3D Shape Estimation in Video Sequences Provides High Precision Evaluation of Facial Expressions ⋆

László A. Jeni [a], András Lőrincz [b], Tamás Nagy [b], Zsolt Palotai [c], Judit Sebők [b], Zoltán Szabó [b], Dániel Takács [b,d]

[a] *University of Tokyo, Japan*

[b] *Eötvös Loránd University, Hungary*

[c] *Sparsense Inc., USA*

[d] *Realeyes Data Services Ltd., UK*

Abstract

Person independent and pose invariant estimation of facial expressions and action unit (AU) intensity estimation is important for situation analysis and for automated video annotation. We evaluated raw 2D shape data of the CK+ database, used Procrustes transformation and the multi-class SVM leave-one-out method for classification. We found close to 100% performance demonstrating the relevance and the strength of details of the shape. Precise 3D shape information was computed by means of Constrained Local Models (CLM) on video sequences. Such sequences offer the opportunity to compute a time-averaged '3D Personal Mean Shape' (PMS) from the estimated CLM shapes, which – upon subtraction – gives rise to person independent emotion estimation. On CK+ data PMS showed significant improvements over AU0 normalization; performance reached and sometimes surpassed state-of-the-art results on emotion classification and on AU intensity estimation. 3D PMS from 3D CLM offers pose invariant emotion estimation that we studied by rendering a 3D emotional database for different poses and different subjects from the BU 4DFE database. Frontal shapes derived from CLM fits of the 3D shape were evaluated. Results demonstrate that shape estimation alone can be used for robust, high quality pose invariant emotion classification and AU intensity estimation.

Key words: Constrained Local Model, shape information, action unit recognition, emotion classification, CK+, BU-4DFE

Email addresses: `laszlo.jeni@ieee.org` (László A. Jeni),

1 Introduction

Our everyday communication is highly influenced by the emotional information available to us about those whom we communicate with. Facial expression and body language are the main sources of this information. Thus, recognition of facial expression is highly relevant for human-computer interaction and may gain broad applications in video annotation, situation analysis of social interactions.

In the last decade many approaches have been proposed for automatic facial expression recognition. We are experiencing a breakthrough in this field due to two factors; (1) high quality databases that have been made available to everybody, like the marked Cohn-Kanade Extended Facial Expression Database (CK+) [1], its enhanced version [2], and the dynamic 3D facial expression database [3] as well as (2) the advance of learning algorithms, most notably the advance of constrained local models (CLM) [4,5]. Recently, very good results have been achieved by means of textural information [6,7]. On the other hand, shape of the face extracted by active appearance models (AAM) [8] showed relatively poor performance (see, e.g., [1], and references therein).

Line drawings, however, can express facial expressions very well, so shape information could also be a good descriptor of emotions. Shape – as opposed to texture – is attractive for facial expression recognition since it should be robust against rotations and may be robust against light conditions. We studied facial expression recognition using all available landmarks of the shape. We found close to 100% performance, indicating that the compression inherent in AAM was responsible for the relatively poor performance. We then used the more expressive CLM method and studied the behavior of CLM fits with respect to head pose directions. Our main result is that shape information extracted from CLM based automated marker generation gives rise to, sometimes surpasses state-of-the-art performance and it could be improved further with more precise automated marker identification. Beyond the theoretical interest that shape alone may give rise to 100% performance, we note that one may safely replace the estimated personal AU0 normalization by the normalization with an easily measurable descriptor set, the personal mean shape. We show here that results outperform known results of texture and shape AAM method [1], texture based CLM method [7] and 2D shape based method that makes use of the Grassmann manifold [9]. In turn, we suggest shape identification for a standard benchmark. This paper reviews some results of our previous works [10,11] and extend those by a more thorough study of pose (in)dependence

andras.lorincz@elte.hu (András Lőrincz), bkil.hu@gmail.com (Tamás Nagy), palotai@sparsense.com (Zsolt Palotai), dymorse@gmail.com (Judit Sebők), szzoli@cs.elte.hu (Zoltán Szabó), takacsd@gmail.com (Dániel Takács).

of emotion classification as well as with our new results on Action Unit (AU) intensity estimation.

The paper is built as follows. Theoretical components and the datasets are reviewed in Section 2. Experimental results on emotion classification and comparisons with previous work are detailed in Section 3. Results on AU intensity estimation are described in Section 4. Both Section 3 and Section 4 treat pose dependence. Discussion and a summary conclude the paper (Section 5).

2 Methods

Our proposed method contains two main steps. First, we estimate 3D landmark positions on face images using the CLM method. We describe the details of this technique in Section 2.1. Then we remove the rigid transformation from the acquired 3D shape, project it to 2D and perform SVM-based multi-class classification or SVM regression using the different emotions classes or AU values, respectively. SVM based classification and regression methods are reviewed in Section 2.2. Procrustes method, AU0 and Personal Mean Shape (PMS) normalization are detailed in Section 2.3. Section 2.5 is about the datasets used in the tests and Section 2.6 lists the features that we extracted.

2.1 Constrained Local Models

CLM methods are generative parametric models for person-independent face alignment. In this work we were using a 3D CLM method, where the shape model is defined by a 3D mesh and in particular the 3D vertex locations of the mesh, called landmark points. Consider the shape of a 3D CLM as the coordinates of 3D vertices of the M landmark points:

$$\mathbf{x} = (x_1, y_1, z_1, ..., x_M, y_M, z_M)^T, \tag{1}$$

or, $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T)^T$, where $\mathbf{x}_i = (x_i, y_i, z_i)^T$. We have N samples: $\{\mathbf{x}^{(n)}\}_{n=1}^N$. CLM models assume that – apart from the global transformations; scale, rotation, and translation – all $\{\mathbf{x}^{(n)}\}_{n=1}^N$ can be approximated by means of the linear principal component analysis (PCA) forming the PCA subspace.

In the next subsection we briefly describe the 3D Point Distribution Model and the way CLM estimates the positions of the landmarks.

2.1.1 Point Distribution Model

The 3D point distribution model (PDM) describes non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

$$\mathbf{x}_i(\mathbf{p}) = s\mathbf{P}\mathbf{R}(\bar{\mathbf{x}}_i + \mathbf{\Phi}_i\mathbf{q}) + \mathbf{t}, \tag{2}$$

where $i = 1, \ldots, M$, $\mathbf{x}_i(\mathbf{p})$ denotes the 2D location of the i^{th} landmark subject to transformation $\mathbf{p}$, and $\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{q}, \mathbf{t}\}$ denotes the parameters of the model, which consist of a global scaling s, angles of rotation in three dimensions ($\mathbf{R} = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma)$), translation $\mathbf{t}$ and non-rigid transformation $\mathbf{q}$. Here $\bar{\mathbf{x}}_i$ is the mean location of the i^{th} landmark averaged over the database, i.e. $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1^T; \ldots; \bar{\mathbf{x}}_M^T]$, $\bar{\mathbf{x}}_i = [\bar{x}_i, \bar{y}_i, \bar{z}_i]^T$, $\bar{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^{(n)}$, and similarly, for $\bar{y}_i$ and $\bar{z}_i$. Matrix $\mathbf{\Phi}_i$ $(i = 1, \ldots, M)$ is a $3 \times d$ piece in $\mathbf{\Phi} \in \mathbb{R}^{3M \times d}$ and corresponds to the landmarks. Columns of $\mathbf{\Phi}$ form the orthogonal projection matrix of principal component analysis and its compression dimension is d. Finally, matrix $\mathbf{P}$ denotes the projection matrix to 2D:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{3}$$

and thus $\mathbf{x}_i(\mathbf{p}) \in \mathbb{R}^2$ $(\forall i)$.

