LIKELIHOOD-ENHANCED BAYESIAN CONSTRAINED LOCAL MODELS

Pedro Martins, Rui Caseiro, João F. Henriques, Jorge Batista

Institute of Systems and Robotics, University of Coimbra, Portugal

ABSTRACT

This paper addresses to the problem of aligning images in unseen faces. The Constrained Local Models (CLM) are popular methods that combine a set of local landmark detectors whose locations are constrained to lie in a subspace spanned by a linear shape model. The CLM fitting is usually based on a two step approach: locally search, using the detectors, producing response maps (likelihood) followed by a global optimization strategy that jointly maximize all detections at once. In this paper, we mainly focus on the first stage: improving the detectors reliability. Usually the local landmarks detectors are far from perfect. Most often are designed to be fast, having a small support region and are learnt from limited data. As consequence, they will suffer from detection ambiguities. Here we propose to improve the detectors performance by considering multiple detection per landmark. In particular, we propose a joint learning of the detectors by clustering of their training data. Afterwards, the multiple likelihoods are combined using a nonlinear fusion approach. The performance evaluation shows that our (extended) approach further increases the fitting performance of the CLM formulation, when compared with recent state-of-the-art methods.

Index Terms— Face Alignment, Constrained Local Models.

1. INTRODUCTION

Facial alignment, also known as facial registration, is a fundamental problem in computer vision with applications in several tasks such as tracking, recognition, pose estimation, video compression, etc. A widely used approach consists on seeking the parameters of a linear model (a Point Distribution Model - PDM) that best represents the face in a target image. Traditionally, these deformable fitting methods can be divided in two major categories: the generative (holistic) and/or discriminative (patch-based) approaches. The generative paradigm describe the appearance of a face using all its image pixels, typically using an eigen-based representation. The Active Appearance Models (AAM) [1, 2, 3, 4] are probably the most popular generative method, achieving an impressive registration quality. However, this representation generalizes poorly beyond unseen data, when the target individuals are not included in the training dataset.

Presently, there has been a growing interest on discriminative-based methods, such as the Constrained Local Models (CLM) [5, 6, 7, 8], as it avoids several of the drawbacks of generative methods by improving the generic face representation. In this paradigm, both appearance and shape are combined by constraining a set of local feature detectors to lie within the subspace spanned by the PDM. In general, all instantiations of CLM are composed by a two phase fitting strategy. The first phase generates a response map for each landmark (a likelihood map) using the local detectors. The second phase consists in a global optimization strategy that estimates the PDM parameters that jointly maximizes all the response maps at once. Most optimization strategies aim to approximate the responses maps by simple parametric forms (Weighted Peak Responses [5], Gaussians Responses [8, 9], Mixture of Gaussians [10]) or non-parametrically by a Kernel Density Estimator (KDE) [11, 12]. Recently, a new paradigm emerged aiming to solve the global optimization [9, 13, 14]. This new strategy suggests to formulate the global alignment as a Bayesian inference problem. The patch responses are embedded into a Bayesian framework, where the posterior distribution is inferred in a maximum a posteriori sense (MAP) [9, 13, 14].

The main focus of this paper relates to the first phase of CLM fitting, i.e. the likelihood generation using local landmark detectors. Traditionally, these detectors are design to be simple and to operate as fast as possible. Most times they are build from limited data and often have small local support, therefore resulting in ambiguous detection, i.e. they are unable to discriminate correct from incorrect locations. We aim to overcome this limitation by including multiple sets of local detectors, per landmark, further improving the reliability of the detector, and consequently, in the overall alignment. This belief was also pursuit in [15] and [16]. The main difference with respect to our work is that we seamlessly integrate multiple detection within the CLM formulation without a specially designed global optimization and with a very simple fusion technique of response maps. In particular, we propose to learn the multiple detectors by clustering the training data patches examples, obtaining highly specialized filters in a small range of appearance variation. Afterwards, the multiple responses are then combined using a nonlinear fusion approach resulting a single response map that includes all the detections. Following this stage, highly optimized state-of-the-art global strategies, such as [11] or [14], can be used. Our likelihood model, includes the mixture of all 'view-based' specialized detectors for a given landmark, effectively dealing with large appearance changes caused by the natural rigid (head pose) and nonrigid motion of a face.

