Pattern Recognition 60 (2016) 318–333

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Face alignment by robust discriminative Hough voting

Xin Jin [a,b], Xiaoyang Tan [a,b,*]

[a] Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, #29 Yudao Street, Nanjing 210016, PR China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form
8 April 2016
Accepted 7 May 2016
Available online 24 May 2016

Keywords:
Face alignment
Hough voting
Constrained Local Models

ABSTRACT

This paper presents a novel Hough voting-based approach for face alignment under the extended exemplar-based Constrained Local Models (CLMs) framework. The main idea of the proposed method is to use very few stable facial points, i.e., anchor points, to help reduce the ambiguity encountered when localizing other less stable facial points by Hough voting. A less studied limitation of Hough voting-based methods, however, is that their performance is typically sensitive to the quality of anchor points, especially when only very few (e.g., one pair of) anchor points are used. In this paper, we mainly focus on this issue and our major contributions are three-fold: (1) We first propose a novel method to evaluate the goodness of anchor points based on the diagnosis of resulted distribution of their votings for other facial points; (2) To deal with the remaining small localization errors, an enhanced RANSAC method is presented, in which a sampling strategy is adopted to soften the range of possible locations of the chosen anchor points, and the top ranking exemplars are then selected based on a newly-proposed cost-sensitive discriminative objective; (3) Finally, both global voting priors and local evidence are fused under a weighted least square framework. Experiments on several challenging datasets, including LFW, LFPW, HELEN and IBUG, demonstrate that the proposed method outperforms many state-of-the-art CLM methods. We also show that the performance of the proposed system can be further boosted by exploring the deep CNN technique in the RANSAC step.

1. Introduction

Localizing feature points in face images, which is usually called face alignment, is crucial for many face-related applications, such as face recognition, gaze detection and facial expression recognition. Given an image with faces detected, face alignment is usually performed and serves as an important and essential intermediary step for the subsequent automatic facial analysis. As the web and personal face photos explosively increase nowadays, an accurate and efficient face alignment system is in demand.

However, such a task is very challenging due to the complexity of appearance variations possibly exhibited in the patch centered at each facial point, caused by the change of lighting, pose, occlusion, expression and so on. Numerous approaches have been proposed in recent decades, among which a popular and very successful approach is the family of methods coined Constrained Local Models [1] that independently train a specific local detector for each feature point, and use a parameterized shape model to regularize the detection of these local detectors.

In this paper, we extend the range of CLM framework by including the nonparametric (exemplar-based) shape models, with a novel Hough voting-based approach to improve the efficiency and accuracy of feature points localization. The main idea of our method is to first localize very few stable facial points (priority is given to eyes in our implementation)[1] from the given face image, then use the locations of these to help reduce the ambiguity encountered when locating other less stable facial points by Hough voting.

However, a less studied limitation of Hough voting-based methods is that their performance is typically sensitive to the quality of anchor points. To address this issue, most of the aforementioned models are either complex in inference or using many anchor points so as to provide reliable constraints. For example, seven anchor points are needed in [3] and eleven in [4], while Belhumeur et al. [5] choose to sample anchor points randomly from peak points of the response maps, for which an exhaustive search by local detectors must be performed first. Furthermore, while Asthana et al. [6] propose a discriminative regression method to fit the parameterized (PCA) shape model within the

* Corresponding author.
E-mail address: x.tan@nuaa.edu.cn (X. Tan).

[1] We locate the eyes using the method introduced in [2] (with codes provided by the authors).

http://dx.doi.org/10.1016/j.patcog.2016.05.017
0031-3203/

CLM framework, to our knowledge, the power of discriminative learning beyond local detector training has not been exploited for the nonparametric CLM framework.

Considering these and starting from the work of Belhumeur et al. [5], the goal of this paper is mainly to develop a novel and efficient approach to incorporate the prior knowledge of very few anchor points into the task of face alignment, while being robust against localization errors of anchor points. Furthermore, inspired by Asthana et al. [6], we aim to improve the fitting results of the extended exemplar-based CLM framework by exploiting the power of discriminative learning. Due to these, we refer the proposed method as Robust Discriminative Hough Voting (RDHV) method.

In particular, this paper is an extension of our previous work [7]. Compared to [7], this paper has two novel contributions, mainly focusing on inaccurate anchor points handling. Firstly, we propose to evaluate the goodness of anchor points based on the diagnosis of resulted distribution of their votings for other facial points, which effectively helps to avoid using anchor points with large localization errors. Secondly, to deal with the remaining small localization errors, an enhanced RANSAC method is presented, in which a sampling strategy is adopted to soften the range of possible locations of the chosen anchor points, and the top ranking exemplars are then selected based on a newly-proposed cost-sensitive discriminative objective. Furthermore, to mitigate the problem of ambiguity of local detectors, a deep CNN-based score function is also proposed. Finally, following [7], both global voting priors and local evidence are fused under a discriminative weighted least square framework. We show that the proposed method outperforms many state-of-the-art CLMs methods on several challenging face alignment datasets (c.f., Fig. 1). We also show that the deep CNN-based score function for selecting top ranking exemplars can significantly boost the performance of the proposed system.

The paper is structured as follows. The next section discusses the background and related work. After that, we describe our robust discriminative Hough voting method and how to use it for face alignment in Section 3. Section 4 gives some implementation details and Section 5 shows experimental results on four publicly available datasets. A final discussion in Section 6 concludes our work.

2. Background

Large amount of research has resulted in significant progress for face alignment over the last decades [8,1,5,6,9–15]. According to whether a method designs a special local model (detector, regressor or part template) for each feature point and use it for independent prediction or matching, we roughly divide the face alignment methods into two categories, i.e., *holistic methods* and *local methods*.

In the following, we first briefly introduce the holistic methods as well as some notable examples. Then we focus on the local methods, especially the exemplar-based CLMs that our method belongs to. Last but not least, we present some discussions about exemplar-based CLMs, which motivate this work.

2.1. Holistic methods

The common characteristic of holistic methods is that they consider all the feature points as a whole, rather than treat them as conditionally independent. The most well-known holistic methods are the Active Appearance Models (AAMs), which simultaneously models the intrinsic variation in both appearance and shape as a linear combination of basis models of variation. The shape updating of AAMs, i.e., the fitting procedure, is usually a function of the error between the warped image and the model instantiation, measured in a canonical reference frame. To tackle the AAM fitting problem, both generative [8] and discriminative [16,17] strategies have been developed and have obtained the certain success.

Among others, Explicit Shape Regression (ESR) [9], Supervised Descent Method (SDM) [11], Ensemble of Regression Trees (ERT) [18] and Local Binary Feature (LBF) [13] are four representative state-of-the-art holistic methods in face alignment. All of them are performed under the cascaded shape regression framework using shape-indexed features. ESR directly learns a regression function to infer the shape from a sparse subset of pixel intensities indexed relative to current shape estimate, while ERT substitutes the weak fern regressor in ESR with a regression tree which further

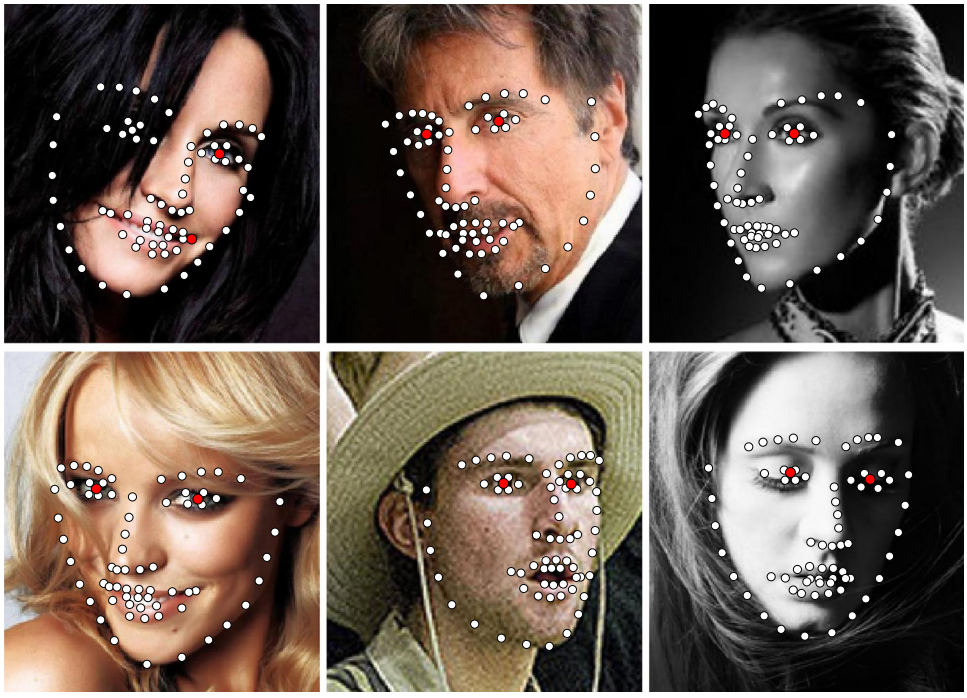

Fig. 1. Some aligned images from the IBUG dataset by the proposed system, where the red filled dots denote the pre-located anchor points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

improves the performance. SDM employs a cascaded linear regression to estimate the shape based on hand-designed SIFT feature, while LBF learns a set of highly discriminative local binary features for each feature point independently, and then uses the learned features jointly to learn a linear regression for the final prediction, which is highly efficient and achieves very accurate performance.

2.2. Local methods

Local methods follow the strategy of dividing and conquering, independently training a special local model (detector, regressor, or part template) for each feature point. These pre-trained local models are used to make independent prediction or matching for each feature point, which are then used in the optimization of the global shape model. Notable local methods include ASM [19], CLM [1], LEAR [20], Tree Structure Part Model (TSPM) [10] and CoE [5] to name a few.

Since space is limited, here we focus on local detector based methods. The seminal work by [1] refers these methods, which usually contain two parts, i.e., local detectors and parametric shape model, collectively as constrained local models (CLMs). However, many nonparametric shape models also achieved promising performance in face alignment, including the Markov random field model [20], the tree-structured model [10] and the exemplar model [5,21,22]. In what follows, we extend the range of CLMs by unifying both the parametric (PCA-based) and nonparametric (exemplar-based) shape models into a generic probabilistic CLM framework.