By applying PCA on the $\mathbf{x}_i$ points we get an estimate of the prior of the parameters:

$$p(\mathbf{p}) \propto N(\mathbf{q}; \mathbf{0}, \mathbf{\Lambda}), \tag{4}$$

that is CLM assumes a normal distribution with $\mathbf{0}$ mean and $\mathbf{\Lambda}$ variance for parameters $\mathbf{q}$. $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1; \ldots; \lambda_d) \in \mathbb{R}^{d \times d}$ in (4) is provided by the PCA and the parameter vector assumes the form $\mathbf{p} = [s; \alpha; \beta; \gamma; \mathbf{t}; \mathbf{q}]$.

2.1.2 Formalization of Constrained Local Models

CLM is constrained through the PCA of PDM. It works with *local experts*, whose opinion is considered independent and are multiplied to each other:

$$J(\mathbf{p}) = p(\mathbf{p}) \prod_{i=1}^{M} p(l_i = 1 | \mathbf{x}_i(\mathbf{p}), \mathcal{I}), \tag{5}$$

where $l_i \in \{-1, 1\}$ is a stochastic variable, which is 1 (-1) if the i^{th} marker is (not) in its position, $p(l_i = 1 | \mathbf{x}_i(\mathbf{p}), \mathcal{I})$ is the probability that for image $\mathcal{I}$ and for marker position $\mathbf{x}_i$ determined by parameter $\mathbf{p}$, the i^{th} marker is in its position.

Local experts are built on Logit Regression and are trained on labeled samples.

The functional form of Logit is

$$p(l_i = 1|\mathbf{y}_i, \mathcal{I}) = \frac{1}{1 + e^{\mathbf{w}_i^T N(\mathcal{I}(\mathbf{y}_i)) + b_i}}, \tag{6}$$

where $N(\mathcal{I}(\mathbf{y}_i))$ is a normalized image patch around point $\mathbf{y}_i$, $\mathbf{w}_i$ and b_i are parameters of the distribution to be learned from samples. Positive and negative samples for the right corner of the right eye are shown in Fig. 1(a).

Local expert's response – that depend on the constraints of the PDM and the response map of the local expert in an appropriate neighborhood – can be used to express $J(\mathbf{p})$ in (5) (Fig. 1(b)):

$$J(\mathbf{p}) = p(\mathbf{p}) \prod_{i=1}^{M} \sum_{\mathbf{y}_i \in \Psi_i} p(l_i = 1|\mathbf{y}_i, \mathcal{I})\, p(\mathbf{y}_i|\mathbf{x}_i)\, \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I}) \tag{7}$$

where CLM assumes $\mathbf{y}_i = \mathbf{x}_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \rho\mathbf{I})$, $\rho = \sum_{i=d+1}^{3M} \frac{\lambda_i}{(3M-d)}$, λ_i is the i^{th} eigenvalue of $Cov(\mathbf{x})$, the covariance matrix of stochastic variable $\mathbf{x}$ and where we applied Bayes'rule and the tacit assumption [5] that $\sum_{\mathbf{y}_j \in \Psi_i} \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I})$ is a weak function of the parameters to be optimized was accepted.

(a) (b)

Fig. 1. (a): Landmarks on the face. (b): response map for the right corner of the right eye.

2.1.3 Expectation Maximization for CLM

Expectation maximization (EM) can be used to find the (local) optimum. For the E-step, the probability of parameter $\mathbf{p}$ given the current parameter estimation $\mathbf{q}^{(t)}$ is

$$Q(\mathbf{p}|\mathbf{p}^{(t)}) = E_{q(y)}\left[-\log\left\{p(\mathbf{p}) \prod_{i=1}^{M} p(l_i = 1, \mathbf{y}_i|\mathbf{x}_i(\mathbf{p}), \mathcal{I})\right\}\right]$$

where

$$q_{(y)} = \prod_{i=1}^{M} p(\mathbf{y}_i | l_i = 1, \mathbf{x}_i(\mathbf{p}^{(t)}), \mathcal{I})$$

and thus

$$Q(\mathbf{p}|\mathbf{p}^{(t)}) \propto ||\mathbf{q}||^2_{\Lambda^{-1}} + \sum_{i=1}^{M} \sum_{\mathbf{y}_i \in \Psi_i} \frac{w_{\mathbf{y}_i}}{\rho} \|\mathbf{x}_i - \mathbf{y}_i\|^2$$

where Ψ_i is the set of pixel positions of the i^{th} response map and

$$w_{\mathbf{y}_i} = p(\mathbf{y}_i | l_i = 1, \mathbf{x}_i, \mathcal{I}) = \frac{p(l_i = 1|\mathbf{y}_i, \mathcal{I}) \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} p(l_i = 1|\mathbf{z}_i, \mathcal{I}) \mathcal{N}(\mathbf{z}_i; \mathbf{x}_i, \rho\mathbf{I})} \tag{8}$$

For the parameter optimization M-step the task is to minimize $Q(\mathbf{p}|\mathbf{p}^{(t)})$, i.e.,

$$Q(\mathbf{\Delta p}|\mathbf{p}) = \|\mathbf{p} + \mathbf{\Delta p}\|^2_{\tilde{\mathbf{\Lambda}}^{-1}} + \sum_i \sum_{\mathbf{y}_i \in \Psi_i} \frac{w_{\mathbf{y}_i}}{\rho} \|(\mathbf{x}_i + \mathbf{J}_i \mathbf{\Delta p}) - \mathbf{y}_i\|^2 \rightarrow \min_{\mathbf{\Delta p}} \tag{9}$$

giving rise to

$$\mathbf{\Delta p} = -\left(\rho \tilde{\mathbf{\Lambda}}^{-1} + \mathbf{J}^T \mathbf{J}\right)^{-1} \left(\rho \tilde{\mathbf{\Lambda}}^{-1} \mathbf{p} - \mathbf{J}^T \mathbf{v}\right)$$

where $\mathbf{J}$ is the Jacobian of the cost function, $\tilde{\mathbf{\Lambda}} \in \mathbb{R}^{(6+d)\times(6+d)}$ is the diagonal matrix with the PDM eigenvalues and with leading zeros in its diagonal (the extension of the $\mathbf{\Lambda}$ diagonal matrix with as many leading 0s in its diagonal as many rigid parameters we have, which is 6 in our case), $\mathbf{v}$ is concatenated from the mean shift values $\mathbf{v} = [\mathbf{v}_1; \ldots; \mathbf{v}_M]^T$ where

$$\mathbf{v}_i = \left(\sum_{\mathbf{y}_i \in \Psi_i} \frac{p(l_i = 1|\mathbf{y}_i, \mathcal{I}) \mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho\mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} \pi_{\mathbf{z}_i} \mathcal{N}(\mathbf{x}_i; \mathbf{z}_i, \rho\mathbf{I})} \mathbf{y}_i \right) - \mathbf{x}_i. \tag{10}$$

and $\pi_{\mathbf{z}_i}$ is the shorthand for $p(l_i = 1|\mathbf{z}_i; \mathcal{I})$.

2.2 Support Vector Machines for Emotion Classification and AU Estimation

Support Vector Machines make another building block of the algorithms, since after we estimate the 3D landmarks with the CLM method and remove the rigid transformation from the acquired 3D shape, we project it to 2D and perform an SVM-based multi-class classification using the different emotions as the class labels and a regression procedure for AU intensity estimation.

Support Vector Machines (SVMs) are very powerful for binary and multi-class classification as well as for regression problems. They are robust against outliers. For two-class separation, SVM estimates the optimal separating hyperplane between the two classes by maximizing the margin between the hyperplane and closest points of the classes. The closest points of the classes are

called support vectors; the optimal separating hyper-plane lies at half distance between them.

