Exemplar-based Graph Matching for Robust Facial Landmark Localization

Feng Zhou
Carnegie Mellon University
Pittsburgh, PA 15213
http://www.f-zhou.com

Jonathan Brandt, Zhe Lin
Adobe Research
San Jose, CA 95110
{jbrandt, zlin}@adobe.com

Abstract

Localizing facial landmarks is a fundamental step in facial image analysis. However, the problem is still challenging due to the large variability in pose and appearance, and the existence of occlusions in real-world face images. In this paper, we present exemplar-based graph matching (EGM), a robust framework for facial landmark localization. Compared to conventional algorithms, EGM has three advantages: (1) an affine-invariant shape constraint is learned online from similar exemplars to better adapt to the test face; (2) the optimal landmark configuration can be directly obtained by solving a graph matching problem with the learned shape constraint; (3) the graph matching problem can be optimized efficiently by linear programming. To our best knowledge, this is the first attempt to apply a graph matching technique for facial landmark localization. Experiments on several challenging datasets demonstrate the advantages of EGM over state-of-the-art methods.

1. Introduction

Facial landmark localization (*a.k.a.*, face alignment) is a critical component in many computer vision applications such as face recognition [34], face reconstruction [20], expression recognition [25] and expression re-targeting [19]. In the past decade, many approaches have been proposed with varying degrees of success on benchmark datasets composed by mostly frontal faces in controlled setting. However, accurately localizing facial landmark points in real-world, cluttered images is still a challenging problem due to the large variability in pose and appearance, and the existence of occlusions. Given the image shown in Fig. 1a, how can we accurately localize the facial landmarks in the chin area even though it is partially occluded?

Conventional algorithms for face alignment typically proceed by fitting a joint shape model to regions around each feature point. Following the pioneering work on the active shape model (ASM) [7], a number of shape models have been proposed for face alignment. Among them, parametric models such as point distribution model (PDM) have been shown to be effective in governing the layout of facial landmarks. Unfortunately, the formulation based on these models is non-convex and in general prone to local minima.

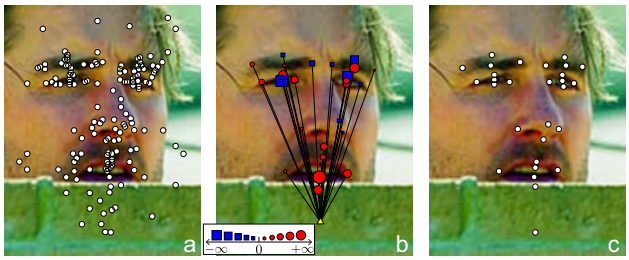


Figure 1. Localizing facial landmarks by exemplar-based graph matching (EGM). Despite the fact that the chin area is partially occluded, EGM still accurately locates the facial landmarks. EGM first finds similar exemplars through a RANSAC step. These exemplars are then used to generate (a) candidate positions for landmarks and to learn (b) an affine-invariant shape constraint, where the position of each landmark (*e.g.*, the chin) is modeled as a weighted linear combination of the other landmarks. By combining these two sources, EGM solves a graph matching problem to obtain (c) the optimal landmark positions.

In this paper, we present exemplar-based graph matching (EGM), a robust framework for facial landmark localization. Unlike previous face alignment algorithms, EGM models the layout of the facial landmarks as a graph in a non-parametric way. For instance, Fig. 1b illustrates the sub-graph centered at the chin area. Compared to conventional methods, EGM has three advantages: (1) the shape constraint (Fig. 1b) estimated from exemplars is invariant to affine transformation and adaptive to the test image, thereby making the system more robust to pose variations; (2) the optimal matching between candidate points and the shape constraint is modeled as a graph matching problem; (3) the graph matching problem can be efficiently solved using linear programming (LP). We compare EGM with state-of-the-art methods and validate its effectiveness on several

challenging benchmark datasets.

2. Related work

Early work on facial landmark localization [12] often treated the problem as a special case of the object part detection problem. However, general detection methods are not suitable in detecting facial landmarks because only a few salient landmarks (*e.g.*, eye centers, mouth corners) can be reliably characterized by their image appearances. Therefore, shape constraints or support from nearby areas are essential for augmenting weak local detectors. According to the type of shape constraint inherently imposed, previous work can be categorized into two groups: parametric methods and non-parametric methods.

Active shape model (ASM) [7] and active appearance model (AAM) [5] are the two most representative face alignment models using parametric shape constraints. In ASM, a point distribution model captures the shape variation of a set of landmark points. In AAM, the appearance is globally modeled by PCA on the mean shape coordinates. The shape parameters are locally searched using a linear regression function on the texture residual. In the past decade, various strategies [10, 24, 16] have been proposed for improving the performance of ASM and AAM. For instance, constrained local model (CLM) [9, 26, 30] extends ASM by modeling the non-rigid face as an ensemble of low dimensional independent patch experts. Due to the robustness of patch detectors to global illumination variation and occlusion, CLM have been widely used in detecting and tracking facial landmarks in challenging cases. Although the great flexibility in constraining facial landmarks, parametric shape models are difficult to optimize due to the non-convex nature of the problem. Therefore, the performance of most approximation methods (*e.g.*, the Lucas-Kanade method [1] and the Nelder-Mead simplex method [9]) largely depends on the effectiveness of the initialization step.

