ELSEVIER

Pattern Recognition Letters 18 (1997) 827–833

Pattern Recognition Letters

SESAM: A biometric person identification system using sensor fusion

U. Dieckmann [*], P. Plankensteiner [1], T. Wagner [2]

Fraunhofer-Institute for Integrated Circuits, Am Weichselgarten 3, D-91058 Erlangen, Germany

Abstract

In the present paper we describe the person authentication system SESAM. Person identification and verification still is a very difficult task. Using one biometric feature, i.e. the photograph or the sound of the voice, leads to good results, but there is no reliable way to verify the classification. In order to reach robust identification and verification we are combining three different biometric cues. These cues are dynamic, i.e., the sound of the voice and the lip motion, and static, i.e., the fixed image of the face. Each branch is preprocessed and classified separately and the results are combined, e.g., in a 2-from-3 manner. The recognition of persons may be used for pure identification or can be varied to a verification system. For both cases we have done a field test to show that this approach leads to a reliable person authentication system. © 1997 Elsevier Science B.V.

Keywords: Biometric person identification/verification; Synergetic computer; Sensor fusion; Face recognition

1. Introduction

This paper describes the person authentication system SESAM. [3] The concept of SESAM includes three different biometric cues from two different data sources: One static cue derived from an image of the face and two dynamic cues, the spectrum of the sound and the lip motion of a person saying its name in front of the system.

These cues are analysed independently and used for the training of three separate classifiers. The results of these classifiers are combined using a 2-from-3 approach. This combination has proved to be very reliable and robust against changing lighting conditions (sun movement, clouds, changing electric lights) and against a noisy environment. If one cue is disturbed the other two still guarantee a safe classification.

The combination of static and dynamic biometric features of a person is also a sophisticated instrument against faking and criminal attacks. It is fairly easy to fake static features, i.e. the static image of a person. But it is much more complicated to fake dynamic features like voice or lip movement.

Experiments at the entrance of the Fraunhofer Institute for Integrated Circuits show that the system is robust and reliable in everyday practice. Changes of lighting conditions during the day and a noisy environment are tolerated by the system. These experiments also show the superior results of the whole

[*] Corresponding author. Email: die@iis.fhg.de.

[1] Email: ppl@iis.fhg.de.

[2] Email: wag@iis.fhg.de.

[3] German: **S**ynergetische **E**rkennung mittels **S**tandbild, **A**kustik und **M**otorik.

0167-8655/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved.
PII S0167-8655(97)00063-9

system in contrast to the results produced by each single cue.

This paper is organised in four sections. Section 2 describes the algorithms for speech analysis, face location and scaling, the optical flow and the classifier. Section 3 describes our experiments and their results. In Section 4 an outlook on further work is presented.

2. Algorithms

2.1. SESAM concept: Sensor fusion

The concept of SESAM is to combine three different static and dynamic biometric cues for identifying persons (Fig. 1). The idea of combining different biometric cues is that we as human beings also use different biometric hints to recognise a person. First we look at the face of a person and in addition unknowingly we use the gesture or the sound of the voice to be absolutely sure of the identification of the person we look at.

To follow this idea we have to record all cues we want to consider. Each cue is recorded separately and preprocessed independently. The analysed data of the three cues is trained by three separate classifiers. For a classification the results of the classifiers is combined. Two of three results must lead to the same person and exceed a given threshold in order to make the classification more reliable. The combination of the results takes place as the last step of the classifying process. If one channel is disturbed, the other two cues may be in good shape for a reliable classification.

2.2. Optical cues

The optical cues are recorded with a gray-level CCD-camera (768 × 572 pixels). A one-second video recording is triggered by the acoustic signal. Now the first two cues are extracted.

2.2.1. Locating the face in a real time system

In order to extract a still image of the face and the lip motion it is necessary to locate exactly the face within the image and the mouth position within the face. Once the positions are found relevant data for the person identification can be extracted from the video sequence. The preprocessing of this sequence is split into the following tasks.