2.2.1. A generic probabilistic CLM framework

To begin, we first introduce some notations. Given an image $\mathcal{I}$, the task of face alignment is to locate I facial feature points $\mathbf{X}=(\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^I)^T$ on the 2D image, where $\mathbf{x}^i=(x^i, y^i)^T$, denoting the x, y-coordinates of the ith facial feature point. Let $l^i \in \{1, -1\}$ be an indicator variable that denotes whether the ith feature point is aligned ($l^i=1$) or misaligned ($l^i=-1$). Our goal is to find a face shape $\mathbf{X}$ that maximizes the probability of its points corresponding to consistent locations of the facial features, i.e.,

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} P(\mathbf{X}|\{l^i=1\}_{i=1}^I, \mathcal{I}). \tag{1}$$

CLMs assume that the face shape $\mathbf{X}$ is managed or generated by a shape model. Let the hidden variable $\mathbf{s}$ denote the shape constraints imposed by the shape model, e.g., parameters in a parametrized shape model or some non-parametric shape constraints such as the exemplar shape. Then, the posterior (1) for CLMs can be expanded as follows:

$$\begin{aligned}\mathbf{X}^* &= \arg\max_{\mathbf{X}} \int_{\mathbf{s}} P(\mathbf{X}, \mathbf{s}|\{l^i=1\}_{i=1}^I, \mathcal{I})d\mathbf{s} \\ &= \arg\max_{\mathbf{X}} \int_{\mathbf{s}} P(\mathbf{X}, \mathbf{s})P(\{l^i=1\}_{i=1}^I|\mathbf{X}, \mathbf{s}, \mathcal{I})d\mathbf{s} \\ &= \arg\max_{\mathbf{X}} \int_{\mathbf{s}} P(\mathbf{X}|\mathbf{s})P(\mathbf{s})d\mathbf{s} \prod_{i=1}^I P(l^i=1|\mathbf{x}^i, \mathcal{I}).\end{aligned} \tag{2}$$

where the first item is the prior *shape model* and the second item is the *global likelihood*. To better exposition in the following sections, we let $S(\mathbf{X})$ and $L(\mathbf{X})$ denote them respectively,

$$S(\mathbf{X}) = \int_{\mathbf{s}} P(\mathbf{X}|\mathbf{s})P(\mathbf{s})d\mathbf{s}, \tag{3}$$

$$L(\mathbf{X}) = \prod_{i=1}^I P(l^i=1|\mathbf{x}^i, \mathcal{I}), \tag{4}$$

where the likelihood of $\mathbf{x}_i$ that the ith feature point is correctly aligned, i.e., $P(l^i=1|\mathbf{x}^i, \mathcal{I})$, is fit by the output of the local detector for the ith feature point.

Next, we focus on the shape model (3), and illustrate how two typical shape models: parameterized PCA model and nonparametric exemplar model, can be derived from the unified Bayesian framework (2), by realizing the hidden variable $\mathbf{s}$ in different ways.

2.2.2. PCA-based CLMs

PCA model is the most widely used shape model for CLMs [23,16,1,6], which models the non-rigid shape variations linearly:

$$\mathbf{X} = s\mathbf{R}(\bar{\mathbf{X}} + \mathbf{\Phi}\mathbf{q}) + \mathbf{t}, \tag{5}$$

where $\mathbf{R}$, s and $\mathbf{t}$ control the rigid rotation, scale and translations respectively while $\mathbf{q}$ controls the non-rigid variations of the shape and $\mathbf{\Phi}$ denote the matrix of the basis of variations. Then all the parameters of the shape model can be denoted as $\mathbf{p}=\{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$, where the rigid transformation parameter $\mathbf{q}$ is often assumed to exhibit a Gaussian distribution while the rigid transformation parameters s, $\mathbf{R}$ and $\mathbf{t}$ that place the model in the image are all assumed uniform distributions. *Since the PCA-based CLMs reconstruct the face shape* $\mathbf{X}$ *from the parameters* $\mathbf{p}$, *the hidden variable* $\mathbf{s}$ *in the generic shape model* (3) *is equivalent to* $\mathbf{p}$ *in the PCA model.*

The objective of PCA-based CLMs is to optimize the shape parameters $\mathbf{p}$ such that the locations of the face shape $\mathbf{X}$ reconstructed from $\mathbf{p}$ correspond to well-aligned parts on the image. We substitute the variable $\mathbf{X}$ for optimization in the generic CLM framework (2) with the parameters $\mathbf{p}$, and let $\mathbf{x}^i=\mathbf{x}^i(\mathbf{p})$, then we can derive the formulation of PCA-based CLMs as follows:

$$\begin{aligned}\mathbf{p}^* &= \arg\max_{\mathbf{p}} P(\mathbf{p}|\{l^i=1\}_{i=1}^I, \mathcal{I}) \\ &= \arg\max_{\mathbf{p}} P(\mathbf{p}) \prod_{i=1}^I P(l^i=1|\mathbf{x}^i(\mathbf{p}), \mathcal{I})\end{aligned} \tag{6}$$

For the optimization of $\mathbf{p}$ in Eq. (16), we refer the reader to [1], which unifies various CLM optimization approaches that differ from each other in the ways that responses of local detectors are used for the optimization of the global shape model.

Although PCA shape model is widely used, the formulation based on these models is non-convex, and hence, they are sensitive to initialization in general and are prone to local minima. [6] follows alternative direction by proposing a novel discriminative regression based approach under the CLM framework, resulting in significant improvement in performance. This reveals that the discriminative learning process is important under the CLM framework.

2.2.3. Exemplar-based CLMs

The exemplar shape model is proposed by *Belhumeur* et al. [5], originally named as consensus of exemplars (CoE), which assumes that the face shape $\mathbf{X}$ in the test image is generated by one of the transformed exemplar shapes (global models). *So, the hidden variable* $\mathbf{s}$ *in the generic shape model* (3) *is equivalent to the global model in the exemplar-based shape model.* Recently there have been many extensions of [5], e.g, the exemplar-based graph matching method [21] and the joint non-parametric face alignment method [22]. But in the following we will focus on the seminal work of [5], showing that its formulation can also be naturally derived from the generic CLM framework (2).

To keep the notations consistent with [5], we let $\mathbf{X}_{k,t}$ ($k=1, \ldots, K$) denote locations of all feature points in the kth of the K exemplars that transformed by some similarity transformation t, and let $\mathbf{x}_{k,t}^i$ denote location of the ith feature points of the transformed exemplar $\mathbf{X}_{k,t}$. [5] refers $\mathbf{X}_{k,t}$ as a global model, and

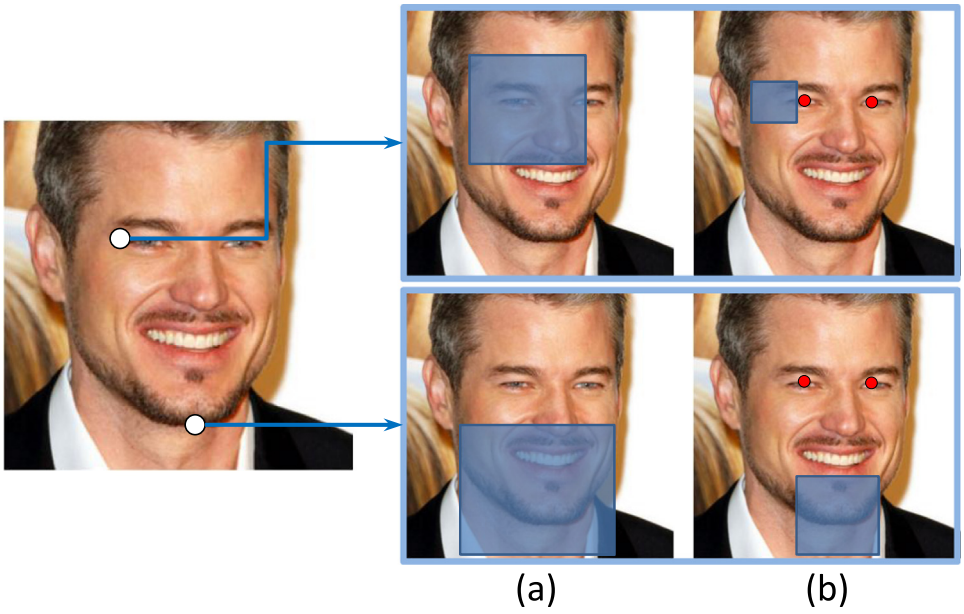

Fig. 2. Comparison of the size of searching windows for different facial points used by various methods: (a) searching regions of traditional exemplar-based CLMs; (b) searching windows defined by the votes casted by the global model candidates transformed from eyes and all exemplar shapes. The size of searching windows used in our system (b) is about 1/5 of that of traditional methods (a).

assumes that conditioned on the global model $\mathbf{X}_{k,t}$, the location of each feature point $\mathbf{x}_i$ is conditionally independent of one another. Then, by substituting the hidden variable $\mathbf{s}$ in (3) with $\mathbf{X}_{k,t}$, we derive the exemplar-based shape model as follows:

$$\begin{aligned} S(\mathbf{X}) &= \sum_{k=1}^{K} \int_{t\in T} P(\mathbf{X}, \mathbf{X}_{k,t}) dt \\ &= \sum_{k=1}^{K} \int_{t\in T} \prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,t}) P(\mathbf{X}_{k,t}) dt, \end{aligned} \tag{7}$$

where $P(\mathbf{x}^i|\mathbf{x}^i_{k,t})$ is modeled as a Gaussian distribution centered at $\mathbf{x}^i_{k,t}$ with the covariances calculated by the exemplars (c.f., [5] for details), and the prior of the global model $P(\mathbf{X}_{k,t})$ is assumed as an uniform distribution.

Combining (2), (3) and (7) yields the objective function of [5] (little difference in notations) as follows:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \sum_{k=1}^{K} \int_{t\in T} \prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,t}) P(l^i = 1|\mathbf{x}^i, \mathcal{I}) dt. \tag{8}$$

To optimize (8), the first step is to retrieve global models $\mathbf{X}_{k,t}$. For this, a RANSAC-like approach is adopted by randomly generating a large number of global models $\mathbf{X}_{k,t}$, evaluating them by setting appropriate value to $\mathbf{X}$ and calculating the value of the objective function (8), and then choosing the top global models.

For given $\mathbf{X}_{k,t}$, the simplest way to maximize (8) is to set $\mathbf{X} = \mathbf{X}_{k,t}$, which maximize $\prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,t})$ in (8). In this setting, the global models $\mathbf{X}_{k,t}$ can be conveniently evaluated by the *global likelihood* (4) $L(\mathbf{X}_{k,t})$, i.e., $\prod_{i=1}^{I} P(l^i = 1|\mathbf{x}^i_{k,t}, \mathcal{I})$. Then, the set $\mathcal{M}$ of m^* top ranking global models (i.e., with large $L(\mathbf{X}_{k,t})$ values), are used to approximate (8) as follows:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \sum_{k,t\in\mathcal{M}} \prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,t}) P(l^i = 1|\mathbf{x}^i, \mathcal{I}). \tag{9}$$

It is worth noting that Simith et al. [24] use a generalized Hough transform framework to score each exemplar image and choose top exemplars, and Yan et al. [25] propose to combine multiple shape hypotheses with a learned score function. However, both of them do not use local detectors, and can not be formulated under the CLM framework.