The remaining of the paper is organized as follows: section 2 briefly explains the basics in CLM design, section 3 describes the multiple detection cluster and learning approach. Section 4 presents the evaluation results and finally, section 5, concludes the paper.

2. BACKGROUND

2.1. Linear Shape Model

The shape $\mathbf{s}$ of a Point Distribution Model (PDM) [17] with v landmarks is represented by a vector with the 2D vertex locations of a mesh $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$. Briefly, the PDM describes a shape by the linear model

$$\mathbf{s} = \mathcal{S}\left(\mathbf{s}_0 + \Phi\mathbf{b}, \mathbf{q}\right) \tag{1}$$

where $\mathbf{s}_0$ is the mean shape (the base mesh), Φ is the shape subspace matrix holding n eigenvectors (or the modes of deformation that re-

This work was supported by the Portuguese Science Foundation (FCT) under the project with reference PTDC/EEA-CRO/122812/2010 and through grants SFRH/BPD/90200/2012 (Pedro Martins), SFRH/BD74152/2010 (Rui Caseiro) and SFRH/BD/75459/2010 (João Henriques), respectively.

tain a given amount of variance, e.g. 95%), $\mathbf{b}$ is a vector of shape parameters and $\mathcal{S}(.,\mathbf{q})$ linearly represents a similarity transformation [2] function of the $\mathbf{q} = [s\cos(\theta) - 1, s\sin(\theta),\ t_x,\ t_y]^T$ parameters (s, θ, t_x, t_y are the scale, rotation and translations, respectively).

2.2. Local Detectors

The appearance model of an CLM consists of an ensemble of v local detectors [18, 11, 13]. The correlation of the i^{th} landmark detector, evaluated at the pixel location $\mathbf{x}_i = (x_i, y_i)$, is given by

$$\mathcal{D}_i(\mathbf{I}(\mathbf{x}_i)) = \mathbf{h}_i^T \mathbf{I}(\mathbf{x}_i) \tag{2}$$

where $\mathbf{h}_i$ is a linear detector and $\mathbf{I}(\mathbf{x}_i)$ is a surrounding $L \times L$ support region (image patch, denoted by $\mathbf{\Omega}_{\mathbf{x}_i}$). Next, the detector score must be converted into a probability value. The simplest solution is to use a logistic function. Defining a_i to be a binary variable that denotes correct landmark alignment, the probability of pixel $\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{x}_i}$ being aligned is given by

$$p_i(\mathbf{z}_i) = p(a_i = 1|\mathcal{D}_i, \mathbf{I}(\mathbf{z}_i)) = \frac{1}{1 + e^{-a_i\beta_1\mathcal{D}_i(\mathbf{I}(\mathbf{z}_i))+\beta_0}} \tag{3}$$

where β_1 and β_0 are the regression coefficient and intercept, respectively. In the previous, $p_i(\mathbf{z}_i)$, is just used as a condensed representation of the response map. Note that a proper probability is used, always non-negative and $p(a_i = 1|\mathbf{I}(\mathbf{z}_i)) + p(a_i = -1|\mathbf{I}(\mathbf{z}_i)) = 1$.

2.3. CLM Fitting - A Bayesian Approach

In a Bayesian setting [13, 9], the optimal shape parameters $\mathbf{b}^*$ are given by the Bayes' theorem, where we seek to maximize the following posterior probability

$$\mathbf{b}^* = \arg\max_{\mathbf{b}} p(\mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b}) \tag{4}$$

with $\mathbf{y}$ being a $2v$ vector that represents the observed shape (measurement), $p(\mathbf{y}|\mathbf{b})$ is the likelihood term (that comes from the response maps) and $p(\mathbf{b})$ is the prior term that defines the knowledge of the model (the PDM). Conditional independence between landmarks is usually assumed, by sampling each landmark independently, hence the overall likelihood becomes the individual contribution for each landmark as $p(\mathbf{y}|\mathbf{b}) \approx \prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b})$.