2.2.2. Reflection suppression

SESAM is equipped with its own infra-red light source which provides a constant illumination of the

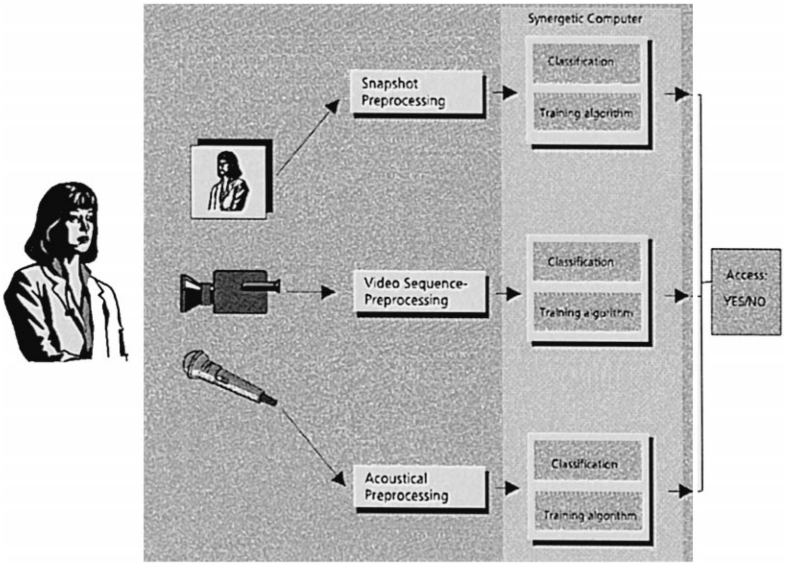

Fig. 1. Sensor fusion leads to a better performance.

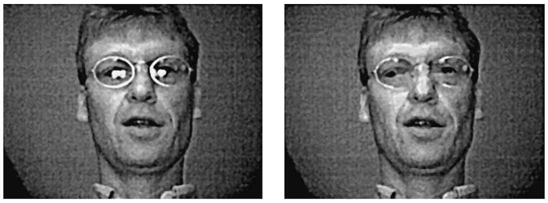

Fig. 2. Before and after the suppression of the reflections.

face in changing lighting conditions. As a side effect bright spots on eyeglasses as a reflection of the infra-red lights are unavoidable (Fig. 2). As we use gray level projections and template matching for locating the face these reflections must be eliminated. The area of the spots must be filled with appropriate pixel values. The average gray value of the image is used to determine a threshold value. Then a scan-line algorithm replaces pixel values above this threshold by bilinear interpolation using the gray levels of the surrounding pixels. This leads to an image without disturbing bright spots.

2.2.3. Horizontal projection

As we have a simply structured background a horizontal projection of the gray level pixel values can be used to localise the vertical edges of the face (Brunelli and Poggio, 1993). The graph representing the sums of the columns of the picture is calculated (horizontal projection). This curve is scanned from the left and from the right for the first high gradient which represent the left and right border of the face (Fig. 3).

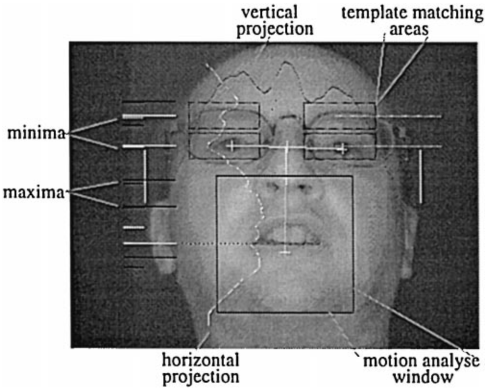

Fig. 3. Statistical means are suitable to determine the shape of the head.

2.2.4. Vertical projection

A second projection curve is calculated representing the sums of the rows within the borders of the face (vertical projection). This curve is scanned for local minima and maxima that represent forehead, eyes, nose, mouth and sometimes the chin. Using statistical means it is possible to assign each minimum and maximum to its associated facial feature.

2.2.5. Template matching

As a next step the exact eye positions and the mouth position must be localised. Template matching has proved to be a reliable tool in pattern recognition (Brunelli and Poggio, 1993; Frischholz et al., 1994). Nevertheless the search area should be restricted to reduce processing time, especially in a real time system like SESAM. Therefore the eye positions are first guessed from the results of the previous steps and then localised in a fine-grain search in the determined areas. This is done by template matching with a prototype eye template. From here it is only a small step to an exact mouth localisation. As a natural constant biometric relation the eye-to-eye distance is very close to the eyes-to-mouth distance. Additionally the middle of the mouth usually can be found on the symmetry axis determined by the two located eye positions. The mouth position can be found by these means (Fig. 3), and is further refined by another local vertical projection. Now both eye and mouth positions are known.

