Exemplar based approaches on Face Fiducial Detection and Frontalization

Thesis submitted in partial fulfillment
of the requirements for the degree of

MS by Research
in
Computer Science & Engineering

by

Mallikarjun B R
201307681
mallikarjun.br@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, India
May 2017

To Family

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Exemplar based approaches on Face Fiducial Detection and Frontalization" by Mallikarjun B R, has been carried out under my supervision and is not submitted elsewhere for a degree.

01-07-2017
Date

Adviser: Prof. C V Jawahar

Acknowledgments

I would like to thank Prof C V Jawahar, Prof P J Narayanan, Prof Jayanthi Sivaswamy and Prof Anoop M. Namboodiri for their dedicated work in fostering a lab of such high standard. This has helped me and many other students to work in a conducive environment and getting the much required exposure in a field with vast diversity. I sincerely thank my advisor, Prof C V Jawahar for not just his guidance in research work, but also for being patient with certain decisions I made during the course of my stay at IIIT. He has guided me when I failed and helped me develop a healthy attitude towards research. I would like to thank my research mentor, Dr Visesh Chari for all the helpful advices and discussions which helped me get a better hang of research work.

I have to mention about the friends I made here, Brij, Anirudh, Sarthak, Raghu, Anurag. I would like to thank them for all those useless yet fun discussions, for helping me through quarter-life crisis, for their company to innumerable biryani and DLF late night visits. I will cherish these moments for life. Most importantly, I like to thank Appa and Amma for giving me the freedom to choose what I want and their unwavering belief and support. I like to thank Akka for being the constant friend out there.

Abstract

Computer vision solutions such as face detection and recognition, facial reenactment, facial expression analysis and gender detection have seen fruitful applications in various domains such as security, surveillance, social media and animation. Many of the above solutions have common pre-processing steps such as fiducial detection, appearance modeling, face structural modelings *etc*. These steps can be considered as fundamental problems to be solved in building any computer vision solutions concerning face images.

In this thesis, we propose exemplar based approaches to solve two fundamental problems, such as face fiducial detection and face frontalization. Exemplar based approaches have been proved to work in various computer vision problems, such as object detection, image impainting, object removal, action recognition, gesture recognition. This approach directly utilizes the information residing in the examples to achieve a certain objective, instead of coming up with a model representing all the examples and has shown to be effective.

Face fiducial detection involves detecting key points on the faces such as eye corner, nose tip, mouth tips *etc*. It is one of the main pre-processing step done for face recognition, facial animation, gender detection, gaze identification and expression recognition systems. Number of different approaches like active shape models, regression based methods, cascaded neural networks, tree based methods and exemplar based approaches have been proposed in the recent past. Many of these algorithms only address part of the problems in this area. We propose an exemplar based approach which takes advantage of the complimentarity of different approaches and obtain consistently superior performance over the state-of-the-art methods. We provide extensive experiments over three popular datasets.

Face frontalization is the process of synthesizing frontal view of the face given a non-frontal view. Method proposed for frontalization can be used in intelligent photo editing tools and also aids in improving the accuracy of face recognition systems. Methods previously proposed involve estimating the 3D model or assuming a generic 3D model of the face. Estimating an accurate 3D model of the face is not a completely solved problem and assumption of generic 3D model of the face results in loss of crucial shape cues. We propose an exemplar based approach which does not require 3D model of the face. We show that our method is efficient and performs consistently better than other approaches.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation and Objectives

Computer vision is a field that includes making inference out of images and videos. There has been decades of research happening in this field. We have seen fruitful applications deployed in various domains. Be it reconstruction of 3D world, which helps in automatic car navigation, analysis of medical images for early detection of diseases, analysis of astronomical images to study universe in general, bio-metric systems for security purposes using finger print or face images. Even though there has been decades of research effort, problems mentioned above are not completely solved as there is demand for more precision and efficient solutions. Technological advancement from other ends such as camera capabilities, computational capacity aids in solving problems which were not feasible earlier and also in improving existing solutions.