2.4. The Likelihood Term

In general, the likelihood term follow the Gaussian form

$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))^T\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))\right) \tag{5}$$

where $\Sigma_{\mathbf{y}}$ is the uncertainty of the spacial localization of the landmarks (being a $2v \times 2v$ block diagonal covariance matrix due to the conditional independence assumed). Most of the CLM fitting approaches differ from each other by the way that the shape measurement $\mathbf{y}$ and its uncertainty $\Sigma_{\mathbf{y}}$ are obtained from the response maps. In fact, these methods can be considered as *local optimization strategies* and the most used are the Active Shape Models (ASM) [5], the Convex Quadratic Fitting (CQF) [8] and more recently the Subspace Constrained Mean-Shifts (SCMS) [11]. The last approximates the response maps by a non-parametric representation using a Kernel Density Estimator (KDE) [19]. Maximizing over the KDE

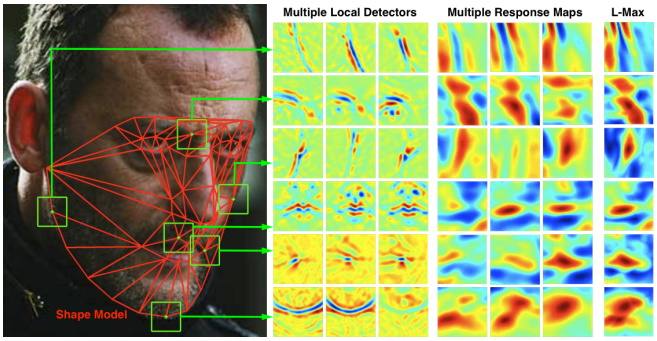


Fig. 1. The CLM combine an ensemble of local landmark detectors (one per landmark) whose locations are regularized by a linear shape model. Our approach, aims to further increase the likelihood model, using multiple detectors for the same landmark. The image shows the detectors and their individual responses as well as the overall combined response map, for the highlighted landmarks, respectively.

is typically achieved by the mean-shift algorithm [20]. In a bit more detail, the i^{th} landmark observation is given by

$$\mathbf{y}_i^{\text{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i\ p_i(\mathbf{z}_i)\ \mathcal{N}(\mathbf{y}_i^{\text{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2\mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\ \mathcal{N}(\mathbf{y}_i^{\text{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2\mathbf{I}_2)} \tag{6}$$

where $\mathbf{I}_2$ is a two-dimensional identity matrix, $\sigma_{h_j}^2$ represents the decreasing bandwidth schedule, $\mathbf{y}_i^c$ is the centered location of the search region the superscript (τ) accounts for the mean-shift iterations. The KDE uncertainty error consists on computing the weighted covariance of the form

$$\Sigma_{\mathbf{y}_i}^{\text{KDE}} = \frac{1}{d-1}\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\text{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\text{KDE}})^T. \tag{7}$$

2.5. The Prior Term

By definition [21], the shape parameters $\mathbf{b}$, follow a multivariate Gaussian distribution $\mathbf{b} \propto \mathcal{N}(\mathbf{b}|\mathbf{0}, \Lambda)$, with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where λ_i denotes the PCA eigenvalue of the i^{th} mode of deformation. The prior term is then defined as $p(\mathbf{b}) \propto \mathcal{N}(\mathbf{b}|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ where $\mu_{\mathbf{b}} = \mathbf{0}$ and $\Sigma_{\mathbf{b}} = \Lambda$. The pose parameters (similarity) are modeled using a non-informative (uniform) prior.

2.6. Global MAP Solution

When the likelihood and the prior terms are both Gaussian distributions, the Bayes' theorem for Gaussian variables [22] states that the posterior is also a Gaussian distribution. This straightforward inference was consider in [9]. Later, in Discriminative Bayesian Active Shape Models (DBASM) [13], this approach was extended to include second order estimates for the shape and pose parameters (i.e. the covariance of the parameters). The global alignment was formulated in terms of a Linear Dynamic System (LDS). The LDS recursively computes a Gaussian posterior probability using incoming (also Gaussian) measurements and a linear model process. The state and measurement equations can be written as

$$\mathbf{b}_k = \mathbf{I}_n\mathbf{b}_{k-1} + q \tag{8}$$

$$\mathbf{y} - \mathbf{s}_0 = \Phi\mathbf{b}_k + r \tag{9}$$

where is assumed that previous shape estimated parameters $\mathbf{b}_{k-1}$ are connected to the current parameters $\mathbf{b}_k$ by an identity relation $\mathbf{I}_n$ with noise (subscript k represents the iteration number). $q \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{b})$ is the additive dynamic noise, $(\mathbf{y} - \mathbf{s}_0)$ is the observed shape deviation from the base mesh (related to the shape parameters by the linear relation Φ in eq.1) and r is the additive measurement noise following $r \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{y})$. The LDS inference accounts with an adaptive prior, where the posterior distribution follow

