

TDK-dolgozat

Tamás Nagy, Judit Sebők

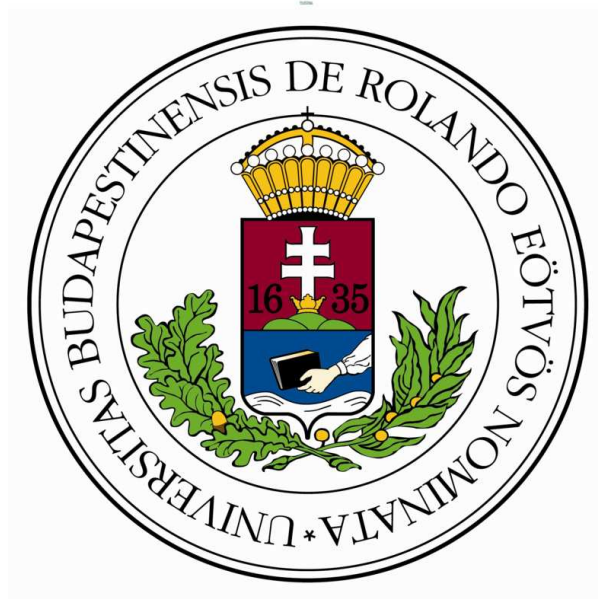
3D Constrained Local Model and its use in Facial Expression Recognition

Tamás Nagy (MSc) and Judit Sebők (MSc)
thomas.nagy@bkil.hu / dymorse@gmail.com

8th December 2011

Supervisors:
András Lőrincz
andras.lorincz@elte.hu
Senior research fellow, head of the Neural Information Processing Group
Eötvös Loránd University, Hungary

László A. Jeni
laszlo.jeni@ieee.org
PhD Candidate
University of Tokyo, Japan



Eötvös Loránd University
Faculty of Informatics
Department of Software Technology and Methodology
2011.

Contents

1	Introduction	3
2	Methods	4
2.1	Constrained Local Models	4
2.1.1	Point Distribution Model	4
2.1.2	Constrained Local Model	5
2.1.3	Expectation Maximization for CLM	6
2.2	Procrustes' transformation	6
2.3	Datasets	7
2.3.1	Datasets based on the Cohn-Kanade Facial Expression Database . .	7
2.3.2	FaceGen Database	7
2.3.3	The BU-4DFE Dynamic Facial Expression Database	7
2.4	Extracted Features	8
2.5	AU0 Normalization	8
2.6	Multi-class Support Vector Machine for Emotion Classification	8
3	CLM experimental Results	9
3.1	2D vs 3D models	9
3.2	Training the classifiers	10
3.3	Training the PDM	10
4	Emotion tracking experimental Results	10
4.1	Experiment on the CK+ dataset with Procrustes' method and original landmarks	10
4.2	Experiment on the CLM-tracked CK+ dataset	11
4.3	Experiments on the CLM-tracked BU-4DFE dataset	12
4.4	Experiment on Pose Dependence using BU dataset	13
5	Discussion and Conclusions	15
5.1	Future work	17

Abstract

Person independent and pose invariant facial emotion classification is important for situation analysis and for automated video annotation. We started from the raw 2D shape data of the CK+ database and used a simple Procrustes transformation and the multi-class SVM leave-one-out method for classification. We found close to 100% performance demonstrating the relevance of details in shape space. Precise 3D shape information can be computed by means of constrained local modes (CLM). In the next step, we used this method, fitted 3D CLM to CK+ data, and derived the frontal views of the 2D shapes. Performance reached and sometimes surpassed state-of-the-art results. In the third step, we rendered 3D emotional database with different poses using BU 4DFE database, fitted 3D CLM, transformed the 3D shape to frontal pose and evaluated the outputs of our classifier. Results show that the high quality classification is robust against pose variations. Our approach enables both situational analysis where non-frontal head poses are typical and the utilization of robust temporal methods like [1], which can give rise to further performance gains.

1 Introduction

Our everyday communication is highly influenced by the emotional information available to us about our partner. 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+) [2], its enhanced version [3], and the dynamic 3D facial expression database [4] as well as (2) the advance of learning algorithms, most notably the advance of constrained local models (CLM) [5, 6]. Recently, very good results have been achieved by means of textural information [7, 8]. On the other hand, shape of the face extracted by active appearance models (AAM) [9] showed relatively poor performance (see, e.g., [2], 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 and CLM based automated marker generation gives rise to, sometimes surpasses to 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 easily measurable personal mean shape normalization and – as we show here – results outperform known results of texture and shape AAM [2], texture based CLM [8] and 2D shape based method using the Grassmann manifold [10]. In turn, we suggest shape identification for a standard benchmark. Preliminary studies concerning the emotion classification of this paper have been submitted to a conference.

