Face Fiducial Detection by Consensus of Exemplars

Mallikarjun B R[1] Visesh Chari[1] C V Jawahar[1] Akshay Asthana[2]
1. KCIS, IIIT Hyderabad 2. Seeing Machines, Inc
{mallik.jeevan,visesh}@gmail.com jawahar@iiit.ac.in akshay.asthana@seeingmachines.com

Abstract

Facial fiducial detection is a challenging problem for several reasons like varying pose, appearance, expression, partial occlusion and others. In the past, several approaches like mixture of trees [32], regression based methods [8], exemplar based methods [7] have been proposed to tackle this challenge.

In this paper, we propose an exemplar based approach to select the best solution from among outputs of regression and mixture of trees based algorithms (which we call candidate algorithms). We show that by using a very simple SIFT and HOG based descriptor, it is possible to identify the most accurate fiducial outputs from a set of results produced by candidate algorithms on any given test image. Our approach manifests as two algorithms, one based on optimizing an objective function with quadratic terms and the other based on simple kNN. Both algorithms take as input fiducial locations produced by running state-of-the-art candidate algorithms on an input image, and output accurate fiducials using a set of automatically selected exemplar images with annotations. Our surprising result is that in this case, a simple algorithm like kNN is able to take advantage of the seemingly huge complementarity of these candidate algorithms, better than optimization based algorithms.

We do extensive experiments on several datasets, and show that our approach outperforms state-of-the-art consistently. In some cases, we report as much as a 10% *improvement in accuracy. We also extensively analyze each component of our approach, to illustrate its efficacy.*

An implementation and extended technical report of our approach is available www.sites.google.com/site/wacv2016facefiducialexemplars.

1. Introduction

Facial fiducial detection is an important problem with applications in facial expression recognition, gaze identification, face recognition etc. The task of identifying *several* locations for different components of a face in an image like ears, nose, mouth etc. becomes very daunting considering that each part might have a much more non-distinctive appearance profile than an entire face, and could also be subject to complete occlusion (Figure 1, second row, eyes), drastic appearance and illumination variation (Figure 1, third row, pose) or expression variation (Figure 1, first row, mouth). Though there is no consensus yet on even the number of fiducial points assigned to a face [21], there is a broad realization among recent papers for the necessity to reduce failure rates and increase the accuracy of fiducial detection in a wide variety of challenging examples [7, 8, 14, 21, 27, 31], since it automatically lends to better performance of systems that rely on fiducial detection.

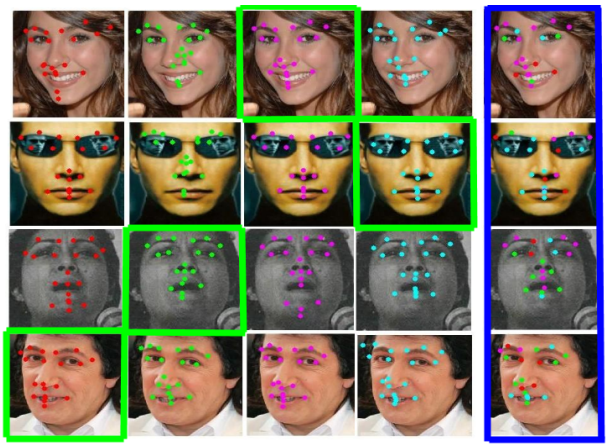


Figure 1: Fiducial detection of Chehra [3](red points), Zhu *et al*. [32](green points), Intraface [24](magenta points) and RCPR [8](cyan points) can be observed in column 1, 2, 3 and 4 respectively. Output selection by kNN is highlighted in green boxes. Last column shows the output selection by optimization highlighted in blue box. Best viewed in color.

While a number of different approaches like active shape models [15], regression based methods [27], cascaded neural networks [28], tree based methods [32] and exemplar based approaches [7] have been proposed in the recent past, many of these algorithms only address part of the problems in this area. Since datasets available today like COFW [8] , LFPW [7] (Figure 4) and AFLW [14] offer images vary-

ing widely in appearance, pose, expression, illumination and occlusion, each of these algorithms demonstrate their strengths in specific areas like occlusion handling [8], or robust performance in the case of profile views [32]. Indeed, while regression based approaches are better suited to perform well on metrics that measure pixel-wise accuracy of detection [15, 27], exemplar or mixture-of-trees based approaches [7, 32] are better suited to be more robust to pose change.

The surprising finding of our work is that many of these algorithms show decent complementarity in performance, which could be identified and exploited. In this paper, we present two algorithms that build on top of recent results in this space. Our kNN based algorithm is simple and effective, while our optimization algorithm provides a flexible framework to incorporate complicated models. Specifically, our algorithms use several state-of-the-art candidate algorithms [3, 8, 22, 24, 32] to generate fiducial points on a given image, and pose the detection problem as one of *selecting* the best result from the obtained outputs. By using several candidate algorithms, we ensure that we have access to the output of different approaches to fiducial detection, and thus reduce our problem to that of *classifying* between accurate and inaccurate fit to the data.

