BAYESIAN CONSTRAINED LOCAL MODELS WITH DEPTH DATA

Pedro Martins, João Faro, Patrick Brandão, Jorge Batista

Institute of Systems and Robotics, University of Coimbra, Portugal

ABSTRACT

This paper proposes an extended Constrained Local Model (CLM) formulation for aligning faces using depth information. The CLMs are popular methods that were initially designed to locate facial features in regular intensity images. Briefly, they combine a set of local detectors, one for each landmark, whose locations are regularized by a linear shape model. Fitting a CLM is usually framed as a two step approach: locally search, using the detectors, producing response maps (likelihood maps) followed by a global optimization strategy that jointly maximize all detection scores while enforcing an appropriate shape. Including depth data could be simply posed as adding additional likelihood sources to the main formulation. The paper discusses several likelihood fusion techniques and propose to jointly learn a multi-dimensional correlation filter as a more reliable solution. Moreover, we propose to learn the local detectors, in the Fourier domain, effectively augmenting the training set with virtual samples. Besides improving the detections reliability, this approach is particular important when applied to depth data, as no additional processing is required (such as fill missing information). The performance evaluation shows that our extended approach further increases the fitting performance (accuracy) effectively proving the benefit of using depth data in facial alignment tasks.

Index Terms— Non-rigid face alignment, facial feature localization, Constrained Local Model (CLM), RGBD face alignment.

1. INTRODUCTION

Non-rigid face alignment is a fundamental task in many computer vision applications such as tracking, recognition, pose estimation, video compression, video editing, etc. Accurately retrieving facial information still remains a challenging problem, mainly to the huge variability that a human face can exhibit, such as changes in identity, expression, occlusion, illumination and 3D orientation.

The Active Appearance Model (AAM) [1, 2], since its introduction, had become a quite popular strategy to align faces, therefore locating the desired facial features. The AAMs are generative methods that rely in a shape model (Point Distribution Model) that describes the spatial localization of all facial landmarks and an holistic appearance model that encodes its texture (captured by PCA). The AAMs are indeed able to achieve an impressive registration quality, however, their generative nature generalizes poorly beyond unseen data (individuals not included in the training dataset).

Later, the Constrained Local Model (CLM) [3, 4, 5, 6, 7] was proposed. The CLM has greatly improved the generic face representation by using a discriminative based appearance model. It combines a set of local feature detectors, one for each landmark, whose search locations are regularized by a linear shape model. Fitting a CLM is usually framed as a two step approach: locally search, using the detectors, producing response maps (likelihood maps) followed by a global optimization strategy that jointly maximize all detection scores while enforcing an appropriate shape. Most CLM strategies aim to approximate the responses maps by simple parametric forms (Weighted Peak Responses [8], Gaussians Responses [4, 9], Mixture of Gaussians [10]) or non-parametrically by a Kernel Density Estimator [5]. However, the most recent CLM optimizations have been reformulated in terms of Bayesian inference [9, 11, 12, 6] where the local detectors responses and the shape model behave as building blocks for the likelihood and prior terms.

Unfortunately, even leading CLMs techniques [4, 5, 6, 7] can struggle with some problems, such as extreme illumination changes. In addition, the well-know aperture problem can sometimes lead to wrong detector estimates in some landmarks (with strong responses along the edges, p.e. nose ridge and chin regions). Fortunately, with the recent developments in RGBD sensors, depth data can be easily and affordably obtained. Actually, depth data might be a potential solution to relieve some of these drawbacks.

In this paper we present an extended CLM formulation that will allow to take advantage of both color/intensity and depth data, simultaneously. Essentially, considering depth data could be simply posed as including an additional likelihood source to the main formulation. In fact, previous attempts were made [13][14], however, only basic likelihood data fusion was achieved. In particular, only two different sources, originating from grey intensities and depth, were considered and both had a 50/50 contribution for the overall likelihood map. Here we discuss several more alternatives, that can deal with more than two sources (p.e. RGBD with 4 channels), and propose to jointly learn a multi-dimensional correlation filter as a more reliable solution. Moreover, we efficiently learn these detectors, in the Fourier domain, where training data is effectively augmented with virtual samples, making the local detections much more reliable. In addition, these detectors when applied to depth data do not require any extra processing, such as fill missing information.