We are given sample and label pairs $(\mathbf{x}^{(k)}, y^{(k)})$ with $\mathbf{x}^{(k)} \in \mathbb{R}^m$, $y^{(k)} \in \{-1, 1\}$, and $k = 1, ..., K$. Here, for class '1' and for class '2' $y^{(k)} = 1$ and $y^{(k)} = -1$, respectively. We also have a set of feature vectors $\boldsymbol{\phi}(= [\phi_1; \ldots; \phi_J]) : \mathbb{R}^m \rightarrow \mathbb{R}^J$, where J might be infinite. The support vector classification seeks to minimize the cost function

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{k=1}^{K}\xi_k \tag{11}$$

$$y^{(k)}(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}^{(k)}) + b) \geq 1 - \xi_k, \;\; \xi_k \geq 0. \tag{12}$$

where ξ_k $(k = 1, \ldots, K)$ are the so called slack variables that generalize the original SVM concept with separating hyper-planes to soft-margin classifiers that have outliers that can not be separated.

We used multi-class classification, where decision surfaces are computed for all class pairs, i.e., for k classes one has $k(k-1)/2$ decision surfaces and then applies a voting strategy for decisions. Multi-class SVM is considered competitive to other SVM methods [12] and we found differences when compared it to the one-against all procedure when the number of positive samples was small. In all cases, we used only linear classifiers.

Support vector regression (SVR) has a very similar form to support vector machine. For details on SVR techniques, the interested reader is referred to the literature, e.g., to [13] and the references therein. In our case, the CK database had only two points for function approximation; those where the function takes the two extreme values, the zero value and the value of 1. This case can be approximated with the L_2-loss L_2-regularized least-squares SVM (LS-SVM). The intermediate cases of the Enhanced CK database annotated frame-by-frame served as our test points. This least squares SVM formulation for the linear case modifies (11) and (12) to a loss function

$$J(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\mu}{2}\sum_{k=1}^{K}\left(y^{(k)} - \left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}^{(k)}) + b\right)\right)^2 \tag{13}$$

that should be minimized.

2.3 Normalization

Our motivation to use shape for the characterization of emotions comes from the surprisingly good results on marked 2D databases [10]. Normalization is

the key concept to reach such good performance.

2.3.1 Procrustes Normalization

For any shape $\mathbf{x}$, Procrustes transformation applies translation, uniform scaling and rotation to match the reference shape $\mathbf{x}_r$ in Euclidean norm. The minimum of this cost function is called the Procrustes distance. We applied it in our 2D transformations.

2.4 Using AU0 and Personal Mean Shape for Normalization

For the classification task, we used the so called AU0 normalization; we computed the differences between the features of the actual shape and the features of the first (neutral) frame. AU0 normalization is crucial for facial expression recognition, however it is person dependent and it is not available for a single frame.

Since we are interested in situation analysis, we can safely assume that we have videos (frame series) about the subject like and we can compute the Personal Mean Shape (PMS). We found that the mean shape is almost identical to the neutral shape, i.e., to AU0 especially since the assumed neutral shape differs slightly for the same person from video to video. We assume that time averaged shapes are either close to neutral, or if not, the difference should be taken into account. We computed the PMS for every subject using two different methods: (1) we averaged over the neutral facial expressions of the different emotional series of the same subject and (2) we averaged over both the neutral and over the extreme facial expressions of the different emotional series of the same subject. In our experiments we replaced the unavailable AU0 information by the PMS values computed by means of the CLM and compared the two PMS versions. The datasets are reviewed below.

2.5 Datasets

2.5.1 Datasets based on the Cohn-Kanade Facial Expression Database

In our simulations we used the Cohn-Kanade Extended Facial Expression (CK+) Database [1]. This database was developed for automated facial image analysis and synthesis and for perceptual studies. The database is widely used to compare the performance of different models. The database contains 123 different subjects and 593 frontal image sequences. From these, 118 subjects are annotated with the seven universal emotions (anger, contempt, disgust,

fear, happy, sad and surprise). The image sequences are annotated with 68 landmark points. Action units are also provided with this database for the apex frame. The original Cohn-Kanade Facial Expression Database distribution [14] had 486 FACS-coded sequences from 97 subjects. CK+ has 593 posed sequences with full FACS coding of the peak frames. A subset of action units were coded for presence or absence. To provide more ground-truth data for the original database, the RPI ISL research group manually re-annotated the original CK dataset [2] frame-by-frame. This annotation contains temporal segments of 14 Action Units. The presence of 14 AUs were annotated with three intensities: absent, present but with low intensity, and present. This dataset is referred to as the CK Enhanced Dataset.

2.5.2 The BU-4DFE Dynamic Facial Expression Database

For our studies on pose dependence we used the BU-4DFE dataset [3]. This dataset is a high-resolution 3D dynamic facial expression database. It contains 3D video sequences of 101 different subjects, with a variety of ethnic/racial ancestries. Each subject was asked to perform six prototypic facial expressions (anger, disgust, happiness, fear, sadness, and surprise), therefore the database contains 606 3D facial expression sequences. In the pose invariant experiment we marked a neutral frame and an apex frame of each sequence and rendered short video sequences with different yaw rotations. For these sequences the landmarks were provided by the CLM tracker itself.

2.6 Extracted Features

We used the CLM software developed at CMU [5], which was kindly provided to us by Jason Saragih. There is a slight difference between the set of the CK+ marker points and the set of 3D CLM marker points: the latter dropped marker points #60, #64 corresponding to the inner points of the left and right corners of the mouth, respectively. That is, we had 2*68=136 dimensional vectors for classification. We used the Procrustes method to compute the mean shape and to normalize all shapes to this mean.

We also used the 3D CLM for the fitting of the facial expression series of the CK+ database. This way we collected good estimates of the rigid transformation parameters. Then we extracted the normalized 2D shape parameters by removing the rigid transformation and then projecting to 2D.

In another set of experiments, we used the BU-4DFE dataset [3], rendered 3D facial expressions, and rotated those to different poses. This procedure was followed by the 3D CLM analysis and we extracted the 2D shape parameters alike in the previous case.

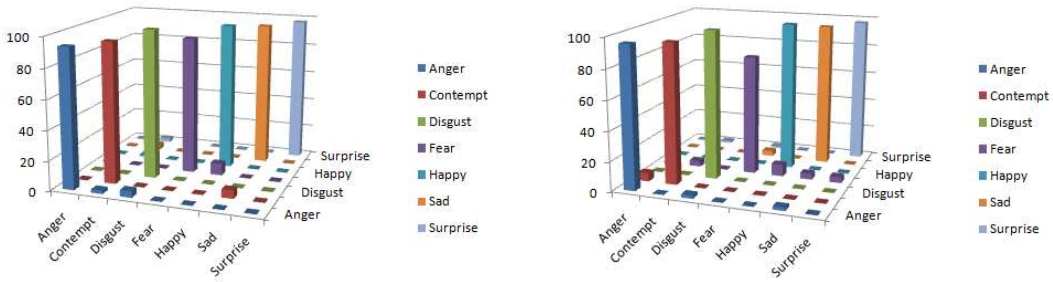
3 Experimental Results on Emotions

We executed a number of evaluations to judge the strength of shape based methods. Studies concern (i) the AU0 normalization versus the personal mean shape one, (ii) the performances of shape based method on the different databases for emotion classification for frontal and for rotated cases and for AU regression.

3.1 Experiment on the CK+ dataset with Procrustes method and original landmarks

In this experiment we used the CK+ dataset with the original 68 CK+ landmarks. First, we calculated the mean shape using Procrustes method. Then we normalized all shapes by minimizing the Procrustes distance between individual shapes and the mean shape. We trained a multi-class SVM using the leave-one-subject-out cross validation method. The result of the classification is shown in Figs. 2(a) and 2(c): emotions with large distortions, such as disgust, happiness and surprise, gave rise to nearly 100% classification performance.

Even for the worst case (fear), performance was 92%. Such high values are comparable to human performance on single frames [15].



