Constrained Local Model-based tracking for pose estimation and expression classification

F. Javier Sánchez Rois
Daniel González Jiménez
J. Luis Alba Castro

***Abstract*—Constrained Local Models (CLM) have gained much attention over the last years due to their good performance results for face alignment. In the context of face analysis, CLM systems exhibit great robustness in uncontrolled conditions and for person-independent tasks. In this work we address the effective implementation of a CLM as well as the design and test of CLM-based applications for automatic pose estimation and facial expression classification.**

***Index Terms*—Constrained Local Model (CLM), Action Unit (AU), expression recognition, pose estimation**

I. Introduction

FACIAL expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications [1]. Facial expression analysis has been an active research topic for behavioral scientists since the XIX century [2], whilst much progress has been made to build computer systems to help us understand and use this natural form of human communication [3]. The automatic analysis of facial expression involves the design of computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from visual information.
The usual approach to automatic facial expression analysis distinguishes between the tasks of feature extraction and expression classification. The extraction stage involves both adquiring the face image and obtaining the very discriminative features suitable for classification [4]. While face detection algorithms offer a coarse but valid approach to face feature extraction, different algorithms for precise landmark location and face alignment have arisen in the last decade.

Head movement and orientation are also key components in human communication. Deliberate head pose movement has as important meaning as a form of gesturing during direct conversations, as particular head movements can point out emotional states or reactions in a person. Moreover, head pose has a crucial role when it comes to analyze the visual focus of attention of an individual. Due to the relevance of the pose in human interaction, people display the ability to unconciously interpret head pose[1] from an early age. The ease and involuntariness this task is performed with suggest that the basis behind pose interpretation has to be relatively simple. However, the difficulty of this problem has challenged Computer Vision for decades, as it involves inferring the orientation of the head from digital imagery [5].

[1]Onwards, both movement and orientation of the head will be referred simply as *head pose*.

Point Distribution Models (PDM) have gained much attention in the last decades, because of their potential to model and describe deformable objects. Particularly linked with face analysis, the evolution of this family of models for representation through the years is motivated by an increasing interest for computer vision and more complex face analysis systems. One of the most influential concepts in face alignment can be found in the early work of Cootes [6], [7], who considers the combination of statistic linear models for both shape and appearance variation, leading to Active Appearance Models (AAM). During subsequent years much of the work done in the area of face alignment has been directly based on the concept of AAM, resulting in different approaches. Some of the most important developments in this respect include the use of linear regression [8], optimization by gradient descent [9] and the use of 3D models [10]. Moreover, the adjustment based on AAM has served to develop pose detection systems and gaze tracking interfaces [11]–[13].
An alternative approach to AAM-based face alignment is given by the use of local descriptors associated with the shape landmarks. Rather than employing a holistic model of the texture confined by the points, Constrained Local Models (CLM) consider the use of independent local descriptors *tied together* by the statistical model of shape variation. Although the idea of CLM was initially conceived by Cristinacce [14], [15] and the concept of a PDM using non-holistic descriptors have been also introduced by Cootes' Active Shape Models (ASM) [16], this approach has remained in background due to a greater development in AAM-based methods during the last decade. However, the improvement of the initial CLM formulation in recent years [17]–[20] has led to new systems which generally overtake the holistic approaches both in terms of computational cost and generalization.

In this work, we analyse and develop the practical implementation of a Constrained Local Model for face alignment, with special interest in the image descriptors employed to model each landmark's local appearance, as well as the optimal subset of points chosen among the possible candidates. We also train our CLM fitting scheme in the fashion of [21], where the facial alignment is based on local mean-shift displacements constrained by the global shape model. On the other hand, we extend the analysis of our face

alignment scheme to specific application environments. The usability of our CLM-based system for emotion detection and pose estimation is evaluated in specific image databases.

This paper is structured as follows: Section II explains in detail the fundamentals of CLM, focusing on its training and the implemented alignment algorithm. Section III describes the applications deployed from CLM, focusing on pose estimation and facial expression recognition. Section III outlines the evaluation process performed both for training and for CLM expression recognition applications and shows the achieved results. Finally, section IV presents the conclusions to the results and some future lines posed by this work.

II. CLM System Design

Constrained Local Models provide an efficient and accurate solution to facial feature localization by coupling a set of local detection experts and applying global constraints over the resulting individual detections. This family of models are particularly suited to non-rigid alignment and registration, as they combine shape and texture modelling. As holistic representation of texture is avoided, the CLM approach tends to outperform Active Appearance Models (AAM) [18], [21], specially due to its robustness to occlusion and changes in appearance.