In this thesis, we concentrate on approaches and applications concerning face images. Vision community has been working on problems such as face detection, face recognition, expression detection, gaze identification, face reenactment *etc*., for a long time now. Also, we see these systems deployed in various domains for practical use [1], [2]. Most of the above solutions depends on pre-processing steps such as representation, face fiducial detection, structural modeling *etc*., which influences the effectiveness of the overall system to a great extent. We want to further improve the effectiveness of fiducial detection and structural representation of face which in-turn help in improving the accuracies of the above mentioned systems.

Most challenging aspect of face recognition or verification system is to handle the variation in pose. Assume if there is a method to perfectly align all the faces to one space such that the eyes, ears, nose, mouth *etc*., of all the faces fall on the same location, recognition and verification can be done with a simple KNN approach. Schroff *et al*. [36] relies on a purely data driven approach to attain invariance to pose. This method is suitable when there is large data. Zhenyao *et al*. [61] employ a deep network to *warp* faces into a canonical frontal view and then learn CNN that classifies each face. Sun *et al*. [43]

relies on extracting features from different face patches to counter pose variation. Facial fiducials help in aligned feature representation as we could extract feature corresponding to various key points on the face independent of overall pose and hence achieve invariability to pose.

Facial expression recognition systems such as Chew *et al*. [14], Valstar *et al*. [47] use face fiducial detection as their starting point of the system. Statistical distribution facial landmarks with each other helps in predicting accurate expression. Similarly the gender detection system would need facial landmarks as their initial point as fiducials are differently distributed for different gender in general.

We also address the problem of face frontalization. This approach helps in improving the accuracy of face recognition system. Data is expensive. To obtain all possible poses of a given person to come up with a model for recognition system is a tedious task. Models generated with fewer images of the person which predominantly includes frontal pose faces wouldn't perform well on profile view test image. Approach proposed for face frontalization can also be used for face reenactment where video of a person can be mimicked by replacing the face of another person.

1.2 Problem Statement

In this section, we briefly describe the problem of face fiducial detection and face frontalization along with related work.

1.2.1 Face Fiducial Detection

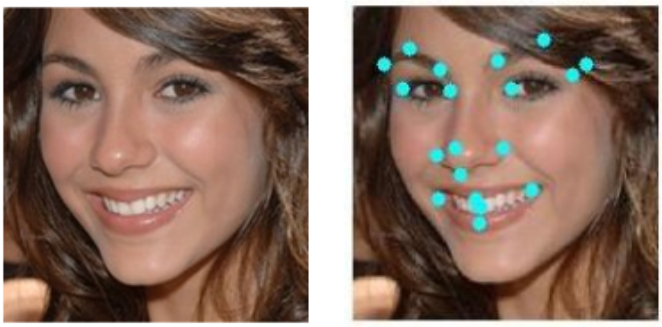


Figure 1.1: Left Image: Input to face fiducial detection algorithm. Right Image: Output of the algorithm with co-ordinates of the landmarks.

Face fiducial detection is a problem of detecting key points on the face like eye corners, nose tip, mouth corners *etc*., given a face image (refer Figure 1.1). It is a challenging problem considering the influencing factors which include physical phenomena like camera distortion, projective geometry, mul-

tisource lighting, biological appearance, facial expression, and the presence of accessories like glasses and hats.

Face fiducial detection is intrinsically linked with the head pose estimation, visual gaze identification and emotion recognition. Head pose estimation can be considered as a more coarser level problem compared to that of fiducial detection as it involves only inferfing the orientation of human head from the image. Head pose estimation comes for free if the fiducial detection problem is solved as it just a mapping of location to orientation as they are highly correlated. Also, perceived gaze identification directly depends on the head pose at coarser level. Finer level estimation of eye direction would require location of eyes, which can be inferred from fiducial detection.

