# Latent Semantic Analysis of Facial Action Codes for Automatic Facial Expression Recognition

Beat Fasel
D-ITET/BIWI
ETH Zurich
Zurich, Switzerland
bfasel@vision.ee.ethz.ch

Florent Monay
IDIAP Research Institute
Martigny, Switzerland
monay@idiap.ch

Daniel Gatica-Perez
IDIAP Research Institute
Martigny, Switzerland
gatica@idiap.ch

## ABSTRACT

For supervised training of automatic facial expression recognition systems, adequate ground truth labels that describe relevant facial expression categories are necessary. One possibility is to label facial expressions into emotion categories. Another approach is to label facial expressions independently from any interpretation attempts. This can be achieved via the facial action coding system (FACS). In this paper we present a novel approach that allows to automatically cluster FACS codes into meaningful categories. Our approach exploits the fact that FACS codes can be seen as documents containing terms -the action units (AUs) present in the codes- and so text modeling methods that capture co-occurrence information in low-dimensional spaces can be used. The FACS code derived descriptions are computed by Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA). We show that, as a high-level description of facial actions, the newly derived codes constitute a competitive alternative to both basic emotion and FACS codes. We have used them to train different types of artificial neural networks.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*

## General Terms

Algorithms, Theory

## Keywords

Latent Semantic Analysis, Automatic Facial Expression Recognition


*MIR'04*, October 15–16, 2004, New York, New York, USA.


## 1. INTRODUCTION

Automatic facial expression recognition is an active research area with applications in human-computer interfaces and human emotion analysis [15][10]. The training of automatic facial expression systems relies on the availability of face image databases of sufficient size, with high-quality ground truth labels. Unfortunately, labeling facial expressions is not only a tedious endeavor, but also prone to errors, even for trained human annotators.

Many facial expression systems in the literature use basic emotion labels. Basic emotions were first postulated by Ekman and Friesen [5], each one possessing a distinctive content together with a unique facial expression, and encompassing joy, surprise, anger, fear, disgust and sadness. However, coding facial expressions directly into basic emotion categories has several drawbacks: (1) emotion categories can only describe a subset of all facial expressions, (2) emotions constitute not a neutral description of facial actions but an interpretation, and (3) the judgement of different coders can vary a great deal.

An alternative to basic emotion codes is the FACS coding approach. Defined by Ekman and Friesen [6], FACS allows for interpretation-independent description of facial actions, and has been successfully used for various applications, notably in the field of psychology.

As a discrete representation, FACS codes feature a great number of different action unit combinations in order to allow for a comprehensive description of facial actions. Theoretically, the number of possible AU combinations amounts to $AU_{comb}(n,k) = n^k$, where $n$ corresponds to the number of facial expression intensities and $k$ to the number of AUs. FACS encompasses 44 facial action and 8 head pose action units at 5 asymmetric intensity levels, co-occurring in various combinations. Even by assuming a single facial expression intensity, and by employing the 44 AUs that describe facial actions, the number of possible AU combinations is $2^{44}$. Hereby, $n$ was set to 2, comprising a neutral face and a facial action display state. With five intensity levels and a neutral face state, the total number of different categories amounts to $6^{44}$. Of course, in reality, not all of these action unit combinations can occur due the fact that some AUs describe similar or overlapping face areas. The same is true for facial expression intensities, where only one intensity level can be perceived for a given AU at any time. Furthermore, some AU combinations are physiologically not possible. All the same, the number of possible AU combinations is important. So far, about 7000 valid AU combinations have been

identified within the FACS framework [4].

Handling FACS directly thus overlook two important points. In the first place, FACS codes often reveal unnecessary details that can hamper data-driven facial expression recognition approaches. The sheer number of combinations can lead to a bad generalization performance as it is virtually impossible to have access to a training database that covers all possible AU combinations while featuring a sufficient number of instances of specific facial expressions. In the second place, most AUs do not appear independently, but in combinations. Facial expressions are indeed characterized by the co-occurrence of AUs and AU intensities. Viewed as a collection of discrete data, and using a clear analogy with the vector-space representation for text documents in information retrieval, a FACS database is suitable for the use of latent space models, well-known for capturing co-occurrence information in a low-dimensional semantic space [3, 12].

In this paper, we propose two approaches based on latent semantic analysis for the automatic clustering of AUs, in order to cope with relatively small FACS-labeled databases. Our methods allow to group FACS codes into semantically close facial expression categories. We do not cluster the images themselves, but the text-based image descriptions of facial actions. This allows us to significantly reduce the number of expression categories, and thus alleviate the problem of the large number of possible AU combinations, while achieving semantic clustering of facial expressions without using image-based facial actions. We demonstrate the application of the clustered facial expression codes for the training of neural networks for automatic facial expression recognition, showing that they represent an competitive alternative to both basic emotion and FACS codes.

The rest of the paper is organized as follows. For sake of completeness, Section 2 briefly reviews the latent space models used in our work. Section 3 describes the strategies to map latent space representations to facial expression categories. Section 4 describes the architectures used for automatic recognition. Section 5 describes experiments and discusses the results. Section 6 provides some concluding remarks.