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 on this technique in Section 2.1. Second, we remove the rigid transformation from the acquired 3D shape, project it to 2D and perform an SVM-based multi-class classification on it using the different emotions as the class labels. We show this method in Section 2.2.

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:

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

or, $\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_M^T)^T$, 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 all $\{\mathbf{x}^{(n)}\}_{n=1}^N$ can be approximated by means of the linear principal component analysis (PCA) forming the PCA subspace apart from scale, rotation, and translation.

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 describes non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

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

where $i = 1, \dots, M$, $\mathbf{x}_i(p)$ denotes the 2D location of the i^{th} landmark subject to transformation p , and $p = \{s, \alpha, \beta, \gamma, \mathbf{q}, \mathbf{t}\}$ 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. $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1^T; \dots; \bar{\mathbf{x}}_M^T]$, $\bar{\mathbf{x}}_i = [\bar{x}_i, \bar{y}_i, \bar{z}_i]^T$, $\bar{x}_i = \frac{1}{N} \sum_{n=1}^N x_i^{(n)}$, and similarly, for $\bar{y}_i$ and $\bar{z}_i$. Matrix Φ_i ($i = 1, \dots, M$) is a $3 \times d$ piece in $\Phi \in \mathbb{R}^{3M \times d}$ and corresponds to the landmarks. Columns of Φ 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:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (3)$$

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

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

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

that is CLM assumes a normal distribution with $\mathbf{0}$ mean and $\mathbf{\Lambda}$ variance for parameters $\mathbf{q}$. $\mathbf{\Lambda} = \text{diag}(\lambda_1; \dots; \lambda_d) \in \mathbb{R}^{d \times d}$ 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 Constrained Local Model

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

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

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

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

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)):

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

where CLM assumes $\mathbf{y}_i = \mathbf{x}_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \rho \mathbf{I})$, $\rho = \sum_{i=d+1}^{3M} \frac{\lambda_i}{(3M-d)}$, and λ_i is the i^{th} eigenvalue of $Cov(\mathbf{x})$, the covariance matrix of stochastic variable $\mathbf{x}$.

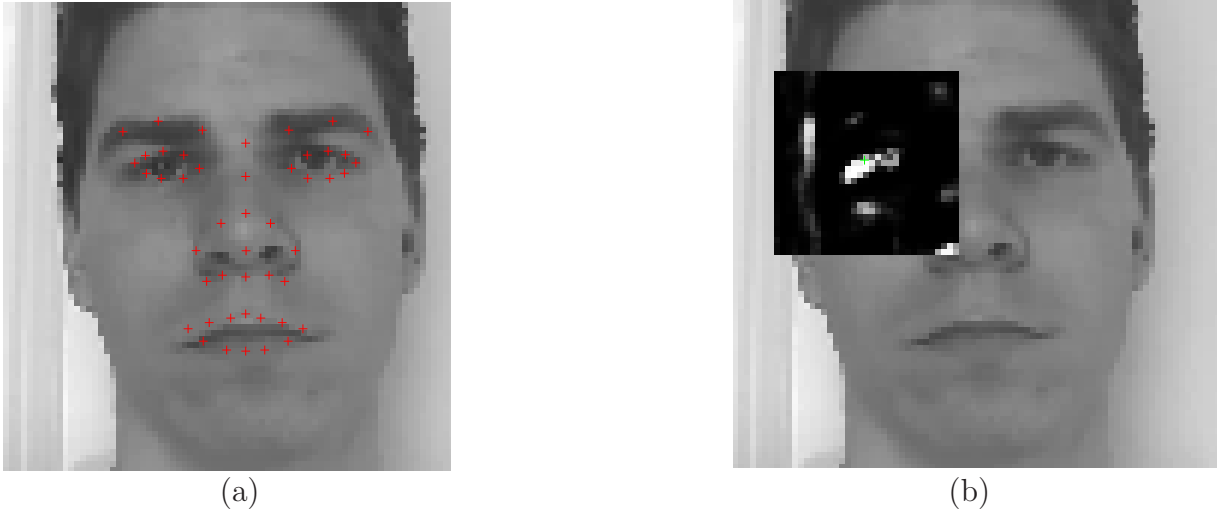


Figure 1: (a): Landmarks on the face. (b): response map for the right corner of the right eye.