More formally, we propose an *initialization-insensitive*, *pose/occlusion and expression-robust* approach to face fiducial detection with the following characteristics

- Our approach attempts the problem of fiducial detection as a *classification* problem of differentiating between the best vs the rest among fiducial detection outputs of state-of-the-art algorithms. To our knowledge, this is the first time such an approach has been attempted.

- Since we only focus on selecting from a variety of solution candidates, this allows our pre-processing routine to generate outputs corresponding to a variety of face detector initialization, thus rendering our algorithm insensitive to initialization unlike other approaches.

- Combining approaches better geared for sub-pixel accuracy and algorithms designed for robustness leads to our approach outperforming state-of-the-art in *both* accuracy and robustness.

The outline of our paper is as follows. In section 2, we review related work with a perspective to distill out complimentary advantages of different approaches to fiducial detection. This is followed in section 3 by the formulation in section 3.1 and outline of our approach with focus on exemplar selection (section 3.3), output selection (section 3.4 for the kNN algorithm, section 3.5 for the optimization algorithm) and implementation details (section 3.6). We then follow up with an extensive experimental section 4, where we first show results on all the popular datasets like AFLW, COFW, LFPW and in each case present both mean part-wise pixel accuracy and failure-rate comparisons of our approach with the state-of-the-art. We finally conclude with a summary of our approach and future extensions in section 5.

2. Related Work

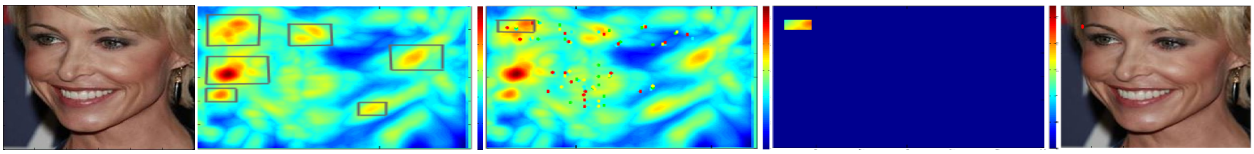
In this section, we categorize recent facial fiducial detection algorithms and discuss their advantages in brief.

Active Appearance Models (AAM): The AAM framework has existed for almost two decades [6, 12] and the traditional AAM based methods have not been suitable for fiducial detection *in the wild* [13, 17]. However, some recent methods that deviate from the traditional pixel-value based texture model have shown new promise [1, 5].

Constrained Local Models (CLMs): The CLM framework has existed for a decade [11, 18] and has been shown to be more capable of handling *in the wild* settings. In short, CLM is a part-based approach that relies on the locally trained detectors to generate response maps for each fiducial point followed by a simple Gauss-Newton method based optimization [18] for facial shape estimation. A regression based strategy for CLM optimization has also been proposed recently [2].

Exemplar Methods: Exemplar based approaches have been popular since Belhumeur *et al*.'s work [7]. Zhao *et al*. [30] use gray scale pixel values and HOG features to select k-nearest neighbor training faces, from which they construct a target-specific AAM at runtime. Smith *et al*. [20] and Shen *et al*. [19] perform Hough voting using k-NN exemplar faces, which provides robustness to variations in appearance due to occlusion, illumination and expression. Finally, Zhou *et al*. [31] combine an exemplar-based approach with graph-matching for robust facial fiducial localization. Since, we build upon outputs of candidate algorithms, we take inherent advantage of the shape based regularization schemes employed by individual approaches and thus either side-step this problem (section 3.4) or *smoothen* candidate outputs using optimization (Figure 5) in our algorithms.

Cascaded Regression Based Methods: Cascaded regression based methods are considered to be the current state-of-the-art for facial fiducial detection [4, 9, 16, 22, 25]. All these methods are capable for robustly handling *in the wild* settings in real-time. In general, the training strategy is to synthetically perturb each of the ground truth shapes and extract robust image features (SIFT or HOGs) around each of the perturbed fiducial points. The regression is then used to learn a mapping from these features to the shape perturbation w.r.t the ground truth shape. Generally, a cascaded



(a) Input Image (b) Distance from Exemplars (c) Output of Fiducials (d) Constrained Distance (e) Final Result

Figure 2: An example of fiducial detection of eye corner in a test image. Best viewed in color.

regression based strategy is adopted to learn this mapping and has been shown to converge in 4-5 iterations [4, 25].

A recent work of Smith *et al*. [21] addresses the problem of analyzing the quality of facial fiducial results using an exemplars based approach. However, several difference exist between our approaches. Firstly, they work on a completely different problem of *aggregating* fiducials from different datasets and transferring them to a target dataset through Hough based feature *detection* [19], while the goal of the work presented in this paper is to *select* the best locations for each fiducial among the candidate locations provided by various candidate algorithms on every image. Secondly, they use algorithms like graph matching to ensure that the detected fiducials resemble a face [31], while we either side-step such issues (section 3.4) or handle them using optimization (section 3.5).