# 2. LATENT SEMANTIC ANALYSIS

In the context of text information retrieval, retrieving or clustering documents that are semantically close by exact keyword matching is prone to fail due to *synonymy*, where two different words may refer to the same concept, and *polysemy*, where the same words may refer to different concepts, depending on the context. In the same way, two facial expression displays can be similar from a semantic point of view even though their visual appearance (e.g. described by FACS AUs) is not exactly the same. On the other hand, the same AUs can be shared amongst distinct facial expression displays. We therefore attempt to tackle this problem with latent semantic analysis (LSA) [3], and probabilistic latent semantic analysis (PLSA) [12]. So far, these techniques have found various applications in information retrieval and natural language processing. However, we are not aware of any similar work that has applied the same concept to our problem.

## 2.1 Standard Latent Semantic Analysis

The key idea behind LSA is to map high-dimensional term count vectors, such as the ones arising in vector space representations of text documents, to a lower dimensional representation in a so-called *latent semantic space*. LSA, decomposes the $n \times m$ term-by-document matrix $\mathbf{A}$ into three matrices via a truncated singular value decomposition (SVD),

$$\tilde{\mathbf{A}} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T, \tag{1}$$

where $\mathbf{U}_k \in \mathbb{R}^{n \times k}$, $\mathbf{\Sigma_k} \in \mathbb{R}^{k \times k}$, $\mathbf{V}_k \in \mathbb{R}^{m \times k}$ and $k < r$. Dimensionality reduction is then performed in a next step by thresholding the matrix $\mathbf{\Sigma}$. The new basis is a linear combination of terms of the original document space, which is supposed to describe with sufficient accuracy the latent topics expressed in the analyzed corpus. Thus, the reduced SVD representation is assumed to capture the major associative relationships between terms and documents.

In order to attribute documents to certain categories, or cluster FACS coded expressions, we can employ queries. Hereby, documents or facial expression codes that are semantically close to a given query $\mathbf{q}$ can be retrieved via following projection

$$\hat{\mathbf{q}} = \mathbf{q}^T \mathbf{U}_k \mathbf{\Sigma}_k^{-1}, \tag{2}$$

where $\mathbf{q}$ is a vector in a $n$-dimensional space. It corresponds to a set of words that are indicative for documents of interest. The result of the query, $\hat{\mathbf{q}}$, is a vector in a $k$-dimensional space and can then be compared (via the cosine distance measure) to all existing document vectors. Hereby, the documents are ranked by their similarity with regard to the query. Therefore, given a few labeled documents or FACS codes, we can index the reminding documents.

## 2.2 Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis [12] is a relatively novel statistical technique for the analysis of co-occurrence data. In contrast to standard latent semantic analysis, which stems from linear algebra and performs a singular value decomposition of co-occurrence tables, PLSA is based on a mixture decomposition derived from a latent class model. It can be described as follows

$$p(w_j, d_i) = p(d_i) \sum_{l=1}^{k} p(z_l \mid d_i) p(w_j \mid z_l), \tag{3}$$

where $w_j$ refers to the word $j$ and $d_i$ to the document $i$. Hereby, in the context of FACS representations, the words correspond to the AUs and the document to the FACS code. This so called *aspect model* in asymmetric parameterization is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = \{z_1, ..., z_k\}$with each observation. Model fitting is performed by expectation maximization (EM). In order to allow for a better generalization performance, early stopping training, a document perplexity measure as well as a tempered EM algorithm according to [12] were deployed, see also [11].

# 3. FROM ASPECTS TO FACIAL EXPRESSIONS

Before we can employ the latent semantic analysis approaches described in the previous section for the retrieval and clustering of facial expression aspects, we first have to introduce a technique that allows to measure how well aspects map to ground truth facial expression categories. Furthermore, we discuss a strategy, that allows to select PLSA

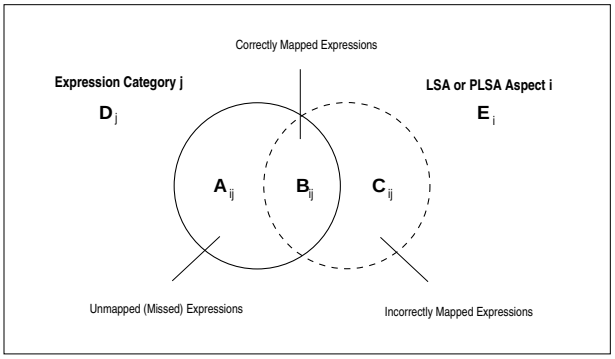


**Figure 1: Aspect to Expression Category Mapping: $A_{ij}$ indicates the set of missed, $B_{ij}$ the set of correctly mapped and $C_{ij}$ the set of incorrectly mapped facial expressions. The current target facial expression category is denoted by $j$ and the current LSA or PLSA aspect by $i$.**

computed aspects that can be mapped to a meaningful facial expression display.