(a) (b)

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Avg.
AU0 normalization	93.3	94.4	100	92	98.5	96.4	97.4	96
Personal Mean Shape	95.6	94.4	100	80	100	96.4	97.4	94.8

(c)

Fig. 2. Performance of two normalization methods for the expert annotated case. Results are averaged for the 118 subjects of the CK+ database. (a): Confusion matrix of the Procrustes method with AU0 normalization, (b): Confusion matrix of personal mean shape based normalization. (c): Comparison of the classification rates of the two methods. For more details, see text.

Personal mean shape was computed by averaging over neutral facial expressions for the one case and by averaging over both the neutral over extreme facial expressions for the other. Performances were similar for the expert coded CK+ database, but – intriguingly – they were better for the latter when the CLM was tested, so we decided to use this one. Replacing AU0 normalization by personal mean shape slightly decreases average performance: recognition on the CK+ database drops from 96% to 94.8% (Figs. 2(b), 2(c)). When the personal mean shape was computed by averaging over neutral facial expressions only, performance for contempt drops from 94.4% (AU0 normalization) to 77.7% (personal shape normalization), whereas performance on fear increases from 80.0% to 92,0%. Average performance is 93.6% (94.1%) if averaged with (without) contempt.

3.2 Experiment on the CLM-tracked CK+ dataset

In this experiment we studied the performance of the multi-class SVM on the CK+ dataset using the CLM method. First, we tracked facial expressions with the CLM method and annotated all image sequences starting from the neutral expression to the peak of the emotion. Comparing to the ground truth 2D landmarks, the average root-mean-square (RMS) error on the whole CK+ dataset is 4.84 RMSE unit (1 pixel error for all landmarks corresponds to 1 RMSE unit). Different parts of the face have different fitting characteristics. The RMS error of the different regions is depicted in Fig. 3.

3D CLM estimates the rigid and non-rigid transformations. We removed the rigid ones from the faces and projected the frontal view to 2D. We also applied the Procrustes normalization in this experiment. We did not find significant differences between the two cases and report the results for the CLM based normalization.

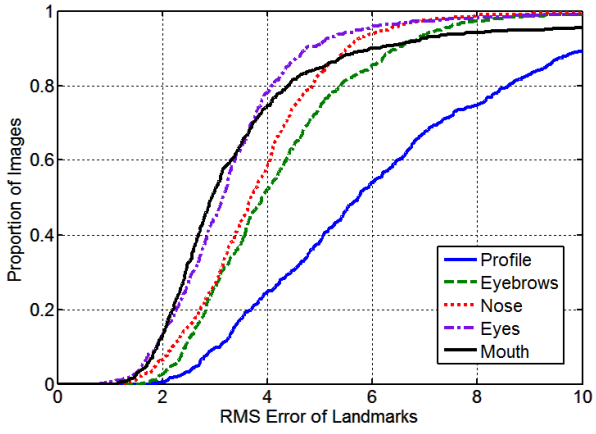
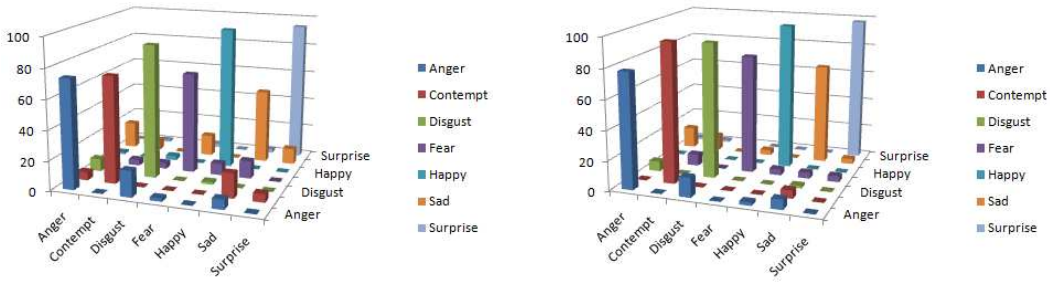


Fig. 3. The fitting curve of the CLM method on the CK+ dataset. The different curves depict the error characteristics of different facial parts.

The recognition performance of the system can be seen in Figs. 4(a) and 4(c): classification performance is affected by the imprecision of the CLM tracking. Emotions with large distortions can be still recognized in about 90% of the cases, whereas more subtle emotions are sometimes confused with others.



(a) (b)

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Avg.
AU0 normalization	73.3	72.2	89.8	68	95.7	50	94.0	77.6
Personal Mean Shape	77.8	94.4	91.5	80	98.6	67.9	97.6	86.8

(c)

Fig. 4. Performance of two normalization methods for the CLM annotated case. Results are averaged for the 118 subjects of the CK+ database. (a): Confusion matrix of the CLM method with AU0 normalization, (b): Confusion matrix of the CLM method with personal mean shape based normalization, (c): Comparison of the classification rates of the two methods. For more details, see text.

Table 1

Comparison on the CK+ dataset. Results of different methods on the 7 main facial expressions. S: shape based method, T: texture based method. S+T: both shape and texture based method.

Authors	Method	T/S	An.	Co.	Di.	Fe.	Ha.	Sa.	Su	Avg.
Lucey et al. '10	AAM+SVM	S	35.0	25.0	68.4	21.7	98.4	4.0	100.0	50.3
	AAM+SVM	T	70.0	21.9	94.7	21.7	100.0	60.0	98.7	66.7
	AAM+SVM	T+S	75.0	84.4	94.7	65.2	100.0	68.0	96.0	83.3
Chew et al. '11	CLM + SVM	T	70.1	52.4	92.5	72.1	94.2	45.9	93.6	74.4
This work	CLM + SVM (AU0 normalization)	S	73.3	72.2	89.8	68.0	95.7	50.0	94.0	77.6
	CLM + SVM (Personal Mean Shape)	S	77.8	94.4	91.5	80.0	98.6	67.9	97.6	86.8

We evaluated the personal mean shape normalization (Figs. 4(b), 4(c)). We found that this method improves performance and compensates for the estimation error of the CLM method; the CLM makes the same estimation error for the same person. Correct classification percentage raises from 77.57% to 86.82% for the CLM tracked CK+. This result is better than the available best AAM result that uses texture *plus* shape information [1] and the best CLM result that utilizes *only* textural information [7], see our comparisons in Table 1.

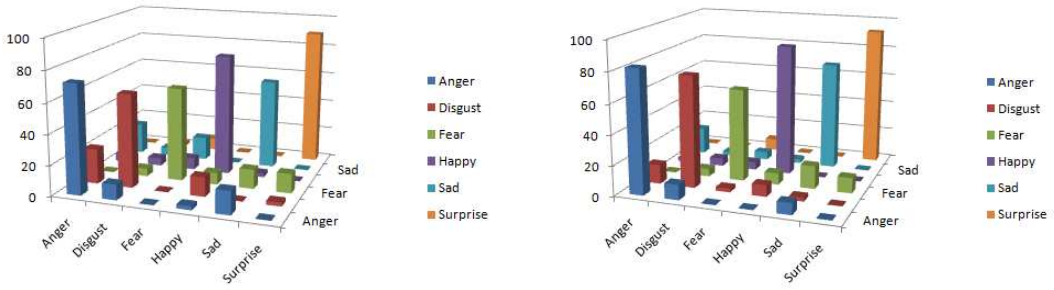
Different learning methods, including SVMs and the Grassmann manifold, have been elaborated in the literature. The comparison is in favor of the Grassmann manifold [9]. We note that the results in [9] were reported on the first version of the CK dataset with different emotion labels so direct comparison with our results is hard, but one may expect further gains by replacing the linear SVM with Grassmann manifold based methods. Table 1 gives an overview of the results using SVM.