Based on the sampled m^* global models, the location of each feature point $\mathbf{x}^i$ is approximately optimized by combining the predictions of global models with the responses of the local detector,

$$\mathbf{x}^{i*} = \arg\max_{\mathbf{x}^i} \sum_{k,t\in\mathcal{M}} P(\mathbf{x}^i|\mathbf{x}^i_{k,t}) P(l^i = 1|\mathbf{x}^i, \mathcal{I}). \tag{10}$$

In conclusion, the optimization process of existing exemplar-based CLMs can be divided into two steps: (1) *RANSAC* step to retrieve top global models, (2) *fusion* step to combine the information of the global models and response maps. Although achieving promising performance, there still exists some limitations in existing exemplar-based CLMs, as discussed in below.

2.3. Discussions about exemplar-based CLMs

Exemplar-based CLMs employ a RANSAC procedure to randomly select among different feature points as the anchor points for facial calibration, which effectively improves their tolerance to partial occlusions. Furthermore, unlike conventional iterative algorithms (e.g., PCA-based CLMs) whose performance depends on good initialization, exemplar-based CLMs use the sampled global models to infer the locations of feature points, which naturally bypasses the problem of bad initialization.

Despite these advantages, traditional exemplar-based CLMs do have their limitations.

- *High computational cost*: existing exemplar-based CLMs use a greedy searching procedure to generate a response map for each feature point, from which the peak points are randomly sampled as anchor points. However, these methods usually do not have a good strategy to control the size of the searching window. Fig. 2 shows the searching windows for two different feature points with and without spatial constraints posed by the pre-located anchor points, estimated by resizing all selected training faces to the same size and computing the minimal bounding box that covers the locations of the same feature

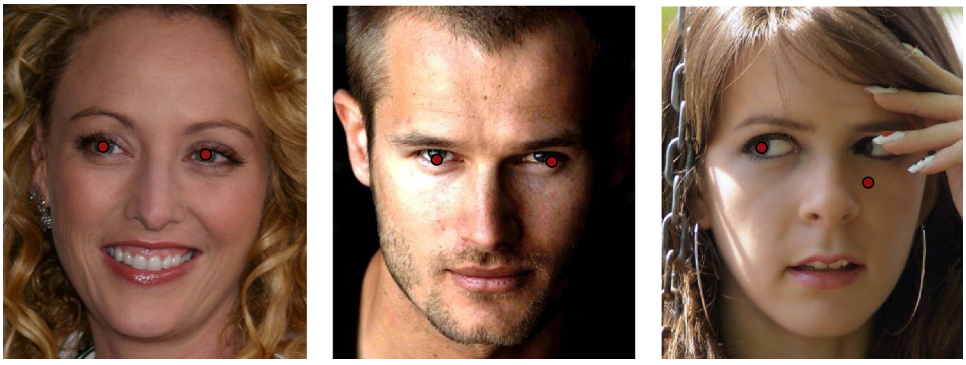

Fig. 3. Illustration of the precise and imprecise eyes (default anchor points) localized by auto eye detector [2], where (a) shows the ground truth while (b) is with the small localization error and (c) with the large localization error.

points from different images.

- *Sensitive to inaccurate anchor points*: inaccurate localization of anchor points may lead to large variation of voting maps, which significantly increases the difficulty of subsequent processing (see Fig. 3(a)). Actually, small localization errors of anchor points are almost inevitable in practice, while large localization error of anchor points, once happen, will definitely lead to large prediction errors of other feature points, especially when only very few (e.g., one pair of) anchor points are used. While randomly sampling different anchor points helps to alleviate them, the bias errors remain most of the time.
- *Ignoring the use of more supervision*: existing exemplar-based CLMs consider the global likelihood as the score function in the RANSAC step. However, this likelihood may not be the best choice for scoring – it on one hand ignores the supervision from the ground truth shapes in the training data, while on the other hand does not take into account the reliableness differences between feature point detectors. Also, in the formulation of [5], the voting distribution map and response map are fused by simple multiplication operation, which lacks supervision from ground truth and is sensitive to ambiguous local responses.

3. Robust discriminative Hough voting for face alignment

In this section, we first give an overview of the proposed system, then describe our robust discriminative Hough voting method for face alignment in detail.

3.1. Overview of the proposed system

Fig. 4 gives the overall pipeline of the proposed system, which mainly consists of following components:

- **Anchor points localization**: here we employ an off-the-shelf eye detector to detect eye as our default anchor points. This is because the eyes are arguably the most salient facial features that can be reliably localized. However, when the eyes are partially occluded, the localization becomes less reliable. In this case, we need to find another pair of facial points as anchor points.
- **Enhanced RANSAC** and **Anchor point validation**: first we use a sampling strategy to enrich the possible locations of anchor points, which are subjected to further validation check based on the diagnosis of resulted distribution of their votings for other facial points. Consequently, only those anchor points with small localization errors are allowed. Furthermore, we select most useful exemplars for Hough voting using a discriminative model.
- **Information fusion**: the retrieved top ranking global models cast votes and construct a special voting map for each facial point. Then, the global voting priors and local evidence are fused with a multi-output regression method to give the final feature locations.

The enhanced RANSAC, anchor point validation and information fusion are the main contributions of the proposed Robust Discriminative Hough Voting (RDHV) method. Among them,

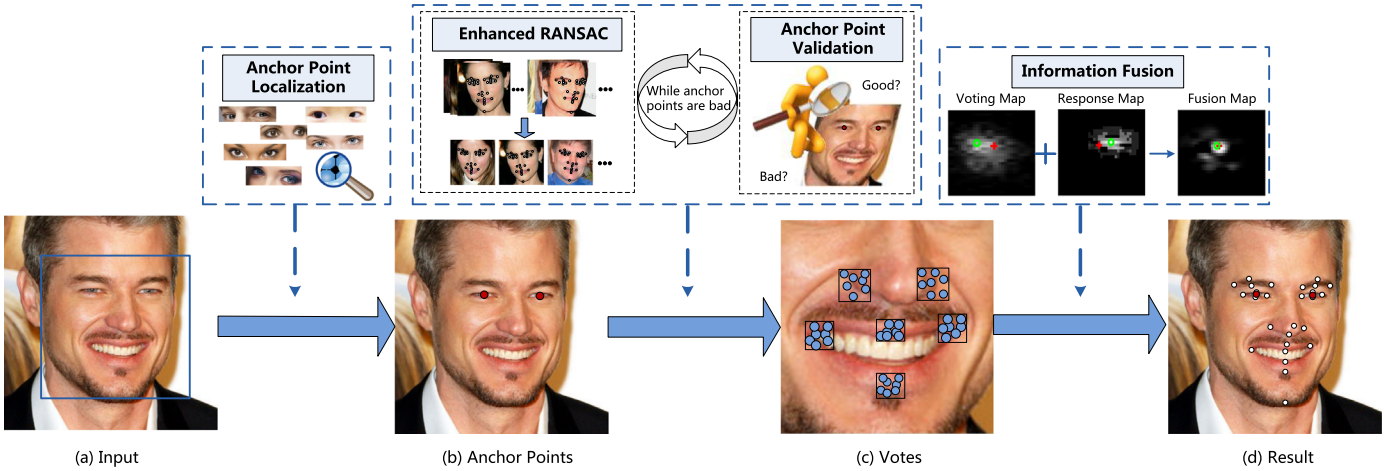

Fig. 4. Pipeline of the proposed face alignment system. This figure is best viewed in the electric form.

anchor point validation and anchor point softening of the enhanced RANSAC work together in combination to make our system very robust to inaccurate anchor points. To mitigate the problem of ambiguity of local detectors, we also propose a deep CNN-based score function to select top ranking exemplars, which achieves much better performance than the proposed baseline discriminative model. Note that in our system the anchor point validation step follows along with the enhanced RANSAC step, as it relies on the output of enhanced RANSAC, i.e., the resulted voting distribution.

It also worth noting that, the enhanced RANSAC and information fusion steps do not share the same objective. The global models produced by enhanced RANSAC are considered as weak shape predictors to generate a local voting map for each facial point, which allows us to obtain a compact local response map for each point. Intuitively and empirically, we show that the discriminative map-based fusion strategy is more robust and effective than the greedy pixel-based fusion in [5]. Hence, the fusion step in our system does not fit in the exemplar-based CLM framework, but can be seen as a robust post-processing method of the exemplar-based CLM.

3.2. Enhanced RANSAC

Suppose that we are given the locations $\mathbf{z}$ of the anchor point pair, and with these references locations, an exemplar $\mathbf{X}_k$ is transformed to $\mathbf{X}_{k,\mathbf{z}}$ under some similarity transformation T, i.e., $\mathbf{X}_{k,\mathbf{z}} = T(X_k|\mathbf{z})$. Given the global model $\mathbf{X}_{k,\mathbf{z}}$, the locations of feature points $\mathbf{x}^i$ in the test image can be treated as conditionally independent of each another. Then, the generic shape model (3) in our system can be rewritten as follows:

$$S(X) = \sum_{k=1}^{K} \sum_{\mathbf{z}} \left(\prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,z}) \right) P(\mathbf{z}) P(\mathbf{X}_k). \tag{11}$$

where the prior of the anchor points $P(\mathbf{z})$ is introduced into the framework of the exemplar-based CLM and $P(\mathbf{x}^i|\mathbf{x}^i_{k,\mathbf{z}})$ is modeled as 2D Gaussian distribution centered at $\mathbf{x}^i_{k,\mathbf{z}}$, as in [5]. We further assume a uniform distribution for the exemplar $\mathbf{X}_k$. Then, combing (2), (3) and (11) yields

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \sum_{k=1}^{K} \sum_{\mathbf{z}} \left(\prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,\mathbf{z}}) P(l^i = 1|\mathbf{x}^i, \mathcal{I}) \right) P(\mathbf{z}). \tag{12}$$

To solve this objective function, one common method is to approximate it by sampling some top ranking global models $\mathbf{X}_{k,\mathbf{z}}$. In [5], the *global likelihood* (4) $L(\mathbf{X}_{k,\mathbf{z}})$ is used to evaluate the scores of global models. However, as mentioned before, this ignores useful supervision information available in the training data. We hence propose a new scoring function which alleviates this problem, based on the local evidence from the face shapes.

Particularly, let $\mathcal{R}(\mathbf{X})$ be the I-dimensional vector consisting of the responses at all I locations in the face shape $\mathbf{X}$, and our desired score function is then denoted as $\mathcal{S}(\mathcal{R}(\mathbf{X}))$. How to learn this score function will be delayed to the next section, and in particular, we will show that the deep CNN technique [26] can be conveniently borrowed to designed a more powerful score function to boost the performance. By substituting the global likelihood $\prod_{i=1}^{I} P(l^i = 1|\mathbf{x}^i, \mathcal{I})$ in (12) with $\mathcal{S}(\mathcal{R}(\mathbf{X}))$, we have,

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \sum_{k=1}^{K} \sum_{\mathbf{z}} \left(\prod_{i=1}^{I} P(\mathbf{x}^i|\mathbf{x}^i_{k,\mathbf{z}}) \right) \mathcal{S}(\mathcal{R}(\mathbf{X})) P(\mathbf{z}), \tag{13}$$

The RANSAC process according to (13) proceeds as follows:

a) Generate a large pool of r global models $\mathbf{X}_{k,\mathbf{z}}$, transformed linearly from a random exemplar $\mathbf{X}_k$ based on anchor points $\mathbf{z}$ sampled according to $P(\mathbf{z})$;
b) Calculate the score for each global model $\mathbf{X}_{k,\mathbf{z}}$ using the discriminatively trained score function $\mathcal{S}(\mathcal{R}(\mathbf{X}_{k,\mathbf{z}}))$;
c) Choose m^* top ranking global models and record the pairs of $\mathbf{z}$ and k in a set $\mathcal{M}$.