## 3.1 Mapping Performance Measurements

For the comparison of the automatically and manually (i.e. real) derived correspondences, we can employ the following quality measures that stem from information retrieval:

$$precision(i,j) = \frac{|B_{ij}|}{|B_{ij}| + |C_{ij}|} = \frac{|B_{ij}|}{|E_i|} \tag{4}$$

and

$$recall(i,j) = \frac{|B_{ij}|}{|A_{ij}| + |B_{ij}|} = \frac{|B_{ij}|}{|D_j|} \tag{5}$$

where $|\cdot|$ denotes the number of elements, $A_{ij}$ the set of missed (and thus unmapped) facial expressions, $B_{ij}$ the set of correctly assigned facial expressions and $C_{ij}$ the set of incorrectly assigned facial expressions in category $j$ with aspect $i$. Figure 1 illustrates the relationship between $A_{ij}$, $B_{ij}$ and $C_{ij}$. Furthermore, we define $D_j$ as the set of facial expressions contained in the ground truth category $j$ and $E_i$ the set of facial expressions in the aspect $i$, where $|D_j| = \sum_i (|A_{ij}| + |B_{ij}|)$ and $|E_i| = \sum_j (|B_{ij}| + |C_{ij}|)$.

*Precision* is the percentage of correctly assigned expressions in relation to the total number of aspects, while *recall* is the percentage of correctly assigned expressions in relation to the total number of expressions. Hence, *precision* and *recall* do measure different properties and we therefore need a combined quality measure in order to determine the best matching aspect to expression category mappings. The so called *F-measure* $fm$ computes the harmonic mean of *precision* and *recall* and allows to take into account both properties at the same time:

$$\begin{aligned} fm(i,j) &= \frac{2 \cdot precision(i,j) \cdot recall(i,j)}{precision(i,j) + recall(i,j)} \\ &= \frac{2\,|B_{ij}|}{(|D_j|) + (|E_i|)} \end{aligned} \tag{6}$$

Global measurements of aspect to ground truth expression category mapping encompass the *overall precision* defined as

$$precision_{ov} = \sum_j \frac{|D_j|}{|D|} \max_{i \in M} [precision(i,j)], \tag{7}$$

the *overall recall*

$$recall_{ov} = \sum_j \frac{|D_j|}{|D|} \max_{i \in M} [recall(i,j)] \tag{8}$$

and the *overall F-measure*

$$fm_{ov} = \sum_j \frac{|D_j|}{|D|} \max_{i \in M} [fm(i,j)], \tag{9}$$

where $fm(i,j)$ denotes the F-measure of aspect $i$ and ground truth target facial expression category $j$. $M$ is the set of aspects $i$ that are mapped to the target facial expression category $j$. This can either be single aspects or multiple aspects assigned to the same target expression category, see also Subsection 5.3. The overall measures weight the different scores according to the number of facial expressions contained in category $j$. Note that the overall recall $recall_{ov}$ is also known as *accuracy*. For more information on information retrieval we refer to [1].

## 3.2 Mapping PLSA Aspects to Expression Categories

Our goal is to find PLSA aspect that represent meaningful facial expression categories. Unlike query-extracted LSA aspects that directly correspond to ground truth facial expression categories, unsupervised clustered PLSA aspects have to be mapped to the closest facial expression ground truth categories. This was achieved by calculating *precision*, *recall* and *F-measures* for every possible aspect - expression category pair $(i,j)$, indicating a mapping operation of a given aspect $j$ to a ground truth facial expression category $i$. We chose the *maximum overall F-measure* $fm_{ov}$, described in formula 9, to be the decision factor determining which facial expression category a given aspect should be mapped to. For the actual mapping procedure, we propose the algorithm described in Algorithm 1 that allows to map multiple aspects to single facial expression categories. Hereby, all aspects that lead to a maximum precision for a given facial expression category are included into the aspect-to-expression category mapping matrix $M$ if they improve the current overall F-measure $f_{old}$, i.e. $f_{new} > f_{old}$. Note that aspects mapping with a low *overall F-measure* to an expert category (ground truth expression category) can only be accepted if the *F-measure* between them exceeds a certain threshold value. This ensures that a minimum degree of similarity between a target facial expression category and an aspect exists before allowing an association. We have used a threshold value $f_{thresh}$ of 0.20. This value was chosen based on observations of how *F-measures* relate between different aspect - expression categories pairs with varying degrees of similarity.

# 4. AUTOMATIC FACIAL EXPRESSION RECOGNITION

Automatic facial expression recognition approaches should be able to cope with pose variations of subjects, large intra-class facial expression variations and be robust to environ-

**Algorithm 1 Multiple PLSA Aspects to Expression Category Mapping.** Using all Aspects that Increase Expression Category Specific Overall F-measures.