Results show that our shape based emotion estimation is better than previous shape based methods and compares favorably with state-of-the-art procedures using texture or shape and texture information combined. It may be worth noting that FACS coders work by visual inspection and make use of textural information [16].

3.3 Experiments on the CLM-tracked BU-4DFE dataset

In order to proceed, we characterized the BU-4DFE database by using the CLM technique. First, we selected a frame with neutral expression and an apex frame of the same frame series. We used these frames and all frames between them for the evaluations. We applied CLM tracking for the intermediate frames in order, since it is more robust than applying CLM independently for each frames. We removed the rigid transformation after the fit and projected the frontal 3D shapes to 2D. We applied a 6 class multi-class SVM (this database does not contain contempt) and evaluated the classifiers by the leave-one-subject-out method. We compared the normalization using the CLM estimation of the AU0 values with the normalization based on PMS. Note that for the expert annotation, i.e., for the CK+ database, performance drop for personal mean shape was only 1% in average. For the BU-4DFE database, however, we found an 8% improvement on the average in favor of the PMS method (Fig. 5).

We also compared the two databases and executed cross evaluations (Fig. 6). In this case we used the CK+ as the ground truth, since it seems more precise than the BU-4DFE; in the case of CK+ the target expression for each



(a) (b)

	Anger	Disgust	Fear	Happy	Sad	Surprise	Avg.
AU0 normalization	71.8	61.5	61.5	79.5	59.0	89.5	70.5
Personal Mean Shape	82.1	74.4	61.5	87.2	71.8	92.1	78.2

(c)

Fig. 5. Performance of two normalization methods for the BU database using the CLM method. (a): Confusion matrix of the CLM method with AU0 normalization, (b): Confusion matrix of the CLM method with personal mean shape based normalization, (c): Comparison of the classification rates of the two methods. For more details, see text.

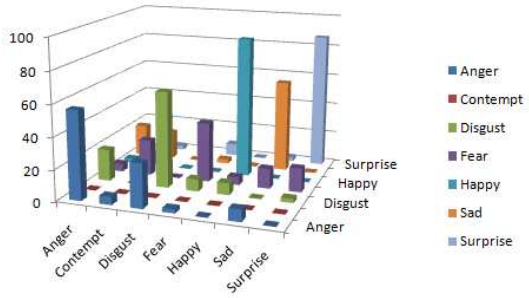


Fig. 6. Confusion matrix of CK+ and BU-4DFE datasets. Training: CK+, testing: BU-4DFE.

sequence is fully FACS coded, emotion labels have been revised and validated, and CK+ utilizes FACS coding based emotion evaluation and this method is preferred in the literature [1]. We note however, that both the CK+ and the BU-4DFE facial expressions are posed and not spontaneous. We will return to this point in the next section. Our results depicted in Fig. 6 show considerable discrepancies between the two databases.

Results from different sources on facial expression estimation using the BU dataset are shown in Table 2. However, the dataset is used in different manner in the different works. We used *only* the information available from single 2D images and estimated the 3D parameters by means of the CLM method. In contrast, the other works exploited the range information provided with the

Table 2

Comparison of results on BU dataset. Temp: temporal information is exploited. G: ground truth 3D range information used, E: 3D range information estimated. Note that we used only shape information for emotion estimation. In our approach the temporal characteristics of the expressions were not exploited.

Authors	Method	Temp	Range	An.	Di.	Fe.	Ha.	Sa.	Su	Avg.
Sun et al. '08	R2D-HMM	✓	G	92.44	87.58	85.40	97.81	80.32	99.34	90.48
Sun et al. '10	4D-HMM	✓	G	94.12	94.09	94.45	94.52	93.87	95.02	94.35
Le et al. '11	UMB-HMM	✓	G	-	-	-	95.00	91.67	90.00	92.22
This work	CLM+SVM		E	82.10	74.40	61.50	87.20	71.80	92.10	78.18

dataset and/or temporal information [17,18,19,20,21]. In our application, we could also take advantage of various temporal methods, including temporal Independent Component Analysis that has been studied previously [22].

3.4 Experiment on Pose Dependence using BU dataset

The main goal of our studies is to develop methods for situation analysis when the head pose can change and may considerably differ from the frontal view. In turn, it is crucial for us to study pose dependence. We used the BU dataset for these experiments, since it is excellent from our point of view. In this experiment we studied CLM's performance as a function of pose, since we are interested in pose invariant emotion recognition for situation analysis. We rendered 3D faces with six emotions (anger, disgust, fear, happiness, sadness, and surprise) using the BU-4DFE dataset: only these expressions are available in the database. We randomly selected 25 subjects and rendered rotated versions of every emotion. We covered anti-clockwise rotation angles between 0 and 44 degrees around the yaw axis.

As illustrated in Fig. 7, CLM based classification is robust against large pose variations, including the hard cases like anger. However, the composition of the misclassified emotions changes as a function of angle.

We computed the error of the estimated landmark position as a function of pose angle for all landmarks and all samples that we generated. The red curve of Fig. 8 shows the results: as the angle of rotation increases, the error of the estimation accumulates and may reach 10 RMSE unit on average (1 pixel error for all landmarks corresponds to 1 RMSE unit). This error influences emotion recognitions only slightly. We also computed the same error but only within a 4 degree window and the error was computed relative to the mean angle within

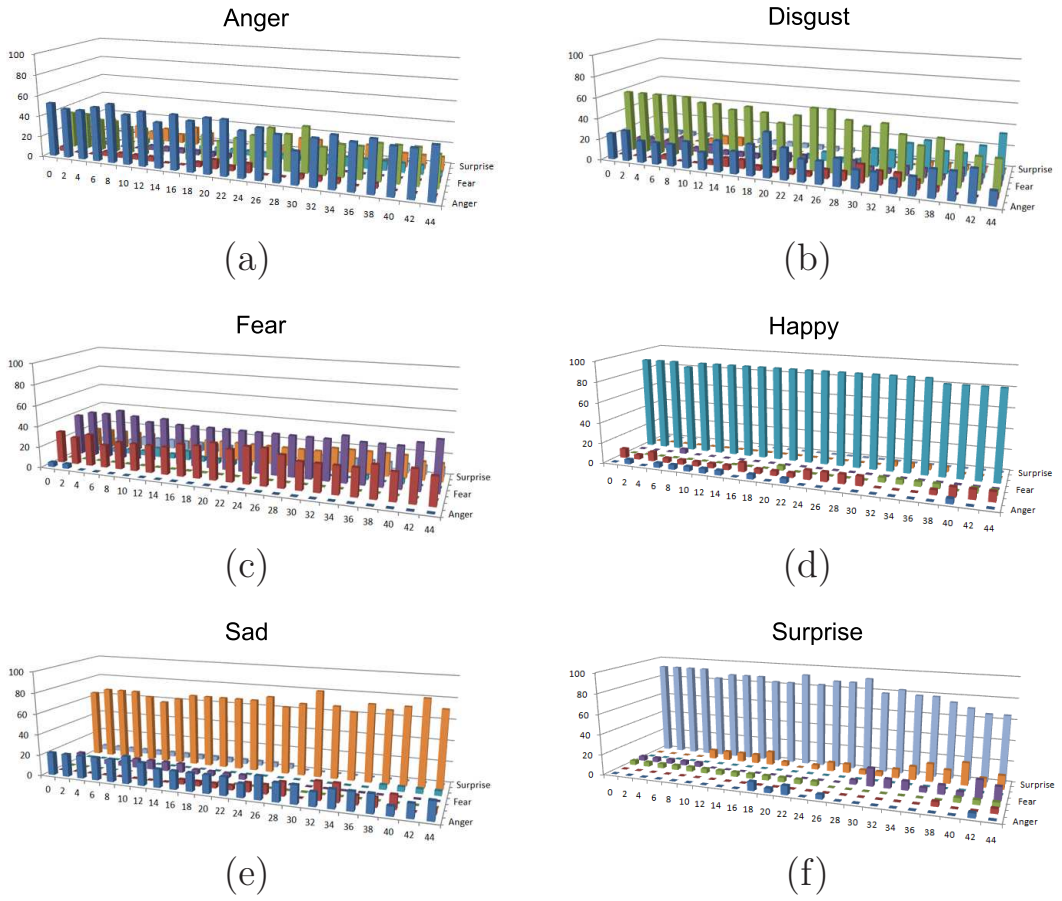


Fig. 7. Measurements on pose dependence using the BU-4DFE database. The graphs show the classification of different emotions as a function of rotational degree.

