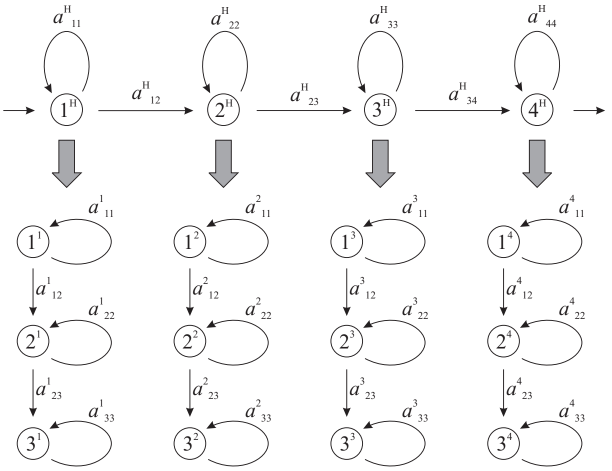

Figure 2. Pseudo 2-D Hidden Markov Model

The Baum-Welch Method is used to train the model for each person based on the maximum likelihood optimization. The Forward-Backward Algorithm calculates the probability $P(\mathbf{O}|\lambda_i)$ of the observation array $\mathbf{O}$ for the model λ_i. For the recognition the model with the maximum $P(\mathbf{O}|\lambda_i)$ is selected as result. The Bayes' rule

$$P(\lambda_i|\mathbf{O}) = \frac{P(\mathbf{O}|\lambda_i)P(\lambda_i)}{P(\mathbf{O})} \tag{3}$$

calculates the posterior probability. The unconditional probability $P(\mathbf{O})$ is constant for one image and for face recognition the prior probability $P(\lambda_i)$ is constant for all people. Therefore *argmax*$[P(\lambda_i|\mathbf{O})] =$ *argmax*$[P(\mathbf{O}|\lambda_i)]$. By using the maximum selector a lot of information, that is provided by the Forward-Backward Algorithm, is unused. This information can be used for the rejection of unknown people and the detection of recognition errors.

3 Confidence based on posterior probability

The posterior probability derived by the Bayes decision theory (Equation 3) provides a confident measure that is optimal [5, 6], if $P(\mathbf{O})$ can be calculated. In the area of speech recognition most confidence measures are based on the posterior probability. The recognition result is rejected, if the posterior probability $P(\lambda_i|\mathbf{O})$ is below a threshold τ, that has to be determined by experiments. In face recognition the prior probability $P(\lambda_i)$ is equal for all classes (people) and is set to 1.0. The main problem is that the unconditioned likelihood $P(\mathbf{O})$ cannot be calculated exactly because the number of classes is almost infinite. Therefore approximates of $P(\mathbf{O})$ has to be used for the rejection of unknown people.

3.1 Normalization

The normalization approximates the unconditional likelihood with the sum over the probabilities of all classes.

$$P(\mathbf{O}) \approx \sum_{j=1}^{C} P(\mathbf{O}|\lambda_j)P(\lambda_j) \tag{4}$$

For the detection of recognition errors this calculation of $P(\mathbf{O})$ should be optimal, because the classes are known and no unknown people (classes) have to be rejected.

For the HMM based face recognition the dynamic of the probabilities exceeds the precision ranges of the computer. Therefore Equations 3 and 4 have to be reformulated to get an alternative expression for the rejection:

$$\left(\frac{1}{\tau} - 1\right)^{-1} = \tau' > \frac{P(\mathbf{O}|\lambda_i)}{\sum\limits_{j=1, j\neq i}^{C} P(\mathbf{O}|\lambda_j)} \tag{5}$$

Equation 5 shows that the rejection rule can be approximated as the probability ratio of the recognized class and the second best class, if the likelihoods of the first and second best class are much greater than the likelihood of the other classes. This measure is called ”two-best” confidence measure in [7].