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

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

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

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

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

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

giving rise to

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

where $\mathbf{J}$ is the Jacobian of the cost function, $\tilde{\Lambda} \in \mathbb{R}^{(6+d) \times (6+d)}$ is the diagonal matrix with the PDM eigenvalues and with leading zeros in its diagonal (the extension of the Λ 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 $\mathbf{v} = [\mathbf{v}_1; \dots; \mathbf{v}_M]^T$ where

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

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

2.2 Procrustes' transformation

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.

2.3 Datasets

2.3.1 Datasets based on the Cohn-Kanade Facial Expression Database

In our simulations we used the Cohn-Kanade Extended Facial Expression (CK+) Database [2]. 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 [11] 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 [3] 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.

Since this database provides 2D landmark coordinates only, our models trained on it was not accurate enough for faces with rotation. It also did not usually work well on images with lighting conditions other than standard indoor lighting.

2.3.2 FaceGen Database

In order to improve our CLM models, we used the FaceGen SDK to generate different faces annotated with landmark points. This database not only includes the 2D landmark point coordinates of the images, but also the corresponding 3D points of the face. It provides a variety of facial expressions and lighting positions. We used 52 landmark points, each of them corresponding to one of the landmark points of the CK+ database. Our training experiences however concluded that the landmark positions are not accurate enough which makes the tracking very noisy.

2.3.3 The BU-4DFE Dynamic Facial Expression Database

For our studies on pose dependence we used the BU-4DFE dataset [4]. 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.

We have additionally manually marked a neutral and multiple apex frames in each sequence, resulting in a total of 1678 precisely mapped 3D models. These marked face models could then be rendered in OpenGL in order to obtain images of rotated faces with different lighting conditions. This way we are able to generate the needed images for the patch-experts and also gain 3D landmark points for training the PDM.

2.4 Extracted Features

We 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.

We also used the FaceGen dataset, rendered 3D facial expressions, and rotated those to different poses. Changed the lighting intensity and pose. This procedure was followed by the 3D CLM analysis and we extracted the 2D shape parameters alike in the previous case.

In another set of experiments, we used the BU-4DFE dataset [4], 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.

2.5 AU0 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. We assume that we have videos (frame series) about the subject like in the case of the BU-4DFE and we can compute the personal mean shape. 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. Since the BU-4DFE contains highly distorted shapes of different kind, we believe that time averaged shapes are either close to neutral, or if not, the difference should be taken into account. Quantitative analysis of this conjecture with naive subjects, however, is to be conducted in future studies.

2.6 Multi-class Support Vector Machine for Emotion Classification

After we estimated the 3D landmarks with the CLM method, we removed the rigid transformation from the acquired 3D shape, projected it to 2D and performed an SVM-based multi-class classification on it using the different emotions as the class labels.

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 hyper-plane between the two classes by maximizing the margin between the hyper-plane 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 $k = 1, \dots, K$. Here, for class ‘1’ and for class ‘2’ $y^{(k)} = 1$ and $y^{(k)} = -1$, respectively. We also have a set of feature vectors $\phi(= [\phi_1; \dots; \phi_J]) : \mathbb{R}^m \rightarrow \mathbb{R}^J$, where J might be infinite. The support vector classification seeks to minimize the cost function

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

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

where ξ_k ($k = 1, \dots, K$) 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

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

that should be minimized.

3 CLM experimental Results

For shape based emotion tracking, we needed a CLM model that is accurate and robust enough. Since on a video we monitored the position changes of the landmark points, we needed a model that promptly follows the face when the expression is changing but staying still when no change is happening. We tried training on different datasets with various characteristics. These were the CK+, FaceGen, BU-4DFE and MultiPIE. We trained a number of different CLM models and evaluated them on different datasets. The accuracy of CLM fitting can be automatically evaluated if we have the ground-truth for the dataset (BU-4DFE, CK+, FaceGen), or it can be evaluated manually for example on videos recorded with web-cameras.

3.1 2D vs 3D models

In the first experiments we used the CK+ dataset with the original 68 CK+ 2D landmarks. These models did not track the rotated faces well enough. The solution seemed to be to switch from 2D to 3D models. For that purpose we generated our training databases with FaceGen. This produced a noisy fitting on videos, meaning that the position of the landmarks changed frame-to-frame even if no movement happened. This can make the emotion tracker detect false values. Our next experiment was to use the BU-4DFE database. Manually marking some of the 2D images led to the possibility of using those images to render 3D meshes with more accurate landmark positions. This way we are

able to obtain these faces from any angle and with different lighting settings. Using 3D models have the advantage of being able to rotate the model around the y-axis allowing us to follow the pitch and yaw rotation of the head better.