A. Learning Constrained Local Models

A 2D shape is described by a set of N landmark points, represented as vector

$$\mathbf{s} = \{x_1, y_1, \ldots, x_N, y_N\}$$

To model 2D shapes, most of PDM-based algorithms consider the use of a linear generative shape model. This involves the representation of a given shape vector as the combination of a set of shape modes. Rigid transformations are often considered in the form of 2D similarity (or affine) variations, which add to the model the notions of scaling, rotation and translation. With this considerations, a shape $\mathbf{s}$ is then given by

$$\mathbf{s} = \alpha R(\bar{\mathbf{s}} + V\mathbf{q}) + \mathbf{t} \tag{1}$$

where α represents a scaling factor, R is a rotation matrix and $\mathbf{t}$ represents 2D translation. Vector $(\bar{\mathbf{s}} + V\mathbf{q})$ is the *nonrigid* component of $\mathbf{s}$, parametrized by $\mathbf{q}$ in the space spanned by the rows in V (shape modes) and the mean shape $\bar{\mathbf{s}}$.

The usual way of constructing the shape space (i.e. obtaining $\{\bar{\mathbf{s}}, V\}$) involves first aligning a set of training vectors to a common coordinate frame (usually by means of Generalized Procrustes Analysis [22]) and then applying Principal Component Analysis (PCA) to obtain the shape modes representing the non rigid shape variation. Given a subspace spanned by $\{V, \bar{\mathbf{s}}\}$, each shape is uniquely characterized by the set of parameters $\mathbf{p} = \{s, R, t, \mathbf{q}\}$, which is precisely the essence of Point Distributed Models.

While the statistical modelling of shape variation adopted in this scheme is common to many of PDM-based solutions, the form of texture representation employed in the CLM differentiates it from other methodologies. While the early Constrained Local Model derived the existing AAM conception of a PCA-based texture model, the subsequent proposed CLM variations have included discriminative local classifiers to characterize each landmark point neighbour region in the image. Even though generative models are desirable for regression and gradient descent-based solutions [14], [23] local classifiers allow exhaustive local search-driven localization [17] and provide a greater robustness to oclussions and changes in illumination conditions.

In particular, we opted for training an SVM classifier [24] to obtain individual scores of alignment/misaligment for every landmark in the point subset, employing positive and negative samples of alignment in each case. The positive samples were taken from image patches centered in the fiduciary points of the training set of images, while the negative ones were obtained by shifting the positive samples away from the true locations. Given the resulting support vectors $\{\gamma_i\}_{i=1}^{N_{SV}}$ and bias β, the alignment score of a new given patch centered at $\mathbf{x}$ is defined as

$$\hat{f}(\mathbf{x}') = \sum_{i=1}^{N_{SV}} \alpha_i \varphi(T(\mathbf{x}; \mathcal{I}), \gamma_i) + \beta \tag{2}$$

where $T(\mathbf{x}, \mathcal{I})$ is an image patch from image $\mathcal{I}$ centered at the point $\mathbf{x}$ and φ denotes the chosen kernel operator. If we consider a linear SVM kernel and evaluate the alignment score for all neighbor locations from a reference point $\mathbf{x}$, i.e. for all $\mathbf{x}' = (\mathbf{x} + \Delta\mathbf{x}) \in \kappa_{\mathbf{x}}$, the score obtained for each displacement is given by

$$\begin{aligned} \hat{f}(\mathbf{x}') &= \sum_{i=1}^{N_{SV}} T(\mathbf{x}; \mathcal{I}) \alpha_i \gamma_i + \beta \\ &= T(\mathbf{x}; \mathcal{I}) \left(\sum_{i=1}^{N_{SV}} \alpha_i \gamma_i) \right) + \beta \end{aligned} \tag{3}$$

Notice that the advantage of this particular scheme is that the template $\left(\sum_{i=1}^{N_{SV}} \alpha_i \gamma_i)\right)$ can be precalculated and thus obtaining the score map for each point in a neighbor region $\kappa_{\mathbf{x}}$ will only involve a dot product between each patch and the template. For this scheme to be effective, the misalignment error is considered to depend only in the patch displacement, so that scale changes must be first removed from the image.

B. Constrained Local Model Fitting

The Constrained Local Model fitting has been posed as the search of a set of parameters $\mathbf{p}$ that minimize the *misalignment error* function [21]:

$$\mathcal{E}(\mathbf{p}) = \sum_{i=0}^{n-1} \mathcal{D}(\mathbf{x}_i; \mathcal{I}) + \mathcal{R}(\mathbf{p}) \tag{4}$$

where $\{\mathbf{x}_i = (x_i, y_i)\}_{i=0}^{n-1}$ denotes a set of landmark points and $\mathcal{D}(\mathbf{x}_i; \mathcal{I})$ is the misalignment of the i-th landmark for a given image $\mathcal{I}$. The term $\mathcal{R}(\mathbf{p})$ is coined *regularization*, as it constraints the parameter estimates according to a joint (global) model of variation. Although this expression (optimization..),

From a probabilistic interpretation as the one adopted by several authors [25]–[27], minimizing equation (4) can be also viewed as maximizing the likelihood of the global parameters which result in the set of landmarks aligned with their true locations. If we assume conditional independence between each point alignment, this can be expressed as:

$$p(\mathbf{p}|\{\text{aligned}_i = 1\}_{i=0}^{n-1}, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=0}^{n-1} p(\text{aligned}_i = 1|\mathbf{x}_i, \mathcal{I}) \tag{5}$$