The remaining of the paper is organized as follows: section 2 briefly explains the basics in the CLM formulation, section 3 describes the multiple likelihood fusion strategies, section 4 presents the evaluation results and finally, section 5 draws the conclusions.

2. BACKGROUND

Briefly, the Constrained Local Model (CLM) consist of a collection of v local detectors, one for each landmark, denoted here as $\{\mathbf{h}_i\}_1^v$ whose search locations are regularized by a linear shape model.

2.1. Shape Model

The shape ($\mathbf{s}$) with v landmarks is usually represented by a vector with their 2D locations $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$. A Point Distribution Model (PDM) [15] essentially describes a shape by

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

This work was supported by the Portuguese Science Foundation (FCT) through the grant SFRH/BPD/90200/2012 (Pedro Martins).

where $\mathbf{s}_0$ is the mean shape, Φ is the shape subspace matrix holding n eigenvectors (that resulted from applying Principal Components Analysis on a set of normalized training shapes), $\mathbf{b}$ is a vector of shape parameters representing the mixing weights and $\mathcal{S}(.,\mathbf{q})$ linearly represents a similarity transformation [2] function of the pose parameters $\mathbf{q} = [s\cos(\theta) - 1, s\sin(\theta),\ t_x,\ t_y]^T$ where s, θ, t_x, t_y are the scale, rotation and translations, respectively.

2.2. Local Detectors

The score $\mathcal{D}_i$ 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 image patch (with $L \times L$ support region centered at $\mathbf{x}_i$, denoted by $\mathbf{\Omega}_{\mathbf{x}_i}$). Note that, by now, only single channel images and detectors are assumed. In a probabilistic framework, the detector score $\mathcal{D}_i$ must be converted into a probability value. The common 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

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

where β_1 and β_0 are the regression coefficient and intercept, respectively. Eq. 3 use $p(\mathbf{z}_i)$ as a short notation for the response map.

2.2.1. MOSSE Filters

Several kinds of local detectors have been used within the CLM framework [16][17][4][18][11]. Probably, the most popular is the linear SVMs when trained with aligned (positive) vs. misaligned (negative) image patch examples. Recently, correlation filters have been used [11][7][6], in particular the Minimum Output Sum of Squared Error (MOSSE) filter [19]. Compared with the previous, it has several advantages: (1) it extends the linear SVM with real valued labels (meaning that a large amount of virtual samples are incorporated in the training [20]), (2) it allows discriminative learning using only aligned (positive) data, (3) it maintains a linear nature, and (4) it can outperform others [6].

Briefly, finding each MOSSE filter $\mathbf{h}_i$, consists of solving the following linear regression problem

$$\min_{\mathbf{h}_i} \sum_{j=1}^{N} \left(\mathbf{h}_i * \mathbf{I}_j - \mathbf{g}_j\right)^2 + \lambda||\mathbf{h}_i||^2 \tag{4}$$

where $(*)$ is the correlation operator, $\mathbf{I}_j$ is the j^{th} training patch (N in total), $\mathbf{g}_j$ the desired target correlation (usually set to be a 2D Gaussian with σ_h standard deviation) and λ is a regularization parameter. Eq. 4 can be efficiently solved in the Fourier domain (where convolutions become products), with the solution given by

$$\mathbf{h}_i = \mathcal{F}^{-1}\left\{\frac{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{g}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^\dagger}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^\dagger + \lambda}\right\}^\dagger \tag{5}$$

where $\mathcal{F}$ represents the 2D Fourier transform, the $\odot$ symbol the Hadamard product and $(^\dagger)$ the complex conjugate. Note that, the MOSSE filter is particularly useful when applied to depth data because it finds a low pass filter (with null dc value), thus avoiding the need to an additional processing stage (such as excluding or filling missing information [13]). Depth data 'holes' are simply filtered out.