3.2 Training the classifiers

One of the crucial point of the CLM training is the training of the classifiers. If the classifier is trained on a too large database, it can-overlearn, however if it did not see enough examples, it might not be able to follow the features of the face accurately enough. In our experiments we first used databases from one separate database. Each of the databases has its own limits. The CK+ could not follow the face under different lighting conditions, also it was not able to follow the rotation of the head well enough. The FaceGen database produced a noisy output, it also did not follow accurately enough some of the crucial points needed for emotion tracking such as the eyebrow points. In our experiments using the BU-4DFE database gave us better results than the others. Images generated from it with random lighting was able to track the faces nearly as accurately as the CMU CLM.

3.3 Training the PDM

An other crucial point in the training of a CLM model is the training of the PDM model. If the trained model is not flexible enough, it will not follow the features. During our experiments, it turned out that the 1678 different shape data of the BU-4DFE seems enough to train a good PDM model. If the classifiers are not strong enough, than the model can not work properly, but if the PDM model is not flexible enough than it can happen that even a good classifier can not move the landmark point to its place, because of a too strong shape constraint. It is therefore important to find the right amount of images to train the shape and the classifier. In our experiments using 1678 different images for the shape training was enough, using more data did not gave us any better results. As for the classifier training we used 3000-5000 images.

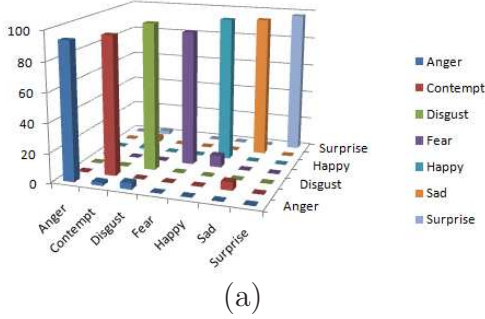
4 Emotion tracking experimental Results

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

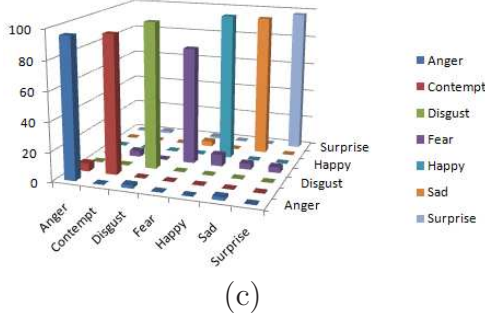
In this experiment we used the CK+ dataset with the original 68 CK+ landmarks. First, we calculated the mean shape using Procrustes' method. Then we normalized all shapes by minimizing the Procrustes distance between individual shapes and the mean shape. We trained a multi-class SVM using the leave-one-subject-out cross validation method. The result of the classification is shown in Figs. 2(a) and 2(b): 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 [14].

Personal mean shape was computed by averaging over neutral facial expressions for the one case and 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.



	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
Anger	93.3	2.2	4.4	0	0	0	0
Contempt	0	94.4	0	0	0	5.6	0
Disgust	0	0	100	0	0	0	0
Fear	0	0	0	92	8	0	0
Happy	0	1.5	0	0	98.5	0	0
Sadness	0	3.6	0	0	0	96.4	0
Surprise	0	2.6	0	0	0	0	97.4



	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
Anger	95.6	0	2.2	0	0	2.2	0
Contempt	5.6	94.4	0	0	0	0	0
Disgust	0	0	100	0	0	0	0
Fear	0	4	0	80	8	4	4
Happy	0	0	0	0	100	0	0
Sadness	0	0	0	3.5	0	96.4	0
Surprise	0	1.3	0	1.3	0	0	97.4

Figure 2: Performance of two normalization methods for the expert annotated case. Results are averaged for the 118 subjects of the CK+ database. *Upper row*: Procrustes method with AU0 normalization. (a): confusion matrix, (b): same in tabular form. *Lower row*: personal mean shape based normalization. (c): confusion matrix, (d): same in tabular form. Note the results are similar. For more details, see text.

Replacing AU0 normalization by personal mean shape slightly decreases average performance: recognition on the CK+ database drops from 96% to 94.8% (Figs. 2(c), 2(d)). When the personal mean shape was computed 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.