In the second case, the global layout of facial landmarks is constrained in a non-parametric manner. Among a number of non-parametric shape priors, Markov random field (MRF) [8] is perhaps the most natural way to govern the geometrical configuration of a point set. For instance, Liang *et al*. [23] proposed a constrained MRF by regularizing the shape with a PCA-based prior. Valstar *et al*. [32] combined the support vector regression with MRF to drastically reduce the time needed to search for point location. Unfortunately, globally optimizing MRF is intractable and thus much effort has focused on devising more accurate and efficient approximations. As an alternative choice to the graph-based MRF, tree-structured models [14] have been explored in detecting general object parts. A major benefit of using a tree model is the existence of an efficient dynamic programming algorithms [15] for finding globally optimal solutions. It has been recently discovered [13, 36, 31] that tree-structured models are surprisingly effective at capturing global elastic deformation of human faces. Both MRF and tree-structured models encode the shape in pair-wise geometric relations between parts. To leverage other relations, regression-based methods [4, 6, 11, 28, 35] directly predicts the shape parameters from the image.

The most relevant work to our method is the exemplar approach [2], where RANSAC was employed to efficiently sample exemplar shapes. A major limitation of [2] is that the position of each landmark is independently inferred by a greedy fusion procedure. In contrast, our method estimates all the landmarks jointly with a shape constraint learned online. We formulate this inference problem as graph matching and proposed an efficient solution based on linear programming. Our solution is globally optimal and satisfies the global shape constraints automatically. The shape constraints we use here are learned online from similar exemplars, hence they are adaptive to the pose of the test face.

3. Overview of the proposed system

In this section, we describe the proposed system for localizing facial landmarks. As shown in Fig. 2, our system consists of the following five steps.

Training: In the first offline step, we train individual landmark detectors based on support vector regressor (SVR). The positive and negative patches are sampled from the training images with manually labeled landmarks.

Sliding window: Given a test image, we run the detectors in a sliding-window manner to generate a response map for each landmark. For instance, Fig. 3b illustrates the response map for the mouth-top landmark.

RANSAC: We search for a set of similar exemplars in the training dataset to generate candidates positions for landmarks based on a RANSAC algorithm similar to [2]. Fig. 3c illustrates a subset of candidates generated for the test image. To augment the candidate set, we also included the top-five peak points from the response map.

Learning: Given the similar exemplars (Fig. 3d) by the RANSAC, we solve an efficient quadratic programming problem to obtain a shape constraint adaptively for the test face. This constraint is affine-invariant, making the system more robust to pose variation.

Matching: By combining the candidate position and the learned shape constraints, we solve an efficient graph matching problem based to find the optimal landmark positions using linear programming.

The last two steps of learning and matching are the main contributions of the proposed exemplar graph matching (EGM) algorithm. Due to the limited space, we skip the implementation details of first three steps, whose details can be found in the supplementary material.

In this paper, we adopt the annotation scheme used in the LFPW dataset [2], where each image is manually labeled

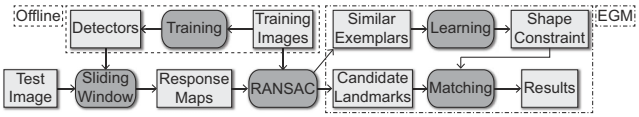
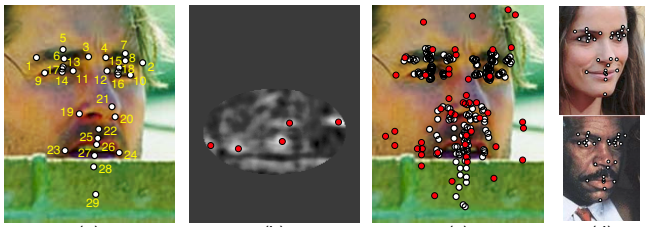


Figure 2. Pipeline of the proposed system for detecting facial landmarks.



(a) (b) (c) (d)

Figure 3. Example of response map and candidates generations. (a) A testing image labeled with 29 landmarks. (b) The response map for the mouth-top area, where the red circles are the top 5 peak points. (c) A $1/7$ portion of all the candidates. The white circles indicates the candidates that are transformed from exemplars, while red ones are the peaks of the response maps. (d) The top two exemplars found by the RANSAC step.

with 29 landmarks. See Fig. 3a for the landmark positions of an example image. Throughout the rest of the paper, we will denote (see notation[1]) the number of landmarks as k (*e.g.*, $k = 29$ in LFPW) and the number of exemplars as m. The coordinates of landmarks on the training image and test image are denoted as $\mathbf{p} \in \mathbb{R}^2$ and $\mathbf{q} \in \mathbb{R}^2$ respectively.

4. Exemplar-based graph matching

This section describes exemplar-based graph matching (EGM), the main component of our system for localizing facial landmarks. Given k sets of landmark candidates (Fig. 3c) and m exemplar faces (Fig. 3d), EGM aims to find the optimal subset of candidates in two steps: (1) learning an affine-invariant shape constraint online from the retrieved similar exemplars and (2) solving a graph matching problem to find the optimal candidates.

4.1. Learning

As mentioned before, use of a shape constraint is crucial for face alignment because the detector is usually not reliable and the local response may vary due to the change in pose and the existence of occlusions. A common choice of shape constraint is the point distribution model (PDM), in which the variances of facial landmarks are jointly modeled by a covariance matrix. However, there are two limitations in PDM: (1) it is sensitive to pose change; (2) it usually leads to a non-convex problem. To overcome these limitations, we adopt an affine-invariant shape constraint (AISC) originally proposed in [22] for object matching. Compared to PDM, AISC has two advantages: (1) the constraint is affine-invariant, making the system more robust to pose variation; (2) based on AISC, the matching step can be formalized as a graph matching problem, which can be efficiently solved by LP.