Employing the log-likelihood form of this equation we finally rewrite equation (4) as

$$\mathcal{E}(\mathbf{p}) = -\sum_{i=0}^{n-1} \log(p(\text{aligned}_i = 1|\mathbf{x}_i, \mathcal{I})) - \log\{p(\mathbf{p})\} \tag{6}$$

Notice that the regularization term is now defined as $\mathcal{R}(\mathbf{p}) = -\log(p(\mathbf{p}))$ and the misaligment of each landmark as $\mathcal{D}(\mathbf{x}_i; \mathcal{I}) = -\log(p(\text{aligned}_i = 1|\mathbf{x}_i, \mathcal{I}))$. We also consider modelling the alignment likelihood with a sigmoid function as proposed in [18]:

$$p(\text{aligned}_i = 1|\mathbf{x}, \mathcal{I}) = \frac{1}{1 + e^{\alpha \hat{f}(\mathbf{x}) + \beta}} \tag{7}$$

where $\hat{f}(\mathbf{x})$ is the output of the alignment classifier for an image patch centered at $\mathbf{x}$ (Eq. 2). Fitting a logistic regression [24] to the output of the SVM in this manner ensures that the alignment likelihood follows approximately a probability distribution.

Even though (4) posed the CLM fitting as an optimization problem, general-purpose strategies and off-the-shelf optimization methods are rarely employed in this scenario. Most authors are inclined to use domain specific algorithms, being the most common trend the one which first performs exhaustive local search on the candidate locations for each point of the CLM, to then restrict the results by using the regularization term. Under this approach, some authors have proposed different approximations of the response map given by the evaluation of (7) in the candidate positions for each landmark. Whilst the quadratic aproximation [18], [26] enhances the performance of the CLM compared to using directly the noisy response maps [17], this approach often fails when the response map is strongly multimodal [20]. The Gaussian Mixture Model (GMM) solution proposed in [19] achieves better fitting results, but it becomes more computationally expensive and it requires the estimation of the GMM parameters from the response maps. An efficient KDE-based approximation has been proposed more recently [20]. It does not only achieve more accurate results than the quadratic and GMM solutions, but its computational cost is suitable for real time video applications. We have chosen this particular CLM fitting scheme, coined Subspace Constrained Mean Shift (SCMS)[2], for our system .

C. Subspace Constrained Mean Shift

As we have presented a PCA-based solution for the shape modelling of our CLM, we can expect an approximation error between the true location of the landmark points, $\{y_i\}_{i=0}^{n-1}$, and their model reconstruction, $\{x_i\}_{i=0}^{n-1}$, so that

$$\mathbf{y}_i = \mathbf{x}_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(\varepsilon_i; \mathbf{0}, \rho\mathbf{I}) \tag{8}$$

The error variance ρ will be determined by the truncation level chosen in the shape model training [28]. If we also introduce the finite set of candidate locations for each landmark, $\{\kappa_i\}_{i=0}^{n-1}$, we can marginalize out the true locations from the alignment likelihood (Eq. 7):

$$\begin{aligned} p(\text{aligned}_i = 1|\mathbf{x}, \mathcal{I}) &= \sum_{\mathbf{y}_i \in \kappa_i} p(\text{aligned}_i = 1|\mathbf{y}_i, \mathcal{I}) p(\mathbf{y}_i|\mathbf{x}_i) \\ &= \sum_{\mathbf{y}_i \in \kappa_i} p(\text{aligned}_i = 1|\mathbf{y}_i, \mathcal{I}) \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I}) \end{aligned} \tag{9}$$

Also, if we exploit the fact that the non rigid parameters obtained from a PCA model are Gaussian-distributed, we can obtain an informative prior over our CLM parameters [21]

$$p(\mathbf{p}) \propto \mathcal{N}(\mathbf{q}; \mathbf{0}, \Lambda) \tag{10}$$

where Λ is a diagonal matrix sorting the corresponding eigenvalues for each mode of the shape subspace. Assuming this prior we can formulate a Maximum a-*posteriori* (MAP) estimation of the model parameters ensuring the correct alignment. Substituting (9) in (5) and denoting $\pi_{\mathbf{y}_i} = p(\text{aligned}_i = 1|\mathbf{y}_i, \mathcal{I})$ we get:

$$p(\mathbf{p}|\{\text{aligned}_i = 1\}_{i=0}^{n-1}, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=0}^{n-1} \sum_{\mathbf{y}_i \in \kappa_i} \pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I}) \tag{11}$$

This expression can be maximized by means of the EM algorithm, as proposed in [21]. The solution for the parameter update results as

$$\Delta\mathbf{p} = -(\rho\Lambda^{-1} + \mathbf{J}^T\mathbf{J})^{-1}(\rho\Lambda^{-1}\mathbf{p} - \mathbf{J}^T\mathbf{v}) \tag{12}$$

where $\mathbf{J}$ is the PDM Jacobian[3] and $\mathbf{v} = [\mathbf{v}_0; \ldots; \mathbf{v}_{n-1}]$ is formed by the mean shift vectors obtained for each landmark,

$$\mathbf{v}_i = \left(\sum_{\mathbf{y}_i \in \kappa_i} \frac{\pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I})}{\sum_{\mathbf{z}_i \in \kappa_i} \pi_{\mathbf{z}_i} \mathcal{N}(\mathbf{z}_i; \mathbf{x}_i, \rho\mathbf{I})} \right) - \mathbf{x}_i^0 \tag{13}$$

The complete Subspace Constrained Mean Shift algorithm is summarised in Algorithm 1.