4.2 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. 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. The recognition performance of the system can be seen in Figs. 3(a) and 3(b): 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.

We evaluated the personal mean shape normalization (Figs. 3(c), 3(d)). 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

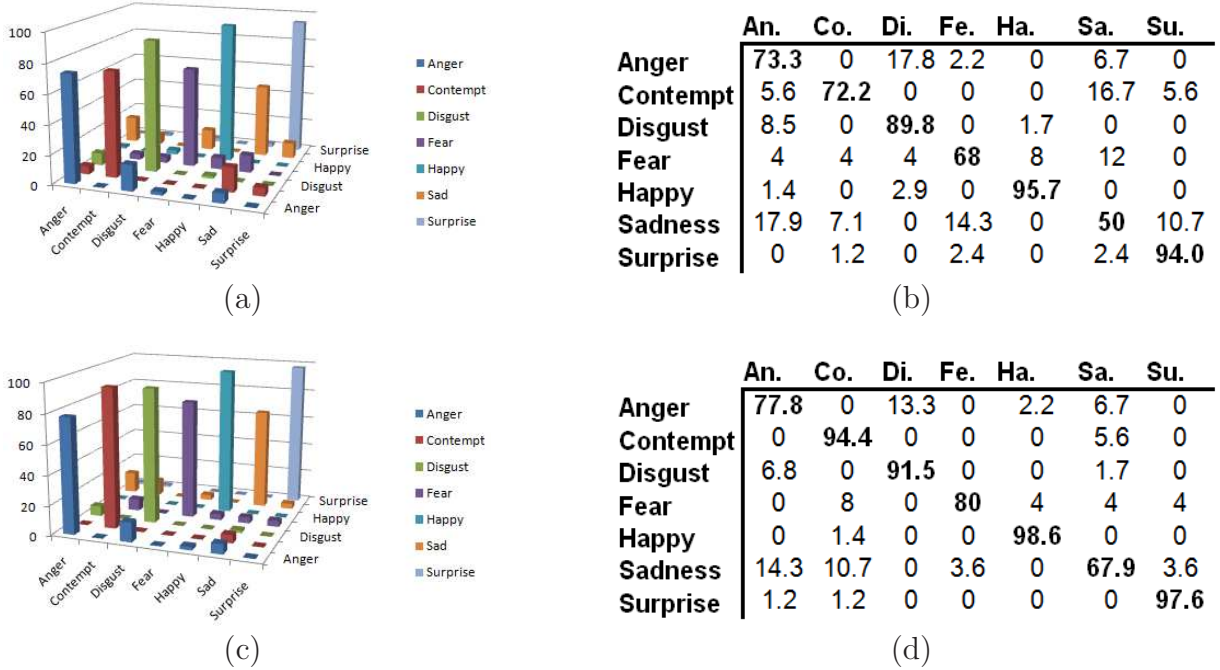


Figure 3: Performance of two normalization methods for the CLM annotated case. Results are averaged for the 118 subjects of the CK+ database. *Upper row*: CLM method with AU0 normalization. (a): confusion matrix, (b): same in tabular form. *Lower row*: CLM method with personal mean shape based normalization. (c): confusion matrix, (d): same in tabular form. Note the results are similar. For more details, see text.

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 [2] and the best CLM result that utilizes *only* textural information [8], see our comparisons in Fig. 4.

4.3 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 the personal mean shape. 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 personal mean shape method (Fig. 5). We executed cross evaluations (Fig. 6). We used the CK+ as the ground truth, since it seemed more precise than the BU-4DFE: the target expression for each sequence is fully FACS coded, emotion labels have been revised and validated, and CK+ utilizes FACS coding based emotion

Authors	Method	T/S	An.	Co.	Di.	Fe.	Ha.	Sa.	Su	Avg.
Original CK dataset										
Taheri et al. '11	E-LDA	S	60.0	-	76.3	70.2	91.3	75.0	96.1	78.2
	G-LDA	S	68.0	-	80.5	74.4	88.9	78.2	97.3	81.2
	G-KLDA	S	65.7	-	86.8	83.0	95.1	85.7	98.6	85.8
	E-SVM	S	62.8	-	78.9	74.4	91.3	80.3	97.2	80.8
	G-SVM	S	65.7	-	78.9	74.5	95.0	85.7	97.2	82.8
CK+ dataset										
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

Figure 4: Comparison of results on CK dataset. Original CK dataset: results for the 6 main facial expressions: comparison with results for the same facial expressions plus contempt is troublesome. CK+ dataset: results for the 7 main facial expressions. S: shape based method, T: texture based method.