*Requirements*
- $\mathbf{B}_{(a\times e)}$Aspect count matrix *(a: # aspects, e: # expressions)*
- $\mathbf{F}_{(a\times e)}$F-Measure matrix
- $\mathbf{P}_{(a\times e)}$Precision matrix
- $a > e$

*Initialization*
$\mathbf{M} := \mathbf{0}_{(a\times e)}$ *Mapping matrix*
$f_{thresh} := 0.2$ *F-Measure threshold*

*Algorithm*
**for** j := 1 to e
*Current aspect has max precision for current expression?*
$s_{old} := 0$
$f_{old} := 0$
**for** i :=1 to a **do**
$[p, ind] = \max_{k=1..e} \mathbf{P}(i,k)$
*Adding the current aspect improves the F-measure?*
**if** (ind = j **and** $\mathbf{F}(i,j) \geq f_{thresh}$) **then**
$s_{new} := s_{old} + \mathbf{B}(i,j)$
$t_{new} := t_{old} + E(i)$
$f_{new} := \frac{2 \cdot s_{new}}{t_{new} + D(j)}$
**if** $f_{new} > f_{old}$ **then**
$\mathbf{M}$(i,j) := 1
$f_{old} := f_{new}$
$s_{old} := s_{new}$
$t_{old} := t_{new}$
**endif**
**endif**
**endfor**
**endfor**


**Figure 2: Sample images from the Cohn-Kanade DFAT-504 facial expression database. Shown are the six basic emotions on the left hand side as well as the mouth opening display on the right hand side.**

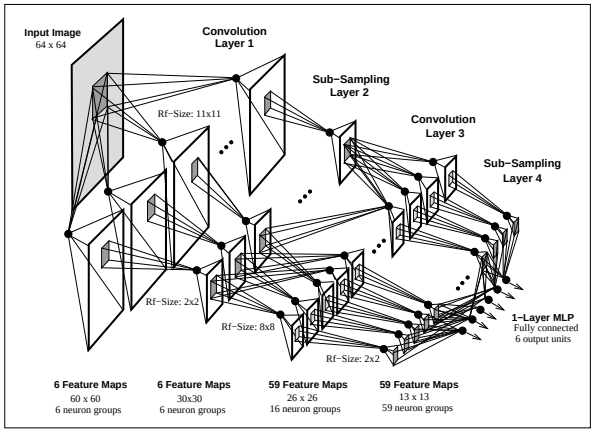


**Figure 3: Convolutional Neural Network Architecture for the Recognition of Facial Expressions. Shown is a network featuring 2 convolutional, 2 subsampling and a fully connected MLP layer. In contrast to many conventional neural networks such as Multilayer Perceptrons, Convolutional Neural Networks (CNNs) are not fully connected and process information locally via receptive fields. Like many other neural networks, CNNs are trained with an adapted version of the back-propagation algorithm.**

mental variations due to changing backgrounds and lighting. Ideally, no manual intervention should take place during both training and deployment, including segmentations and initializations. In this paper we compare two neural network architectures that were trained with the facial expression aspects computed by the above mentioned information retrieval methods. The first is a standard Multilayer Perceptron (MLP) and the second a convolutional neural network (CNN) [14], also known as Neoperceptron (NP), see Figure 3. The latter network architecture is more robust to pose variations, as it allows for partial translation and deformation invariance due to local information processing via receptive fields and massive weight-sharing over different locations in input images. Weight-sharing allows to reduce the number of weights and therefore free parameters that have to be learned, alleviating the requirements concerning the size of necessary training databases.

# 5. EXPERIMENTAL RESULTS

In this section we first describe the characteristics of the employed database. We then present the results of query-based facial expression aspect retrieval using LSA, and the results of unsupervised clustering of FACS AUs by PLSA. Both LSA and PLSA aspects are compared to hand-labeled facial expression categories, in order to determine the quality of the aspect-to-expression category mapping, and also to select the best matching aspects in the case of PLSA. Finally, we demonstrate the application of of the computed aspects for the recognition of facial expressions using neural networks.

## 5.1 Employed Database

For our experiments we employed the Cohn-Kanade DFAT-504 facial expression database [13] that is provided complete

with FACS labels. See Figure 2 for some sample images. This database is one of the largest publicly available facial expression databases. While the database contains image sequences, we worked only with still images. Of the totally 481 sequences we retained all FACS labels for the latent semantic analysis and used 432 still images for the training, evaluation and testing of our neural networks. 49 labels and the associated images were retained for the evaluation set for the training of the PLSA aspects and were neither used for the LSA-based aspect retrieval in order to allow for a common basis that allows for comparing these two approaches. Facial expression intensities were not taken into account. Furthermore, in order to compare the extracted LSA aspects with a reference code, we labeled the DFAT-504 database into 7 broad facial expression categories that encompass the 6 basic emotions as well as a mouth opening category. Note that while the DFAT-504 database seems to be small, it is fairly representative for the task of describing the afore mentioned 7 broad facial expression categories. Other datasets used in the literature are about the same size, and often even smaller, e.g. the JAFFE facial expression database [7].

## 5.2 LSA-based Aspect Retrieval

LSA aspects were computed by issuing FACS code queries and retrieving the closest matching FACS action codes. We have formulated 112 EMFACS queries for the computation of 7 LSA facial expression aspects. The queries are based on the basic emotion prototypes, as defined in the EMFACS framework, see the FACS investigator's guide [8]. The 6 prototype emotion categories were completed with a mouth opening category that occurs in the DFAT-504 database. Table 1 lists the 112 EMFACS basic emotion prototype expressions. Note that all queries were simplified. We took into account neither facial expression intensities (3 or 5 intensity levels) nor the location of certain facial action units (bottom, top, left, right).