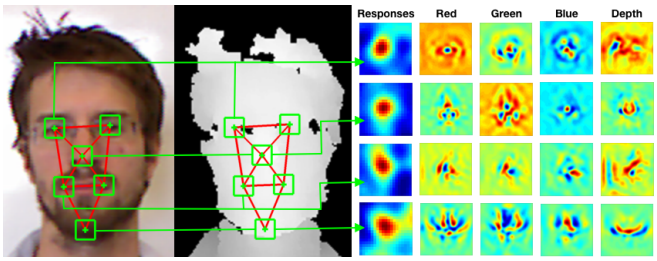


Fig. 1. The Constrained Local Model (CLM) combine a set of local detectors whose locations are regularized by a linear shape model. The proposed extension jointly learns multi-dimensional local detectors that include color and depth data simultaneously. The image shows RGBD scan regions, followed by a column of response maps and each RGBD local detectors, respectively.

2.3. CLM Fitting

Under a Bayesian paradigm [9, 6, 7], the optimal shape parameters $(\hat{\mathbf{b}})$ are provided by the Bayes' theorem, where the following posterior probability is maximized

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} p(\mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b}) \tag{6}$$

with $\mathbf{y} \in \Re^{2v}$ being the observed shape vector (shape measurement), $p(\mathbf{y}|\mathbf{b})$ is the likelihood term (which is extracted from the response maps) and $p(\mathbf{b})$ is the prior term that defines the current knowledge of the shape model. Usually in CLMs, conditional independence between landmarks is assumed (sampling each landmark independently), leading to an overall likelihood that becomes the individual landmark contribution as $p(\mathbf{y}|\mathbf{b}) \approx \prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b})$.

2.4. The Likelihood Term

The likelihood term can be expressed by [6]

$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))^T \Sigma_{\mathbf{y}}^{-1} (\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))\right) \tag{7}$$

where $\Sigma_{\mathbf{y}}$ represents the uncertainty in the localization of the landmarks ($2v \times 2v$ block diagonal covariance matrix due to the conditional independence assumed). Several strategies have been proposed to extract the shape measurement $(\mathbf{y})$ and its uncertainty $(\Sigma_{\mathbf{y}})$ from the response maps. The most popular are the Active Shape Models (ASM) [8], the Convex Quadratic Fitting (CQF) [4] and, more recently, the Subspace Constrained Mean-Shifts (SCMS) [5]. SCMS approximates the response maps by a non-parametric representation using a Kernel Density Estimator (KDE) [21]. Maximizing over the KDE is typically accomplished by using the mean-shift algorithm [22]. Formally, the i^{th} landmark observation is given by

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i\, p(\mathbf{z}_i)\, \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_j^2\mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p(\mathbf{z}_i)\, \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_j^2\mathbf{I}_2)} \tag{8}$$

where $\mathbf{y}_i^c$ represents the center location of the search region, $\mathbf{I}_2$ is a 2D identity matrix, σ_j^2 defines the decreasing bandwidth schedule and the superscript (τ) is the iteration number. The uncertainty localization error $(\Sigma_{\mathbf{y}})$ consists of computing the weighted covariance, centered at $\mathbf{y}_i^{\mathrm{KDE}}$, which is given by

$$\Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{\sum_{\mathbf{z}_i} p(\mathbf{z}_i) - 1} \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T. \tag{9}$$

2.5. The Prior Term

The shape parameters ($\mathbf{b}$) follow a multivariate Gaussian distribution with zero mean and a diagonal covariance $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where λ_j denotes the j^{th} PCA eigenvalue [23]. According, the prior term becomes $p(\mathbf{b}) \propto \mathcal{N}(\mathbf{b}|\mathbf{0}, \Lambda)$. The pose parameters ($\mathbf{q}$) are typically modeled using an uniform prior.