AISC relies on a similar geometric intuition used in the local linear embedding [29]. Suppose that a shape consists of k landmarks denoted by $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_k]$ and the c^{th} landmark $\mathbf{p}_c$ can be reconstructed by the linear combination of its neighbors as, $\mathbf{p}_c = \mathbf{P}\mathbf{w}_c$, where $\mathbf{w}_c \in \mathbb{R}^k$ denotes the weights of the other $k-1$ landmarks to reconstruct $\mathbf{p}_c$. Then the relation always holds, $\tau(\mathbf{p}_c) = [\tau(\mathbf{p}_1), \cdots, \tau(\mathbf{p}_k)]\mathbf{w}_c$, for any affine transformation $\tau(\mathbf{p}) = \mathbf{V}\mathbf{p} + \mathbf{b}$.

In this paper, we extend AISC for face alignment and we formalize the problem of learning $\mathbf{w}_c$ as follows. Recall that m exemplar faces, $\{\mathbf{P}^i\}_{i=1}^m$, are returned by the RANSAC step described in the supplementary material. Each exemplar consists of k landmarks, $\mathbf{P}^i = [\mathbf{p}_1^i, \cdots, \mathbf{p}_k^i]$, where $\mathbf{p}_c^i$ is the 2-D coordinate of the c^{th} landmark from the i^{th} exemplar. For each landmark $c \in \{1, \cdots, k\}$, we aim to find the optimal weight vector $\mathbf{w}_c$ that minimizes the sum of reconstruction errors:

$$\min_{\mathbf{w}_c \in \mathbb{R}^k} \quad \sum_{i=1}^{m} \|\mathbf{P}^i\mathbf{w}_c - \mathbf{p}_c^i\|_2^2 + \eta\|\mathbf{w}_c\|_2^2, \qquad (1)$$

$$\text{s.t.} \quad \mathbf{w}_c^T\mathbf{1}_k = 1 \text{ and } w_{cc} = 0,$$

where $\eta\|\mathbf{w}_c\|_2^2$ is a regularization term that penalizes the sparsity of the weight vector. In other words, we prefer the weight to be distributed uniformly across all the landmarks. This is beneficial especially in the case shown in Fig. 4a, where a large area around chin is occluded and few confident landmarks exist below the top of the mouth. By increasing η, larger weights could be assigned to non-local landmarks (*e.g.*, nose and eyes) that also carry important information to infer the position of the mouth-top landmark. In the extreme case, when $\eta \to \infty$, all landmarks are of equal importance. In the experiments, we found $\eta = 10^3$ produced consistently good results. After independently solving each of the k landmarks, we compose the joint weight matrix as, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k] \in \mathbb{R}^{k \times k}$. Eq. 1 is a convex quadratic problem in small size and the MATLAB QP solver can find $\mathbf{W}$ in less than one second.

4.2. Matching

Given the generated landmark candidate sets from the RANSAC step, we aim to select a single candidate for each landmark such that the corresponding global configuration best fits to the shape constraint $\mathbf{W}$ learned from the exemplars. This section proposes a graph matching algorithm to efficiently approximate this combinatorial problem.

[1]Bold capital letters denote a matrix $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$. $\mathbf{x}_i$ represents the i^{th} column of the matrix $\mathbf{X}$. x_{ij} denotes the scalar in the i^{th} row and j^{th} column of the matrix $\mathbf{X}$. All non-bold letters represent scalars. $\mathbf{1}_{m\times n}, \mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ are matrices of ones and zeros. $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix. $\|\mathbf{x}\|_p = \sqrt[p]{\sum_i |x_i|^p}$ and $\|\mathbf{X}\|_p = \sqrt[p]{\sum_{ij} |x_{ij}|^p}$ denote the p-norm for vector and matrix respectively.

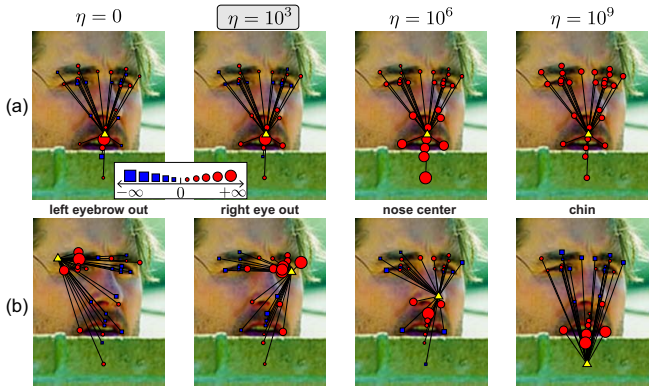


Figure 4. Visualization of weights for reconstructing landmark (yellow triangle). The size of landmark is proportional to its contribution in reconstruction. (a) Weights learned for the mouth-top landmark with different settings of η. (b) Weights learned for other landmarks using $\eta = 10^3$.

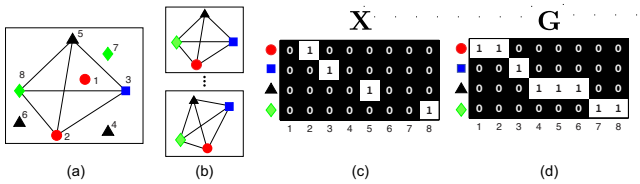


Figure 5. A synthetic example for graph matching. (a) Test image with 8 candidates for 4 landmarks. (b) Exemplar graphs. (c) The correspondence matrix ($\mathbf{X}$) for the matching defined in (a). (d) The landmark-candidate association matrix ($\mathbf{G}$), where each candidate (column) is only associated to one landmark (row).