2.2.6. Generating the still image

In order to do a classification or verification of the still images the extracted faces should all have the same size and should be aligned such that eyes and mouth lie on the same positions for all images. This can be achieved in the following way: first a rotation angle is determined to align both eyes horizontally. Then a horizontal scaling factor is specified depending on the eyes-to-mouth distance and a vertical scaling factor depending on the eye-to-eye distance. The extraction of the face in one of the images of the video sequence can then be performed by an algorithm that does rotation, cutting and rescaling by bilinear interpolation in one step. The resulting gray-level image is used as a feature vector for the synergetic computer and consists of 31329 floating point values.

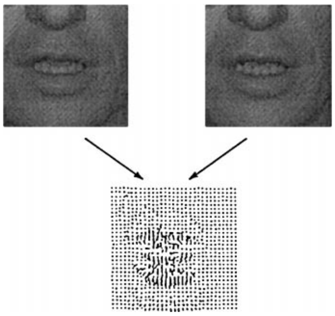

Fig. 4. Two snapshots from an image sequence and the corresponding optical flow.

2.2.7. Generating the mouth sequence

For motion analysis of the lip movements a sequence of sub-images containing only the mouth area is extracted from the video sequence. A 128 × 128 pixel window centered on the located mouth positions in each of the original images of the video sequence is used. For this step neither scaling nor rotating is required.

2.2.8. Lip motion in mouth sequences

An optical flow analysis using the method of Horn and Schunk (Horn and Schunk, 1981) is applied to the generated mouth sequence. This algorithm extracts the motion in a image sequence in a quick and robust manner. The optical flow is calculated between each two consecutive frames and stored in 16 vector fields of 32 × 32 vectors. Fig. 4 shows the corresponding optical flow field between two frames from an image sequence of a person speaking the word ‘‘Dieckmann’’. In order to guarantee invariance with respect to spatial and temporal shifts, the power spectrum from the three-dimensional motion field is calculated. The resulting feature vector contains 16384 floating point values.

2.3. Voice specific features

The voice is recorded by a microphone pointing to the person speaking. The recording unit, video and audio, is always active. To trigger the analysing process a person standing in front of the system has to speak his/her name. As soon as the volume of the voice reaches a certain threshold both recording branches are triggered and stopped after 800 ms. The data sampled 200 ms before the trigger is also included in the preprocessing.

2.3.1. Acoustic preprocessing

In order to avoid spectral leakage the acoustical data is first windowed by a Hann-window and then mapped into the time-frequency domain by a short time Fourier transformation. Using 32 windows with 1024 samples in each window ensures a high frequency-resolution. Fig. 5 shows the time-frequency domain of the spoken word ‘‘Dieckmann’’. A two-dimensional Fourier transformation guarantees invariance to spatial and frequency shifts. Again, using the power spectrum leads to a reduction of the data size.

Due to the nature of the human voice, lower frequencies have to be emphasised and higher frequencies have to be united. This is accomplished by a special power function which compresses the coefficients of the higher frequencies. A secondary objective of this function is further data reduction. The

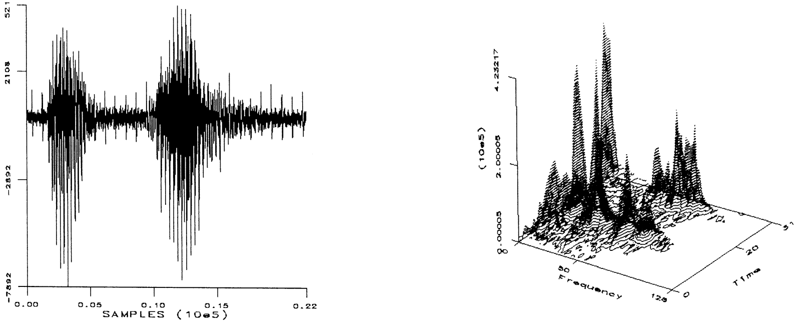

Fig. 5. Acoustic signal of the word ‘‘Dieckmann’’ (top) and its short time Fourier transformation (bottom).

final feature vector used for the classification consists of 4096 floating point values only.