[2]Also named Regularised Landmark Mean Shift (RLMS) [21].

[3]We will consider the Jacobian as $\mathbf{J} = \mathbf{V}$ for non rigid movement, while for scale/rotation/translation changes the Jacobian is constructed from an auxiliary subspace, as in [9].

Algorithm 1 Subspace Constrained Mean Shift [21]

Require: current shape parameters $\mathbf{p}$ and image $\mathcal{I}$
1: Calculate patch response maps (Eq. 7)
2: while (not converged) **do**
  3: Construct shape from current parameters (Eq. 1)
  4: Calculate mean-shift displacements (Eq. 13)
  5: Compute parameter update (Eq. 12) and $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$
6: end while
7: return $\mathbf{p}$

Algorithm 2 CLM - driven pose estimation

Require: pre-built 3D rigid shape model $\mathbf{s}_{3D}$ and associated 2D CLM, image $\mathcal{I}$
1: Apply CLM fitting (Algorithm 1) to obtain parameters $\mathbf{p}$.
2: Construct 2D shape $\mathbf{s}$ from parameters $\mathbf{p}$ (Eq. 1)
3: Run POSIT algorithm [29] with current shape $\mathbf{s}$ and reference $\mathbf{s}_{3D}$ to obtain pose parameters: scale, translation and 3D pose angles.
4: **return** pose parameters

III. Applications design

A. CLM-based Pose Estimation

As it has been introduced, head pose estimation is still a challenge for Computer Vision, so that many different approaches exist in the literature. For instance, head pose estimation has been addressed as a template matching problem, as a nonlinear regression issue or as a feature tracking scheme [5]. In particular, PDM-based approaches to pose estimation, as by using Active Appearance Models [13] [10], are specially interesting as they exploit the capabilities of such complex models and open a different perspective of the problem.

From the monocular point of view, estimating the head (or any other *object*) pose consists of recovering the camera position and relative orientation to a known set of 3D points. Numerous algorithms exist which, given a set of 3D points and their corresponding 2D projections in a image, perform pose estimation with considerable accuracy. Two examples are Pose from Orthography and Scaling with ITerations (POSIT) [29] and Perspective-n-Point (PnP) [30]. In principle, any of those solutions to the camera pose estimation can be used efficiently over plain face imagery, even when they all involve the error derived from assuming head to be a rigid structure.

Pose from Orthography and Scaling with Iterations is a fast and accurate iterative algorithm by DeMenthon and Davis [29] for finding the 6 degrees-of-freedom pose of a 3D model with respect to a camera, given a set of 2D image points and the 3D object correspondences. In a similar way as proposed in [31] with AAM, we employ our CLM-based fitting to obtain the 2D positions of the landmark points in each new image. Then, we employ the POSIT algorithm to obtain the pose angles and scale of the aligned 2D shape with respect to a previously built rigid 3D face model. Our pose estimation scheme is summarized in Algorithm 2. Further details on the POSIT algorithm can be found in [29], while our implementation details and experiments can be found in Section III.

B. CLM-based Expression Recognition

Advances in affective computing in the last years include an active research in face expression recognition, a topic which has motivated different approaches to this issue from Computer Vision. The creation of a periodical challenge on the matter has contributed to the appearance of new proposed solutions and has established a unified dataset and test protocol, easing the evaluation of the different methodologies in the state of the art.

Among the different methods presented to the I Facial Expression Recognition and Analysis (FERA) Challenge [3], some of them [32] employed a CLM-based face alignment schemes to detect a series of landmark points in the images and perform feature extraction over the image texture bounded by the CLM points. Although this approach may seem one of the most complete schemes, the performance this PDM-guided scheme were far from attaining the results achieved by those which employed rough face detection and avoided complex models. Despite of this fact, we choose to employ the CLM fitting results to feed our expression recognition given its representational power and good alignment results. However, we opt for different features to perform classification, as the ones presented in [32] did not exhibit significant advances which justify the computational cost required for their extraction.

Most of the existing work regarding emotion recognition through face analysis is grounded in the research work of Ekman [33]. Its studies have defined the six basic emotions [34] adressed by many automatic systems for classification [3]. Moreover, to establish the basis for an appropriate categorization of complex emotional states, Ekman also proposed the Facial Action Coding Sistem (FACS) [1]. This system is based on detecting 32 Action Units (AU), muscle movements that occur on the face, which combined give rise to different emotions. Unlike the early automated systems for facial expression detection, new proposed methodologies aim to detect the occurrence of AUs in images, thus incorporating this new categorization and conception of *emotion coding*.