To make the illustration more convenient, we introduce a global coordinate matrix $\mathbf{Q} = [\mathbf{Q}^1, \cdots, \mathbf{Q}^k] \in \mathbb{R}^{2\times n}$, where $\mathbf{Q}^c \in \mathbb{R}^{2\times n_c}$ denotes the candidates for c^{th} landmark and $n = \sum_{c=1}^{k} n_c$. Observe that each of the n candidate points is known to be associated with one of the k landmarks in the RANSAC searching. We encode this prior relation in a binary association matrix $\mathbf{G} \in \{0,1\}^{k\times n}$, where $g_{ci} = 1$ if the i^{th} point belongs to the c^{th} landmark. To simplify the discussion, let us consider a synthetic graph shown in Fig. 5a. Each of the 8 points can be considered as one facial landmark candidate, and each of the 4 colors denotes one landmark label. The landmark-candidate association is defined by the matrix shown in Fig. 5d. In addition, we denote the feature cost by $\mathbf{A} \in \mathbb{R}^{k\times n}$, where $a_{ci} = -\log(r_c(\mathbf{q}_i))$ indicates the cost of assigning i^{th} candidate point to c^{th} landmark. Please refer to the supplementary material for the details of computing $r_c(\mathbf{q}_i)$.

Given the candidates ($\mathbf{Q}, \mathbf{G}, \mathbf{A}$) and the shape constraint ($\mathbf{W}$), the problem consists of finding the optimal correspondence ($\mathbf{X}$) that minimizes the following error:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \lambda \operatorname{tr}(\mathbf{A}\mathbf{X}^T) + \|\mathbf{Q}\mathbf{X}^T(\mathbf{I}_k - \mathbf{W})\|_1, \qquad (2) \\ \text{s.t.} \quad & \mathbf{X}\mathbf{1}_n = \mathbf{1}_k, \mathbf{X} \in \{0,1\}^{k\times n}, \\ & x_{ci} = 0, \quad [c,i] \in \{[c,i] | g_{ci} = 0\}, \end{aligned}$$

where the second term in the objective measures the self-reconstruction error ($\|\mathbf{Y}(\mathbf{I}_k - \mathbf{W})\|_1 = \|\mathbf{Y} - \mathbf{Y}\mathbf{W}\|_1$) of the k selected candidates ($\mathbf{Y} = \mathbf{Q}\mathbf{X}^T \in \mathbb{R}^{2\times k}$) with respect to the shape constraint ($\mathbf{W}$). Instead of using an l_2 norm, the reconstruction error is defined in l_1 because of its efficiency and robustness. λ is a regularization weight to trade-off between the feature cost and the reconstruction error. In the experiment, we always set $\lambda = 100$ and we found the final result was not sensitive to small change of this weight. The first constraint enforces $\mathbf{X}$ to be a many-to-one mapping. According to the second constraint, each row of $\mathbf{X}$ can only select an optimal candidate from the corresponding candidate set defined by $\mathbf{G}$. For instance, Fig. 5c illustrates the optimal $\mathbf{X}$ for a synthetic problem.

Due to the integer constraint on $\mathbf{X}$, optimizing Eq. 2 is NP-hard. To approximate the problem, we relax the integer constraint with a continuous one, $\mathbf{X} \in [0,1]^{k\times n}$. Unfortunately, the presence of the non-smooth l_1 norm ($\|\cdot\|_1$) in Eq. 2 makes it impossible to directly apply LP. Therefore, we re-formulate the problem using the trick [18, 22] as:

$$\min_{\mathbf{X},\mathbf{U},\mathbf{V}} \lambda \operatorname{tr}(\mathbf{A}\mathbf{X}^T) + \mathbf{1}_2^T(\mathbf{U} + \mathbf{V})\mathbf{1}_k, \qquad (3)$$

$$\text{s.t. } \mathbf{Q}\mathbf{X}^T(\mathbf{I}_k - \mathbf{W}) = \mathbf{U} - \mathbf{V}, \mathbf{U} \geq \mathbf{0}_{2\times k}, \mathbf{V} \geq \mathbf{0}_{2\times k}, \qquad (4)$$

$$\mathbf{X} \in [0,1]^{k\times n}, \; x_{ci} = 0, \; [c,i] \in \{[c,i] | g_{ci} = 0\}, \qquad (5)$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2\times k}$ are two auxiliary variables introduced for replacing the non-smooth l_1 norm with a smooth term and the linear constraint defined in Eq. 4.

Although both the objective and constraints are linear, a direct LP solution would still be slow because the number of variables is $O(kn)$. Inspired by the idea of lower convex hull proposed in [18], we simplify the optimization task by removing the ineffective x_{ci}s. Fortunately, due to the special structure defined on $\mathbf{X}$ by Eq. 5, we can more easily identify the x_{ci}s of interest by checking whether $g_{ci} = 1$ or not. As a result, the number of variables in the reduced LP is proportional to the number of non-zero elements in $\mathbf{G}$, *i.e.*, $O(n)$. After the reduction, MATLAB LP solver can find the optimal solution in less than one second for a large-scale problem, where $n > 3000$ and $k = 29$.