3.2 Filler model

The second approximation uses a filler model to calculate $P(\mathbf{O})$.

$$P(\mathbf{O}) \approx P(\mathbf{O}|\lambda_{\text{filler}}) \tag{6}$$

In [8, 9] filler model are used for the spotting of spoken words, and filler models were successfully applied to the rejection of unknown gestures in the gesture recognition [10]. The filler model should be trained on a large amount of faces that are not part of the training set for the recognizer.

3.3 Fix probability

The third approximation is related to the use of Hidden Markov Models for the classification, but can be used for other classifiers, too. It is used in online signature verification [11]. The observation **O** is regarded as the output of a process with a uniform output probability p. Then $P(\mathbf{O})$ for an observation **O** of the size $X \times Y$ is:

$$P(\mathbf{O}) \approx p^{X \cdot Y} \tag{7}$$

p cannot be determined analytically, but in the case of a constant image size, it can be included into the threshold for the rejection which has to be determined by experiments. Equation 7 is inserted into 3 and solved for $t \cdot p^{X \cdot Y}$ and gives the new threshold and rejection rule:

$$\tau \cdot p^{X \cdot Y} = \tau' > P(\mathbf{O}|\lambda_i) \tag{8}$$

It should be noted that a constant image size is assumed for all the approximations of $P(\mathbf{O})$.

4 Confidence based on ranking of results

The posterior probability uses only the probabilities of the different classes to calculate a confidence measure. In this section the ranking of the classes for the probability $P(\mathbf{O}|\lambda_i)$ is used for confidence measures. The basic idea is that every person looks similar to other people and can be identified on these similar people. This can be illustrated by the observation that humans are able to recognize a unknown person if they know a brother or sister of the unknown person. For each person of the training set a ranking list is created by recognizing the training images with the face recognition system and sorting the people by the probabilities $P(\mathbf{O}|\lambda_i)$. After the recognition of a test image, the ranking list of the recognized person, which was selected by the maximum $P(\mathbf{O}|\lambda_i)$, is compared to the ranking of the test image. Two different methods are used for the evaluation of the ranking.

4.1 Sum of ranking difference

This algorithm uses the top N classes of the training image and calculates the sum of the absolute differences of the positions of the classes in the training ranking to the test ranking. The following table shows an example for the calculation:

rank	test ranking	training ranking	absolute difference
1	a	a	$\|1-1\| = 0$
2	b	c	$\|3-2\| = 1$
3	c	d	$\|4-3\| = 1$
4	d	e	$\|5-4\| = 1$
5	e	b	$\|2-5\| = 3$
6	f	f	$\|6-6\| = 0$
			$\Sigma = 6$

A weighting factor can be used to emphasize the top results.

4.2 Levenshtein Distance

The second method uses the top N classes of the training and test ranking and calculates the Levenshtein Distance [12]. The Levenshtein Distance Algorithm is normally used for a fuzzy comparison of the strings, but can be used for this evaluation, if the rankings are considered as strings and each class as a different character.