With the aim of detecting basic expressions and specially Action Units in face images, we design and implement a recognition pipeline guided by CLM face aligment and SVM for classification of the extracted data. Among the possible features to train the classifiers, we choose to employ (a) the nonrigid parameters $\mathbf{q}$ yielded by the CLM after the alignment process or (b) Local Binary Pattern (LBP) [35] vectors from cropped images, with a normalization driven by the points that result from the alignment. The first features have been chosen by their extraction simplicity (because they require no further processing after the CLM alignment) and because they code the face deformations by nature. The point-guided

shape normalization and LBP feature extraction solution has been chosen for the idea of aligning the input face images to a common frame without defining too complex warps. Both feature types are employed for both AU detection and expression classification, as outlined in Algorithm 3.

Algorithm 3 CLM - driven AU detection / expression classification

Require pre-built CLM, pre-trained SVM AU/expression classifiers, image $\mathcal{I}$
1: Apply CLM fitting (Algorithm 1).
2: Extract feature vector $\mathcal{F}$,
(a) Nonrigid shape parameters $\mathbf{q}$
or
(b) LBP from image cropped from $\mathcal{I}$ and point subset $\mathbf{s}' \subset \mathbf{s}$
3: Run the SVM classifiers with the extracted feature vector $\mathcal{F}$
return obtained classification scores

Notice that the only difference between the expression and AU classification schemes is the chosen SVM in the classification step. Notice also that a *subset* $\mathbf{s}'$ of points is chosen instead of the full point vector $\mathbf{s}$ for warping the image. We do so to preserve the maximum expression information in the resulting normalized images. The main drawback of not employing the totality of the CLM points (as in [32]) is that other information aside from the expression remains in the obtained images. However, since similar methods (as eye position-guided warping) have been widely employed for face recognition [36] and expression detection [37], we think of the CLM as a way to achieve similar or better classification results than these systems. Further discussion about the subset of points chosen for the image warp can be found in Section III.

IV. Experiments

A. CLM Training

Even though the main aspects regarding the training process of SCMS Constrained Local Models have been outlined in Section II, there are some issues in its effective implementation which need a deeper analysis. The first one addresses the question of what landmark points to choose to be part of the shape model, as it results in a better fitting performance of the model while keeping most of the representational power. The second issue refers to what image descriptors use when training local experts for landmark detection. This work addresses both questions and evaluate the different options to achieve the best fitting results in terms of alignment accuracy.

To train the shape model (as explained in II-A) we have used the Multi-PIE face database [38]. This dataset contains more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. In particular, a subset of 5300 images were chosen for shape model training. While the images were annotated with 68 landmarks, some of these points show to be less informative than others, and can even lead to bad fitting results due to their lack of robustness to pose and illumination changes. As it has been discussed in [39], the points located in the jaw and face contour are prone to induce errors, while the points located in the eyes, eyebrows, nose and mouth are more distinctive and spatially informative. Thus, we focus our analysis in the performance of those point configurations which include (1) all of these points, (2) some of them or (3) none of them at all.

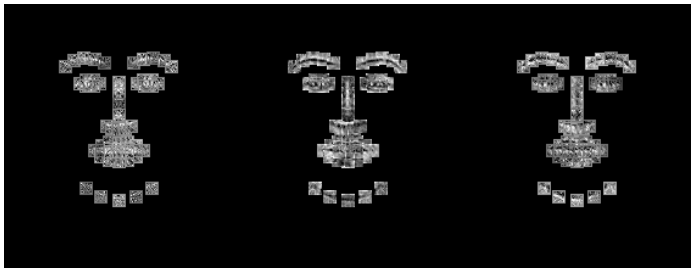

Fig. 1. Trained local experts using pixel, lbp and gradient features. All the employed image descriptors maintain the spatial information.

As explained in Section II, the CLM fitting algorithms generally perform local detections before introduce global constraints to obtain the PDM parameter update. The training of the classifiers associated to each landmark is then crucial to achieve good alignment results. For that reason we extend our CLM training analysis to choose the best image descriptors available. Notice that, although the idea of employing different image descriptors was not introduced in the previous section, we only consider those that preserve the spatial information. This makes possible that the alignment error depend only on the spatial displacement of the estimates w.r.t. the true point locations. Hence, the notation in (3) and (11). In particular, we have chosen to test the training of local experts based in LBP, image gradients, sobel operator magnitudes and the raw image patches (i.e. greyscale pixels). Also, we considered the use of combined detectors to form the alignment score (Eq. 7) as the product of the single feature detectors. Figure 1 show patch experts trained with different local descriptors.