Observe that an integer rounding step is necessary to discretize the continuous $\mathbf{X}$. Similar to [18], we gradually make $\mathbf{X}$ to be discrete by taking a successive refinement. More specifically, a trust region centered at $[\mathbf{Q}\mathbf{X}^T]_i \in \mathbb{R}^2$ is initialized for each landmark. We gradually shrink the size of the trust region and remove the candidates ($\mathbf{q}_i$s) outside the region. This procedure was repeated five times in the experiment. Finally, we optimize Eq. 2 by ICM [3] to discretize $\mathbf{X}$. See [18] for more details about the trust-region shrinking and ICM.

4.3. Difference from [2] and [22]

The proposed EGM is similar to [2] in the RANSAC step for candidate generation. However, EGM significantly dif-

fers from [2] in the step of inferring the final landmark positions. In [2], the final position of each landmark is independently obtained by a weighted averaging of the candidate points. This greedy approach is sensitive to the outliers existed in the exemplar and candidate set. In contrast, EGM jointly infers the position for all the landmarks by solving a graph matching problem with an affine-invariant shape constraint learned online from similar examplars. Due to the robustness of the shape constraint and the effectiveness of the graph matching step, EGM obtained much more accurate landmarks than [2] did in all the experiments.

Although similar in spirit, our shape constraint differs from [22] in three important aspects: (1) In [22], the weights are learned from a single exemplar. Without sufficient constraints, however, there are infinite choices of weights for reconstructing one landmark by more than 3 neighbors. In addition, the weights learned from one exemplar might not generalize well to a non-rigidly deformed face (*e.g.*, expression). In contrast, Eq. 1 is designed to jointly optimize $\mathbf{w}_c$ for multiple exemplars. The $\mathbf{w}_c$ is not only unique but also more robust in capturing various non-rigid facial poses contained in the exemplar set. (2) The object shape in [22] is represented by a sparse graph, where the landmark is influenced by its nearby neighbors. However, in many cases, the local structure of a landmark can be distorted by noise and occlusion. Instead of using a sparse graph, our model adopts a fully-connected graph. (3) With proper regularization, all the other $k-1$ landmarks make important contribution in the reconstruction of each landmark. Therefore, EGM is less susceptible to local noise and occlusion than [22].

5. Experiments

This section compares EGM against several state-of-the-art algorithms on three public datasets.

5.1. LFPW dataset

The LFPW dataset [2] consists of images downloaded from internet and the images contain a wide range of poses, lighting conditions and facial expressions. The original dataset contained 1132 training images and 300 test images. Unfortunately, many URLs have become expired and we were only able to download 868 images for training and 228 images for testing.

According to [2], the bounding box of labeled faces were given by a commercial face detector. To mimic the experimental setting, we initialized the face bounding box estimated from the ground-truth landmarks. For instance, Fig. 6a shows the results of EGM on some example faces cropped by the estimated bounding box. Observe that we did not specifically constrain the scale of the bounding box in our system and in general any face detector (*e.g.*, OpenCV) is suitable for initializing EGM. Unlike conventional iterative algorithms (*e.g.*, ASM) depending on a good initialization, EGM computes the facial landmarks by directly solving a combinatorial problem. Due to this reason, EGM achieves great stability with respect to the position of the bounding box even if applying some spatial or scale perturbation around this box.

To establish a baseline, we implemented the consensus of exemplar method proposed in [2]. To be fair in comparison, we fixed the parameter setting in the RANSAC step and used the same set of candidates and exemplar images for both EGM and [2]. Fig. 6b shows the quantitative comparison between EGM and [2]. Overall, EGM improved [2] in localizing all of 29 landmarks. In particular, EGM outperformed [2] by a large margin in the landmarks around the nose tip ($19 \sim 21$) and the chin ($27 \sim 29$), where appearance features are frequently unreliable. For these landmarks, the geometrical support from non-local parts becomes more crucial. Because of the affine-invariant shape constraint and the global LP-based optimization, EGM more robustly handles these areas than the greedy fusion method proposed in [2]. For the landmarks around eyebrows ($1 \sim 8$), our method outperformed [2] by a small margin. This is because for these landmarks, the detectors play a more important role than the shape constraint. Due to the less available training data, we were not able to train the detectors with same accuracy and reproduce the results in [2]. We expect that training the landmark detectors using the author's original training data would further boost the performance of our method significantly.

Fig. 6c shows the time cost of our system. The most expensive step was the RANSAC, taking about 15 secs for each image. However, this step can be largely sped up in a parallelized implementation, which is shared by [2]. The second step of learning the affine invariant constraint was very efficient since it solved 29 independent small-size QP problems. Based on the Matlab function $linprog$, the matching step took 9 secs for selecting the optimal landmarks from more than 3000 candidates. Recall that we repeated to solve 5 linear programming problems to successively discretize $\mathbf{X}$. Therefore, each linear programming was taking less 2 seconds. In addition, this step can be largely sped up using a more efficient LP solver.