2.6. The Posterior Term

According to the Bayes' theorem for Gaussian variables [24], when the likelihood and the prior terms are both Gaussian distributions, comes that posterior distribution is also a Gaussian. In fact, this basic inference was used in [9]. Later in, Bayesian CLM Revisited (BCLM) [11, 6], the previous approach was extended to include second order estimates (the covariance) of shape parameters. In essence, the global CLM alignment was formulated in terms of a Linear Dynamic System (LDS) where it recursively estimates a Gaussian posterior distribution using Gaussian shape measurements and a linear process. According, the state and measurement equations can be written as

$$\mathbf{b}_l = \mathbf{I}_n \mathbf{b}_{l-1} + q \quad (10)$$

$$\Delta\mathbf{y} = \Phi \mathbf{b}_l + r \quad (11)$$

where $q \sim \mathcal{N}(\mathbf{0}, \Lambda)$ is the dynamic transition noise, $\Delta\mathbf{y} = \mathbf{y} - \mathbf{s}_0$ is the observed shape deviation from the mean, $r \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{y})$ is the measurement noise and the subscript (l) represents the iteration number. The state transition (eq. 10) relates $\mathbf{b}_{l-1}$ to $\mathbf{b}_l$ by an identity relation $\mathbf{I}_n$ with additive noise [6] and the measurement step (eq. 11) simply expresses the likelihood term from eq. 7. The LDS infers the posterior distribution according to

$$p(\mathbf{b}_l|\mathbf{y}_l, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_l|\boldsymbol{\mu}_l^\mathbf{F}, \boldsymbol{\Sigma}_l^\mathbf{F}) \quad (12)$$

with the mean $\boldsymbol{\mu}_l^\mathbf{F}$ and covariance $\boldsymbol{\Sigma}_l^\mathbf{F}$ given by the Kalman Filter equations [6, 25]. Finally, the shape parameters that maximize the goal, in eq. 6, are given by the expectation of the posterior distribution which is $\hat{\mathbf{b}} = \boldsymbol{\mu}_l^\mathbf{F}$. In summary, fitting a BCLM is an iterative procedure that requires to generate a shape using the model (eq. 1), evaluate the response maps around each landmark, extract the likelihood parameters $(\mathbf{y}, \Sigma_\mathbf{y})$ and then infer a new shape using the LDS.

3. LIKELIHOOD FUSION STRATEGIES

In a Bayesian framework, multiple likelihood observations (shape measurements) can be considered by just updating the posterior distribution (eq. 12) using several times the LDS update steps [25]. However, the LDS relies in Gaussian inference techniques, where each likelihood source contributes (weighted by the error covariance) to a final solution. When facing low quality response maps (weak detectors or noisy estimates), the LDS will over smooth the solution, leading the overall CLM fitting into a loss of accuracy.

In this work, we overcome this limitation, by moving to a solution that relies in the fusion of several response maps into one, making it still suitable to be used with regular CLM fitting algorithms.

In the following sections it is assumed that, (k) multiple local detectors $\mathbf{h}_i^{(k)}$ exist for the i^{th} landmark. These local detectors can be learnt from several features/image channels (such as R-G-B colors or depth), by using eq. 5 in each data channel, independently. Every detector then produce a likelihood map (response map), according to eq. 3, given by $p(a_i|\mathcal{D}_i^{(k)}, \mathbf{I}(\mathbf{x}_i)^{(k)})$ where the superscript (k) represents the k^{th} image channel. The next sections describe some possible response map fusion strategies.

3.1. Average Fusion

The simplest likelihood fusion strategy is to take the mean value across all available D response maps. According to eq. 3, comes

$$p(\mathbf{z}_i)^{\text{AVG}} = \frac{1}{D}\sum_{k=1}^{D} p(a_i|\mathcal{D}_i^{(k)}, \mathbf{I}(\mathbf{z}_i)^{(k)}) \quad (13)$$

where $\mathbf{I}(\mathbf{z}_i)^{(k)}$ is the k^{th} feature channel (taken from the i^{th} landmark support region) and $\mathcal{D}_i^{(k)}$ the score produced by each individual detector. In fact, this strategy extends the approach proposed in [13] and [14], when grey and depth channels are used ($D = 2$). In both works it was considered that the overall response has 50/50 contribution between grey intensity and depth data.

3.2. Max Fusion

A different non-linear approach, proposed in [26], consists of using the maximum norm (or L-infinity norm) in each response map

$$p(\mathbf{z}_i)^{\text{MAX}} = \max_{\mathbf{z}_i} p(a_i|\mathcal{D}_i^{(k)}, \mathbf{I}(\mathbf{z}_i)^{(k)}). \quad (14)$$

The resulted combined response map, simply holds the highest score values, preserving all modes of every individual response.