Recently, some promising attempts have also been made to approach the problem of facial fiducial detection in the deep-learning framework [28]. However, most of the proposed deep-learning based models work on low resolution images [28, 29]. This prevents us from getting accurate fiducials on actual data. In this paper, we present a fully-automatic and principled approach for selecting the best fiducial location by combining results from multiple candidate algorithms for every image.

3. Face Fiducial Detection

In this section, we first outline our formulation in section 3.1, followed by our algorithm for fiducial detection. Briefly, given an input image, candidate algorithms return vectors of locations of various fiducials for that image. Given the output of each of the candidate algorithms, our task is to identify a set of fiducials that best represent the face in the input image. This can be done by either selecting the *entire* output of *one* of the candidate algorithms, or by selecting *individual* fiducials from the various outputs of candidate algorithms to form a facial structure of our own. In order to do this, we first identify a set of *exemplars* from the training dataset, that serve as guidelines on how a face should look like, both in shape and appearance. Our approach is to then *match* candidate algorithm outputs to exemplars from the training dataset, in order to *select* the best output for the given image. Our algorithm has two main components: *exemplar selection* (section 3.3) and *output selection* (section 3.4, section 3.5). A flowchart of our approach is illustrated in Figure 3.

3.1. Formulation

Let $X = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ be a variable that represents the n locations of a set of fiducials. Let $\hat{X}$ denote the true locations of fiducial features in any given image I, while X_k refer to ground truth fiducials in the exemplar set used in our algorithm, where $k = 1 \ldots K$ indexes into the set of exemplars in consideration. In this paper, we consider $K = 20$, & $n = 20$ since that is the set of common fiducials detected by algorithms presented in recent literature [3, 8, 22, 24, 32]. Note that recent approaches [21] offer a way to increase the number of common fiducial locations, and thus our assumption is not restrictive. Let $R = \{\mathbf{r}^1, \ldots, \mathbf{r}^m\}$ represent features extracted at m pixels on the image. We would like to optimize the following function to obtain the fiducial locations at the current image

$$X^* = \arg\max_{\tilde{X}} P(X \mid R) \tag{1}$$

Note that $\tilde{X}$ is the space of all possible sets of fiducial locations. It is a huge (40 dimensional) space, and sampling all of it is impractical. Instead, let us assume that we have been given some candidate locations where probability of a correct result is higher, and assume we will pick X from one of these locations. Let us depict these locations with the variable $\mathcal{X} = \{\bar{X}_1, \ldots, \bar{X}_l\}$, where $\bar{X}_i, i = 1 \ldots l$ are the number of candidates we have selected. We can now re-write equation 1 as

$$X^* = \arg\max_{\tilde{X}} P(X \mid R, \mathcal{X}) = \arg\max_{i} P(\bar{X}_i | R) \tag{2}$$

where we assume that the probability of selecting fiducials not represented by candidate algorithms is negligible. Using Bayes rule, and adopting a similar strategy of marginalizing over exemplars used in [7], equation 2 can now be elaborated as

$$P(\bar{X}_i \mid R) \propto P(R \mid \bar{X}_i) \tag{3}$$

$$\propto \sum_{k \in K} P(R \mid X_k, \bar{X}_i) P(X_k \mid \bar{X}_i) \tag{4}$$

where we marginalize over all exemplars X_k. Note that equation 4 *splits* the probability into comparison between

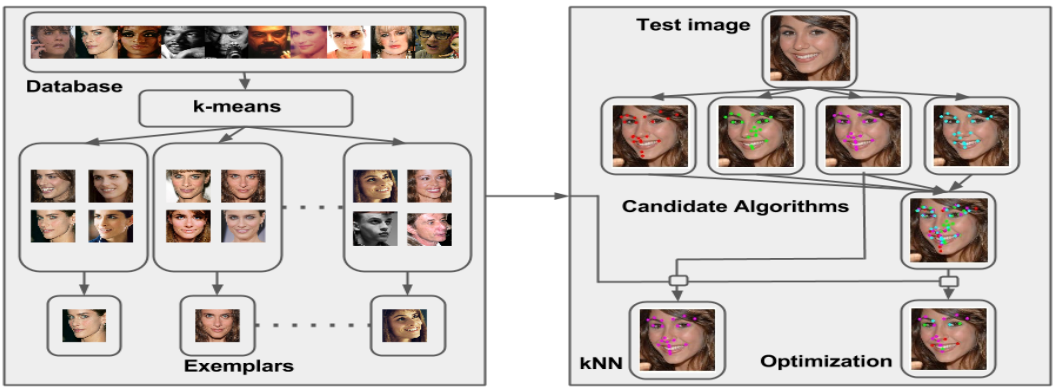


Figure 3: Left box pictorially represents exemplars selection. Right box represents our two algorithms for output selection. One by using kNN approach and other using optimization. Best viewed in color.

appearances of our candidates and exemplars (first term), and comparison between their shapes (term 2). Further, given structure is preserved in the way these two sets of candidates are generated, we can breakdown the above equation into parts

$$P(\bar{X}_i \mid R) \propto \sum_{k \in K} \prod_j P(R \mid \mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) P(\mathbf{x}_k^j \mid \bar{\mathbf{x}}_i^j) \qquad (5)$$