$$p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k^\mathbf{F}, \boldsymbol{\Sigma}_k^\mathbf{F}) \tag{10}$$

with the mean $\boldsymbol{\mu}_k^\mathbf{F}$ and covariance $\boldsymbol{\Sigma}_k^\mathbf{F}$ given by the well-known Kalman Filter equations [13, 23]. The LDS equations are iteratively reused, along with the response maps evaluated at the new updated locations, until convergence.

3. ENHANCING LOCAL DETECTORS

The appearance model of most CLM approaches [5, 18, 10, 11] consists of a single detector for each of the landmarks. Usually these detectors are learnt, in a training stage, by a linear classifier built from aligned (positive) and misaligned (negative) grey level patch examples [6, 8]. They are mainly chosen by their efficient evaluation of the response maps. However they typically have a limited representation power (detection ambiguities). Recently, the Minimum Output Sum of Squared Error (MOSSE) [24] filters have been successfully used in the CLM framework. Sharing the same efficient evaluation and built only with aligned data, they have been proven to perform better in face alignment tasks (see section 4.2 from [13]).

As previously mentioned, we propose to enhance the discriminative power of the appearance model. This could be done by including a set of multiple local detectors per landmark (i.e. multiple likelihood sources), further increasing the specificity in the detection. Under a Bayesian setting, multiple shape observations (measurements) can be considered by just updating the posterior distribution $\mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k^\mathbf{F}, \boldsymbol{\Sigma}_k^\mathbf{F})$ using multiple times the LDS correction steps [23]. Although this is a valid and interesting approach, it has a major drawback. The fusion of several detections consists of sequential Gaussian corrections. If one of the detectors is weak, producing wrong or very noisy estimates, the final solution is therefore very poor as all detections contribute equally. In fact, when at least one of the detectors is 'bad' the fitting performance will be far worse than simpler CLM formulations based on single detectors.

Here we consider a different approach. We still use multiple detectors per landmark, all based in MOSSE filters, but they are learnt from aligned patch examples which are previously grouped by clusters. The goal is to specialize each detector in a small, more focused, range of visual appearance variation. This is accomplished in a training stage where a combined clustering and detector learning step is used. Afterwards, at test time, all multiple response maps are combined into a single response using a nonlinear approach, making it possible to use standard, highly optimized, global optimization strategies (SCMS [11] or DBASM [13]). The following sections describe the basic detector, the combined clustering and detector learning technique and the nonlinear fusion of multiple response maps.

3.1. Local Detector - MOSSE Filter

The MOSSE filter, recently proposed in [24], finds the filter $\mathbf{H}$ (in the Fourier domain) that minimizes the SSD between the actual output and the desired output of the correlation across a set of N training images, by $\min_{\mathbf{H}^*} \sum_{j=1}^{N} \left(\mathcal{F}\{\mathbf{I}(\mathbf{x}_j)\} \odot \mathbf{H}^* - \mathbf{G}_j\right)^2$. The $*$ symbol represents the complex conjugate, $\mathbf{I}(\mathbf{x}_j)$ is the j^{th} training example and $\mathbf{G}$ is the desired correlation output (usually a 2D Gaussian). Solving for the filter $\mathbf{H}^*$ yields the closed form solution

$$\mathbf{H}^* = \frac{\sum_{j=1}^{N} \mathbf{G}_j \odot \mathcal{F}\{\mathbf{I}(\mathbf{x}_j)\}^*}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}(\mathbf{x}_j)\} \odot \mathcal{F}\{\mathbf{I}(\mathbf{x}_j)\}^* + \epsilon} \tag{11}$$

where ϵ is a regularization parameter. The MOSSE filter maps all aligned training patch examples to an output, $\mathbf{G}$, centered at the feature location, producing highly stable correlation filters. The linear detector for the i^{th} landmark, in eq.2, is given by $\mathbf{h}_i = \mathcal{F}^{-1}\{\mathbf{H}_i^*\}$.