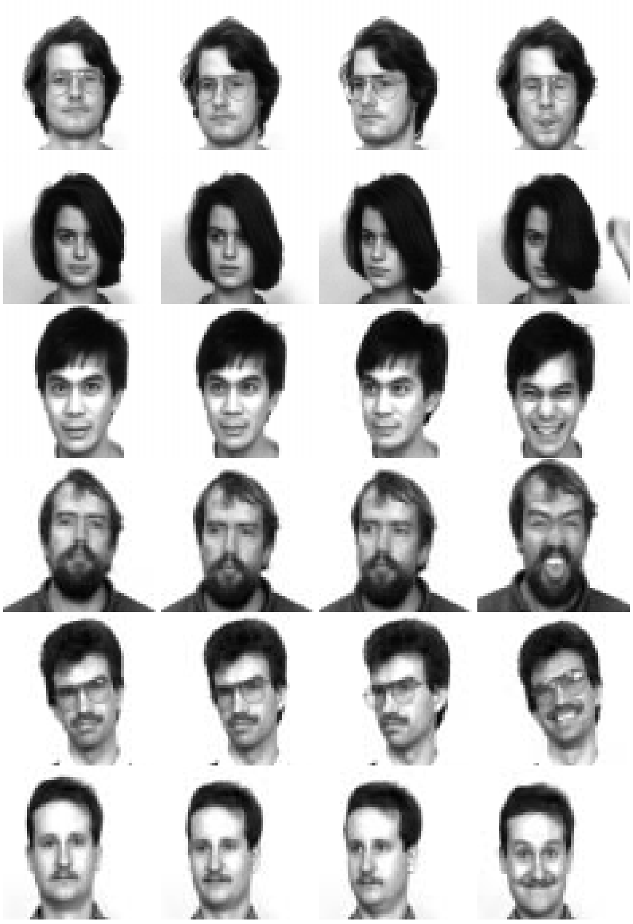

Figure 3. Images of first six people of Bochum database

5 Experiments and results

The experiments are carried out on the Bochum database [13]. It consists of 111 people with four images of each person. Each image is of the size 128×128 pixels. One image of each person, which shows a frontal view, is used for the training of the pseudo 2-D HMM for each person. Three images of each person are used for testing. The images show a head rotation of 11° and 22° and a different facial expression (fb). Figure 3 shows examples of the database.

The face recognition system presented in Section 2 was used with parameters that gave a recognition rate of 94 % to get 20 recognition errors (Figure 4). It is noteworthy that the best result of our system on the Bochum database is 99.4 % (fb: 98 %, 11°: 100 %, 22°: 100 %), but this would give too few recognition errors to test the confidence measures. All images shown in Figure 4 are correctly recognized by our best system. Additionally to the people in the training set the Bochum database provides images of people that are not in the training set. We added some images from the NIST mugshot database and some images collected from the internet to get a set of 40 images of unknown people/objects. 30 images contain people and ten images contain other objects, like a balloon or a bulb (Figure 5).

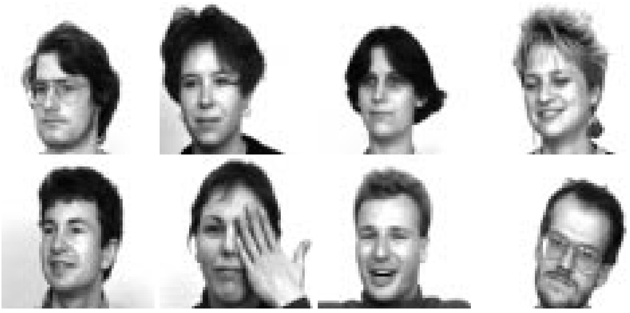

Figure 4. Images of falsely recognized people

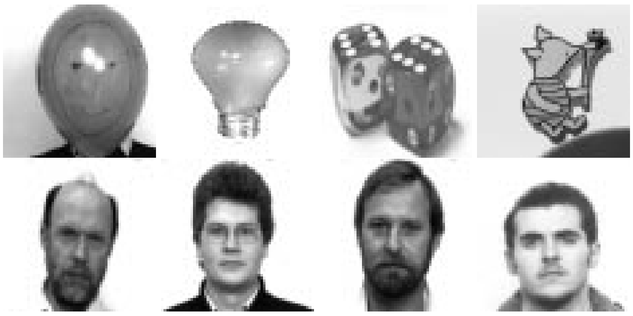

Figure 5. pic:Images of unknown people/objects

Figure 6 shows the results of the different measures as a parametric plot of the threshold. The false alarm rate is the percentage of correct recognized people that are rejected. The undetected error + intruder rate is the percentage of unknown people and false recognitions that are accepted. The plot shows that the evaluation of the ranking based on the sum of the absolute rank difference gives the best result. The ranking based on the Levenshtein Distance is inferior. Both ranking measures evaluated the ranking of the 35 best classes, but showed similar results for other numbers of best classes N. The normalization give the best results of the posterior probability based measures, but the approximation of the unconditioned likelihood with a fix probability can reject up to 20 % of the recognition errors and unknown people without any false alarms. A closer look at this effect shows that most of these rejections are other object than faces. Therefore the fix probability can be used for a face/non-face decision. The filler model gives the worst results. This is caused by training the filler model on the