The quality of how well the computed LSA aspects describe ground truth facial expression categories was analyzed with different measures, see Table 2. Listed are the overall *precision*, *recall* and *F-measures,* as described in equations 7, 8 and 9. As can be seen, the resulting precision and recall values lead to high overall F-measures (Ov). This shows that the EMFACS prototypes are indeed well defined and lead to very accurate basic emotion retrieval results.

## 5.3 PLSA-based Aspect Clustering

In contrast to LSA query-based aspect computation, clustered PLSA aspects are determined in an unsupervised manner. One parameter that has to be defined in advance is the number of desired aspects. As we detected variations of the basic emotion categories contained in the DFAT-504 database (two types of fear and anger) and therefore cover these variations more accurately we chose 9 and 15 aspects, slightly greater than the number of the 7 target facial expression categories. We computed a total of 5 sets with 9 aspects (sets 1.2-1.5) and 5 sets with 15 aspects (sets 2.1-2.5). These sets feature different combinations of FACS codes for the training and evaluation sets. The obtained PLSA aspects were then mapped to ground truth facial expression categories. The *overall precision*, *recall* and *F-measures* are given in Table 3. Listed are the aspects (A), the number of valid mappings (M), and the mappings retained to train

| B. Emotions | EMFACS Prototypes | Queries |
|---|---|---|
| Disgust | 9 | 1 |
| (Di) | 9+16+25, 26 | 2 |
| | 9+17 | 1 |
| | 10* | 1 |
| | 10*+16+25, 26 | 2 |
| | 10+17 | 1 |
| Surprise | 1+2+5B+26,27 | 2 |
| (Su) | 1+2+5B | 1 |
| | 1+2+26,27 | 2 |
| | 5B+26,27 | 2 |
| Anger | (4+5*+7+10*+22+23+25,26)** | 10 |
| (An) | (4+5*+7+l0*+23+25,26)** | 8 (10) |
| | (4+5*+7+23+25, 26)** | 8 |
| | (4+5*+7+17+23,24)** | 8 |
| | (4+5*+7+23,24)** | 8 |
| Happiness | 6+12* | 1 |
| (Ha) | 12C/D | 1 |
| Sadness | (1+4+11+15B +/- 54+64)+/-25,26 | 6 |
| (Sa) | (1+4+15* +/- 54+64)+/-25,26 | 6 |
| | (6+15* +/- 54+64)+/-25,26 | 6 |
| | (1+4+11 +/- 54+64)+/-25,26 | 6 |
| | (1+4+15B +/- 54+64)+/-25,26 | 0 (6) |
| | (1+4+15B+17 +/- 54+64)+/-25,26 | 6 |
| | (11+15B +/- 54+64)+/-25,26 | 6 |
| | 11+17 +/- 25,26 | 3 |
| Fear | 1 +2+4+5*+20*+25, 26, or 27 | 3 |
| (Fe) | 1+2+4+5*+25, 26, or 27 | 3 |
| | 1+2+4+5*+L or R20*+25, 26, or 27 | 0 (3) |
| | 1+2+4+5* | 1 |
| | 1+2+5Z, +/- 25, 26, 27 | 0 (4) |
| | 5*+20* +/- 25, 26, 27 | 4 |
| Mouth op. | 25,26 or 27 | 3 |
| (Mo) | | |

**Table 1: LSA Queries based on EMFACS Emotion Prototypes. * In this combination the AU may be at any level of intensity. ** Any of the prototypes can occur without any one of the following AUs: 4, 5, 7, or 10. The right-most column lists the number of emotion prototype queries. Note that we do not take into account intensities or locations of AUs. Therefore, some EMFACS prototypes resulted in zero LSA queries.**

| Di | Su | An | Ha | Sa | Fe | Mo | Ov |
|---|---|---|---|---|---|---|---|
| Precision | | | | | | | |
| **1.00** | **0.97** | **0.88** | **0.99** | **0.96** | **0.89** | **1.00** | **0.95** |
| Recall | | | | | | | |
| **0.87** | **1.00** | **0.82** | **0.99** | **0.99** | **0.98** | **1.00** | **0.96** |
| F-Measure | | | | | | | |
| **0.93** | **0.99** | **0.85** | **0.99** | **0.97** | **0.93** | **1.00** | **0.94** |