3.3. Multi-Dimensional Filters

As an alternative, we propose to jointly learn a multi-dimensional correlation filter that uses all available data (from all channels) at once. This approach is based in a multi-channel extension of the MOSSE filters [27][28][29]. Briefly, eq. 5 is now extended into a minimization across all D channels

$$\min_{\mathbf{h}_i^{(1)}, \ldots, \mathbf{h}_i^{(D)}} \sum_{j=1}^{N}\sum_{k=1}^{D} \left(\mathbf{h}_i^{(k)} * \mathbf{I}_j^{(k)} - \mathbf{g}_j\right)^2 + \lambda \sum_{k=1}^{D} ||\mathbf{h}_i^{(k)}||^2 \quad (15)$$

where (once again) N is the number of training images and λ a regularization parameter. Eq. 15 finds the multi-dimensional filter $\{\mathbf{h}_i^{(k)}\}_{k=1}^{D}$ that minimizes the correlation between the actual output $(\mathbf{h}_i^{(k)} * \mathbf{I}_j^{(k)})$ and the desired correction $(\mathbf{g}_j)$ across all multi-dimensional samples $(\mathbf{I}_j^{(k)})$, simultaneously. The solution becomes

$$\{\mathbf{h}_i^{(k)}\}_1^D = \mathcal{F}^{-1}\left\{\left(\lambda\mathbf{I} + \sum_{j=1}^{N} \boldsymbol{\Xi}_j^H \boldsymbol{\Xi}_j\right)^{-1} \sum_{j=1}^{N} \boldsymbol{\Xi}_j^H (\mathbf{1} \otimes \mathcal{F}\{\mathbf{g}_j\})\right\}^\dagger \quad (16)$$

with $\boldsymbol{\Xi}_j = \left[\text{diag}\left(\mathcal{F}\{\mathbf{I}_j^{(1)}\}\right), \ldots, \text{diag}\left(\mathcal{F}\{\mathbf{I}_j^{(D)}\}\right)\right]$, $\mathbf{1}$ is a D dimensional vector with ones, the $\otimes$ symbol is the Kronecker product and, in this case, $(^H)$ stands for the conjugate transpose. Note that, $\boldsymbol{\Xi}_j$ is mostly sparse, therefore, we can take advantage of this structure to efficiently compute the term that requires inversion in eq. 16. Please refer to [28][29] for additional details.

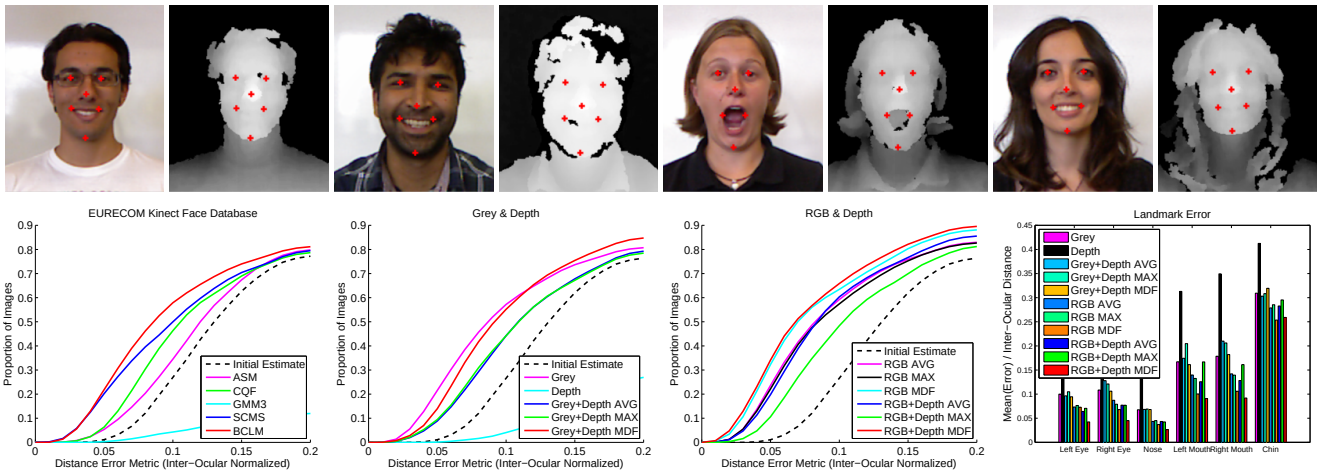
Finally, the overall correlation simply aggregates the score of each feature channel as