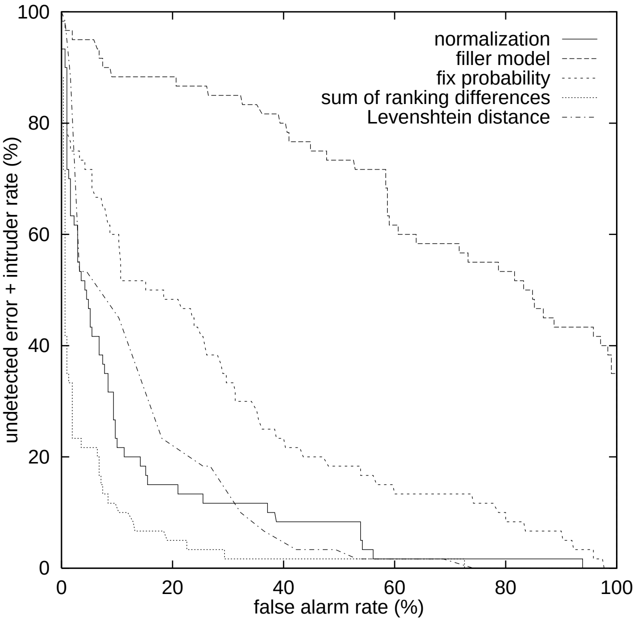

Figure 6. Rejection of undetected errors and unknown people

images of the training set instead of using an independent set of faces.

The separation of the results in unknown people and recognition errors allows a further evaluation of the results. Figure 7 shows the undetected intruder rate and Figure 8 shows the undetected error rate. The ranking based on the sum of the absolute rank difference is the best measure for the detection of unknown people, and should therefore be used in applications where unknown people are an issue. The posterior probability based on normalization gives the best results, as expected from the theory, for the detection of recognition errors.

6 Conclusions

Five different confidence measures for the rejection of recognition errors and unknown people are compared. Three measures were based on the posterior probability using the Bayes' rule and two evaluated the likelihood ranking of the classes for the test image. Experiment expectedly showed that the posterior probability based on normalization gives the best results for the detection of recognition errors. For the rejection of unknown people the unconditioned likelihood $P(\mathbf{O})$ cannot be calculated, because of an almost infinite number of classes, and has to be approximated. The evaluation of the raking with the sum of absolute rank differences resulted in a better rejection of unknown people than the posterior probability using any approximation of