**Table 2: Assessment of the LSA Aspect to Expression Mappings. Shown are the expression specific (Di-Mo) and overall (Ov) precision, recall and F-measure for the 112 EMFACS LSA queries for 7 facial expression categories.**

| Test | | 7 expressions cat | | | | Best |
|---|---|---|---|---|---|---|
| Sets | # A | OvP | OvR | OvFM | # M | |
| **S1.1** | 9 | 0.75 | 0.70 | 0.71 | 6 | - |
| **S1.2** | 9 | 0.80 | 0.72 | 0.73 | 6 | - |
| **S1.3** | 9 | 0.86 | **0.80** | **0.82** | **7** | - |
| **S1.4** | 9 | 0.76 | 0.69 | 0.72 | 6 | - |
| **S1.5** | 9 | **0.89** | 0.78 | **0.82** | **7** | x |
| Mea | | 0.81 | 0.74 | 0.76 | | |
| Stdv | | 0.06 | 0.05 | 0.06 | | |
| **S2.1** | 15 | 0.89 | **0.75** | **0.82** | **7** | - |
| **S2.2** | 15 | 0.84 | 0.56 | 0.66 | **7** | - |
| **S2.3** | 15 | **0.96** | 0.73 | **0.82** | **7** | x |
| **S2.4** | 15 | 0.74 | 0.61 | 0.63 | 6 | - |
| **S2.5** | 15 | 0.88 | 0.70 | 0.78 | **7** | - |
| Mea | | 0.87 | 0.68 | 0.76 | | |
| Stdv | | 0.07 | 0.08 | 0.09 | | |

**Table 3: Summary of PLSA aspect to ground truth facial expression category mapping. Shown are the overall precision (OvP), recall (OvR) as well as F-measure (OvFM) for the mapping of 9 and 15 aspects to 7 expression categories.**

| Net | Emo | LSA | | PLSA-9 | | PLSA-15 | |
|---|---|---|---|---|---|---|---|
| | | Asp | Em | Asp | Em | Asp | Em |
| **MLP** | 50% | 49% | 48% | 42% | 38% | 43% | 40% |
| **NP** | 68% | 66% | 65% | 65% | 60% | 67% | 65% |

**Table 4: Average Categorical Facial Expression Recognition Results for the employed DFAT-504 Database Set using Different Labeling Schemes (Emo: Emotion, LSA: Latent Semantic Analysis, PLSA-9/15: Probabilistic Latent Semantic Analysis using 9 and 15 aspects) and two different neural network classifiers (MLP: Multilayer Perceptron and NP: Neoperceptron)**

networks (Best). Furthermore, we determined the mean and standard deviations in order to illustrate how well PLSA aspects map to target facial expression categories. Note that with 15 aspects we obtain a better mean *precision* than with 9 aspects. However, *recall* is worse with 15 aspects than is the case with 9 aspects. This behavior is reasonable as with more total aspects, facial expressions tend to be represented by more aspects, leading to a greater *precision* as fewer individual expressions are assigned to the same cluster. On the other hand, *recall values* drop as expressions are spread over more aspects. Not all of these aspects are retained (as they can lead to F-measures below the minimum threshold of 0.20) and thus we lose expressions that could otherwise improve the *recall values*.

Synonymy and polysemy of FACS codes are illustrated in Figures 4, 5 and 6. Our experiments showed that both LSA and PLSA were capable to cope synonymy and polysemy as shown in these Figures.

## 5.4 Facial Expression Recognition

We have employed the LSA and PLSA extracted facial action codes for the supervised training of CNNs and MLPs via the back-propagation algorithm. Table 4 lists the average recognition results for 432 images of the DFAT-504