In our current system, we set $r = 10,000$ and $m^* = 40$ by cross validation.

We name the above process as enhanced RANSAC. There are three major differences between the RANSAC procedure in [5] and ours: (1) in [5], various anchor points are tested, while only one pair of anchor points (e.g., two eyes) are used in our system; (2) in [5], the locations of anchor points are fixed, while we sample anchor points according to their prior distribution $P(\mathbf{z})$; (3) in [5], a global likelihood function is used as scoring function for exemplar selection, while we use a specially trained scoring function $\mathcal{S}(\mathcal{R}(\mathbf{X}))$. Furthermore, due to the small patch support and large variation during training, the independent detection responses $\mathcal{R}(\mathbf{X})$ are plagued by the problem of ambiguity. Hence, we also propose a deep CNN-based score function that does not rely on the detection response. All these will be detailed in the subsequent sections.

3.2.1. Softening the localizations of anchor points

Previous works [4,3,27,5] implicitly assume that the locations of anchor points are correct. However, such a hypothesis is seldom true. Actually, small localization errors of anchor points are almost

Fig. 5. Illustration of the anchor point softening for RANSAC. (a) Small error anchor points localized by [2]. (b) Gaussian distribution (the translucent red filled circle) assumed for the locations of the anchor points, where the radius equals the standard deviation of the Gaussian distribution. (c) Anchor points in some top ranking global models sampled from Gaussian distribution, which are more accurate than the initial locations in (a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

inevitable in practice (see Fig. 5(a)). Consequently, the votes casted from the error anchor points to the target points tend to have bias error, especially when the target points are close to the anchor points. Although randomly sampling different anchor points helps to alleviate them, the bias errors remain most of the time.

Our idea to handle this issue is built on a simple intuition: If we randomly sample the anchor points from the region nearby the pre-located anchor points, then through r (r=10000) times sampling, we will with high probability have a few anchor points whose locations are close enough to the ground truth, and the global models transformed from these "good"anchor points are more likely to have higher scores than those transformed from "bad" anchor points.

Particularly, we assume a Gaussian distribution for the location of each anchor point rather than treat it as fixed (see Fig. 5(b)). Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)^T$ be a random vector for the locations of an anchor point pair, and let $\mathbf{z}^* = (\mathbf{z}_1^*, \mathbf{z}_2^*)^T$ denote the pre-located locations of the anchor point pair, we have:

$$\mathbf{z}_1 \sim \mathcal{N}(\mathbf{z}_1^*, \sigma^2\mathbf{I});\ \mathbf{z}_2 \sim \mathcal{N}(\mathbf{z}_2^*, \sigma^2\mathbf{I}), \tag{14}$$

where σ is set to 8 pixels empirically.

We call this method of relaxation of the locations of anchor points as "softening anchor points", which effectively improves the system's tolerance to small errors in anchor points localization. Fig. 5(c) shows some randomly selected anchor points in the top ranking global models, which are more accurate than the initial locations in Fig. 5(a).

3.2.2. Learning a cost-sensitive discriminative score function

The goal of the score function is to rank the effectiveness of the exemplars in the RANSAC procedure. Intuitively, good exemplars should have small root mean square error (RMSE) when used for feature localization. However, since there is no ground truth available on a test image, one usually bases such judgement on the strongness of local evidence $\mathcal{R}(\mathbf{X})$ collected at the predicted feature locations [5]. That is, strong response at some location is assumed to imply small RMSE.

Unfortunately, the above assumption is not necessarily always true due to the unreliability of local facial detectors, and the localization accuracy may significantly decrease if too many "bad" exemplars are used for RANSAC. Hence, one of the key issue that needs to be considered in the design of a score function is how to filter these "bad" exemplars out while preserving good ones. For this, we propose a cost-sensitive discriminative function.

Particularly, our score function is modeled using a logistic regression function based on the local evidence of $\mathcal{R}(\mathbf{X})$ collected on a test image.

$$\mathcal{S}(\mathcal{R}(\mathbf{X})) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathcal{R}(\mathbf{X}) - b)}, \tag{15}$$

where $\mathbf{w}$ and b are the parameters to be learned. Given N_p positive samples and N_n negative samples, the goal is to learn our score function such that it should make the correct classification but the cost of false positives would be larger than that of false negatives, i.e.,

$$\min_{\mathbf{w}} \sum_{n\in N_p} \alpha_p \log(1 + \exp(-\mathbf{w}^T\mathcal{R}(\mathbf{X}) - b)) + \sum_{n\in N_n} \alpha_n \log(1 + \exp(\mathbf{w}^T\mathcal{R}(\mathbf{X}) + b)) + \lambda \parallel \mathbf{w} \parallel_2 \tag{16}$$

where α_p and α_n respectively denotes the cost of positive samples and negative samples. We empirically found that setting $\alpha_n = 1.5\alpha_p$ and $\lambda = 10^{-4}$ leads to good results.

To prepare the positive samples and negative samples, we follow the following steps:

1. Generate r global models for each training image, where the anchor points are sampled according to (14).
2. Categorize the r global models for each training image into positive and negative sample according to their root mean square error (RMSE) in localization, with a predefined threshold T_g.
3. Randomly sample N_p positive samples and N_n negative samples as our training samples.

In our implemenation, we use $N_p = N_n = 20,000$ training samples, and threshold T_g is set to be 6 pixels empirically - the actual value of T_g is not important but it should be set to be such a value that there are sufficient number of good global models appearing in r candidates for each image. Usually, the value of r is set to be much larger than the number of good global models m^* we actually use.

3.2.3. Learning a deep CNN-based score function

We note that although above cost-sensitive discriminative score function can overcome some drawbacks of the global likelihood-based score function in [5], it still relies on the response of local detectors that are plagued by the problem of ambiguity. Inspired by the great success of deep CNN in the field of image classification [26,28,29], we developed a deep CNN-based score function as a powerful alternative of (15). In particular, we learn this deep CNN-based score function by fine tuning the Alexnet [26] with our image data that denotes "good" and "bad" global models. To obtain the image-based representation of the "good" and "bad" global models, we simply extract local patches centered at the points of the global model, and rearrange them line-by-line to form a 2D image representation of each global model. We use the same setting ($N_p = N_n = 20,000$ and T_g=6) to collect positive and negative training samples, and fine tune the Alexnet [26] with these samples by changing the output of the last layer, similar to [30] (and refer to [30] for details).

Alexnet takes about 1 ms to process one image on the Tesla K20 GPU, that is, it will take about 10 seconds for the deep CNN-based score function to select $m^* = 40$ top ranking global models from $r = 10,000$ candidates. To achieve a good trade-off between accuracy and efficiency, we use the cost-sensitive discriminative score function (15) to roughly select 500 top ranking candidates first, and then use the deep CNN-based function to select $m^* = 40$ best global models from these 500 candidates. When using the deep CNN-based score function, we name the system as Deep Robust Discriminative Hough Voting (D-RDHV). We consider D-RDHV as an improved version of RDHV to show that by incorporating the deep learning technique into the proposed discriminative exemplar-based CLM framework, we can achieve comparable results comparing to the state-of-the-art cascaded regression methods (e.g., LBF [13]).

3.3. Anchor point validation

Although our anchor point softening strategy can effectively alleviate the impact of small anchor point localization errors, it can hardly handle well large errors (c.f., Fig. 3(c)), in which case the chance of sampling "good" anchor points would be very low. Although large anchor point localization error rarely happens using the current state of the art detectors, once happens, it will definitely lead to large performance degradation.

To address this issue, our main idea is to evaluate the quality of anchor points pair before using the votes casted from them, and to use a different pair of anchor points if the quality of current pair is found low. One of the most simple ways to quantify the quality of anchor points is to measure the root mean square error (RMSE) between the locations of each of the two anchor points and their corresponding ground truth locations. If such error is larger than

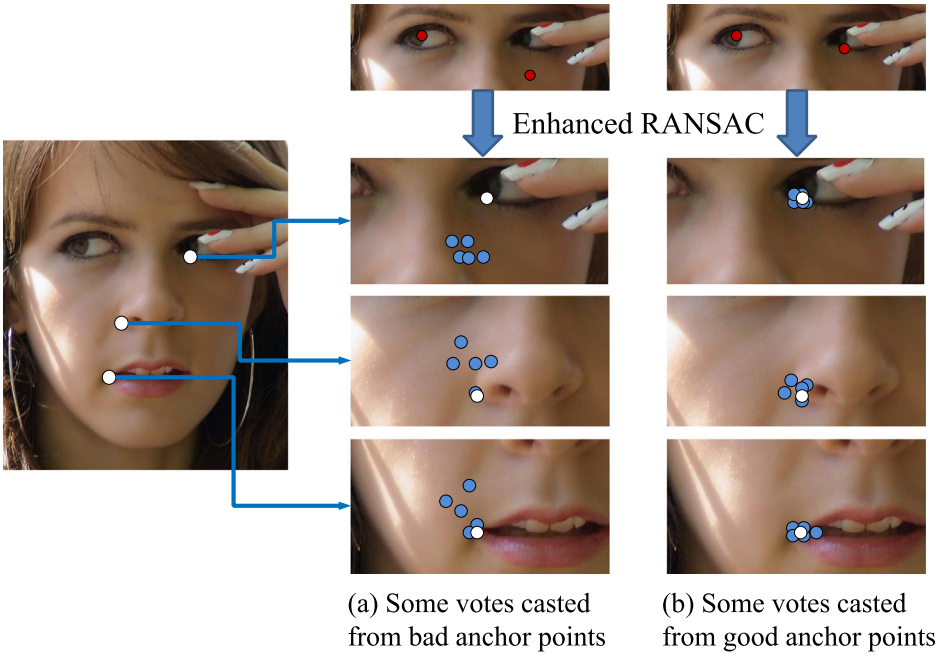

Fig. 6. Illustration of the different behavior of good and bad anchor points, where the red, white and blue points respectively denote anchor points, ground truth locations of the target points and the votes for the target points. One can see that the votes casted from bad anchor points drift farther from the right locations (a) and are distributed in a spatially loose manner while the votes from good anchor points cluster tightly around the target points (b).

some threshold T_a, this pair of anchor pairs could be regarded as "bad". In our current implementation, the value of T_a is set to be 8 pixels.