3.2. Learning Multiple Local Detectors

Defining $\{\mathbf{h}_i^{(m)}\}$ to be a set of M local detectors ($m = 1, ..., M$) dedicated to the i^{th} landmark, the goal of the combined clustering and detector learning stage is to find the group of patch examples ($\mathbf{I}(\mathbf{x}_j)$) to be assigned to each specialized detector $\{\mathbf{h}_i^{(m)}\}$, by maximizing the overall correlation with the selected examples. We seek to find the detectors set that maximize the following expression:

$$\arg\max_{\mathbf{h}_i^{(m)}} \sum_{j=1}^{N} \sum_{m=1}^{M} \mathbf{I}(\mathbf{x}_j) * \mathbf{h}_i^{(m)}. \tag{12}$$

The optimization in (12) is solved iteratively using a two step approach. Starting from a initial estimate for the clustering (e.g. k-means), M detectors are build with the initial cluster assigned samples (using $\mathcal{F}^{-1}\{\text{eq.11}\}$). Then, for each patch example $\mathbf{I}(\mathbf{x}_j)$, test with which filter $\mathbf{h}_i^{(m)}$ the correlation is the highest and move the sample j to its cluster m. Rebuild all detectors with the newly updated examples and repeat this process until no more samples change. The algorithm 1 summarizes the overall method.

for *Landmark* $i = 1$ **to** v **do**
  Define the number of desired detectors M (or clusters)
  Get an initial estimate of the clustering by k-means
  repeat
    Build $\mathbf{h}_i^{(m)}$ using current estimate of the labels (j) (eq.11)
    if $\max\{\boldsymbol{h}_i^{(m)} \mathcal{S}\left(\boldsymbol{I}(\boldsymbol{x}_j), \boldsymbol{q}_j\right)\}$ **then**
      Move label (j) to cluster (m)
    end
  until *Labels* (j) *do not change anymore* ;
  return *the specialized set of detectors* $\boldsymbol{h}_i^{(m)}$
end

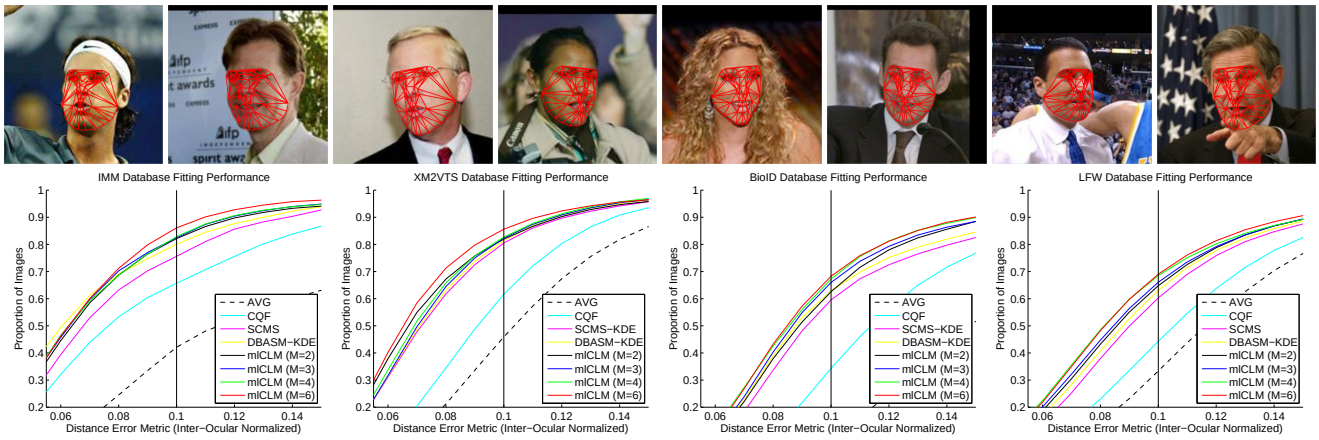
Algorithm 1: Learning the multiple detectors by clustering the training examples (converges in less than 10 iterations).