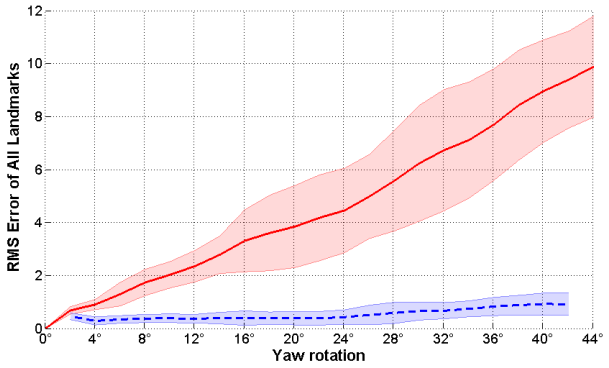


Fig. 8. RMSE of reconstructed CLM estimations of the landmark positions of the 2D meshes in pixels as a function of rotation angle and using every landmark. Red: Error of angle estimation: distortion was compared to the initial frame (0 degrees of yaw rotation). Blue: error of angle estimation relative to the mean within a 4 degree angular domain. Bold lines denotes the mean values. Shaded regions around the mean show the standard deviation of the data. 1 pixel error for all landmarks corresponds to 1 RMSE unit.

the window (blue curve of Fig. 8). Close to linear dependence of the estimated angle (red curve) and small standard deviation around every angle for every samples in a 4 degree angle region indicates that angle correction/calibration is feasible.

Other works have also addressed pose invariant recognition on the BU dataset, like [23,24], however they are using the static version of the database, therefore direct comparison with our results was not made.

4 Estimation of Action Units

A number of works demonstrates that emotion recognition from estimation of AU intensities – i.e., following the practice of FACS coders – is comparable or better than direct emotion recognition. Also, for situation analysis, recognition of facial expressions beyond the basic emotions is of high relevance and such information is encoded into the AU intensities. In turn, we studied CLM's precision for AU estimation.

Table 3

Shape based AU binary classification for the Enhanced CK and the CK+ database using CLM. Notations. N: number of samples. CR: classification rate. AUC: area under curve, ROC for the case when precision and recall are equal. Avg: average performance. wAvg: average performance weighted according to the number of the positive samples.

AU	Enhanced CK				CK+			
	N	CR	AUC	F_1	N	CR	AUC	F_1
1	1641	92.04	90.90	75.43	173	94.77	93.45	82.29
2	1107	96.24	95.99	83.20	116	97.81	96.59	87.18
4	1974	88.44	87.32	74.19	191	88.95	83.88	68.56
5	803	93.37	87.29	64.38	102	94.94	91.61	68.27
6	1292	91.25	93.11	70.61	122	93.84	93.59	69.11
7	1271	91.73	91.00	71.56	119	91.82	90.05	63.87
9	566	95.33	94.56	64.49	74	98.31	97.98	86.67
11	-	-	-	-	33	97.13	87.54	33.33
12	1715	92.42	95.53	80.75	111	96.12	96.93	81.82
15	872	91.34	84.04	52.63	89	91.99	78.98	46.32
17	1803	83.34	82.20	60.27	196	85.58	77.55	59.51
20	-	-	-	-	77	94.94	93.78	58.97
23	586	94.95	88.34	55.73	59	95.36	88.49	50.00
24	730	92.43	84.11	49.32	57	94.27	75.95	40.91
25	3503	87.52	88.92	85.84	287	90.64	90.26	84.31
26	-	-	-	-	48	95.03	80.26	15.25
27	741	97.96	97.65	87.99	81	98.65	99.65	90.24
Avg	-	**92.03**	**90.07**	**69.74**	-	**94.13**	**89.21**	**63.92**
wAvg	-	**90.64**	**89.75**	**73.12**	-	**92.91**	**89.13**	**67.80**

We used the data of the Enhanced CK database and tuned binary linear SVM for deciding if an AU was active or not. We evaluated the binary classifiers with the leave-one-sample-out in order to measure performance: we counted the true (t) and false (f) positive (p) and negative (n) results. We computed the receiver-operating characteristic (ROC) as a function of the SVM parameter: it changes the precision ($P = \frac{tp}{tp+fp}$) and the recall ($R = \frac{tp}{tp+fn}$) values. We also computed the F_1 value ($F_1 = 2 \cdot \frac{P \cdot R}{P+R}$) for the case when recall and precision are equal ($P = R$). Our ROC results for $P = R$ are shown in Table 3 in the column called 'area under curve' (AUC). The same figure lists the other performance measures for our method.

Table 4 shows the results of different methods including ours. Most of the research on automatic AU recognition has been based on individual frames of an image sequence [25,26,27,28,29] and some research efforts toward using temporal information for facial expression recognition [30,31,32].

Table 4

Comparison of AU estimation with different methods. Notations. T (S): texture (shape) based method, CR: classification rate, AUC: area under curve for the case when precision and recall are equal. The CK+ database part (bottom part) contains the average performances weighted according to the number of the positive samples. We also stated the unweighted averages in brackets. Temp: temporal information is exploited. For the sake of comparison we computed performance measures – beyond the 17 AUs availble – for 10 AUs that were studied in [7].

Authors	T/S	Temp	AU	CR	AUC	F1
Original CK dataset						
Bartlett et al '05	T		17	94.80	-	-
Bartlett et al '06	T		20	90.90	92.60	-
Whitehill et al '06	T		11	92.35	-	-
Littlewort et al '06	T		7	92.90	-	-
Valstar et al '06	T	*	15	90.20	-	-
Tong et al '07	T	*	14	93.30	-	-
Koelstra et al '10	T	*	18	89.80	-	72.10
Koelstra et al '10 (15 best AUs)	T	*	15	92.50	-	72.50
CK+ dataset						
Lucey et al '10	S		17	-	90.02 (88.51)	-
Lucey et al '10	T		17	-	91.38 (90.82)	-
Lucey et al '10	T+S		17	-	94.47 (93.45)	-
Chew et al '11	T		10	-	-	73.0 (68.8)
This work	S		17	92.91 (94.13)	89.13 (89.21)	69.7 (63.92)
This work (Chew et all '11 subset)	S		10	92.66 (91.81)	88.16 (88.34)	65.82 (71.12)

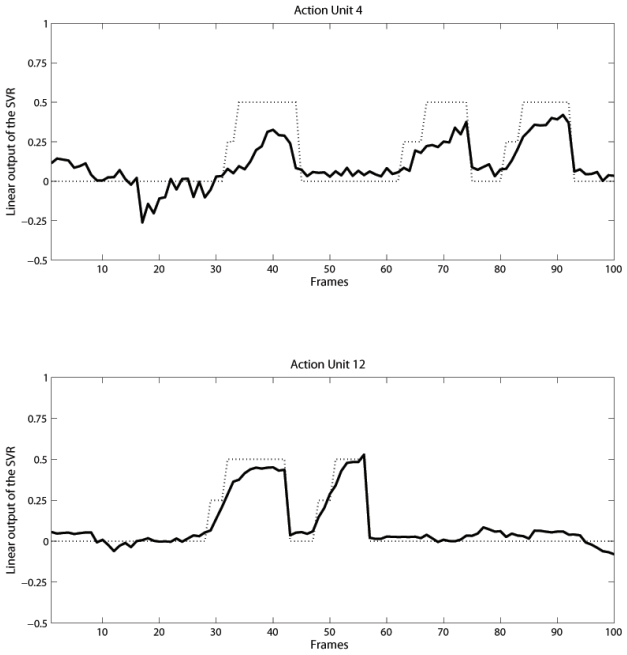


Fig. 9. Output of support regressor for two different AUs. Solid line: SVR output, dotted line: ground truth from Enhanced CK.