However, for a given test image, the ground truth locations of anchor points are unknown to us. One natural way to bypass this is to use the responses of anchor point detectors. Unfortunately, such responses are unstable due to the large variance of local appearance. In this work, we follow an alternative direction and use the information of the top ranking global models to infer the goodness of anchor points. As illustrated in Fig. 6, votes casted from bad anchor points tend to not only drift far away from the right locations, but are distributed in a spatially loose manner. This is because the r global models generated by bad anchor points are too noisy to cast effective votes that cluster tightly in the voting space.

Based on these observations, we use the mean response μ_r^i and the distance deviation σ_d^i of the votes casted for each feature point by the top ranking global models as the feature to evaluate the quality of anchor points. Let $\mathcal{W}(\mathbf{z})$ denote such feature representation, we have:

$$\mathcal{W}(\mathbf{z}) = (\mu_r^1, \ldots, \mu_r^I, \sigma_d^1, \ldots, \sigma_d^I)^T \tag{17}$$

Based on $\mathcal{W}(\mathbf{z})$, we train a linear SVM classifier $C(\mathcal{W}(\mathbf{z}))$ as the evaluation function. For this the positive and negative training samples are split and generated using the anchor point error threshold T_a. In our current implementation, we use 20,000 positive samples and 20,000 negative samples, amongst which 95% samples take the eyes as anchor points and the remaining samples randomly take other facial feature points as anchor points.

Finally, once a pair of anchor points are quantified as "bad" ones, we should generate another pair of anchor points that are good enough for the subsequent RANSAC procedure. The whole pipeline of anchor points evaluation and regenerating process is summarized as follows:

1. Use the classifier $C(\mathcal{W}(\mathbf{z}))$ to evaluate the accuracy of current anchor points. If $C(\mathcal{W}(\mathbf{z})) = 1$, means that current anchor points are good, otherwise go to the next step.
2. Randomly select two local detectors of candidate anchor points to generate response maps, then randomly choose one of two top peaking points in the response maps as the new anchor points.
3. Repeat Step 1 and Step 2, until $C(\mathcal{W}(\mathbf{z})) = 1$.

In practice, the condition of $C(\mathcal{W}(\mathbf{z})) = 1$ meets for most of the eyes localized by [2] (about 95%), while in the remaining cases, the above process usually finishes in less than 4 iterations, thanks for our rather relaxed requirement of anchor point accuracy benefiting from the aforementioned anchor point softening strategy.

3.4. Information fusion

Each global model generated by enhanced RANSAC can be thought of as a weak predictor of the face shape on the test image. The minimal bounding boxes of the votes by these weak predictors naturally define independent voting maps, which constrain the search space of each feature point into a small local region.

However, unless the votes are rich enough, the predictions in the voting map might not contain the ground truth location. One way to address this is to use a non-parametric density method to smooth the voting map. We use a Gaussian kernel for this purpose,

$$p(\mathbf{x}_i|\mathcal{M}) = \frac{1}{m^*}\sum_{m=1}^{m^*}\frac{1}{(2\pi h_i^2)^{1/2}}\exp\left\{-\frac{\| \mathbf{x}_i - \mathbf{x}_{j_m,k_m}^i \|^2}{2h_i^2}\right\} \tag{18}$$

where h_i is standard deviation of the Gaussian components of the ith feature point. h_i is computed in the same way as in [5], estimated by the exemplar shapes.

With the voting and response maps, we adopt a multi-output ridge regression method to fuse the information from them for each feature point. The central idea for this is to learn two linear transforms (rotations) for the voting map and response map respectively, such that after rotation both maps align well with the ground truth map. However, in the training set, the ground truth feature points are not always covered by the voting window. In these cases, we consider the point in the voting window closest to the ground truth as the mimic ground truth, and generate the

ground truth map using a Gaussian kernel with the same deviation h_i of (18).

Then the voting map, response map and ground truth map are normalized to the same size and the score in each map is normalized to behave like a probability by dividing by the sum of the scores in the voting window before aligning them. Mathematically, denote the three maps of sample n as corresponding matrices $\mathbf{V}_n$, $\mathbf{E}_n$, and $\mathbf{G}_n$, respectively. Then what we want to do is to learn two 'rotation' matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ for the voting map $\mathbf{V}_n$ and response map $\mathbf{E}_n$ respectively. In our implementation, we used a vector representation and concatenated the two maps into a single combined vector $\mathbf{u}_n = [vec(\mathbf{V}_n)^T, vec(\mathbf{E}_n)^T]$. Further denote $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]$ and $\mathbf{g}_n = vec(\mathbf{G}_n)$, then our goal can be formulated as a standard multi-output ridge regression objective function,

$$\min_{\mathbf{W}} \sum_{n=1}^{N} \|\mathbf{g}_n - \mathbf{W}\mathbf{u}_n\|_{l_2}^2 + \lambda\|\mathbf{W}\|_F^2 \tag{19}$$

with the closed form solution,

$$\mathbf{W}^* = \left(\sum_{n=1}^{N} \mathbf{g}_n\mathbf{u}_n^T\right)\left(\sum_{n=1}^{N} \mathbf{u}_n\mathbf{u}_n^T + \lambda\mathbf{I}\right)^{-1} \tag{20}$$

The regularization parameter λ is set to be a very small number (10^{-4} in our implementation).

After fusing the voting and response maps, the top-one peak point in the fusion map is regard as the final location of the feature point. As illustrated in Fig. 7, fusing these helps to reduce the ambiguity in searching for the best response. For example, as shown in the second response map in the last row, a facial point located at the face contour has a wide range of response, but such kind of ambiguity is dealt well with the method by combining the information from the voting map and the response map (see the last map in the third row).

4. Implementation details

In this section, we will discuss some implementation details including face detection and normalization, local detector training.

4.1. Face detection and normalization

A consistent aspect of all the following experiments is the face detection, the result of which serves as the starting point for face alignment. For this, we use the Viola–Jones detector [31] to detect faces due to its high efficiency. After we obtained the face region of each image, we rescale the face images to make the inter-ocular of each face about 50 pixels, the computation of which is relative to the size of the face region.

However, the Viola–Jones detector [31] sometimes fails to detect the faces with varying pose and illumination in some images. For instance, 12.05% of all faces of the LFPW [5] databases are missed or incorrectly detected. For these failure cases, we initialized the face bounding box estimated from the ground-truth face shape as follows: (1) we first compute the scale variation of the ground-truth face shape through L_2 fitting to the mean shape, then (2) resize the current face image and ground-truth shape according to the computed scale variation, and (3) shift the center of the mean shape to the center of the resized ground-truth shape to place the face bounding box on current image, finally (4) randomly perturb the estimated face bounding box by 10 pixels for translation to mimic the experimental setting. Similar idea is used in [21,32].

4.2. Local detector training

Local detector is an important component of CLMs. We use two-scale SIFT feature (as used in [5]) and linear SVM to train our local detectors. Note that we found in our early experiments that with sufficient training data, the performance of liner SVM is comparable to RBF kernel SVM used by [5] for face alignment, but is much more efficient. Additionally, we augmented the training images by left-right flip and random rotations so that we have about 6000 training images for each dataset.

5. Experiments

In this section, we present four sets of experiments, i.e., (1) comparison to the baseline CLMs, (2) comparison to the state-of-the-art, (3) running time performance analysis, and (4) algorithm

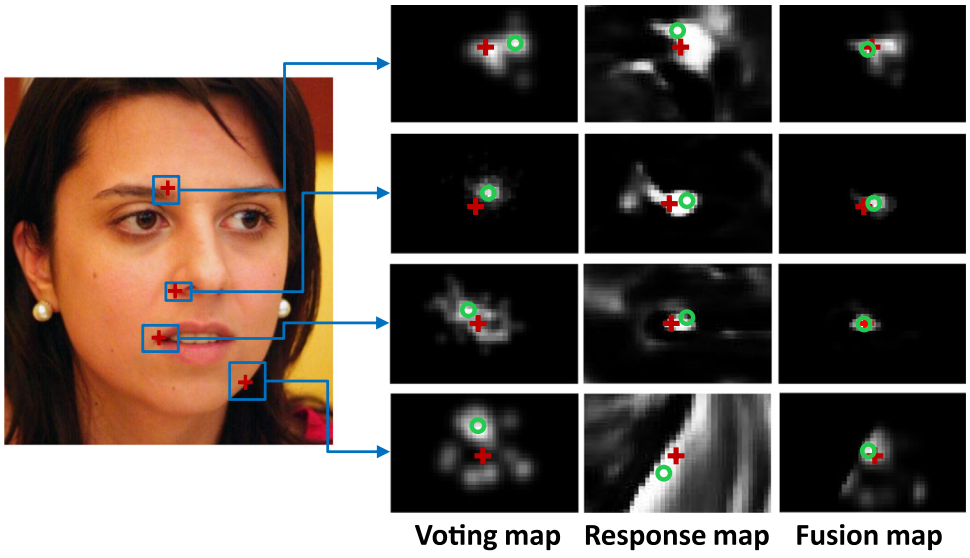

Fig. 7. Illustration of the voting map, response map and fused response map, where the red cross is the ground truth and the blue circle is the location with maximum response of each map. Note that the voting maps for different feature points differ from each other in their size, and for better illustration here we resize them to get the same size. This figure is best viewed in the electric form. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

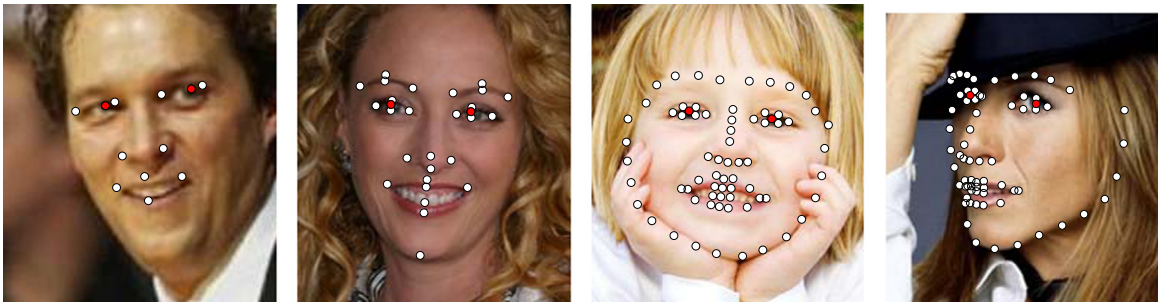

Fig. 8. Illustration of landmarks for sample images from 4 datasets respectively. The red points, i.e., two eyes, are the default anchor points in the proposed system. Note that the original annotations on LFW, HELEN and IBUG do not contain the ground truth eyes, we annotated them manually. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

validation and discussions. Below, we first introduce the wild datasets and the evaluation metric used in our experiments.

Datasets: we briefly introduce the wild datasets used in our experiments. These datasets are challenging due to images with large head pose, occlusions, and illumination variations.