3.3. Combining Multiple Detections

Representing the multiple response maps by $p_i(\mathbf{z}_i)^{(m)}$, these are all combined using the maximum norm (or L-infinity norm), defined as maximum (of the absolute) values of its components, as

$$p_i(\mathbf{z}_i)_\infty = \max_{\mathbf{z}_i}\{p_i(\mathbf{z}_i)^{(1)}, \ldots, p_i(\mathbf{z}_i)^{(M)}\}. \tag{13}$$

Note that each response map $p_i(\mathbf{z}_i)^{(m)} \geq 0$ as it represents the probability of each candidate pixel $\mathbf{z}_i$ is aligned, according to eq.3. This nonlinear metric preserves the modes of all response maps, it is robust to poor detections and it is also very fast to compute. Figure 1 shows both the specialized filters ($M = 3$), the individual response maps and the their combined robust fusion $p_i(\mathbf{z}_i)_\infty$.



(a) IMM [25] database (b) XM2VTS [26] database (c) BioID [27] database (d) LFW [28] database

Reference $e_m = 0.1$ (vertical line)	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)		LFW (13233 images)	
CQF [8]	65.7		61.6		34.3		44.2	
SCMS [11]	75.6	(0)	80.5	(0)	59.6	(0)	60.3	(0)
DBASM-KDE [13]	80.1	(+4.4)	81.5	(+1.0)	62.9	(+3.4)	63.0	(+2.8)
mlCLM (M=2) (our method)	82.2	(+6.6)	82.0	(+1.5)	62.6	(+3.0)	64.8	(+4.5)
mlCLM (M=3) (our method)	82.5	(+6.9)	82.4	(+1.9)	66.0	(+6.4)	66.1	(+5.8)
mlCLM (M=4) (our method)	82.9	(+7.3)	82.7	(+2.2)	67.2	(+7.7)	68.5	(+8.2)
mlCLM (M=6) (our method)	**86.1**	(+10.5)	**85.6**	(+5.1)	**68.3**	(+8.7)	**69.0**	(+8.7)

Fig. 2. Fitting performance curves. The table holds quantitative values taken by setting a fixed error amount ($e_m = 0.1$, i.e. the vertical line in the graphics). Each table entry show how many percentage of images converge with less (or equal) error than the reference.

4. EVALUATION RESULTS

The performance evaluation was conducted in several standard databases, namely the IMM [25] (240 images taken from 40 people annotated with 58 landmarks), the BioID [27] (1521 images from 23 subjects with 20 landmarks), the XM2VTS [26] (2360 frontal images of 295 subjects with 68 landmarks) and the Labeled Faces in the Wild (LFW) [28] (13233 images and 12 landmarks). Both XM2VTS and BioID mainly focuses on variations in identity. Nevertheless, they exhibit large diversity in appearance due to facial hair, glasses, ethnicity and other subtle changes. The IMM is the smallest database, however it presents a large variation in head pose, illumination, and spontaneous facial expressions along several individuals. Unlike the previous, the LFW database is an extremely challenging database, completely taken in wild. Their images are captured under uncontrolled natural conditions presenting changes in pose, illumination, facial expression, occlusion, etc. Both the shape model ($v = 58$ landmarks) and the detectors (MOSSE filters) have been built using training images taken from the IMM [25] combined with images collected at our institution. Several sets of multiple detectors (per landmark) where evaluated using $M = 2, 3, 4$ and 6 clusters using the approach in section 3.2. Each has size of 31×31 and it were used to scan a local region of 25×25. The desired MOSSE correlation output ($\mathbf{G}$) was set to be a 2D Gaussian centered at the each landmark with 3 pixels of standard deviation.

Our method refereed as multiple likelihood-CLM (mlCLM), using $M = 2, 3, 4$ and 6 sets of detectors, was evaluated against the DBASM-KDE [13], which is equivalent to mlCLM with $M = 1$. The CQF [8] and the SCMS [11] were included as baseline. Note that mlCLM uses the same global optimization than DBASM (section 2.6) differing only in the likelihood model. All methods share the same shape model, the initial shape parameters start from zero (mean shape), the pose parameters were initialized by a face detector ('AVG' in the evaluation charts) and the model was fitted until convergence up to a maximum of 20 iterations.