CLM compares favorably with the other methods with the following notes: (a) temporal information should improve CLM's performance and (b) evaluation of pose dependence can not be accomplished on the CK+ database at present. For the sake of comparison, we repeated the study for the subset of AUs used in [7] – 10 AUs out of 17 ones – since the same CLM is applied in this work, but they exploit different features and normalization.

We used LS-SVM for the linear estimation of the intensities of the 14 AUs that were annotated frame-by-frame with three intensities in the Enhanced CK database, namely (a) absent, (b) present but with low intensity, and (c) present. The output of the support regressor for two different AUs can be seen in Fig. 9.

4.2 Experiment on Pose Independent Estimation of Action Units

Pose independent estimation of Action Units is critical for situation analysis. In this experiment we studied the performance of AU recognition as a function of pose.

We rendered 3D faces with six emotions (anger, disgust, fear, happiness, sadness, and surprise) using the BU-4DFE dataset. We randomly selected 25

subjects and rendered rotated versions of every emotion. Alike for the case of facial expressions, we covered anti-clockwise rotation angles between 0 and 44 degrees around the yaw axis.

The BU dataset does not contain ground truth Action Unit labels and intensity information. We used the previously trained Support Vector Classifiers to provide the Action Units for the frontal faces for each subjects, we used these values as ground truth. We selected a subset of AUs with different classification difficulty based on the previous findings (see Table 3) and studied the classification performance for the different poses.

As illustrated in Fig. 10, CLM based classification is robust against large pose variations, including the difficult cases like (AU 4, 6 and 17).

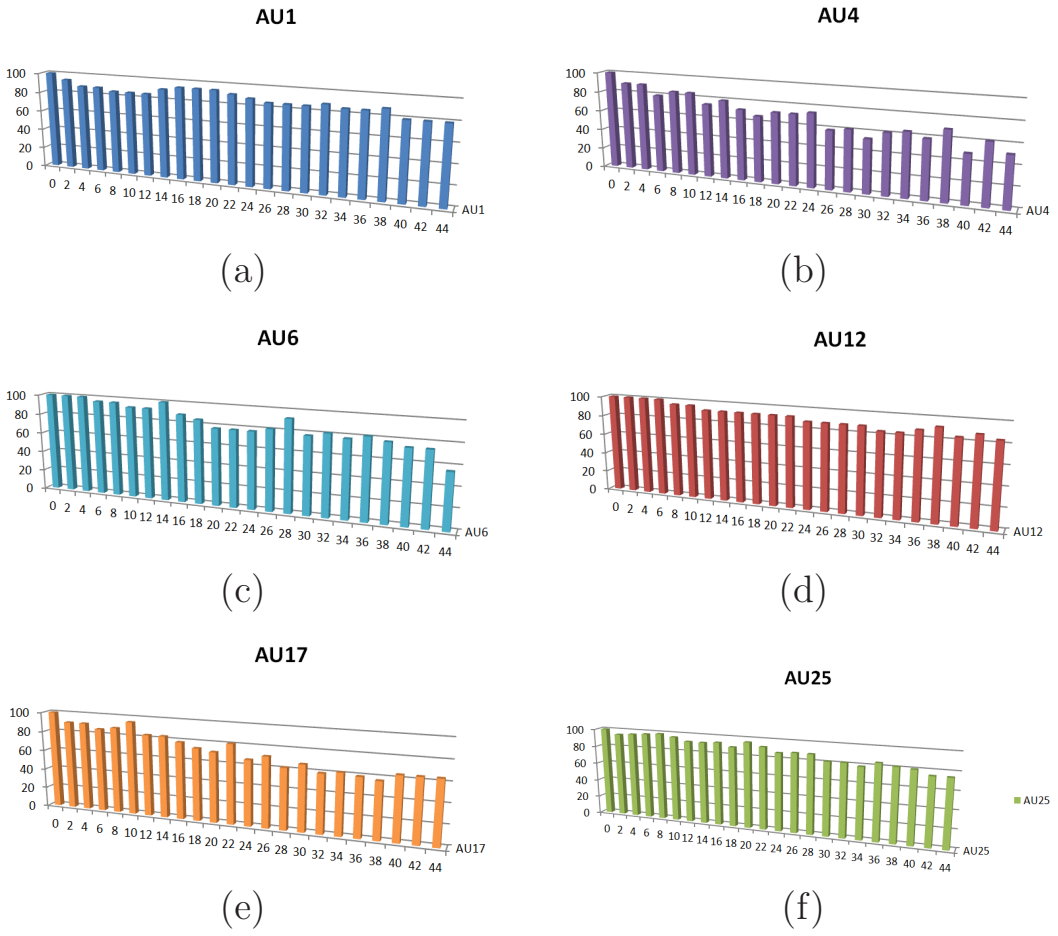


Fig. 10. Measurements on pose dependence for Action Unit classification using the BU-4DFE database. The graphs show the classification of different AUs as a function of rotational degree. (a): AU1: Inner Brow Raiser, (b) AU4: Brow Lowerer, (c) AU6: Cheek Raiser, (d) AU12: Lip Corner Puller, (e) AU17: Chin Raiser (f) AU25: Lips part.

5 Discussion and Conclusions

We are interested in situation analysis and human-computer interaction that call for emotion estimation for different head poses. Since simple line drawings are very expressive, we decided to study shape based emotion and action unit estimations. The importance of this direction is underlined by our real life experiments; the gray level of the faces of our anonymous subjects covered only 3% of the full range (every pixel was above 244 on the 0-255 scale). We were able to extract highly precise shape information usin a CLM, whereas textural information was almost totally lost and useless.

In the present work we conducted a preliminary study: we used the full shape information provided through the 2D landmarks (i.e., we did not apply any PCA compression) and applied Procrustes normalization. This study was very promising, it gave rise to excellent results on the CK+ database. Recently, a highly robust CLM method have been developed in the literature [4] both for 2D and 3D models (see, e.g., [6] and the references therein). This CLM preserves more information than previously applied AAM methods [1], so we wanted to evaluate its performance. In turn, we engaged in studies using the 3D CLM method, which should be robust against head pose differences.

We used a variety of evaluations to study performance of shape representations for facial expression recognition in order to have the ability of comparing results achieved with different methods and on different databases. In all studies, we applied multi-class SVM classification [12]. We used expert annotated frontal databases [1] as well as 3D dynamic datasets [3].

In our studies, we discovered that the neutral shape can be recovered from temporal sequences quickly and we replaced AU0 based normalization with the personal mean shape. For the expert annotated CK+ database results changed only slightly, but for CLM estimations this method gave rise to considerable improvements. We think that the difference is in the noise of CLM based AU0 estimation, which is larger than the discrepancy between AU0 values (as determined by the experts) and mean shape values (as determined by averaging over the shapes of the same person). Note that personal mean shape can be estimated in a number of ways. We found that personal mean shape is better if we average over the neutral shape *and* the shapes of the different facial expressions. This intriguing result has great promises for practice, since it allows for online updates of the mean shape with different time windows allowing for the detection of slight changes in the mood and also for the estimation of the more-and-more precise personal mean shape.