LFW [33]: this dataset consists of 13,233 images which are collected in the wild and vary in lighting condition, pose, expression and background. Moreover, images of LFW have lower quality compared to other datasets. Following [34], a ten-fold cross validation experiment is performed to report our performance on LFW. It worth noting that since our default anchor point (eye) detector [2] is trained on 1000 images from LFW, we carefully remove these images when testing according to the annotation data released by the author. Each image of LFW is annotated with 10 landmarks by [34], and we add two eyes, so we have 12 landmarks for each image, see Fig. 8(a).

LFPW [5]: This dataset consists of 1400 images. Unfortunately, because some URLs are no longer valid, we only collected 833 of the 1100 training images and 232 of the 300 test images. LFPW is a completely wild dataset, i.e. consists of face images with combined variations in pose, illumination and expression. Each face image in LFPW is annotated with 35 points, but only 29 points defined in [5] are used for the face alignment, see Fig. 8(b).

HELEN [35]: This dataset consists of a total of 2330 high-resolution face images, 2000 images for training and 330 for testing. HELEN is an extremely challenging dataset for face alignment due to its large variations in pose, illumination, expression and occlusion. In our experiments, we use the 68-point annotation for HELEN by the *ibug* group[2] rather than the original 194-point annotation, as later dense annotation inherently brings about ambiguity in training local detectors for CLMs. We also add two eyes manually, so we have a total of 70 points for each face image from HELEN, see Fig. 8(c).

IBUG [36]: this dataset is a challenging 135-image subset of the 300-W dataset created for a challenge of face alignment. IBUG dataset is extremely challenging as its images have large variations in face poses, expressions and illuminations. Following the same configuration as in [13], to perform testing on IBUG, we regard all the training samples in 300-W from LFPW, HELEN and the whole AFW as the training set (3148 images in total). Each face in IBUG is manually annotated with 68 points, and we add two eyes to serve as default anchor points. Therefore, we have 70 landmarks for each image in IBUG, see Fig. 8(d).

Evaluation: in most of the following experiments, we use the normalized root-mean-squared error (NRMSE) relative to the ground truth as the error measurement, until otherwise noted. The NRMSE is computed by dividing the root mean squared error by the distance between the two eye centers. When evaluating different algorithms on the same database, we use the facial points from dataset annotation which are common in all the algorithms.

[2] http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/.

5.1. Overview of experiments and results

We consider several state-of-the-art local methods in recent years as the baselines for comparison. They are the consensus of exemplar method (CoE) [5], the tree structure part model (TSPM) [10], the discriminative response map fitting method (DRMF) [6] and the optimized part mixtures method (OPM) [32], amongst which CoE is implemented by ourselves, while TSPM, DRMF and OPM are released by the authors. We compare with them on three commonly-used wild datasets, LFW, LFPW and HELEN.

To further verify the capability of the proposed methods (RDHV and D-RDHV) to handle challenging uncontrolled natural variations, we also test our methods on the extremely challenging IBUG dataset, and compare it to many state-of-the-art holistic cascaded shape regression methods besides the baseline local methods.

Last but not least, we conduct a set of experiments on LFW to verify the effectiveness of the proposed key components of our system, i.e., *enhanced RANSAC*, *anchor point validation* and *Information fusion*.

Overview of results:

1. The proposed RDHV method shows promising results over all baseline local methods consistently on LFW, LFPW and HELEN. In particular, RDHV produces results almost more accurate than human on LFW, and achieves accuracy improvement over 28 of the 29 facial points on LFPW compared to the consensus of exemplar (CoE) method [5]. Fig. 9 also shows that the deep CNN-based score function (D-RDHV) can greatly boost the performance of the proposed RDHV.
2. Overall, the proposed RDHV method achieves good performance on the challenging IBUG dataset. It significantly outperforms the baseline local methods, and shows comparable performance with some cascaded regression methods, but is inferior to recent Local Binary Features Fast (LBF) method [13]. We further verify that great improvement can be gained by incorporating the deep CNN techniques into the step of RANSAC to choose top ranking global models, and the resulting D-RDHV method achieves slightly better performance than LBF.
3. The running time of the proposed RDHV and D-RDHV is about 3–4 time less than the baseline CoE and TSPM methods, since we limit the search space of each point to a small region using the geometrical constraints imposed by the anchor points. It takes about 2.5 s to process an image with our Matlab implementation. Since TSPM is claimed possible to be real-time [10], we expect to push the proposed method real-time by

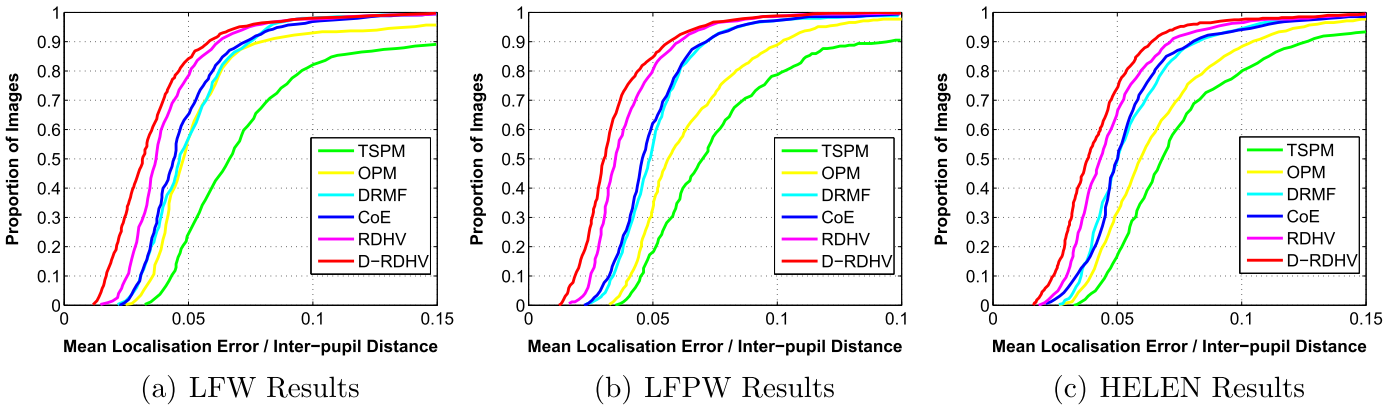

Fig. 9. Comparison to the baselines: cumulative errors distribution (CED) curves of the proposed Robust Discriminative Hough Voting (RDHV) method and Deep Robust Discriminative Hough Voting (D-RDHV) method and four baseline local methods on LFW, LFPW and HELEN.

Fig. 10. Illustration of some aligned images from the LFW dataset, where the red filled dots denote the anchor points.

certain implementing techniques other than Matlab.

4. The algorithm validation experiments on LFW demonstrate the effectiveness of the newly-proposed anchor point validation, enhanced RANSAC and information fusion methods. Specially, the anchor point validation step and anchor point softening incorporated by enhanced RANSAC can make our system very robust against inaccurate anchor points. We also show that it is better to detect eyes as default anchor points by [2] than to sample peak points randomly from the response maps as anchor points.

Figs. 10–13 show our results by RDHV on challenging examples with large variations in pose, expression and occlusion where red filled points are the anchor points.

5.2. Experiment 1: comparison to the baselines

The goal of this experiment is to compare the performance of the proposed RDHV and D-RDHV methods with several baseline local methods, under combined variations of pose, expression, and illumination. In particular, we compared our method with the CoE [5], TSPM [10] DRMF [6] and OPM [32] methods on three widely-used LFW, LFPW and HELEN datasets. We consider CoE and DRMF as the baselines because they are representative methods for non-parametric and parametric CLMs respectively. CoE is the first to introduce the exemplar-based shape model into the face alignment task. We reformulate and cast it into the CLM framework, and consider it as the starting point of our method. DRMF is the first to employ discriminative fitting technique under the CLM framework to improve performance, which also motivates us to incorporate discriminative learning into the exemplar-based CLM. The TSPM presents a unified approach to face detection, pose estimation, and landmark estimation, based on a mixture of tree-structured part models, while OPM is an improved version of it.

The overall accuracy on LFW, LFPW and HELEN is shown by the cumulative errors distribution (CED) curve in Fig. 9. We can see that the proposed RDHV method consistently outperforms other baseline methods with a significant margin on all datasets, while our deep CNN-based D-RDHV method further boosts the performance of RDHV. Furthermore, as Dantone et al. [34] compared their results with the human in their paper, we add our results for comparison. As shown in Fig. 14, RDHV achieves better performance over the Conditional Random Forest (CRF) method proposed in [34], and is almost more accurate than the human. Fig. 15 allows us to have a closer look at the performance of our method and the baseline CoE, showing that 28 from all 29 feature points localized by RDHV are more accurate than CoE, amongst which 5 points, i.e., outside of the eyebrows, eyes and chin, are more than 15% accurate. We credit the eyes localization accuracy improvement to our accurate ad-hoc eye detector [2], while the improvement of the outside of the eyebrows and the chin, where the local appearances are unstable, are mainly due to the virtue of

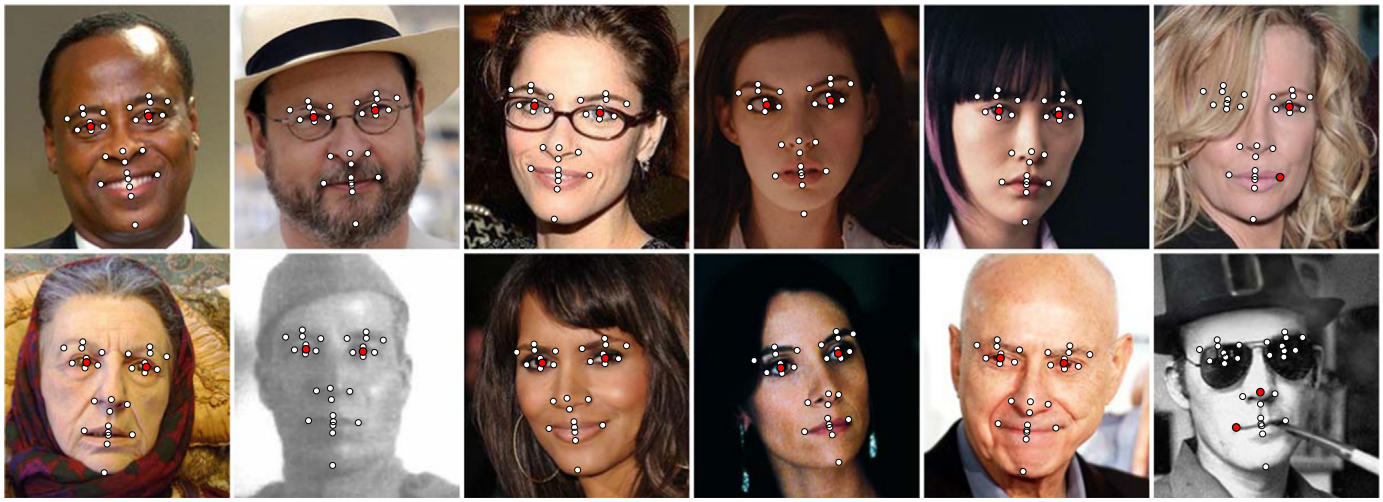

Fig. 11. Illustration of some aligned images from the LFPW dataset, where the red filled dots denote the anchor points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

discriminative learning employed in the enhanced RANSAC and local map fusion steps.