We denote individual probabilities for shape and appearance using the following functions

$$P(R \mid \mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) = (1/\alpha) \exp(-\|F_k^j - F_i^j\|^2) \qquad (6)$$

$$P(\mathbf{x}_k^j \mid \bar{\mathbf{x}}_i^j) = (1/\beta)\ dist(\mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) \qquad (7)$$

where F denotes concatenation SIFT and HOG features, while $dist$ is a scaled inverse Euclidean distance function and α, β are normalization constants to ensure both equations represent valid probabilities. Note that evaluating equation 5 entails summing over SIFT and HOG distances between candidate and exemplar fiducials. Finally, one could alternatively choose to optimize equation 2 using an optimization function as outlined in section 3.5. In this work, candidates are generated using algorithms of Zhu *et al*. [32], Xiong *et al*. [24], Asthana *et al*. [3], Artizzu *et al*. [8], and Tzimiropouluos *et al*. [22].

Example In equation 5, the term $P(R \mid \mathbf{x}_k^j, \bar{\mathbf{x}}_i^j)$ can be seen as the term that *selects* appropriate exemplars given fiducial candidates using a *shape/appearance constraint* represented by equation 6. This is better illustrated with an example. In Figure 2, we show an input image for which the *minimum distance* in SIFT+HOG space from a set of exemplars is shown in Figure 2b, for a single fiducial (eye corner). Note how there are several minima in the distance map (marked by bounding boxes). Running candidate detection algorithms, however, generates eye fiducial candidates only in a specific region (Figure 2c, with bounding box), which is then selected and isolated using equation 7 (Figure 2d), leading to a correct location of the eye fiducial in the final output (Figure 2e).

3.2. Algorithm Outline

As explained earlier, our algorithm is divided into two main sub-parts: *exemplar selection* and *output selection*. The task in exemplar selection is to select a subset of face images with ground truth annotations from the training dataset, that are representative of the *variation of pose, appearance including occlusion, expression etc*. of the dataset in consideration. Algorithm 1 gives an outline of our approach to exemplar selection. Note that while, we could use the entire training dataset annotations as exemplars, it suffices to have this limited set, as we will show in section 4.2.

This subset of annotated images then serve as our basis for differentiating between the various candidate algorithm outputs on any test image. The process of selecting the best fitting fiducials on any test image, given the exemplars, is called output selection.

3.3. Exemplar Selection

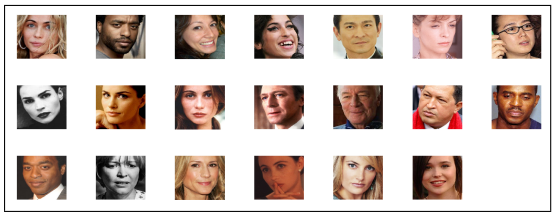
Exemplar selection is the process of selecting a subset of the training images along with fiducial annotations that represent the range of variations in pose/expression/occlusion in the dataset. We term the set of images selected eventually as the *exemplar set*. Ideally we would like the exemplar set to be representative of the training set in that we would like to be able to describe the pose/appearance of all images in the training set as some combinations of images in the exemplar set, in a specific representation space. For example, given annotations of fiducial locations in the training set, we would like have an exemplar set such that the shape of any training image annotation (represented as an ordered list of pixel coordinates of various fiducial points) is a *linear* combination of the annotations in the exemplar set.

Algorithm 1 illustrates our basic exemplar selection algorithm. The function `ComputeClusters` performs the

Algorithm 1 Algorithm for Exemplar Selection (`ComputeDatasetExemplars`)

input Training image data $\mathcal{D}$, fiducials $\mathcal{F}_d$.
  $\mathcal{E} = \varnothing, \mathcal{R} = \varnothing, \mathcal{S} = \varnothing, \mathcal{F} = \mathcal{F}_d$
  $Cntrs$ = `ComputeClusters`($\mathcal{F}$, N_{clus})
  for Each center $C_k \in Cntrs$ **do**
    $[I_i, F_i]$ = `ClosestFiducial`($\mathcal{F}, \mathcal{D}, C_k$)
    $Feat_i$ = `ComputeFeatures`(I_i, F_i)
    $\mathcal{E} = \mathcal{E} \cup \{I_i, F_i, Feat_i\}$
    $\mathcal{F} = \mathcal{F} \setminus F_i$
  end for
  for Each image-fiducial pair (I_i, F_i) in $(\mathcal{D}, \mathcal{F}_d)$ **do**
    $Feat_i$ = `ComputeFeatures`(I_i, F_i)
    $\mathcal{R} = \mathcal{R} \cup Feat_i$
  end for
  $\mathcal{F} = \mathcal{F}_d$
  $Cntrs^{app}$ = `ComputeClusters`($\mathcal{R}$, N_{clus})
  for Each center C_k in $Cntrs^{app}$ **do**
    $[I_i, F_i, Feat_i]$ = `ClosestFeat`($\mathcal{R}, \mathcal{F}, \mathcal{D}, C_k$)
    $\mathcal{S} = \mathcal{S} \cup \{I_i, F_i, Feat_i\}$
    $\mathcal{R} = \mathcal{R} \setminus Feat_i, \quad \mathcal{F} = \mathcal{F} \setminus F_i$
  end for
output $\mathcal{E}, \mathcal{S}$