The Figure 2 shows the fitting performance curves for all the evaluated methods in the four different datasets. These curves, that were widely adopted in [6, 7, 8, 11, 13], are cumulative distribution functions that show the percentage of faces that achieved a given error amount (shown at the horizontal axis). Following common practice [6, 7], the error metric is given by the mean error per landmark as fraction of the inter-ocular distance, d_{eyes}, as $e_m(\mathbf{s}) = \frac{1}{v\, d_{\text{eyes}}} \sum_i^v \|\mathbf{s}_i - \mathbf{s}_i^{\text{gt}}\|$ where $\mathbf{s}_i^{\text{gt}}$ is the location of i^{th} landmark in the shape ground truth annotation. The table presented in the same figure shows quantitative values taken from sampling the curves setting a fixed error metric amount ($e_m = 0.1$, shown as a vertical line in the graphics). As expected, the results show that CQF and SCMS are all outperformed by DBASM-KDE, which is known to be an enhanced global optimization strategy [13]. The main evaluation, in practice, evaluates the effect of using more than one set of detectors per landmark. The results show that using of more sets of filters tend to improve the overall fitting accuracy (more than $M = 6$ only provide marginal improvement) but at the additional computational cost of more evaluations (the complexity scales linearly with M). Still, parallel processing could be used, as response maps are all independent (conditional independence assumed).

5. CONCLUSIONS

This work presents a novel CLM fitting approach that seamlessly is able to include multiple likelihood sources (several detectors per landmark), further improving the fitting performance. Each set of detectors is built with training data constrained by a clustering stage. Finally, the multiple responses are then combined using a nonlinear fusion approach and globally optimized by a state-of-the-art strategy.

6. REFERENCES

[1] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE TPAMI*, vol. 23, no. 6, pp. 681–685, June 2001.

[2] I. Matthews and S. Baker, "Active appearance models revisited," *IJCV*, vol. 60, no. 1, pp. 135–164, November 2004.

[3] P. Martins, R. Caseiro, and J. Batista, "Face alignment through 2.5d active appearance models," in *British Machine Vision Conference*, 2010.

[4] P. Martins, R. Caseiro, and J. Batista, "Generative face alignment through 2.5d active appearance models," *CVIU*, vol. 117, no. 3, pp. 250–268, March 2013.

[5] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J.Graham, "Active shape models-their training and application," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995.

[6] D. Cristinacce and T.F. Cootes, "Boosted regression active shape models," in *BMVC*, 2007.

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

[8] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *IEEE CVPR*, 2008.

[9] U. Paquet, "Convexity and bayesian constrained local models," in *CVPR*, 2009.

[10] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *ECCV*, 2008.

[11] J. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *IEEE ICCV*, 2009.

[12] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shifts," *IJCV*, vol. 91, no. 2, pp. 200–215, 2010.

[13] P. Martins, R. Caseiro, J.F. Henriques, and J. Batista, "Discriminative bayesian active shape models," in *ECCV*, 2012.

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

[15] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting with a mixture of local experts," in *IEEE ICCV*, 2009.

[16] V. Rapp, K. Bailly, T. Senechal, and L. Prevost, "Multi-kernel appearance model," *Image and Vision Computing*, vol. 31, no. 8, pp. 542–554, 2013.

[17] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for computer vision," Tech. Rep., Imaging Science and Biomedical Engineering, Univer. of Manchester, 2004.

[18] D. Cristinacce and T.F. Cootes, "Feature detection and tracking with constrained local models," in *BMVC*, 2006.

[19] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[20] D. Comaniciu and P. Meer, "Mean Shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, pp. 603–619, May 2002.

[21] M. E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *JRSS*, vol. 21, no. 3, pp. 611–622, 1999.

[22] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[23] R.E. Kalman, "A new approach to linear filtering and prediction problems," *TASME*, vol. 82, no. D, pp. 35–45, 1960.

[24] D.S. Bolme, J.R. Beveridge, B.A. Draper, and Y.M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE CVPR*, 2010.

[25] M. Nordstrom, M. Larsen, J. Sierakowski, and M. Stegmann, "The IMM face database - an annotated dataset of 240 face images," Tech. Rep., Technical University of Denmark, DTU, May 2004.

[26] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *AVBPA*, 1999.

[27] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *AVBPA*, 2001.

[28] G.B. Huang, M. Ramesh, T. Berg, and E.L.-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, 2007.