5.3. Experiment 2: comparison to the state-of-the-art

While the proposed RDHV and D-RDHV methods show a significant advantage over the baseline local methods, i.e., CoE [5], TSPM [10] DRMF [6] and OPM [32], further comparison to the state-of-the-art holistic cascaded shape regression methods should also be investigated. Furthermore, some state-of-the-art methods have perhaps reached a saturation point on the commonly-used LFW, LFPW and HELEN. For example, [5] and [34] have reported close to human performance on LFPW and LFW respectively. It is necessary to investigate the performance of our method on more challenging dataset, for example the IBUG dataset which contains a large portion of faces with challenging head pose and facial expression.

For this end, we follow the same dataset configuration as in [13] and test our method on IBUG dataset. In particular, besides the baseline local methods, we further compare the proposed RDHV and D-RDHV methods on IBUG to the Explicit Shape Regression method (ESR) [9], the Supervised Descent Method (SDM) [11], the Robust Cascaded Pose Regression (RCPR) method [37] and the recent Local Binary Features (LBF) method [13]. Table 1 shows the comparison results, from which we can see that the RDHV also significantly outperforms the baseline local methods, and shows comparable performance with ESR, SDM, and RCPR, but is inferior to recent Local Binary Features Fast (LBF) method. We observe that our D-RDHV method achieves slightly better performance than LBF, which shows that great improvement can be achieved by incorporating the deep CNN techniques into the step of RANSAC to choose top ranking global models.

5.4. Experiment 3: running time performance

We implement our system with Matlab code measured on an Intel Xeon E5-2630 CPU (2.60 GHz, 12 core). Table 2 shows the running time performance of our methods (RDHV and D-RDHV) and several CLM baselines on the IBUG dataset. Overall, our running time is about 3–4 time less than the base line CoE method, since RDHV (and D-RDHV) constrains the local search region of each point by incorporating the prior of one pair of anchor points. Meanwhile, our method is faster than the Tree Structured Part Model (TSPM) and its improved version (OPM), but is slower than the Discriminative Response Map Fitting (DRMF) method. We also note that our method is much slower than the cascaded regression

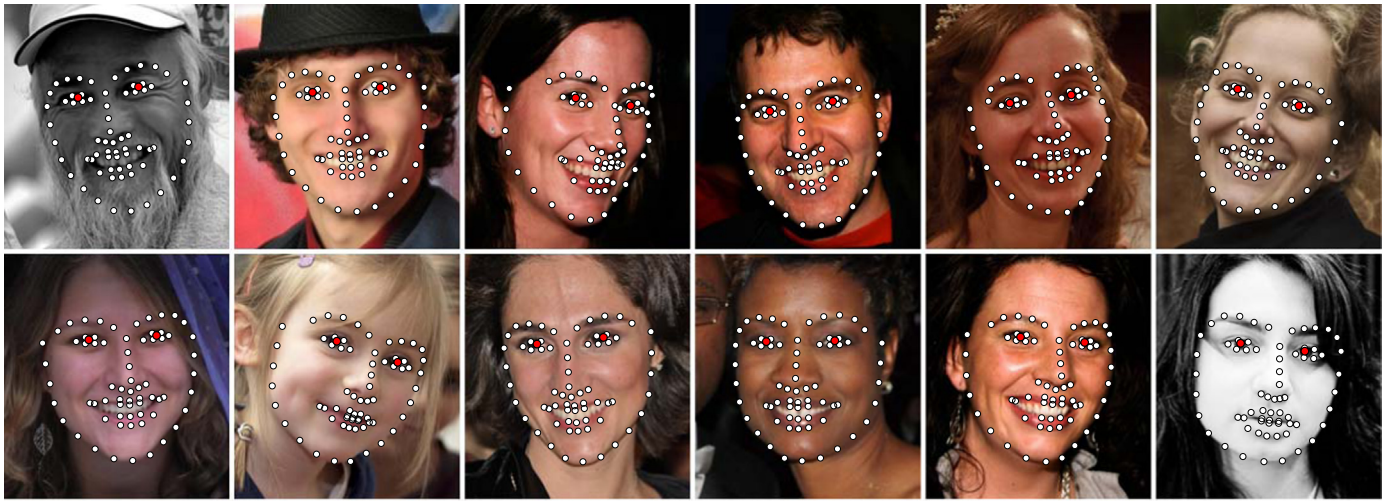

Fig. 12. Illustration of some aligned images from the HELEN dataset, where the red filled dots denote the anchor points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

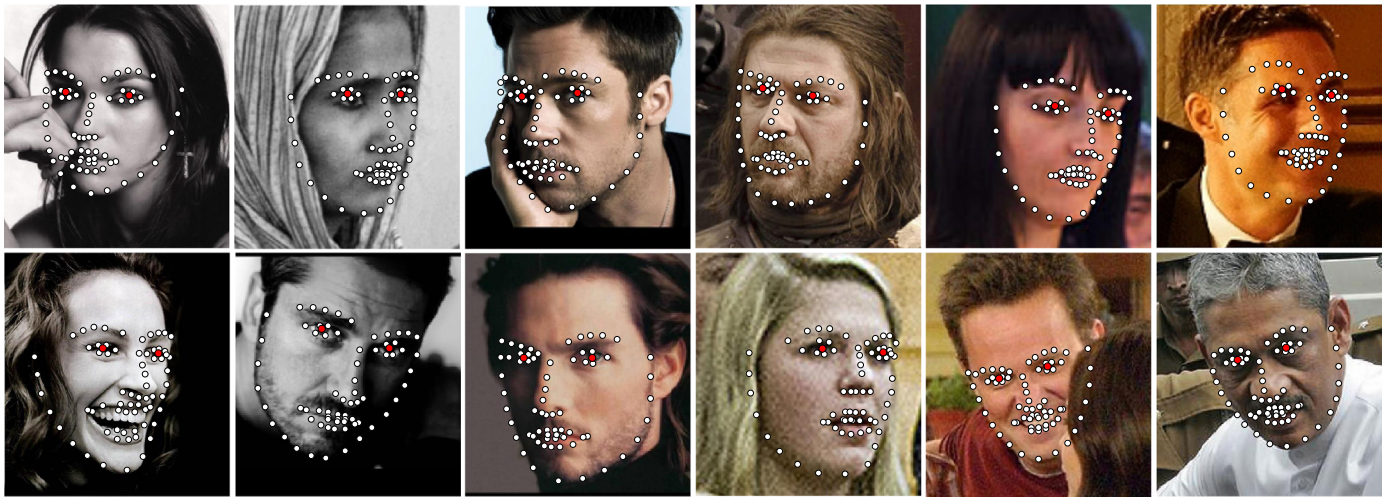

Fig. 13. Illustration of some aligned images from the IBUG dataset, where the red filled dots denote the anchor points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

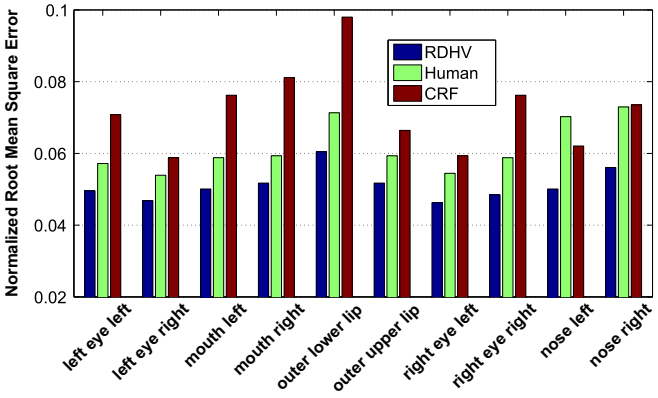

Fig. 14. Normalized root mean square error of 10 individual feature points on LFW. We can see that RDHV achieved better performance than the conditional regression forests method [34], and is almost more accurate than human.

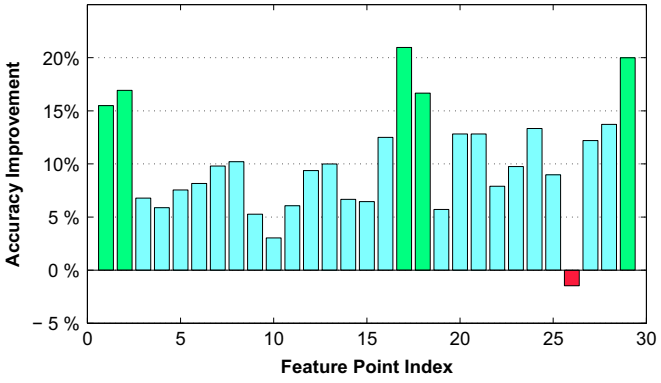

Fig. 15. The accuracy improvement over CoE for 29 individual feature points on LFPW by RDHV. *Green* more than 15%, *cyan* 0–15%, *red* less accurate. Note that our implementation of CoE is worse than the results of [5] reported in their paper due to the lack of training data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

methods [9,13] that bypass the exhaustive local search by directly learning the regression function. However, since TSPM is claimed possible to be real-time [10], in future work we intend to push the proposed method real-time by certain implementing techniques other than Matlab.

Table 1
Comparison to the state-of-the-art: measured by normalized root mean square error (NRMSE) (%) on IBUG.

Algorithm	NRMSE
CoE (Consensus of Examplars) [5]	17.50
ESR (Explicit Shape Regression) [9]	17.00
TSPM (Tree Structure Part Model) [9]	18.33
DRMF (Discriminative Response Map Fitting) [6]	19.79
OPM (Optimized Part Mixtures) [32]	20.43
SDM (Supervised Descent Method) [11]	15.40
RCPR (Robust Cascaded Pose Regression) [37]	17.26
LBF (Local Binary Features) [13]	11.98
LBF fast (Local Binary Features Fast) [13]	15.50
RDHV (Robust Discriminative Hough Voting)	14.45
D-RDHV (Deep Robust Discriminative Hough Voting)	**11.32**

Table 2
Running time performance on IBUG.

	TSPM [10]	OPM [32]	DRMF [6]	CoE [5]	**RDHV**	**D-RDHV**
Time(s)	14.2	5.3	**1.2**	10.5	2.5	3.1

5.5. *Experiment 4: algorithm validation and discussions*

In this section, we conduct a set of experiments for algorithm validation and discussions. We first verify the importance of our automatic eye detector [2]. Then we conduct experiments to verify the effectiveness of the proposed key components of our system, i.e., *enhanced RANSAC*, *anchor point validation*, and *information fusion*, by comparing the proposed RDHV method with the baseline methods that differ in those aspects but remain exactly the same in all other aspects. All the results in the following are computed over a ten-fold cross validation on LFW where the face images are rescaled and the inter-ocular of each face is about 50 pixels.