operation of `kmeans` clustering in the vector space of fiducials, or feature vectors depending upon its input arguments. While the algorithm outputs two datasets for shape based and appearance based exemplars, note that shape based exemplars can be further divided into pose and expression classes and appearance based exemplars can also be tuned to include some examples of occlusion. However, we found that `kmeans` inadvertently does this since it clusters fiducials of the same pose but varying expression (shape clustering) or occlusion (appearance clustering) into one cluster.



(a) LFPW Exemplars

Figure 4: Examples automatically selected by our clustering approach in Section 3.3. Best viewed in color.

3.4. Output Selection by kNN

Once the fiducial detection of the state-of-the-art candidate algorithms are obtained for an input image, we compute appearance vectors for an image patch around each fiducial location. Appearance vectors are represented in HOG and SIFT space. We concatenate these features them to form the feature vector.

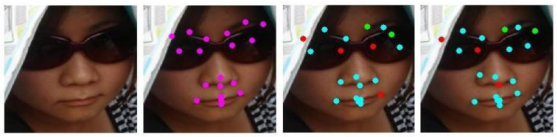


Figure 5: From left to right, we observe input test image, output selection by kNN, output selection by optimization without structural costs and output selection by optimization with structural costs. Observe that the left eye prediction suffers in third image because of not considering structural costs for optimization.

We then compare these candidate algorithm feature vectors to the exemplars chosen from the previous approach, and choose the candidate algorithm-exemplar image output that minimizes the sum of euclidean distance between common features (equation 5). Note that this is a simple kNN based approach, where k=1. Alternatively, we also consider the idea of selecting individual fiducials from various candidate algorithm outputs, to form our own facial structure that minimizes an objective function. This is explained in the following section.

3.5. Output Selection by Optimization

Instead of selecting fiducals from one method for all the parts as explained in earlier section, here we propose a method which selects fiducials for each part from best performing method. We first collect fiducials from all the candidate algorithms on an input image. Our task is now to select a subset of these fiducials for our output.

We propose an optimization framework based on equation 2, where we minimize a function based on appearence and structural costs. The appearance cost forces the areas around the fiducial locations in the input image to "look" like a face, while the structural cost ensures that the outline of fiducial locations resembles a facial structure. We define a quadratic objective function with unary and binary terms that enforce these constraints. Unary terms enforce appearance costs, while binary terms enforce structural costs.

The selection of the j^{th} fiducial from the i^{th} method is represented by the binary variable x_i^j. Let u_i^j be its appearence cost. Let y_{cd}^{ab} be the selection variable which will be 1 when both x_c^a and x_d^b are 1. And, p_{cd}^{ab} define the structural cost when y_{cd}^{ab} is 1. Thus y_{cd}^{ab} is the binary variable that represents *joint selection* of fiducials corresponding to unary variables x_c^a and x_d^b.

Appearance Costs: We would want the fiducial prediction for each part to look similar to the corresponding fiducial of *one* of the exemplars. To do this, we compare the appearance feature vectors (using SIFT and HOG) between the fiducial x_i^j and that of the corresponding fiducials in the

exemplar database. Let $f(x_i^j)$ represent the appearance feature vector corresponding to the j^{th} fiducial produced by the i^{th} method. We define the unary costs as

$$u_i^j = \arg\min_k \|f(x_i^j) - f(\mathcal{E}_k^j)\|^2 \quad (8)$$

where $\mathcal{E}_k^j$ denotes the j^{th} part of the k^{th} exemplar. Let $m(j,i)$ represent the exemplar index that has the fiducial closest in appearance to that of x_i^j. That is, let $u_i^j = \|f(x_i^j) - f(\mathcal{E}_{m(j,i)}^j)\|^2$.

Structural Costs: We would also want to preserve the facial structure while selecting fiducials. This is most naturally enforced in the binary variable cost p_{cd}^{ab}. The importance of this cost is depicted in Figure 5. We enforce structural consistency by ensuring that if two fiducials x_c^a and x_d^b are selected, their corresponding closest exemplars (given by indices $m(a,c)$ and $m(b,d)$ as mentioned above) are as close to each other in shape as possible. Thus, we define the structural cost p_{cd}^{ab} as the euclidean distance between the shape of exemplars $\mathcal{E}_{m(a,c)}$ and $\mathcal{E}_{m(b,d)}$. Note that the structural cost is only defined between two variables that *do not* represent the same fiducial. That is

$$p_{cd}^{ab} = \|s(\mathcal{E}_{m(a,c)}) - s(\mathcal{E}_{m(b,d)})\|^2, \qquad a \neq b \quad (9)$$