5.2. BioID dataset

The BioID dataset [17] contains 1521 images of the frontal faces of 23 different subjects. In our experiment, we trained our landmark detectors on the LFPW dataset and tested EGM on all the 1521 images. Fig. 7a shows the result of running EGM on some images. To evaluate the result, we used 17 landmarks marked for the FGNet project, and used in the me_{17} error measure as defined in [9]. Following the common protocol used in [2, 4], we computed for each landmark a fixed offset by exhaustively matching with the ground-truth label. This offset was fixed for each

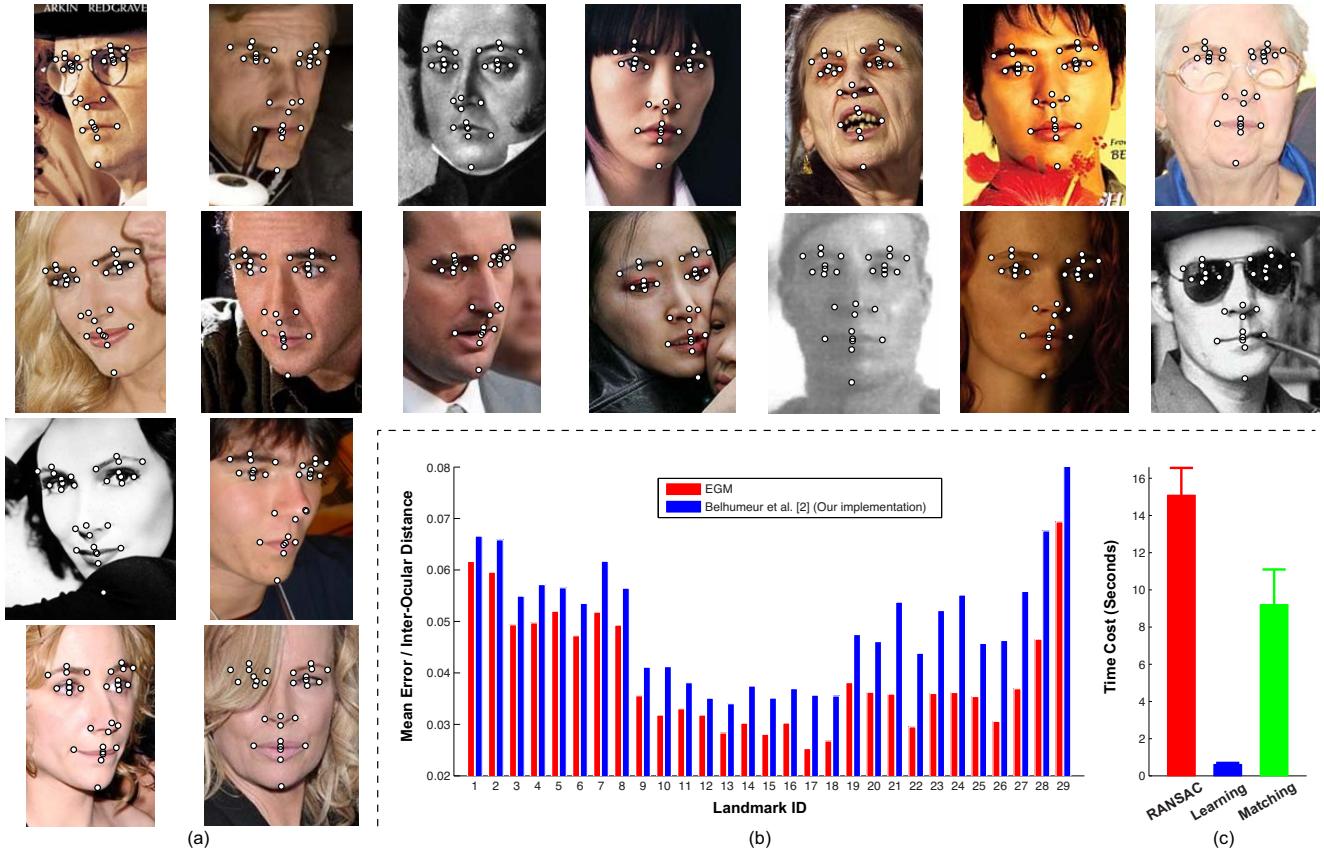


Figure 6. Comparison on the LFPW dataset. (a) Results of EGM on example faces. (b) Mean errors of 29 individual landmarks. Note that since only about 75% of original training data are available, our own implementation of [2] is worse than the result of [2] reported in their paper. (c) Time cost of each step in our implementation of EGM based on Matlab.

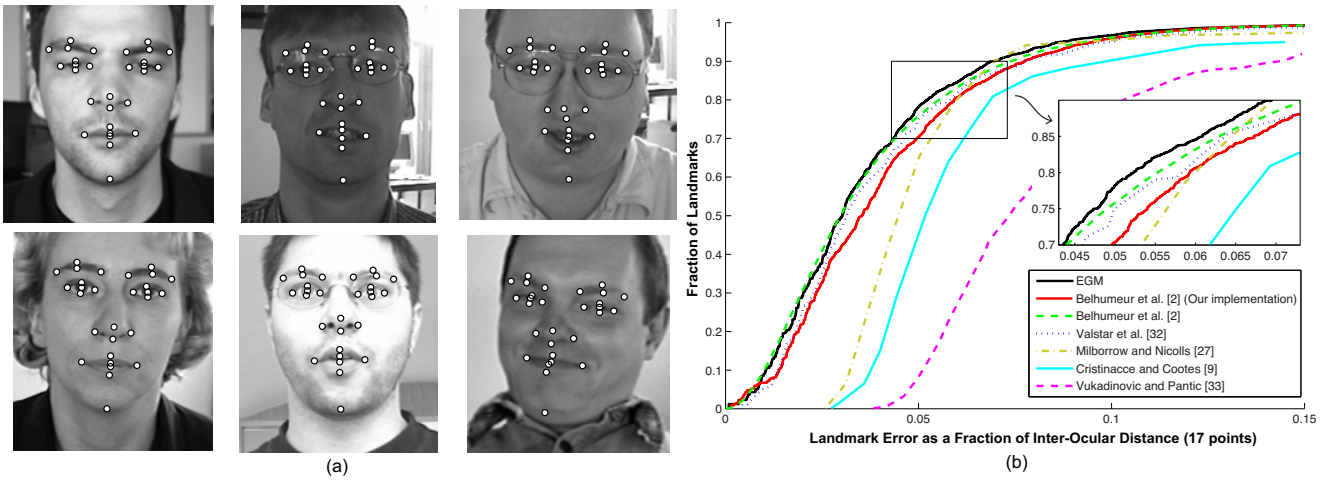


Figure 7. Comparison on the BioID dataset. (a) Result of EGM on example faces. (b) Cumulative error curves.

testing image. This offset is necessary to accommodate the different annotation schemes adopted by BioID and LFPW.