Furthermore, personal mean shape normalization gave rise to very good results for the case when shape information was used exclusively. Our results surpass

performances of the best available AAM methods [1] and CLM [7] that utilize shape *plus* texture and temporal information, respectively. It may be worth noting that FACS coders work from using local textural information [16] and further improvements are expected by using this valuable additional piece of information

We studied the robustness of the CLM method for yaw rotations. We rendered rotated 3D faces using the BU-4DFE database [3] and found that CLM based shape estimation and shape based emotion recognition are highly robust against such pose variations. Our results compare favorably to other methods in the literature, although we applied only shape information for emotion estimation. Further performance gains can be expected if textural information [6] and if temporal information [17,18] are included. Smoothing over time, e.g., by Hidden Markov Models [17,18] seems crucial for sensitive detection of emotions and AU intensities. Practice of FACS coders point to textural information, whereas noise filtering is better if temporal information is exploited.

5.1 Future work

Both the BU-4DFE database and the major part of CK+ contain posed facial expressions that may considerable differ from the natural ones [33]. Our evaluations have been performed on such posed emotions. Spontaneous emotions, especially micro-emotions [34] are highly challenging, especially since (i) these are very important for social interaction, (ii) micro expressions are typically subconscious, and (iii) estimations for different head poses are crucial for situation analysis. The challenge is not only in the pose, but also in the resolution (if the camera is far), the high frequency evaluation required for catching these expressions, and the real time evaluation, which is necessary for human-computer interaction. Present technology is far from sufficient and further work is required, including the development of marked and FACS coded databases for spontaneous emotions. Robustness of AU estimation against pose dependence could be studied, e.g., if the BU-4DFE database were FACS'd.

We suspect that CLM based estimations may also be robust against light conditions due to the strength of the CLM approach that multiples probability estimations of experts [35]. One can envision methods that utilize texture and shape experts for emotion and AU intensity estimation.

In sum, shape information is very efficient for facial expression recognition provided that details of shape changes are spared in the shape representation. This can be of high value in situation analysis, since shape estimation is robust against pose variations as we showed here. Further improvements are expected

for methods that include textural and temporal information. Such improvements are necessary for situation analysis and human-computer interaction.

Acknowledgments

We are grateful to Jason Saragih for providing his CLM code for our work. Research was supported by the European Union and co-financed by the European Social Fund (grant agreement numbers TÁMOP 4.2.1./B-09/1/KMR-2010-0003 and KMOP 1.2-08/1-2008-002).

References

[1] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB10), 2010, pp. 94–101.

[2] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (10) (2007) 1683–1699.

[3] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, A high-resolution 3D dynamic facial expression database, in: 8th IEEE International Conference on Automatic Face Gesture Recognition, IEEE, 2008, pp. 1–6, tracking Number: 66.

[4] D. Cristinacce, T. Cootes, Automatic feature localisation with constrained local models, Pattern Recognition 41 (10) (2008) 3054–3067. doi:10.1016/j.patcog.2008.01.024.

[5] J. M. Saragih, S. Lucey, J. F. Cohn, Deformable model fitting by regularized landmark mean-shift, International Journal of Computer Vision 91 (2) (2011) 200–215.

[6] P. Lucey, J. F. Cohn, K. M. Prkachin, P. Solomon, I. Matthews, Painful data: The UNBC-McMaster Shoulder Pain Expression Archive Database, in: 9th IEEE International Confference on Automatic Face and Gesture Recognition (FG2011), 2011, pp. 57–64.

[7] S. W. Chew, P. J. Lucey, S. Lucey, J. Saragih, J. F. Cohn, S. Sridharan, Person-independent facial expression detection using constrained local models, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), FERA 2011 Workshop on Facial Expression Recognition and Analysis Challenge, 2011, pp. 915–920.

[8] I. Matthews, S. Baker, Active appearance models revisited, International Journal of Computer Vision 60 (2) (2004) 135–164.

[9] S. Taheri, P. Turaga, R. Chellappa, Towards view-invariant expression analysis using analytic shape manifolds, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), 2011, pp. 306–313.

[10] L. A. Jeni, H. Hashimoto, A. Lorincz, Efficient, Pose Invariant Facial Emotion Classification using Constrained Local Model and 2D Shape Information, in: Extended Abstracts, Workshop on Gesture Recognition at CVPR 2011, 2011, extended abstract available at http://clopinet.com/isabelle/Projects/CVPR2011.

[11] L. A. Jeni, D. Takacs, A. Lorincz, High Quality Facial Expression Recognition in Video Streams using Shape Related Information only, in: IEEE International Workshop on Benchmarking Facial Image Analysis Technologies at ICCV, 2011.

[12] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] J. Shawe-Taylor, N. Cristianini., Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[14] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), 2000, pp. 46–53.

[15] J. M. Susskind, G. Littlewort, M. S. Bartlett, J. Movellan, A. K. Anderson, Human and computer recognition of facial expressions of emotion, Neuropsychologia 45 (1) (2007) 152–162.

[16] P. Ekman, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Pres, USA, 2005.

[17] Y. Sun, M. Reale, L. Yin, Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'08), 2011, pp. 1–8.

[18] Y. Sun, L. Yin, Facial expression recognition based on 3d dynamic range model sequences, in: European Conference on Computer Vision, 2008, pp. 58–71.

[19] Y. Sun, X. Chen, M. Rosato, L. Yin, Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis, IEEE Transactions on Systems, Man, and CyberneticsPart A: Systems and Humans 40 (3) (2010) 461–474.

[20] V. Le, H. Tang, T. Huang, Expression recognition from 3d dynamic faces using robust spatio-temporal shape features, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), FERA 2011 Workshop on Facial Expression Recognition and Analysis Challenge, 2011, pp. 414–421.

[21] G. Sandbach, S. Zafeiriou, M. Pantic, D. Rueckert, A dynamic approach to the recognition of 3d facial expressions and their temporal models, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), FERA 2011 Workshop on Facial Expression Recognition and Analysis Challenge, 2011, pp. 406–413.

[22] A. Lőrincz, G. Szirtes, B. Takács, I. Biederman, R. Vogels, Relating priming and repetition suppression, International Journal of Neural Systems 12 (2002) 187–202.

[23] W. Zheng, H. Tang, Z. Lin, T. S. Huang, Emotion recognition from arbitrary view facial images, in: European Conference on Computer Vision, 2010, pp. 490–503.

[24] O. Rudovic, I. Patras, M. Pantic, Coupled gaussian process regression for pose-invariant facial expression recognition, in: European Conference on Computer Vision, 2010, pp. 350–363.

[25] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 568–573.

[26] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully Automatic Facial Action Recognition in Spontaneous Behavior, in: IEEE Conference on Face and Gesture Recognition (FG'06), 2006, pp. 223–230.

[27] Y. Chang, C. Hu, R. Feris, M. Turk, Manifold-based analysis of facial expression, Image and Vision Computing 24 (6) (2006) 605–614.

[28] J. Whitehill, C. Omlin, Haar Features for FACS AU Recognition, in: IEEE Conference on Face and Gesture Recognition (FG'06), 2006, pp. 97–101.

[29] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image and Vision Computing 24 (6) (2006) 615–625.

[30] M. Valstar, M. Pantic, Fully Automatic Facial Action Unit Detection and Temporal Analysis, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2006, pp. 149–156.

[31] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32 (11) (2010) 1940–1954.

[32] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1683–1699.

[33] T. McLellan, L. Johnston, J. Dalrymple-Alford, R. Porter, Sensitivity to genuine versus posed emotion specified in facial displays, Cognition and Emotion 24 (8) (2010) 1277–1292.

[34] P. Ekman, Darwin, Deception, and Facial Expression, Annals of the New York Academy of Sciences 1000 (10) (2006) 205–221.

[35] G. E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Computation 14 (8) (2002) 1771–1800.