2.4. The classifier

As a classifier we use the *Synergetic Computer*. The theory of Synergetics can be found in (Haken, 1993). The Synergetic Computer is represented by a class of algorithms of which we use the MELT-algorithm.[4] Let us just point out the crucial points of the algorithm:

2.4.1. The MELT-algorithm

A set of $\sum_{k=1}^{M} m_k$ learning vectors $\nu_{kj} \in \mathbb{R}^N$, $j \in 1,\ldots,m_k$ form M classes with m_k patterns for each class k. The patterns are normalised to the average 0 and length 1. For these normalised vectors one has to find so called *adjoined prototypes* ν_k^+, following the constraint:

$$\nu_k^+ \nu_{k'j} = \delta_{k'k}, \quad \forall k,k' \in 1,\ldots,M,\ j \in 1,\ldots,m_{k'}.$$

That means the scalar product of all learning vectors of the same class is 1 and all learning vectors of one class are orthogonal to all other learning vectors of the other classes. Finally all learning vectors of one class are summed up into one prototype.

The classification of a test pattern q is achieved by first calculating the Euclidean distance to the prototypes of all classes:

$$\zeta_k = \nu_k^+ q.$$

The maximum ζ_k then determines the class k that forms the result of the classification (winner-takes-all principle).

2.5. Adaptation for person verification

The methods developed in the SESAM system can also be used for the verification of a person. The verification process is slightly different from the identification process. In contrast to identification, where no additional hints from the person are given (only its biometric features) the verification process knows in advance who the person is. The person to be verified has to present a password, a pin number or a similar identification. This hint must be verified against the biometric features of the person.

For the identification task we have to find out which class from the training set of classes represents the person. For the verification task we know in advance which class represents the person and if the person does not meet this class he or she will be rejected. We refer to this fact in the verification task by creating two classes for each trained person. The first class, called the reference-class, contains the biometric data of the person to be verified. The other class, called the union-class, has to represent the ‘‘rest of the world’’ in such a manner that an intruder is classified into the union class. The amount of training data is 10 patterns for the reference class and 30 patterns for the union class. The success of the whole system highly depends on a good selection of the training data for the union class. Before we are going to choose the optimal training data we must provide a measure to judge the quality of a selection.

The efficiency of a verification system is described by the two rates: FRR (False Rejection Rate) and FAR (False Acceptance Rate). Since we consider the minimisation of the FRR as more important for practical use, we combine the two rates in the following way to obtain a criterion G for the quality of our selection:

$$G = \frac{2\mathrm{FRR} + \mathrm{FAR}}{3}.$$

The goal of a selection method is to find a selection that minimises G. To approximately reach this aim we perform a repeated random pattern selection for each class together with each data source (see Section 3.3.1).

3. Experiments and results

3.1. Field test

To examine the performance of the system over a longer period of time, a large field test was carried out at the entrance of our institute. A special recording station had to be developed which allowed an easy acquisition of optical and acoustical data. The

[4] MELT: the prototypes of one class are *melted* into one prototype.

sensor unit consists of a standard CCD camera and a dynamic microphone. An integrated infra-red light source and an optical filter (frequency range down to 950 nm) help to avoid disturbing influences of changing illumination and background. A connection to the buzzer of our entrance door provided access to the institute for 66 staff members by stopping in front of the camera and saying their surname.

The variations to be considered in the field test include both daily perturbations like different daylight conditions and longtime influences like seasons or gradual changes of outer appearance of a person. The data set used for the experiments described below consists of recordings made in the period of July 1995 to May 1996. Altogether we have 2822 recordings. Notice that each of the recordings contains 2.8 MB of data.

3.2. Identification experiments

In this experiment we were doing pure identification. A set of persons was trained into the system. This means that for a person to enter the institute there was no need of additional requirements like PIN number or password. In order to determine the FRR and FAR of the identification we used the IDENT set with a total of 1428 recordings. Each person was trained into the classifiers using 8 training patterns and the remaining patterns from the IDENT set were used as the test patterns.

The results in Table 1 show that the combination of the different cues leads to a higher performance of the system. The recognition rate of 93% in the 2-from-3 test is higher than any recognition rate in the single branches. The FAR rate is very low compared to the single branches; even without any threshold.

Table 1
Results of the identification

Classification IDENT	Correct %	Rejected %	false %
speech	89.6	–	10.4
speech (0.1)	78.2	19.7	2.1
opt. flow	89.0	–	11.0
opt. flow (0.1)	78.2	l20.5	1.3
face image	81.3	–	19.7
face image (0.1)	74.6	18.2	7.2
2-from-3	93.0	6.6	0.4
2-from-3 (0.1)	82.6	17.2	0.2
2-from-3 (0.2)	69.0	30.8	0.2

Table 2
Data sets for verification experiments

Set	Persons	Recordings person	Recordings total
target	15	≥ 110	2084
common	26	≤ 18	240
test	25	20	500

3.3. Verification experiments

For the verification experiments the whole data base was split into three parts as shown in Table 2.