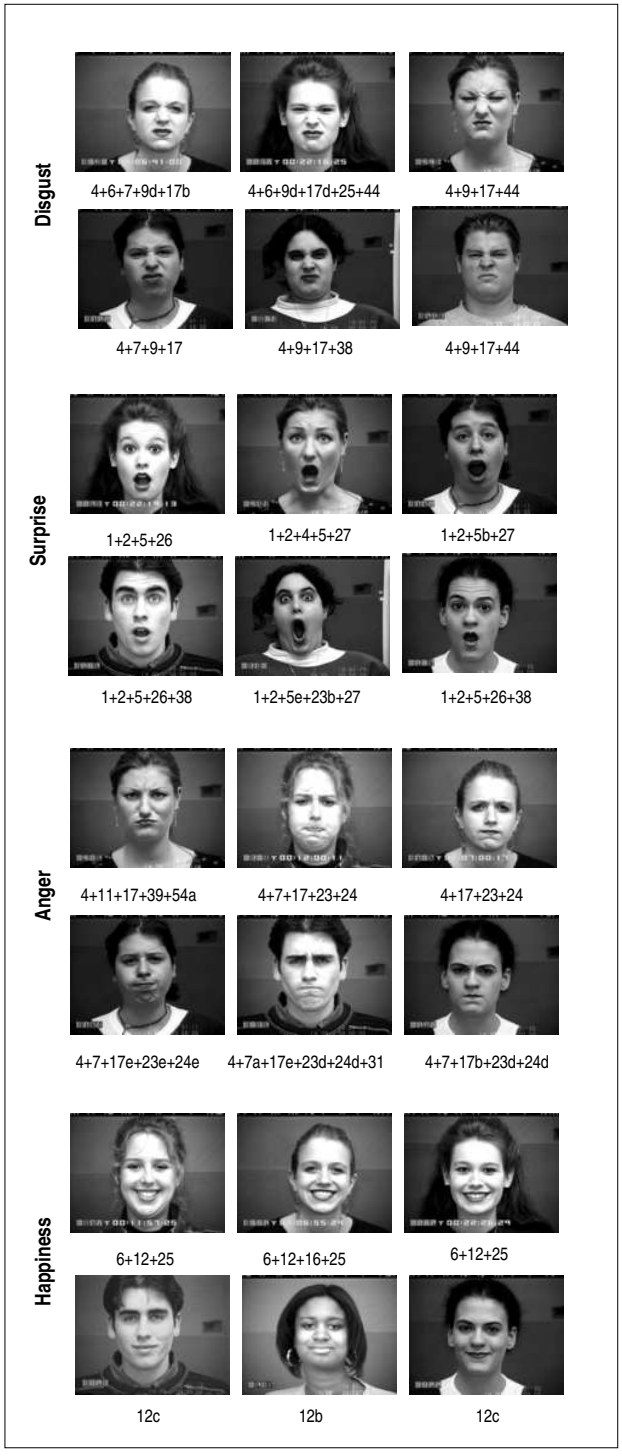


**Figure 4: FACS codes and associated images that illustrate synonymy. Different FACS AUs describe the same facial expression category. Part1: Disgust, Surprise, Anger and Happiness.**

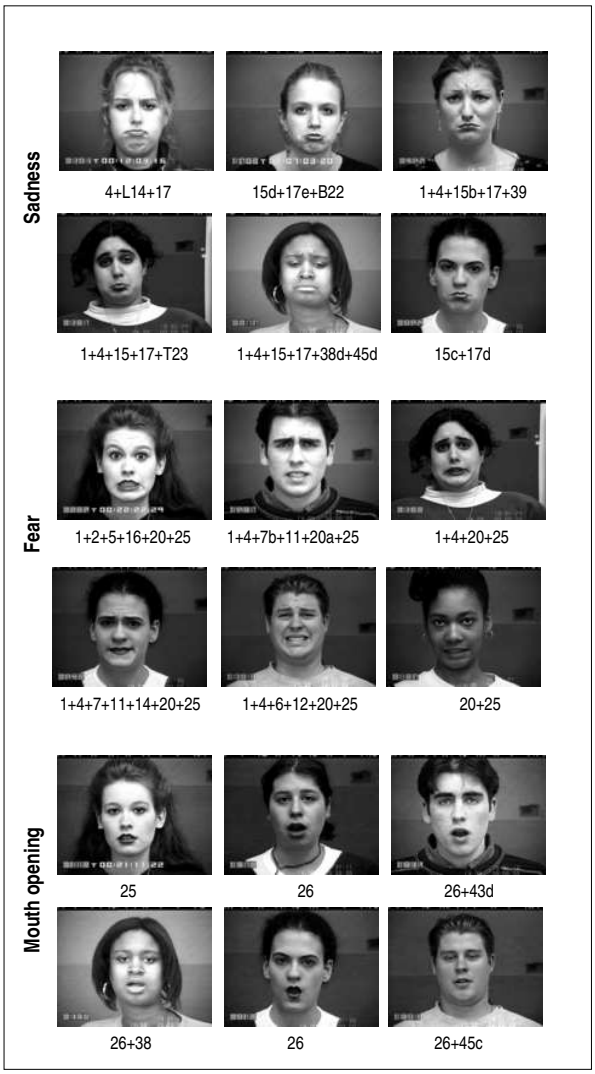


**Figure 5: FACS codes and associated images that illustrate synonymy. Different FACS AUs describe the same facial expression category. Part2: Sadness, Fear and Mouth opening.**

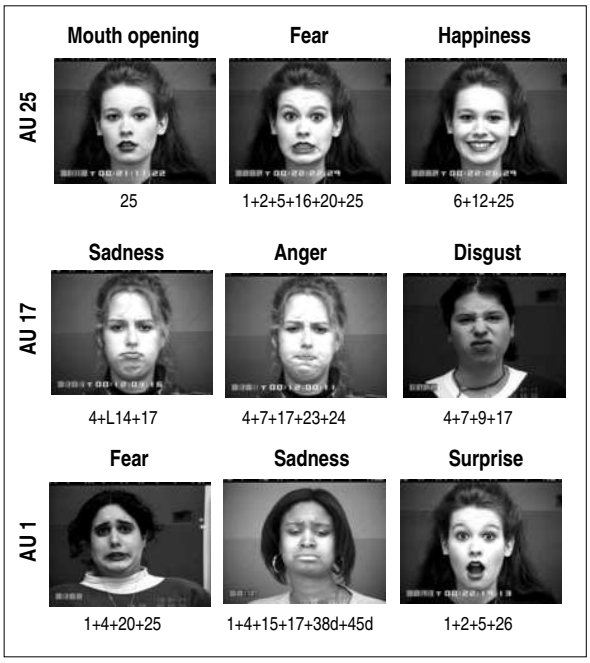


**Figure 6: FACS codes and associated images that illustrate polysemy. The same FACS AUs appear in different facial expression categories.**