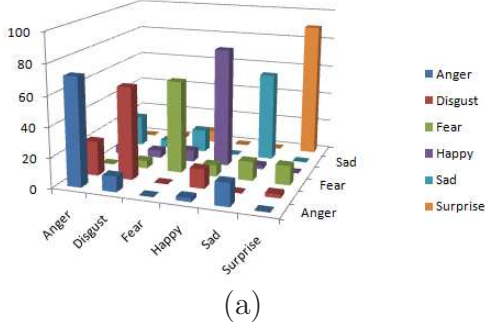
evaluation and this method is preferred in the literature [2]. We note however, that both the CK+ and the BU-4DFE facial expressions are posed and not spontaneous.

Results from different sources on facial expression estimation using the BU dataset are shown in Fig. 7. However, this dataset is used in different manner in the different experiments. Our work used *only* the information of the 2D images, whereas the other works exploited the range information provided with the dataset and/or temporal information [1, 15, 16, 17, 18]. In our application, we could also take advantage of various temporal methods, including temporal Independent Component Analysis that one of us studied previously [19].

4.4 Experiment on Pose Dependence using BU dataset

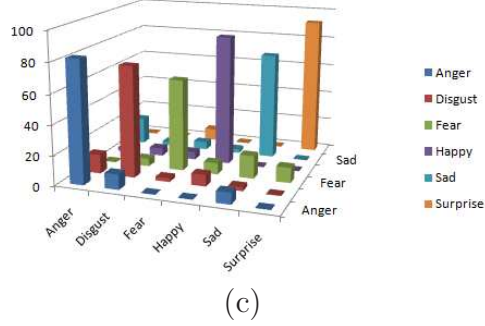
The main goal of our studies is to develop methods for situation analysis, where typically the head pose differs from the frontal view. In turn, it is crucial for us to study pose dependence. We used the BU dataset for these experiments, since is excellent from our point of view. In this experiment we 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 rotation angles between 0 and 44 degrees of anti-clockwise rotation around the yaw axis.

As illustrated in Fig. 8, CLM based classification is robust against large pose variations,



	An.	Di.	Fe.	Ha.	Sa.	Su.
Anger	71.8	10.6	0	2.6	15.4	0
Disgust	23.1	61.5	0	12.8	0	2.6
Fear	0	5.1	61.5	7.7	12.8	12.8
Happy	5.1	5.1	7.7	79.5	2.6	0
Sadness	20.5	5.1	15.4	0	59.0	0
Surprise	0	2.6	7.9	0	0	89.5

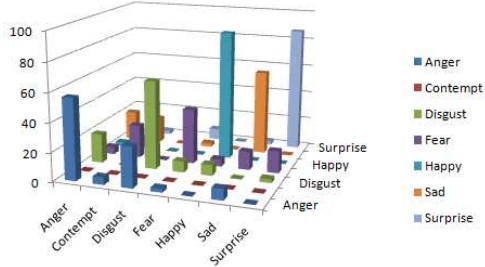
(b)



	An.	Di.	Fe.	Ha.	Sa.	Su.
Anger	82.1	10.3	0	0	7.7	0
Disgust	12.8	74.4	2.6	7.7	2.6	0
Fear	0	5.1	61.5	7.7	15.4	10.3
Happy	2.6	5.1	5.1	87.2	0	0
Sadness	17.9	2.6	5.1	2.6	71.8	0
Surprise	0	0	7.9	0	0	92.1

(d)

Figure 5: Performance of two normalization methods for the BU database using the CLM method. *Upper row*: CLM method with AU0 normalization. (a): confusion matrix, (b): same in tabular form. *Lower row*: CLM method with personal mean shape based normalization. (c): confusion matrix, (d): same in tabular form. Note the results are similar.



	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
Anger	56.4	5.1	28.2	2.6	0	7.7	0
Contempt	-	-	-	-	-	-	-
Disgust	20.5	0	61.5	7.7	7.7	0	2.6
Fear	5.1	23.1	0	38.5	5.1	12.8	15.4
Happy	2.6	7.7	0	0	89.7	0	0
Sadness	20.5	17.9	0	2.6	0	59.0	0
Surprise	0	2.6	0	7.9	0	2.6	86.8

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

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

Figure 7: Comparison of results on BU dataset. Temp: temporal information is exploited. Range: 3D range information used. Note that we applied shape information alone for emotion estimation on the 2D images in the dataset. Our method does not use the range information provided with the dataset, nor the temporal characteristic of the expressions.