In the past decade, the BioID dataset has been widely used as a benchmark for evaluating face alignment algorithms. A number of previous methods have reported their performance on this dataset. Fig. 7b compares the performance of EGM with other five state-of-the-arts methods [2, 9, 27, 32, 33]. The results for these methods were taken from the paper [2]. Even though we have less training data than [2], our method still outperformed the exemplar method [2] as well as the other four. We noticed that the improvement is marginal. This is because the performance

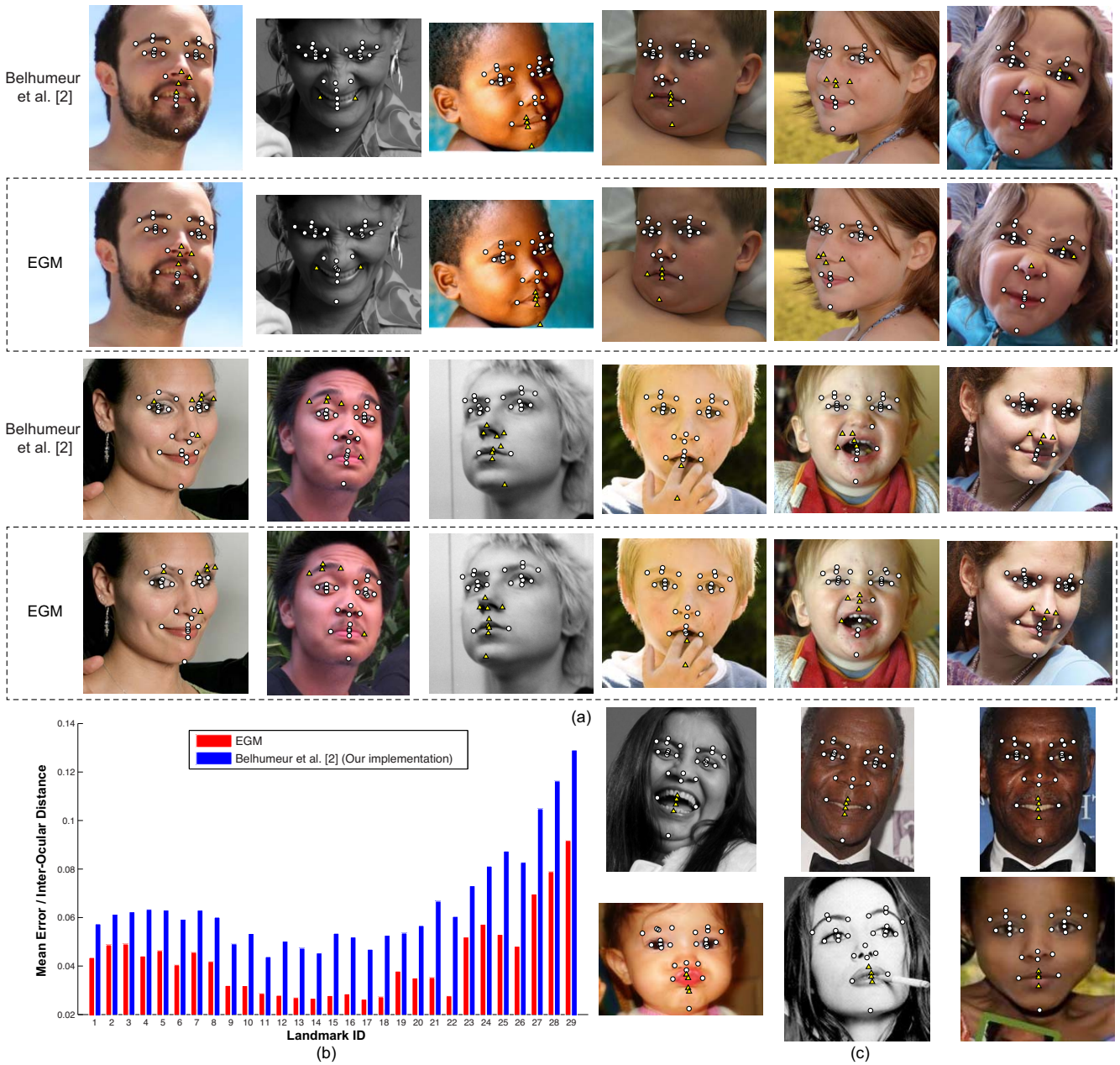


Figure 8. Comparison on the Helen dataset. (a) Results of EGM and the exemplar approach [2] on example faces, where the landmarks denoted as yellow triangles are the ones largely improved by EGM. (b) Cumulative error curves. (c) Two worse examples (in the 1^{st} column), where EGM cannot accurately locate the landmarks denoted as yellow triangles because of very few exemplar images available in LFPW with similar exaggerated expressions. For instance, the top-two most similar exemplars are shown in the 2^{nd} and 3^{rd} columns.

on BioID is nearly saturated due to its simplicity. We also reported the results of [2] implemented by ourselves. With the same set of exemplars and detectors, EGM greatly improved the greedy fusion step proposed in [2].