Each person of the target set is subject to verification. The common set is used for choosing the patterns of the union class and the test set is used to test each chosen configuration of reference class and the union class.

3.3.1. Random pattern selection

A simple approach to the selection problem is a repeated random selection of training patterns for each class. The reference class is randomly filled with 10 patterns of the person (in the target set) who must be verified. The union class is randomly filled with patterns from any person in the common set. Then this selection is tested by verifying all patterns of all persons contained in test set and the remaining patterns of the person trained into the reference class. This random pattern selection is carried out 50 times for each of the three data cues. Then the best result of each branch is combined to perform a final verification making a 2-from-3 decision. A set of two classes for each person in the target set was determined this way. Table 3 shows the mean values of the verification results obtained for all 15 persons in the target set.

Table 3 also shows the results when a 3-from-3 decision (AND combination of the three sources) was made. Row one to three show the results for each data source separately. Row four to six contain the results for the combination using 2-from-3 with no security threshold and with a threshold of 0.1 and 0.4. Finally the last row shows the result based on a 3-from-3 decision.

Table 3
Results of the verification

Classification test/target	1-FRR (%)	1-FAR (%)
speech	99.1	98.1
opt. flow	97.4	96.5
face image	99.5	97.3
2 from 3	99.8	99.7
2 from 3 (0.1)	99.50	99.85
2 from 3 (0.4)	94.32	99.96
3 from 3	97.25	100

Again the results in Table 3 show that the combination of the results is superior to the results in each single branch. The FRR and FAR are very low, 0.21% and 0.33%, respectively.

4. Outlook

The person authentication system SESAM provides good performance in person verification at both low FRR and FAR rates using a 2-from-3 decision and even a minimum FAR rate of 0% using an AND combination of the three data sources. Our next aim is to reduce the amount of storage space of our learning patterns, so that we are able to store the whole required training data on a chip card. The system could then easily be used for tasks like automatic banking (ATM) and identification/verification on computer terminals.

Another goal is implementation of colour in the optical branch. We believe that there is vital information in the colour for improving the systems robustness and reliability.

For further reading see (Dieckmann, 1993; Dieckmann et al., 1995; Haken, 1991; Mase and Pentland, 1991; Schamburger, 1996; Schindel, 1993; Wagner and Boebel, 1993; Wagner and Dieckmann, 1995).

References

Brunelli, R., Poggio, T., 1993. Face recognition: Features versus templates. IEEE Trans. Pattern Anal. Machine Intell. 15 (10), 1042–1052.

Dieckmann, U., 1993. Personenerkennung mit einem synergetischen Computer. Diplomarbeit, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany.

Dieckmann, U., Plankensteiner, P., Wagner, T., 1995. Multisensory pattern analysis for person identification with synergetic computers. In: Proc. Internat. Workshop on Automated Face- and Gesture-Recognition, Zürich, Switzerland.

Frischholz, R.W., Böbel, F.G., Spinnler, K.P., 1994. Face recognition with the synergetic computer. In: Proc. 1st Internat. Conf. on Applied Synergetic and Synergetic Engineering, Erlangen, pp. 100–106.

Haken, H., 1991. Synergetic Computers and Cognition. A Top-Down Approach to Neural Nets. Berlin.

Haken, H., 1993. Synergetics. An Introduction. 3rd ed., Berlin.

Horn, B., Schunk, B., 1981. Determining optical flow. Artificial Intelligence 17, 185–203.

Mase, K., Pentland, A., 1991. Automatic lip reading by optical flow analysis. In: System and Computers in Japan, pp. 67–75.

Schamburger, R., 1996. Entwicklung eines robusten Verfahrens zur Extraktion von Lippenbewegungen aus Videosequenzen. Diplomarbeit, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany.

Schindel, M., 1993. Theorie eines Halbleitersystems zur Realisierung der Ordnungsparameterdynamik eines Synergetischen Computers. Ph.D. Thesis, Stuttgart, 1993.

Wagner, T., Boebel, F., 1993. Testing synergetic computers with industrial classification problems. INNS Neural Networks 44.

Wagner, T., Dieckmann, U., 1995. Sensor fusion for robust identification of persons: A field test. In: IEEE ICIP 95, Washington, USA.