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. Fig. 9 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.

Other works have also addressed pose invariant recognition on the BU dataset, like [20, 21], however they are using the static version of the database, therefore direct comparison with our results is hard.

5 Discussion and Conclusions

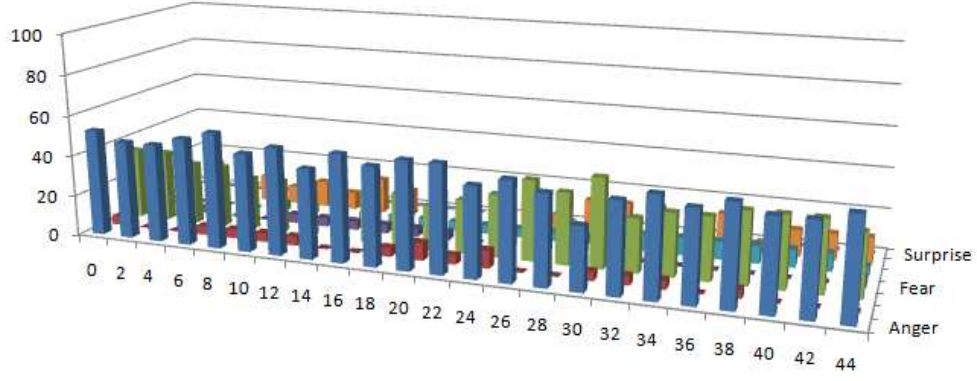
We are interested in situation analysis and human-computer interaction that call for emotion estimation for different head poses. Since simple line drawings are very expressive, we decided to study shape based emotion and action unit estimations. We conducted a preliminary study: we used the full shape 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, which was similar to human performance [14]. Recently, a highly robust CLM method have been developed in the literature [5] both for 2D and 3D models (see, e.g., [7] and the references therein). This CLM preserves more information than previously applied AAM methods [2], 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 [2] as well as 3D dynamic datasets [4].

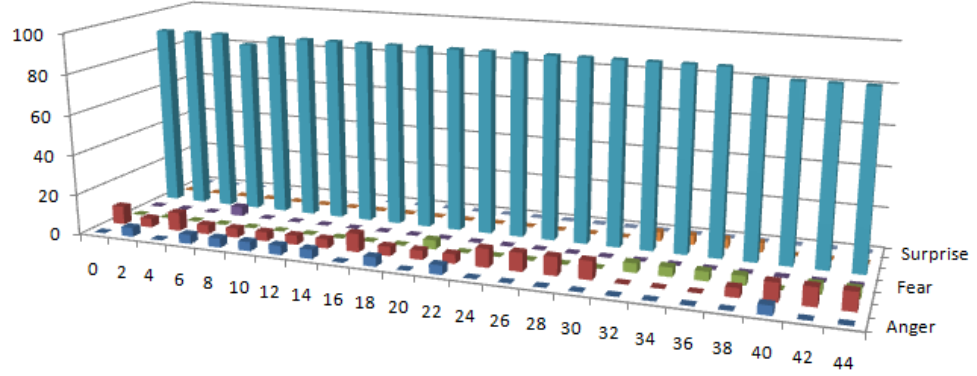
In our studies, we discovered that the neutral shape can be recovered from temporal sequences quickly and we replaced AU0 based normalization with the personal mean shape. For the expert annotated CK+ database results changed only slightly, but for CLM estimations this method gave rise to considerable 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 this mean shape is better if we average over the neutral shape *and* the shapes of the different facial expressions. This intriguing result makes our method useful 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 even that we used shape information alone. Our results surpass performances of the best available AAM methods [2] and CLM [8] that utilize shape *plus* texture and temporal information, respectively. It may be worth noting that FACS coders work from using local textural information [22].