where $s(\cdot)$ is the function that denotes the shape of a set of fiducials (represented as a vector of fiducial locations). Additionally, we also want to enforce the constraint that the same fiducial from different methods should not be simultaneously selected. This is easily enforced by the constraint

$$\sum_i x_i^j = 1 \quad (10)$$

Combining all the above, we want to minimize the following function function,

$$O(X,Y) = \sum_{i=1}^{5}\sum_{j=1}^{20}(x_i^j \times u_i^j) + \sum_{c=1}^{20}\sum_{d=c+1}^{20}\sum_{a=1}^{5}\sum_{b=1}^{5}(y_{cd}^{ab} \times p_{cd}^{ab}) \quad (11)$$

subjected to constraints, $x_i^j \in \{0,1\}$, $y_{cd}^{ab} \in \{0,1\}$, $\sum_{i=1}^5 x_i^j = 1, y_{cd}^{ab} = x_c^a \times x_d^b$

Since the above problem has quadratic constraints and can not be solved in polynomial time, as the solutions are in integers, we relax the constraints [10] to get: $0 \leq x_i^j \leq 1$, $0 \leq y_{cd}^{ab} \leq 1$, $x_c^a \geq y_{cd}^{ab}$, $x_d^b \geq y_{cd}^{ab}$, $x_c^a + x_d^b \leq y_{cd}^{ab} + 1$. Thus, we obtain the final linear optimization problem as

$$\begin{aligned} O(X,Y) \quad = \quad & \sum_{i=1}^{5}\sum_{j=1}^{20}(x_i^j \times u_i^j) + \\ & \sum_{c=1}^{20}\sum_{d=c+1}^{20}\sum_{a=1}^{5}\sum_{b=1}^{5}(y_{cd}^{ab} \times p_{cd}^{ab}) \\ & 0 \leq x_i^j, y_{cd}^{ab} \leq 1, x_c^a \geq y_{cd}^{ab}, x_d^b \geq y_{cd}^{ab} \\ & x_c^a + x_d^b \leq y_{cd}^{ab} + 1 \end{aligned} \quad (12)$$

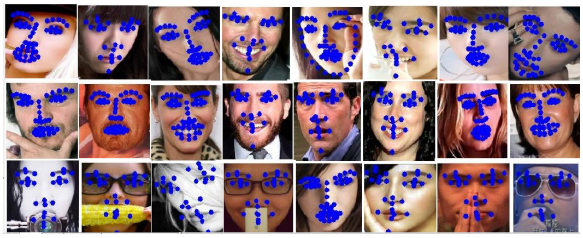


Figure 6: Results with varying pose (Row 1), expression (Row 2) and occlusion (Row 3). Best viewed in color.

We use `MOSEK` wrapper in MATLAB to solve the above optimization problem. Sometimes, because of the non-linear nature of the problem, we get non-integer solutions for x_i^j. In such cases, we take our fiducial location to be the average position of the top two selected outputs for the j^{th} part.

3.6. Implementation Details

In this section, we present some implementation details of the paper along with threshold values.

To compute the appearance vector around each fiducial part, we take 10x10 pixel patches and extract HOG features with a cell size of 3. We also compute the SIFT features around facial fiducial locations at two different scales of 5 and 8 pixels. After concatenating both the features, we obtain a vector of dimension 535 for each part. This is repeated for all the fiducial parts for both candidate algorithms and exemplars. For the experimentation, we used 20 clusters in k-means algorithm to automatically choose the trainnig samples to be used for kNN selection.

We took the author released code for candidate algorithms [3, 8, 22, 24, 32] along with the trained models. Experiments were conducted on the same test split for candidate and our algorithms for all the datasets.

4. Results

Thus far, we have outlined our approaches to fiducial detection in the previous sections. In this sections, we evaluate our algorithms on three state of the art datasets LFPW, COFW and AFLW. Before we present the quantitative result (produced in Table 1) in the remaining part of this section, we describe the 3 datasets in brief below.

We have chosen 3 popular datasets to test the performance of our algorithm for several reasons.

LFPW is the oldest dataset we consider [7], and contains faces of several people in "wild" settings, with lots of occlusions and pose / expression variation. It contains 1035 images, out of which 811 are used for training and 224 are used for testing purposes. Ground truth annotation of training images in the form of 68 fiducial locations for each face is available to us. This dataset has been standard for some

time, but current algorithms give very good performance on it.

COFW is a dataset released by Burgos-Artizzu *et al*. [8], and is specialzed to highlight situations where faces are occluded in a manner that hinders accurate fiducial detection by state-of-the-art algorithms. It contains 1852 images, out of which 1345 are used for training and 507 are used for testing purposes. Ground truth annotation of training images in the form of 29 fiducial locations for each face is available to us. This dataset is relatively new, and moderate performances have been reported on it.