In our experiments, we have defined different combinations of point configurations and image descriptors. The PCA-based shape model has been trained with a 95% fraction of the total energy and employing the full annotated set available in the MultiPIE database. On the other hand, a patch size of 11x11 pixels has been considered for the local detectors. Each local expert has been trained using the patches centered at the true positions as positive samples and taking the ones centered at the neighbor pixels as negatives. Given the amount of negative samples obtained in this manner, only the 25% of the negative samples with the lower sum of squared differences (SSD) to the positive sample has been considered. In the case of the local expert training, a 5-fold cross validation

$$\text{RMSE} \equiv \frac{\sum_{i=0}^{n-1} \sqrt{\|\mathbf{s}^{(i)} - \mathbf{s}_{\text{gt}}^{(i)}\|^2}}{n} \quad (14)$$

Fig. 2. Normalized RMSE error, calculated as the average error obtained for each landmark $\{\mathbf{s}^{(i)}\}_{i=0}^{n-1}$ w.r.t. the ground truth points $\{\mathbf{s}_{\text{gt}}^{(i)}\}_{i=0}^{n-1}$ annotated for each image.

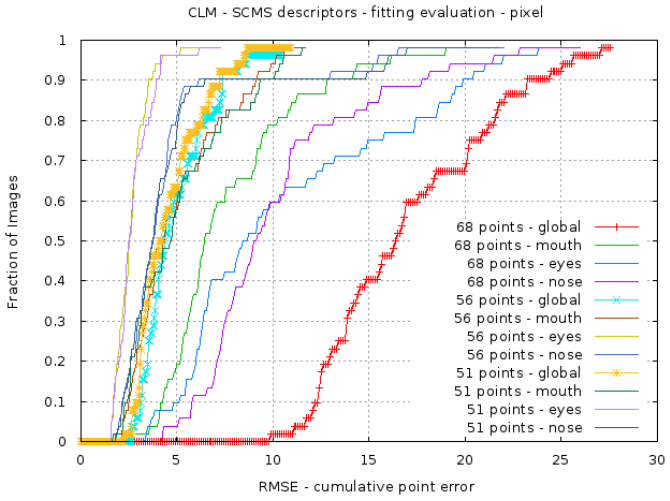

Fig. 3. Cumulative per-point error (in pixels) obtained for different point configurations. Notice that the global error obtained by the point configurations excluding points from the face contour (56 and 51-point) is lower compared to the all-68-point model fitting error. This is true also for the partial error measures in the different face regions.

scheme has been considered to obtain the fitting results over the annotated image set of the MultiPIE database.

To measure the fitting performance of each configuration, we calculate the normalized error proposed in [15], as it allows to compare the performance of models with different number of points (See Eq. 14). Aside from calculating the global error measure with respect to the ground truth annotations, we have also calculated the error corresponding to certain subsets of points: mouth, nose and eyes. Such partial error measures give an idea of how affects the inclusion of the jaw points to the CLM performance regarding other locations. Figures 3 and 4 show the cumulative error of the different proposed solutions.

In the light of the achieved results, we can draw the following conclusions:

- Including the face contour and jaw points into the model hinders the global performance of the CLM. As in our initial guess, it not only increases the error due to the misalignment of the contour points themselves, but it worsen the alignment in the rest of points.

- Regarding to the local descriptors used to train the experts in the CLM, the best individual results are achieved by the pixel descriptor, followed by the LBP and gradient. On the other hand, it is clear that combining descriptors

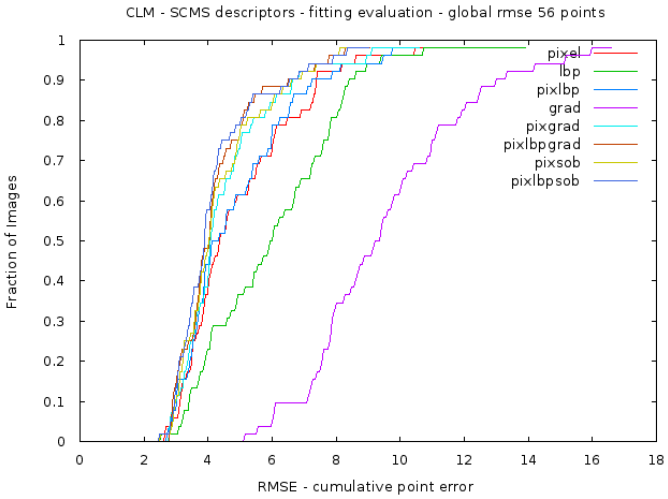

Fig. 4. Cumulative per-point error (in pixels) obtained for different image descriptors and their combinations for local detections.

enhances the overall results, specially in the case of the combinations (pixel, LBP, gradient), (pixel, sobel) and (pixel, LBP, sobel). Figure 4 shows the results obtained for the 56-point model, although similar differences were found in the other point configurations.

B. Pose Estimation

In order to validate the usability of our CLM-driven pose estimation system, it has been tested with pose-labeled images. In particular, we have chosen the HPEG dataset [40], which consists in a collection of videos with the pitch and yaw angles annotated for each frame. As a reference model, we built a 3D reconstruction of an average face by employing Structure-From-Motion techniques [41] over the MultiPIE images. The tests consisted on fitting the CLM to the input images and measure the pitch and yaw angles obtained from POSIT. The resulting average errors were $8.13°$ (pitch) and $6.12°$ (yaw). Figure 4 shows a frame of one of the HPEG sequences during evaluation.