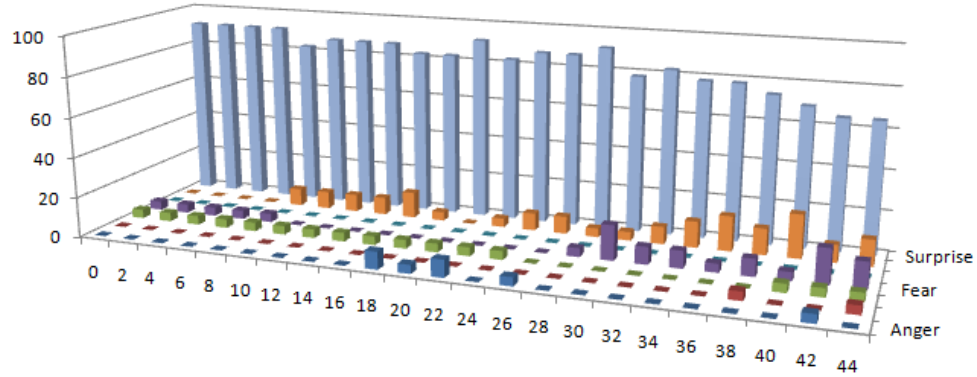
We studied the robustness of the CLM method for yaw rotations. We rendered rotated



(a)



(b)



(c)

Figure 8: Measurements on pose dependence using the BU-4DFE database. Top: classification as a function of rotational degree for angry faces. Middle: Same for happy faces. Bottom: Same for surprised faces.

3D faces using the BU-4DFE database [4] 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 shape information alone for emotion estimation. Further performance gains can be expected if textural information [7] and temporal information [1, 15] are included. Smoothing over time, e.g., by Hidden markov Models [1, 15] seems crucial for sensitive detection

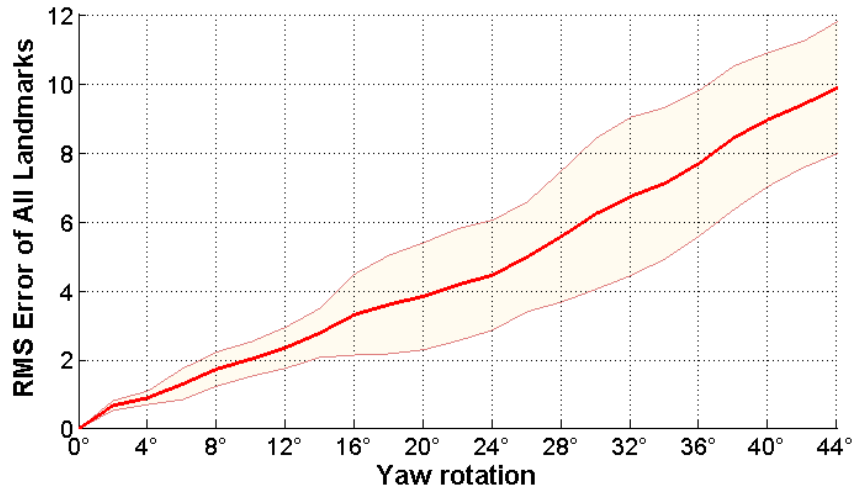


Figure 9: RMSE of reconstructed CLM estimations of the landmark positions of the 2D meshes in pixels as a function of rotation angle and using every landmark. The bold line denotes the mean. Shaded region around the mean shows the standard deviation. Distortion was compared to the initial frame (0 degrees of yaw rotation). 1 pixel error for all landmarks corresponds to 1 RMSE unit.

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 considerably differ from the natural ones [23]. Our evaluations have been performed on such posed emotions. Spontaneous emotions, especially micro-emotions [24] 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 multiplies probability estimations of experts [25]. 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 Zsolt Palotai from Sparsense Inc., USA and Dániel Takács and Zoltán Szabó from Eötvös Loránd University, Hungary for their valuable contributions to the implementation and research.

We are also 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 TAMOP 4.2.1./B-09/1/KMR-2010-0003 and KMOP 1.2-08/1-2008-002).

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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](https://doi.org/10.1016/j.patcog.2008.01.024).
- [6] 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.
- [7] 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 Conference on Automatic Face and Gesture Recognition (FG2011), 2011, pp. 57–64.
- [8] 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.
- [9] I. Matthews, S. Baker, Active appearance models revisited, International Journal of Computer Vision 60 (2) (2004) 135–164.
- [10] 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.
- [11] 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.
- [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] 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.
- [15] Y. Sun, L. Yin, Facial expression recognition based on 3d dynamic range model sequences, in: European Conference on Computer Vision, 2008, pp. 58–71.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] A. Lőrincz, G. Szirtes, B. Takács, I. Biederman, R. Vogels, Relating priming and repetition suppression, *International Journal of Neural Systems* 12 (2002) 187–202.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] P. Ekman, Darwin, Deception, and Facial Expression, *Annals of the New York Academy of Sciences* 1000 (10) (2006) 205–221.
- [25] G. E. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Computation* 14 (8) (2002) 1771–1800.