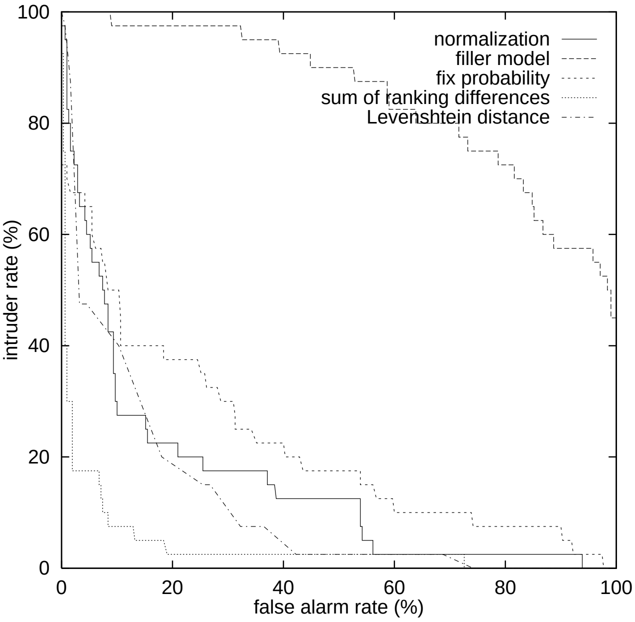

Figure 7. Rejection of unknown people

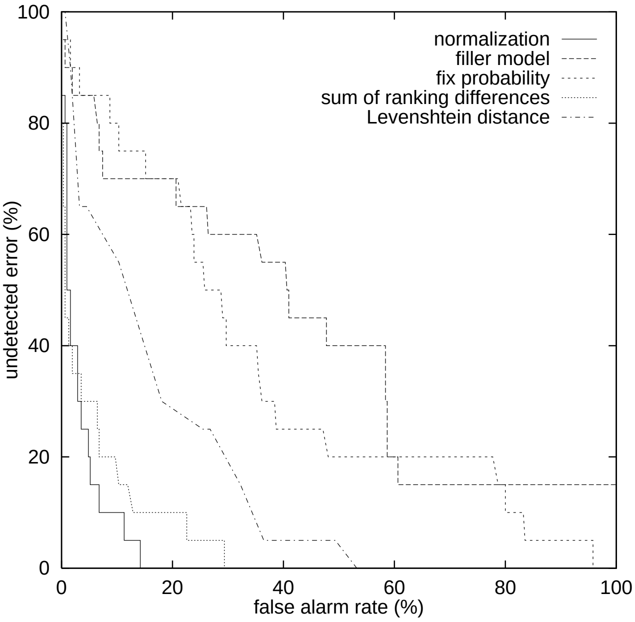

Figure 8. Rejection of undetected errors

the unconditioned likelihood $P(\mathbf{O})$. The rejection based on ranking can be further improved by a person dependent number of similar people N that is determined on additional training data. In the future we will apply the rejection to the main face recognition tasks: access control and video indexing.

References

[1] Gethin Williams. *Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition*. PhD thesis, Department of Computer Science, University of Sheffield, May 1999.

[2] Daniel Willett, Andreas Worm, Christoph Neukirchen, and Gerhard Rigoll. Confidence Measures for HMM-based Speech Recognition. In *5th International Conference on Spoken Language Processsing*, pages 3241–3244, Sydney, December 1998.

[3] Stefan Eickeler, Stefan Müller, and Gerhard Rigoll. Recognition of JPEG Compressed Face Images Based on Statistical Methods. *Image and Vision Computing Journal, Special Issue on Facial Image Analysis*, 2000. to appear.

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

[5] Richard O. Duda and Peter E. Hart. *Pattern Recognition and Scene Analysis*. John Wiley & Sons, New York, 1973.

[6] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[7] J. G. A. Dolfing and A. Wendemuth. Combination of Confidence Measures in Isolated Word Recognition. In *5th International Conference on Spoken Language Processsing*, pages 3237–3240, Sydney, December 1998.

[8] J. Robin Rohlicek, William Russel, Salim Roukos, and Herbert Gish. Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 627–630, Glasgow, May 1989.

[9] Richard C. Rose and Douglas B. Paul. A Hidden Markov Model Based Keyword Recognition System. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 129–132, Albuquerque, April 1990.

[10] Stefan Eickeler, Andreas Kosmala, and Gerhard Rigoll. Hidden Markov Model Based Continuous Online Gesture Recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 1206–1208, Brisbane, August 1998.

[11] Gerhard Rigoll and Andreas Kosmala. A Systematic Comparison Between On-Line and Off-Line Methods for Signature Verification with Hidden Markov Models. In *International Conference on Pattern Recognition (ICPR)*, pages 1755–1757, Brisbane, August 1998.

[12] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

[13] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.