C. Expression Recognition

To evaluate the performance of our system in the tasks of AU detection and basic emotion recognition, we initially chose the GEMEP dataset adopted in the FERA facial recognition challenge [3]. This database is part of the GEMEP corpus, which consists of 7000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 actors who were trained by a professional director. We initially decided to use the training dataset (consisting on 87 annotated videos) to both emotion and AU detection. However, the emotion labels in this dataset are assigned to each entire video and not to the frames showing emotion *peaks*. As this ambiguity would difficult our evaluation of the proposed system, the emotion recognition tests were carried out on the Multi-PIE database

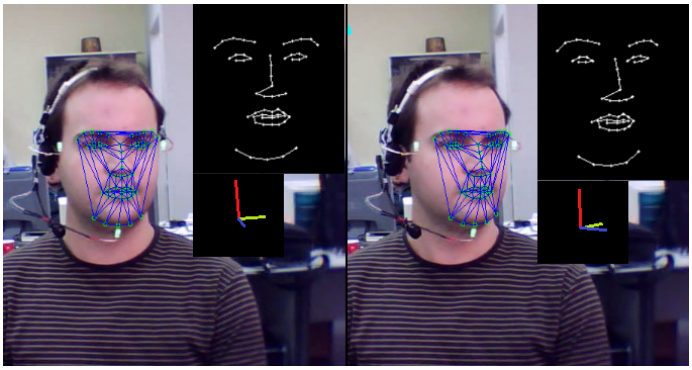

Fig. 5. Sample images from the pose estimation experiments in the HPEG dataset [40]. The white points in the upper right corner of each image show the 3D reference model, while the colour axis show the estimated pose.

Fig. 6. Cropped images for AU and emotion detection. The images in the first row were cropped using the eye positions obtained from the aligned CLM, while the second row shows images cropped with a subset of 5 points.

instead, where images show different annotated expressions (*neutral*, *smile*, *disgust*, *surprise*).

In the case of the shape-guided classification, 27 shape parameters were included in each feature vector. On the other hand, for the point-guided image warping, two different normalization schemes were considered to perform the image alignment: (1) employing the averages of the eye points in the CLM and (2) using a 5-point subset which included the eye outer corners, the nose tip and the mouth corners. See figure 6 for more details. The AU detection tests in the GEMEP database were performed as follows: the videos were divided by identity and each time, six users were chosen to train the SVM classifier and the remaining user to test the results. In turn, for each training one of the six users was *left out* for validation in each test. Different values of the SVM C and bias parameters were tested and PCA was also employed for dimensionality reduction of the feature vectors. As suggested in [42] we chose the results maximizing the F1-score of the detection results.

Table I shows the results of the AU detection tests, after averaging the F1-scores for all the users and seeking for the parameters with the best averaged score. We have also included the results obtained by the baseline system in the FERA challenge [42], as well as the results obtained by systematically (*naive*) detecting the AUs in all the frames. For the expression detection tests in the MultiPIE dataset, a 4 fold cross-validation scheme was carried to test different C and bias values. Table II shows the confusion matrices attained by our emotion detection systems.

AU index	baseline [42]	(A)	(B)	(C)	naive [42]
1	**0.555**	0.4786	0.4161	0.5154	0.4494
2	**0.571**	0.4515	0.4922	0.5405	0.4562
4	0.331	**0.3988**	0.3893	0.3506	0.4016
6	0.702	0.6678	0.4882	**0.7633**	0.4996
7	**0.588**	0.5374	0.5833	0.5542	0.5739
10	**0.624**	0.5901	0.5449	0.5986	0.5472
12	0.719	0.5262	0.6743	**0.7513**	0.6743
15	**0.354**	0.3071	0.3117	0.3504	0.3271
17	**0.282**	0.2174	0.2796	0.2641	0.2723
18	0.253	0.1285	0.1178	**0.2768**	0.1421
25	0.285	0.2452	0.2827	**0.3075**	0.2919
26	0.159	**0.2192**	0.1639	0.1942	0.1885
Av.	0.452	0.3973	0.3953	0.4555	0.4020

TABLE I

ACTION UNIT DETECTION RESULTS (F1-SCORE, GEMEP-FERA DATABASE). THREE TYPES OF FEATURE WERE EVALUATED: **(A)** SHAPE PARAMETERS, **(B)** LBP FROM THE IMAGES WARPED WITH THE EYE LOCATIONS AND **(C)** LBP FROM THE IMAGES WARPED WITH A 5-POINT SUBSET

		neutral	happy	disgust	surprise
(A)	**neutral**	28.45	33.83	**36.28**	1.44
	happy	2.35	**86.23**	5.77	5.65
	disgust	4.34	12.22	**58.87**	24.57
	surprise	2.29	34.73	27.58	**35.4**
(B)	**neutral**	**41.83**	34.61	21.30	2.26
	happy	1.44	**85.32**	0.26	12.98
	disgust	12.14	25.56	**60.62**	1.68
	surprise	2.07	17.86	3.47	**76.6**
(C)	**neutral**	**46.35**	18.15	34.57	0.93
	happy	3.41	**91.47**	0.81	4.31
	disgust	21.05	13.23	**54.01**	11.71
	surprise	7.31	22.34	3.56	**66.79**