AFLW is a dataset released by [14], and contains several annotated face images in extreme settings. It is considered one of the toughest datasets in fiducial detection literature [21, 28], as it has larger pose variations, partial occlusions and illumination variation compared to other datasets. Like [28], we sample 1000 training images and 3000 testing images randomly from the dataset, while ensuring no overlap between the two sets.

4.1. Quantitative Results

In this section, we outline the basis for future experiments detailed in the next sections. Table 1 shows results of our approach on LFPW, COFW and AFLW datasets. To produce these results, we first resize *all* images (training and testing) to a size of 300×300, and compute a set of 20 exemplars for each dataset using Algorithm 1, equally divided between shape and appearance. Figure 4 illustrates our results of exemplar selection on the LFPW dataset. SIFT features for each fiducial are calculated at the scale of 5 and 8 pixels, which roughly translates to 4% and and 6% of the interocular distance. Once this is done, we proceed to the output selection by kNN and optimization based algorithms.

For each test dataset in Table 1, mean errors and failure rates in locating fiducials over the entire dataset are shown. For each fiducial, we first compute the ratio of its Euclidean distance from the ground truth and the interocular distance for that image. We then average this ratio over the entire image and over the entire dataset. Thus the first table represents the *average ratio of fiducial error and interocular distance over the entire dataset*. The failure rate is the fraction of images in the entire dataset, for which this ratio is more than 0.1 (10% error). Thus, while mean error gives an idea of the accuracy of our algorithm, the failure rate gives an idea of its robustness.

A more detailed quantitative comparison of our approach with candidate algorithms is presented in Figure 7. Each point on the x-axis of this figure represents a cut-off threshold, and each corresponding point on the y-axis of this figure represents the fraction of images that have mean normalized error greater than this cut-off. Thus, graphs that dip quickly are more accurate. The mean normalized error is the mean of all interocular distance normalized errors over the entire dataset. We notice that both of our algorithms consistently perform better compared to other five algorithms at almost all cut-off ranges. Figure 6 illustrates some qualitative results using our approach.

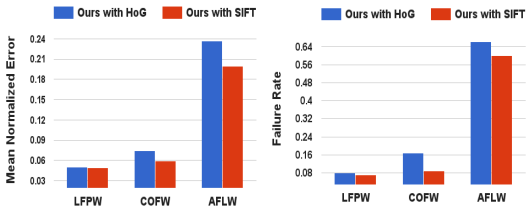


Figure 8: Comparison of mean error and failure rate for SIFT vs HOG experiment. Best viewed in color.

4.2. Experimental Analysis

In the previous section, we outlined our basic algorithm and illustrated its results that show superior performance compared to state-of-the-art on three datasets. In this section we analyze various components of our algorithm to illustrate how our approach performs under different settings. Detailed results are provided in the website.

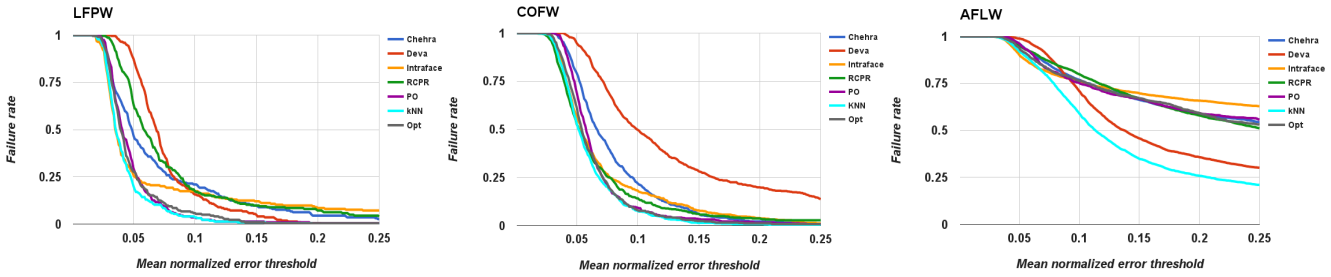
Runtime For both approaches, candidate algorithms can be run in parallel and hence the total time taken by them on an input image is the maximum time of any algorithm. As an overhead, we compute SIFT/HOG based features on the output of these algorithms, which measures in milliseconds since fast GPU based approaches are available for such computations. On top of that, the output selection part uses Euclidean distance computation for kNN, which amounts to 5 (candidate algorithms) x 20 (exemplars) distance computations between 535 dimensional vectors (of SIFT/HOG features). Finally, the optimization algorithm takes 0.4 seconds to converge for a single input image on a Intel(R) Xeon(R) CPU E5-2640 0 @ 2.50GHz system.

SIFT vs HoG In this experiment, we contrast the contribution of SIFT and HOG features for the task of output selection. Results of our experiment comparing mean errors and failure rates on all datasets are shown in Figure 8b. Note that SIFT outperforms HOG, and understandably so since SIFT captures appearance details lost to HOG. We get an improvement of 6% using SIFT and 2% using HOG over competing methods.