database using emotion (Emo), EMFACS based LSA (Lsa) labelling as well as 9 and 15 aspect PLSA based labelling (PLSA-9 and PLSA-15). Hereby, we split the database into three sets of 216 training, 108 evaluation and 108 test images. For each labeling scheme are given the recognition results for the aspects (Asp) as well as the correct recognition of the target facial expression category (Em). As can be seen, LSA and PLSA-based codes lead to recognition results that are comparable with those obtained using emotion labels (Emo). These results are interesting, as the described information retrieval approaches combined with neural networks were able to discriminate facial expressions into meaningful categories, similar to the ground truth emotional expressions provided by human beings. Finally note that the recognition results obtained with the Neoperceptron (NP) are better than those obtained with the Multilayer Perceptron (MLP). The reason for this is that there are slight head pose variations present in the chosen train and test sets to which convolutional neural networks are more robust. For more details about convolutional neural networks see [9]. Note that the small sample size makes the evaluation of the facial expression recognition performance differences between the employed methods problematic. The focus in this paper has been on automatic clustering of FACS coded facial actions. Nonetheless, we would like to compare our facial expression recognition results to the ones obtained in the recent literature. An automatic face detection and facial expression recognition system that has been trained on the DFAT-504 facial expression database was described in [2]. The system employs Gabor filters, Adaboost for feature selection, and Support Vector Machines for classification. In contrast to our proposed approaches, this system relies on a high precision face detection. A generalization performance

of about 85% was reported using cross-validation. Note that the employed protocol is different from ours and therefore we cannot compare the recognition results directly.

## 6. CONCLUSION

In this paper we have shown that LSA and PLSA allow to cluster low-level FACS codes into semantically meaningful facial expression categories, without resorting to manual semantic ratings of facial expression displays that are difficult to achieve and often biased (by meaningful aspects we mean aspects that can be mapped to a distinct facial expressions category). The resulting compact representations are an advantage for the training data-driven methods such as neural networks, as with FACS labels there is usually a great deal of variations, and the training databases are often too small to cover all variations. An advantage of PLSA over LSA, as treated in this paper, is that the former allows to directly cluster aspects in an unsupervised manner, without the need to formulate queries. In addition, PLSA provides class membership probabilities. These can be used to neural networks with non-categorical mixtures of aspects.

In this paper we assumed that each occurring FACS AU is displayed with the same intensity. Future work will take into account individual facial expression intensities of FACS AU components within compound expressions by attributing more weight to a high magnitude facial expression component than to a component with a weaker display. This can be achieved with LSA by allowing for more than one occurrence of a single AU per compound expression. Hereby, the columns in the term-document or AU-expression matrix $\mathbf{A}$ of equation 1 contain the occurrences of each AU in a particular expression $\mathbf{A} = [a_{ij}]$, where $a_{ij}$ denotes the frequency in which the AU $i$ occurs in the compound expression $j$ . The frequency is hereby proportional to the intensity of the AU display, i.e. a frequency of $n$ corresponds to an AU with intensity $n$ and where $n$ is either in the range [1..3] or [1..5] of commonly used facial expression intensity levels. Along this line, there are several simple text weighting techniques (e.g. tf-idf) that could be worth investigating.

## 7. ACKNOWLEDGMENTS

This work was initiated while B. Fasel was with IDIAP, and was funded by the Swiss National Center for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

## 8. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real Time Face Detection and Expression Recognition: Development and Application to Human-Computer Interaction. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 1990.

[4] P. Ekman. Methods for Measuring Facial Actions. In K. Scherer and P. Ekman, editors, *Handbook of Methods in Nonverbal Behaviour Research*, pages 45–90. Cambridge University, 1982.

[5] P. Ekman and W. Friesen. Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[6] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press, Palo Alto*, 1978.

[7] M.J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba. Coding Facial Expressions with Gabor Wavelets In *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara Japan, IEEE Computer Society, pp. 200-205, April 14-16 1998.

[8] P. Ekman, W. V. Friesen, and J. C. Hager. *FACS Investigator's Guide*. A Human Face, 666 Malibu Drive, Salt Lake City UT 84107, 2002.

[9] B. Fasel. Robust Face Analysis using Convolutional Neural Networks. In *Proceedings of the International Conference on Pattern Recognition (ICPR 02)*, volume 2, pages 40–43, Quebec, Canada, aug 2002. IDIAP-RR 01-48.

[10] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259–275, 2003. IDIAP-RR 99-19.

[11] M. Florent and D. Gatica-Perez. On Image Auto-annotation with Latent Space Models. In *ACM Multimedia*, pages 275–278, 2003.

[12] T. Hofmann. Probabilistic Latent Semantic Analysis. In *Proceedings of Uncertainty in Artificial Intelligence (UAI 99)*, Stockholm, Sweden, 1999.

[13] T. Kanade, J. Cohn, and Y. Tian. Comprehensive Database for Facial Expression Analysis. In *IEEE Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition (FG'00)*, March 2000.

[14] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[15] M. Pantic and L. J. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12):1424–1445, Dec. 2000.