TABLE II

EXPRESSION RECOGNITION RESULTS (MULTI-PIE DATABASE). THREE TYPES OF FEATURE WERE EVALUATED: **(A)** SHAPE PARAMETERS, **(B)** LBP FROM THE IMAGES WARPED WITH THE EYE LOCATIONS AND **(C)** LBP FROM THE IMAGES WARPED WITH A 5-POINT SUBSET

Given the results obtained for the detection of AUs, only the third trained systems reaches the baseline level established in the FERA challenge. Although the achieved scores may seem poor for a detection system, the Action Unit detection problem is complex and the results obtained by the different participants in the FERA challenge are not far from ours. Regarding the emotion recognition problem, the classification results obtained by our systems proved the validity of CLM-based scheme. However, future tests should be performed over different datasets which allow us to compare our scheme with other state of the art systems.

V. Conclusions

In this work we have analyzed the basic concepts of Constrained Local Models, as well as the details regarding their construction from training images. We have implemented a CLM-based face alignment algorithm, choosing the parameters allowing to achieve the best fitting results. We have also designed and tested various CLM-based systems for head pose estimation and expression detection. The achieved results show that our proposed method for pose estimation obtains a good approximation in video images. On the other hand, our tests for Action Unit detection and expression recognition show the validity of our approach for these tasks, even though further analysis must be carried to improve our results.

Some possible improvements to our work include employing 3D models as in [21], which would enhance our pose estimation results. Also, a three-dimensional shape model would allow to obtain more precision in terms of non-rigid face movement, which could improve the expression recognition performance as well. Other possible improvements include combining the classification scores obtained by different descriptors and testing new point subsets.

References

[1] J. F. Cohn and P. Ekman, "Measuring facial action," *The new handbook of methods in nonverbal behavior research*, pp. 9–64, 2005.

[2] C. Darwin, "The expression of emotions in animals and man," *Nueva York: Appleton. Traducción*, 1872.

[3] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 921–926.

[4] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[5] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607–626, 2009.

[6] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 300–305.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 681–685, 2001.

[8] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate regression procedures for active appearance models." in *BMVC*, 2011, pp. 1–11.

[9] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[10] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2d+ 3d active appearance models," in *CVPR (2)*, 2004, pp. 535–542.

[11] P. Mittrapiyanuruk, G. N. DeSouza, and A. C. Kak, "Calculating the 3d-pose of rigid-objects using active appearance models," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 5. IEEE, 2004, pp. 5147–5152.

[12] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models."

[13] X. Liu, N. Krahnstoever, T. Yu, and P. Tu, "What are customers looking at?" in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007, pp. 405–410.

[14] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[15] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *BMVC*, vol. 17, 2006, pp. 929–938.

[16] T. F. Cootes and C. J. Taylor, "Active shape modelssmart snakes," in *BMVC92*. Springer, 1992, pp. 266–275.

[17] Y. Wang, S. Lucey, and J. Cohn, "Non-rigid object alignment with a mismatch template based on exhaustive local search," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[18] Y. Wang, S. Lucey, and J. F. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[19] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 413–426.

[20] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1034–1041.

[21] ——, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[22] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[23] X. S. Zhou, A. Gupta, and D. Comaniciu, "An information fusion framework for robust shape tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 1, pp. 115–129, 2005.

[24] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[25] L. Liang, F. Wen, Y.-Q. Xu, X. Tang, and H.-Y. Shum, "Accurate face alignment using shape constrained markov network," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 1313–1319.

[26] U. Paquet, "Convexity and bayesian constrained local models," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1193–1199.

[27] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Let the shape speak-discriminative face alignment using conjugate priors." in *BMVC*, 2012, pp. 1–12.

[28] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 696–710, 1997.

[29] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," in *Computer VisionECCV'92*. Springer, 1992, pp. 335–343.

[30] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[31] P. Martins and J. Batista, "Monocular head pose estimation," in *Image Analysis and Recognition*. Springer, 2008, pp. 357–368.

[32] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 915–920.

[33] P. Ekman, "Facial expression and emotion." *American Psychologist*, vol. 48, no. 4, p. 384, 1993.

[34] ——, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[35] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[36] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[37] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 860–865.

[38] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[39] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable face fitting with soft correspondence constraints," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–8.

[40] S. Asteriadis, D. Soufleros, K. Karpouzis, and S. Kollias, "A natural head pose and eye gaze dataset," in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. ACM, 2009, p. 1.

[41] J. Xiao, J.-x. Chai, and T. Kanade, "A closed-form solution to non-rigid shape and motion recovery," in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 573–587.

[42] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 966–979, 2012.