Varying Number of Exemplars Varying the number of exemplars ideally affects the accuracy of fiducial location, since more exemplars should typically mean that the nearest neighbor should be more similar to the test image. However

Dataset	Mean Error							Failure Rate						
	Chehra	Zhu	Intraface	RCPR	PO	Ours (kNN)	Ours (Opt)	Chehra	Zhu	Intraface	RCPR	PO	Ours (kNN)	Ours (Opt)
LFPW	7.21	7.60	7.79	9.28	**4.82**	**4.31**	4.83	20.98	15.62	17.41	17.41	**3.57**	**3.57**	5.8
COFW	7.95	15.76	7.22	7.30	6.73	**5.98**	**6.28**	21.89	49.70	18.15	14.20	9.27	**7.49**	**7.88**
AFLW	40.44	**25.88**	47.98	39.78	46.67	**19.93**	32.08	80.52	**71.28**	79.80	82.12	75.20	**59.03**	76.30

Table 1: Table shows the mean error and failure rate for three datasets. In each row, top two algorithms are highlighted for both mean error and failure rate. Opt in the table represents output selection by optimization. Observe that both of our algorithms consistently perform better than state-of-the-art algorithms.



(a) (b) (c)

Figure 7: Results of our approach on (a) LFPW, (b) COFW, and (c) AFLW datasets. Drop in failure rate with the change in cut-off threshold of mean error normalized with interocular distance. Lower curve means more accurate results. Best viewed in color.

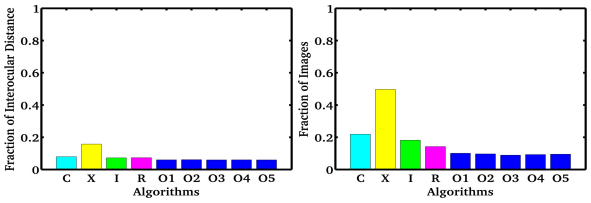


Figure 9: Comparison of mean error and failure rate when the number of exemplars is increased. Results O1-O5 correspond to our algorithm with number of exemplars (20, 30, 40, 50, 60) respectively. Best viewed in color.

if most variations in pose, expression, partial occlusion have been already captured, increasing the number of exemplars will have minimal effect on accuracy. This is precisely what we observe in Figure 9.

Optimization with structural costs In this experiment, we show qualitative result of output selection by optimization with and without structural costs. Structural costs help in optimizing to a solution which looks like face. If only appearance costs are used, it leads to just selecting best looking fiducials individually leading to distortion in facial structure which can be observed in third image of Figure 5.

5. Summary and Conclusion

Facial fiducial detection is challenging in case of severe pose, expression and occlusion variation. We propose two robust algorithms which take care of these conditions with help of consensus of exemplars. We consistently outperform considered state-of-the art algorithms.

Optimization vs kNN Note that one of the major advantages of using optimization is that one could select each fiducial separately, and *enforce* a global shape constraint on the final output. Such a task is not obvious in the kNN based approach, and thus the optimization based approach gives us lot of flexibility. In addition, other models of errors in candidate algorithms or explicit occlusion / expression based constraints can also be included in the optimization framework mentioned above. Finally, note that even with simple global constraints, our optimization algorithm is mostly able to perform as well as the kNN algorithm (Table 1) and better than almost all state-of-the-art algorithms.

References

[1] E. Antonakos, J. A. i medina, G. Tzimiropoulos, and S. Zafeiriou. Hog active appearance models. In *Proceedings of IEEE International Conference on Image Processing (ICIP'14)*, pages 224–228, Paris, France, October 2014. (Top 10

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, Oregon, USA, June 2013.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, 2014.

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *In Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014.

[5] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI). In Press.*, 2015.

[6] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical report, RI, CMU, USA, 2003.

[7] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *PAMI*, 35(12):2930–2940, December 2013.

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

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

[10] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic. On pairwise cost for multi-object network flow tracking. *CoRR*, abs/1408.3304, 2014.

[11] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[12] G. Edwards, C. Taylor, and T. Cootes. Interpreting Face Images Using Active Appearance Models. In *IEEE FG*, 1998.

[13] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *IMAVIS*, 23(12):1080–1093, 2005.

[14] M. Koetsinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *In First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[15] S. Milborrow and F. Nichols. Locating facial features with an extended active shape model. In *CVPR*, 2008.

[16] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1685–1692, June 2014.

[17] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, Nov. 2009.

[18] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, Jan. 2011.

[19] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013.

[20] B. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization. In *CVPR*, 2014.

[21] B. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *ECCV*, 2014.

[22] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015.

[23] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*, 2014.

[24] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[25] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *CVPR*, 2013.

[26] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *ICCV*, 2013.

[27] X. Yu, Z. Lin, J. Brandt, and D. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *ECCV*, 2014.

[28] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning and transferring multi-task deep representation for face alignment. *CoRR*, 2014.

[30] X. Zhao, S. Shan, X. Chai, and X. Chen. Locality-constrained active appearance model. In *ACCV*, 2012.

[31] F. Zhou, J. Brandt, and Z. Lin. Exemplar-based graph matching for robust facial landmark localization. In *ICCV*, 2013.

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