5.3. Helen dataset

The Helen dataset was created by the authors of [21]. This dataset consists of high-resolution images containing large variations in pose, illumination, expression and occlusion. Similar to BioID, we trained our landmark detectors on the LFPW dataset. Helen dataset adopts a highly detailed annotation that is quite different from LFPW. To report a quantitative result, we re-labeled[2] 348 images with the same 29 landmarks as LFPW. We compared EGM with [2]. In order to make a fair comparison, we fixed the RANSAC step and used the same detectors for both methods.

Fig. 8a compares EGM with [2] on several challenged

[2]Available at http://www.f-zhou.com/fa.html

examples. Our method was much more accurate than [2] in detecting facial landmarks (especially the yellow triangles) in challenging images with large variation in pose and expressions. Fig. 8b reports a qualitative comparison, where EGM always achieves the best performance with a large margin. Observe that the relative improvement over [2] is greater than Fig. 6b achieved on the LFPW dataset, presumably because the much more challenges exist in Helen dataset. The result clearly illustrates the benefit of using the proposed graph matching method with affine-invariant shape constraints over the greedy fusion method proposed in [2]. However, EGM performed worse in some images as shown in Fig. 8c with extreme facial expressions. One of the main reasons is due to the limited exemplars available in the LFPW dataset we used as the training set. For instance, the second and third columns in Fig. 8c shows the top-two most similar exemplar images found by RANSAC. With only few similar exemplars, it is very difficult to learn a shape constraint particularly for the mouth of the test image shown in the first column.

6. Conclusions

This paper presents exemplar-based graph matching (EGM), a robust framework for facial landmark localization. Compared to conventional algorithms, the proposed EGM framework has two advantages: (1) the facial shape is enforced by an affine-invariant shape constraint learned online from multiple exemplars for better adaption to the test image; (2) the optimal landmark configuration is obtained by solving an LP-based graph matching problem.

Our experiments have demonstrated these advantages in terms of quantitative comparisons to state of the art. However, we also found that the performance of EGM is directly affected by the quality of the exemplars. Therefore, we conjecture that we can improve EGM by clustering facial shapes to refine the exemplars returned by the RANSAC. In addition, the RANSAC step may be further improved by a component-based facial part matching instead of the matching between entire faces. Since EGM is a general framework, with applicability beyond faces, we are interested in evaluating its performance in other domains, such as body part detection, and human pose estimation.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Comput. Vis.*, 56(3):221–255, 2004. 2

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011. 2, 4, 5, 6, 7, 8

[3] J. Besag. On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Series B Stat. Methodol.*, pages 259–302, 1986. 4

[4] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012. 2, 5

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001. 2

[6] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*, 2012. 2

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995. 1, 2

[8] J. M. Coughlan and S. J. Ferreira. Finding deformable shapes using loopy belief propagation. In *ECCV*, 2002. 2

[9] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006. 2, 5, 6

[10] D. Cristinacce and T. F. Cootes. Boosted regression active shape models. In *BMVC*, 2007. 2

[11] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012. 2

[12] M. Eckhardt, I. R. Fasel, and J. R. Movellan. Towards practical facial feature detection. *IJPRAI*, 23(3):379–400, 2009. 2

[13] M. Everingham, J. Sivic, and A. Zisserman. "Hello! my name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006. 2

[14] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 2

[15] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vis.*, 61(1):55–79, 2005. 2

[16] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *ECCV*, 2008. 2

[17] O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the Hausdorff distance. In *AVBPA*, 2001. 5

[18] H. Jiang, M. S. Drew, and Z.-N. Li. Matching by linear programming and successive convexification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):959–975, 2007. 4

[19] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, and S. M. Seitz. Being John Malkovich. In *ECCV*, 2010. 1

[20] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, 2011. 1

[21] V. Le, J. Brandt, L. Bourdev, Z. Lin, and T. Huang. Interactive facial feature localization. In *ECCV*, 2012. 7

[22] H. Li, X. Huang, and L. He. Object matching using a locally affine invariant and linear programming techniques. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(2):411–424, 2013. 3, 4, 5

[23] L. Liang, F. Wen, Y.-Q. Xu, X. Tang, and H.-Y. Shum. Accurate face alignment using shape constrained Markov network. In *CVPR*, 2006. 2

[24] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008. 2

[25] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. De la Torre, and J. Cohn. AAM derived face representations for robust facial action recognition. In *FG*, 2006. 1

[26] S. Lucey, Y. Wang, J. M. Saragih, and J. F. Cohn. Non-rigid face tracking with enforced convexity and local appearance consistency constraint. *Image Vision Comput.*, 28(5):781–789, 2010. 2

[27] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, 2008. 6

[28] S. Rivera and A. M. Martinez. Learning deformable shape manifolds. *Pattern Recognit.*, 45(4):1792–1801, 2012. 2

[29] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 3

[30] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.*, 91(2):200–215, 2011. 2

[31] M. Uricar, V. Franc, and V. Hlavác. Detector of facial landmarks learned by the structured output svm. In *VISAPP*, 2012. 2

[32] M. F. Valstar, B. Martínez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010. 2, 6

[33] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In *ICSMC*, 2005. 6

[34] L. Wiskott, J.-M. Fellous, N. Krüger, and C. Von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779, 1997. 1

[35] X. Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *CVPR*, 2013. 2

[36] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2