$$\mathcal{D}_i^{\text{MDF}} = \sum_{k=1}^{D} \mathbf{h}_i^{(k)} \mathbf{I}(\mathbf{z}_i)^{(k)} \quad (17)$$

which is then converted to a single response map (using eq. 3) by

$$p(\mathbf{z}_i)^{\text{MDF}} = p(a_i|\mathcal{D}_i^{\text{MDF}}, \mathbf{I}(\mathbf{z}_i)^{(k)}). \quad (18)$$

Note that, the data fusion is intrinsically applied during the detectors learning stage, in the optimization of eq. 15. Figure 1 shows the multi-dimensional filters, and their responses, in RGBD data.



(a) CLM global strategies. (b) Grey & depth fusion. (c) RGB & depth fusion. (d) Individual landmark error.

CLM optimization	Area under cdf curve / total area (%)
Initial Estimate [30]	32.3
ASM [8]	36.8
CQF[4]	40.3
GMM [10]	4.8
SCMS [5]	44.7
BCLM [6]	**48.2**

Fusion strategy	Area ratio (%)
Grey	48.2
Depth	7.8
Grey+Depth AVG	40.1
Grey+Depth MAX	40.2
Grey+Depth MDF	**48.4**

Fusion strategy	Area ratio (%)
RGB AVG	49.8
RGB MAX	48.9
RGB MDF	**54.7**
RGB+Depth AVG	49.8
RGB+Depth MAX	42.5
RGB+Depth MDF	**56.4**

Fig. 2. The fitting performance curves in the Kinect Face Database [31]. (a) CLM optimization strategies evaluation using only grey data, (b) fusion evaluation grey and depth, (c) fusion evaluation RGB and depth and (d) individual landmark error. The tables show a quantitative measure of the ratio between the area below each curve and the total area. The top images show fitting results with BCLM RGB+Depth MDF.

4. EVALUATION RESULTS

The performance evaluation was mainly conducted in the recent Eurecom Kinect Face Database [31]. The dataset provides RGB-D facial images of 52 people (14 females, 38 males) captured by Microsoft Kinect sensor, under different facial expressions, lighting conditions, occlusions and taken in two sessions. Annotations with 6 landmarks are provided for almost all the images (total of $N = 624$).

A initial experiment was designed to evaluate leading CLMs techniques using only standard grey level data. This procedure was aimed to rule out the best approach to further evaluate, in detail, the multiple likelihood fusion strategies. According, some CLM global alignment solutions, ASM [8], CQF [4], GMM [10] with 3 Gaussians (GMM3), SCMS [5] were evaluated against the BCLM [11, 6] (described in sec. 2.6). The local detectors (MOSSE filters in eq. 5) have been built using a 31×31 support region and a desired Gaussian correlation output ($\mathbf{g}$) with a standard deviation $\sigma_h = 2.5$ and regularization $\lambda = 10^{-4}$. All methods share the same shape model ($v = 6$), the initial shape parameters start from the mean shape, the pose parameters were initialized by a face detector [30] and the model was fitted until convergence up to a max of 30 iterations.