5.5.1. *Effect of anchor point detector*

In our system, we employ the automatic eye detector [2] to detect eyes as our default anchor points. Here we investigate the effect of the default anchor point detector by establishing the face alignment results with the anchor points obtained by randomly sampling from the peak points of the response maps, by the output of the automatic eye detector [2], and by ground truth eyes

Table 3
Comparison of the performance with anchor points obtained in different ways.

Anchor points	Mean (pixels)	Median (pixels)	Min (pixels)	Max (pixels)
Peak points	2.75	2.66	0.83	11.23
Eyes detected by [2]	2.56	2.51	0.79	10.95
Ground truth eyes	**2.45**	**2.41**	**0.78**	**10.91**

respectively. As shown in Table 3, our method in using eyes detected by [2] as default anchor points performs better than that using randomly sampled peak points in the response maps, and is only slightly inferior to that using the ground truth eyes as anchor points. We speculate that this is because the carefully designed eye detector [2] is much more accurate than the randomly sampled peak points in the response maps.

5.5.2. Effect of enhanced ransac

Compared to traditional RANSAC in [5], our enhanced RANSAC process mainly has two novelties: (1) anchor point softening and (2) discriminatively trained score function. Below, we will investigate their impacts on the final performance respectively.

To better investigate the tolerance of our anchor point softening strategy to anchor point localization errors, besides using the auto eye detector [2], we also create inaccurately localized anchor points by randomly disturbing the ground truth of the eye by 0, 5, 10 pixels respectively to mimic the experimental setting. For all the experiments in this section, we consider the hard anchor point strategy, which treats the anchor points as fixed, as the baseline for comparison.

From Table 4, we clearly see that in all settings the results using anchor point softening strategy are consistently better than those using hard strategy. As the disturbance range goes larger, the advantage of our soft strategy becomes more obvious. In the extreme setting that the anchor points are randomly disturbed 10 pixels from the ground truth, the percentage of feature points less the 2.5 pixels error using soft strategy is 30.6% while that by hard strategy is only 9.5%. These results highlight the capacity of our anchor point softening strategy to alleviate the impact caused by small localization error of anchor points.

Existing exemplar-based CLMs use the global likelihood as the score function to retrieve good global models, which have some drawbacks. In contrast, we discriminatively train a score function by incorporating supervised information of good/bad global models in the training data. For simplicity, we denote the global likelihood based method as GL and the discriminatively trained cost-sensitive logistic regression model as cLR respectively.

Table 4
Percentage of feature points less than different given RMSE level by different anchor point generation strategies.

RMSE (Pixels)	<2.5	<5	<7.5	<10
Hard-Disturbance-0[a] (%)	53.4	94.9	98.2	99.5
Soft-Disturbance-0[a] (%)	**56.3**	**95.7**	**98.5**	**99.7**
Hard-Disturbance-5[a] (%)	33.1	71.5	93.4	98.5
Soft-Disturbance-5[a] (%)	**51.7**	**90.1**	**97.8**	**99.1**
Hard-Disturbance-10[a] (%)	9.5	41.2	70.5	81.3
Soft-Disturbance-10[a] (%)	**30.6**	**70.7**	**90.3**	**95.4**
Hard-Auto[b] (%)	41.4	78.3	94.6	98.6
Soft-Auto[b] (%)	**53.1**	**94.5**	**97.9**	**99.5**

[a] Hard/Soft-Disturbance-*n* denotes that the anchor points are created by randomly disturbing each of the ground truth eyes by *n* pixels.

[b] Hard/Soft-Auto denotes that the anchor points are localized by our auto eye detector [2] with anchor point validation strategy.

Table 5
Comparison of the performance by different global model retrieval strategies.

Metric	Mean	Median	Min	Max
GL (Pixels)	2.71	2.63	0.81	11.12
cLR (Pixels)	**2.56**	**2.51**	**0.79**	**10.95**

Table 6
Percentage of feature points more than different given RMSE level with and without anchor point validation.

RMSE (pixels)	>5	>7.5	>10	>12.5
No APV (%)	7.8	5.4	3.1	2.9
APV (%)	**5.5**	**2.1**	**0.5**	**0.3**

Table 5 shows the comparison results measured by RMSE of the test images, and we can observe that cLR achieve better performances than GL. This validates the effectiveness of incorporating discriminative learning into the CLM framework, which is consistent with the results of [6].

5.5.3. Effect of anchor point validation

We employ the auto eye detector [2] to localize eyes as our default anchor points. Although the auto eye detector achieves an overall promising performance such that in 94.5% of the images the mean localization error of the two eyes is less than 5 pixels, and our anchor point softening strategy can greatly alleviate the bad impact of small anchor point localization errors, we still need to handle well with large anchor point localization errors. For this end, we designed an anchor point validation strategy, evaluating the anchor point localization result by a discriminatively trained classifier and sampling the peak response locations of the candidate anchor points to substitute the bad anchor points. Over a ten-fold cross validation experiments on LFW, 91% bad localized anchor points (error larger than 8 pixels) are recognized by our classifier and substituted with better anchor points.

Table 6 shows the results obtained with (second row) and without (first row) anchor point validation. We can observe that the rates of large localization errors (more than 10 pixels and 12.5 pixels localization error) are greatly reduced by our anchor point validation strategy.

5.5.4. Effect of information fusion

In our system, we employ a multi-output ridge regression method to fuse the voting map and response map in a discriminative manner. We use the error of the peak point in the local response map as the baseline, then compare the improvement by our ridge regression-based fusion (RRF) method and the greedy fusion (GF) method, i.e., multiplying the value of voting map and response map pixel by pixel.

From the comparison results in Table 7, we can see that RRF in overall outperforms GF. Specially, the results of 19.1% of the feature points are improved by more than 20% using RRF, in contrast to 12.3% using GF. These results reveal that our discriminatively trained fusion models can effectively reduce the ambiguities of local detectors and improve the robustness against appearance variations.

Table 7
Percentage of feature points more than different accuracy improvement level by different local map fusion strategies.

Improvement	>20%	>15%	>10%	>5%
GF (%)	12.3	25.5	35.4	49.1
RRF (%)	**19.1**	**29.5**	**39.6**	**55.6**

6. Conclusion

In this paper we propose a novel Robust Discriminative Hough voting based method for face alignment, under the extended exemplar-based Constrained Local Models framework. Compared to existing exemplar-based CLMs, the proposed method has two main advantages: (1) the robustness of the system against inaccurate anchor points is significantly improved with two newly-proposed methods, including discriminatively measuring the quality of the anchor points, relaxing the locations of anchor points; (2) the power of discriminative training is exploited by developing the cost-sensitive ranking of global models, and least-square based voting map alignment. Extensive experiments demonstrate the advantages of the proposed system: it significantly outperforms many recent local methods consistently across all wild datasets, and shows comparable performance to the state-of-the-art cascaded regression based methods when incorporating the deep CNN technique into the RANSAC step to choose top ranking global models.

Conflict of interest

None declared.

Acknowledgment

This work is partially supported by National Science Foundation of China (61373060), Qing Lan Project, and the Funding of Jiangsu Innovation Program for Graduate Education (KYLX_0289).

References

[1] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[2] X. Tan, F. Song, Z.-H. Zhou, S. Chen, Enhanced pictorial structures for precise eye localization under incontrolled conditions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, 2009, pp. 1621–1628.
[3] M. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2729–2736.
[4] L. Liang, R. Xiao, F. Wen, J. Sun, Face alignment via component-based discriminative search, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 72–85.
[5] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 545–552.
[6] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3444–3451.
[7] X. Jin, X. Tan, L. Zhou, Face alignment using local hough voting, in: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013 IEEE, 2013, pp. 1–8.
[8] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[9] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, in: Proceedings 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2887–2894.
[10] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2879–2886.
[11] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 532–539.
[12] P. Perakis, T. Theoharis, I.A. Kakadiaris, Feature fusion for facial landmark detection, Pattern Recognit. 47 (9) (2014) 2783–2793.
[13] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014.
[14] H. Yang, X. He, X. Jia, I. Patras, Robust Face Alignment Under Occlusion via Regional Predictive Power Estimation.
[15] Z. Zhang, W. Zhang, H. Ding, J. Liu, X. Tang, Hierarchical facial landmark localization via cascaded random binary patterns, Pattern Recognit. 48 (4) (2015) 1277–1288.
[16] J. Saragih, R. Goecke, A nonlinear discriminative approach to aam fitting, in: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, IEEE, 2007, pp. 1–8.
[17] J. Saragih, R. Göcke, Learning aam fitting through simulation, Pattern Recognit. 42 (11) (2009) 2628–2636.
[18] V. Kazemi, S. Josephine, One millisecond face alignment with an ensemble of regression trees, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014.
[19] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[20] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[21] F. Zhou, J. Brandt, Z. Lin, Exemplar-based graph matching for robust facial landmark localization, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 1025–1032.
[22] B. M. Smith, L. Zhang, Joint face alignment with non-parametric shape models, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 43–56.
[23] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models., in: BMVC, vol. 2, 2006, p. 6.
[24] B.M. Smith, J. Brandt, Z. Lin, L. Zhang, Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1741–1748.
[25] J. Yan, Z. Lei, D. Yi, S.Z. Li, Learn to combine multiple hypotheses for accurate face alignment, in: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), IEEE, 2013, pp. 392–396.
[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[27] B. Tiddeman, Facial feature detection with 3d convex local models, in: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 400–405.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[30] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[31] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vision. 57 (2) (2004) 137–154.
[32] X. Yu, J. Huang, S. Zhang, W. Yan, D.N. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 1944–1951.
[33] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, E. : Labeled faces in the wild: A database for studying face recognition in unconstrained environments.
[34] M. Dantone, J. Gall, G. Fanelli, L. Van Gool, Real-time facial feature detection using conditional regression forests, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2578–2585.
[35] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 679–692.
[36] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, IEEE, 2013, pp. 397–403.
[37] X.P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 1513–1520.

Xin Jin received the BSc and MSc degree in the department of computer science and technology from Nanjing University of Aeronautics and Astronautics (NUAA) in 2009 and 2012. He is currently a PhD student at the Department of Computer Science and Engineering, NUAA. His research interests include face recognition and computer vision.

Xiaoyang Tan received his BSc and MSc degree in computer applications from Nanjing University of Aeronautics and Astronautics (NUAA) in 1993 and 1996, respectively. Then he worked at NUAA in June 1996 as an assistant lecturer. He received a PhD degree from Department of Computer Science and Technology of Nanjing University, China, in 2005. From Sept.2006 to OCT.2007, he worked as a postdoctoral researcher in the LEAR (Learning and Recognition in Vision) team at INRIA Rhone-Alpes in Grenoble, France. His research interests are in face recognition, machine learning, pattern recognition, and computer vision. In these fields, he has authored or coauthored over 20 scientific papers.