The Figure 2(a) shows the fitting performance curves for all the evaluated methods in the Kinect Face Database [31]. These curves, that were widely adopted in [3, 4, 5, 6], are cumulative distribution functions that show the percentage of faces that achieved a given error amount (shown at the horizontal axis). Following the usual practice [3, 4, 5, 6], the error metric is given by the mean error per landmark as fraction of the inter-ocular distance, d_{eyes}, as $e_m(\mathbf{s}) = \frac{1}{v\ d_{\text{eyes}}} \sum_i^v \|\mathbf{s}_i - \mathbf{s}_i^{\text{gt}}\|^2$ where $\mathbf{s}_i^{\text{gt}}$ is the location of i^{th} landmark in the ground truth. The table in the same figure shows a quantitative measure of the results, which is defined as the ratio in percentage, between the area bellow the fitting curve and the total area of a ground truth curve (step curve). As expected, the results show that ASM, CQF, GMM3 and SCMS are outperformed by BCLM, which is known to be an enhanced global optimization [6].

The main evaluation was designed to evaluate the effect of the fusion strategies presented in section 3 (while using BCLM fitting technique, previously proven to perform better). The experiments exhaustively evaluate several features (grey, RGB and depth) combined with the likelihood fusion strategies: AVG, MAX and our proposed MDF (sections 3.1, 3.2 and 3.3, respectively). Regarding the MDF strategy, the multi-channel filters used the same settings as before ($L = 31$, $\sigma_h = 2.5$ and $\lambda = 10^{-4}$). Figures 2(b) and 2(c) show fitting performance curves for all combinations. Note that, using only 'grey' features in fig. 2(b) match the curve BCLM in fig. 2(a) and the AVG 'grey+depth' strategy correspond to methods [13, 14] ($D = 2$). Finally, figure 2(d) displays bar charts with the (inter-ocular normalized) average errors in each individual facial landmark, for all evaluated strategies. The results show, in first, that using just raw depth has low accuracy. Grey intensities alone perform better than just adding depth with simple basic fusions. The AVG and MAX strategies have similar results, except when all 4 channels are involved (where AVG seams better). As expected, including full RGB color improves on the results. Adding depth with RGB only produce better results with our MDF strategy. Finally, the best results in every category happen when using our proposed MDF.

5. CONCLUSIONS

This work presents an extended CLM fitting approach that is able to integrate multiple features simultaneously. Several likelihood fusion strategies are described. It is shown that the proposed jointly learning of muti-dimensional correlation filters outperform all methods.

6. REFERENCES

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

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

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

[4] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[5] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shifts," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2010.

[6] P. Martins, J. F. Henriques, R. Caseiro, and J. Batista, "Bayesian constrained local models revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 704–716, April 2016.

[7] P. Martins, R. Caseiro, and J. Batista, "Non-parametric bayesian constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

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

[9] U. Paquet, "Convexity and bayesian constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[10] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *European Conference on Computer Vision*, 2008.

[11] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Discriminative bayesian active shape models," in *European Conference on Computer Vision*, 2012.

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

[13] T. Baltrušaitis, P. Robinson, and L.P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[14] S. Cheng, S. Zafeiriou, A. Asthana, and M. Pantic, "3d facial geometric features for constrained local model," in *IEEE International Conference on Image Processing*, 2014.

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

[16] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *British Machine Vision Conference*, 2006.

[17] T. F. Cootes, M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *European Conference on Computer Vision*, 2012.

[18] T. Baltrušaitis, P. Robinson, and L.P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *IEEE International Conference on Computer Vision Workshops, 300 Faces in-the-Wild Challenge*, 2013.

[19] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, March 2015.

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

[22] D. Comaniciu and P. Meer, "Mean Shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[23] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society*, vol. 21, no. 3, pp. 611–622, 1999.

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

[25] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[26] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Likelihood-enhanced bayesian constrained local models," in *IEEE International Conference on Image Processing*, 2014.

[27] V. N. Boddeti, T. Kanade, and B. V. K. Vijaya Kumar, "Correlation filters for object alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[28] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista, "Beyond hard negative mining: Efficient detector learning via block-circulant decomposition," in *IEEE International Conference on Computer Vision*, 2013.

[29] H. K. Galoogahi, T. Sim, and S. Lucey., "Multi-channel correlation filters," in *IEEE International Conference on Computer Vision*, 2013.

[30] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, July 2002.

[31] R. Min, N. Kose, and J. L. Dugelay, "Kinectfacedb: A kinect database for face recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 44, no. 11, pp. 1534–1548, Nov 2014.