Traditional face fiducial detection methods can be categorized into two types, namely regression based method and template fitting method. Most of the methods iteratievely improvize the initial estimates by regression using image features. Support vector vegression has been employed by Valstar *et al*. [48] and Burgos-Artizzu *et al*. [11] employ cascaded fern regression. Image features like pixel-difference features and Haar-like features have been used. Since most of the regression based methods start with an initial estimate of the locations, they are prone to propogate error with wrong initialization. Template based methods rely on pre-built templates to fit the input images. Part based method [60] build model of each pre-defined parts and come to consensus using voting made by each part model on the input image using a tree representation, which can model the space constraints between parts.

1.2.2 Face Frontalization

Figure 1.2: Left Image: Input to face frontalization algorithm. Right Image: Output with the frontalized version of input.

Face frontalization is a problem of synthesizing frontal view of the face given an non-frontal face image (refer Figure 1.2). Synthesizing novel views of the face has been longstanding challenge in computer vision mainly because of the potential application in graphics domain and recognition sytems. It is a challenging problem considering faces with unconstrained scenarios with occlusions and specularities.

Face recognition methods recently have claimed to reach the accuracy of that of humans even in the unregulated, uncontrolled image settings. These approaches differ in addressing the problem of unconstrained settings. Unconstrained settings includes non-frontal pose, lighting, expression variations and noise. One way to address the problem of non-frontal pose is to align the pose of the face to frontal pose to make the job easy for face recognition system. Pose correction technique can be used to correct the pose of any individual in a group photo if they are not looking towards the camera. Addressing the problem of novel face pose can also be used in face reenactment systems where one can create a mimicry video of a person by replacing the original actor.

General approach to synthesize novel face pose includes estimating a 3D model [52] of the face from a single image and then rendering the 3D model from a different view angle on a 2D image. This approach intuitively seems good, but extracting 3D information out of 2D image is a challenging problem. By relaxing certain constraint, one can think of assuming a generic 3D model [22] of a face and try to get an approximate estimate, but this leads to loss of crucial structural information.

1.3 Methodology

> *Represent all the data with a non-parametric model rather than trying to summarize it with a parametric model, because with very large data sources, the data holds a lot of detail... Now go out and gather some data, and see what it can do.*
>
> – Alon Halevy, Peter Norvig, and Fernando Pereira, The Unreasonable Effectiveness of Data (Google, 2009)

The above statement is more so true when the data at hand is diverse at many fronts. In our case, the data is the face images which can have variations in terms of age, gender, expression, pose, facial structure *etc*. Trying to come up with a parametric model which holds all these information leads to poorly performing systems. Also whenever there is new data, the model has to be updated which is not so efficient. Rather it is better to go with the data driven solutions.

Consider coming up with a generic model which represents an eye. Since the model has to perform in all different scenarios, we need to train the model with all possible variations. The samples could have variations with respect to the type of eyes, open or closed, visible or occluded, in frontal or profile view. When these samples are represented in appearance space, they end up in subspaces far apart. Training a model to capture all the above variations will be difficult. We could think of coming up with separate model for each of the variation above, but coming up with the labeled data for all these variations is tedious task.

Classical Exemplar theory in psychology about the way humans categorize objects state that individual make category decision by comparing the new stimuli with the instances already existing in memory. The instances in memory are called exemplars. Other way to accomplish the same task is based on learnt rules. In this thesis, we explore the exemplar based approach for problems concerning face images because learning a faithful model of such high dimensional data from limited samples is a challenging task. And also to exploit the available semantically annotated data for our advantage.

For face fiducial detection, we employ 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. We also propose two different ways in which the exemplars can be selected and provide analysis of how the performance is affected in choosing between two methods.

For face frontalization, we employ an exemplar based approach to find the transformation that relates the profile view to the frontal view, and use it to generate realistic frontalizations. Our method does not involve estimating 3D model of the face, which is a common approach in previous work in this area. This leads to an efficient solution, since we avoid the complexity of adding one more dimension to the problem.

1.4 Contributions and Novelties

In this thesis, we propose exemplar based approaches for two fundamental problems related to face images. In both the cases, we provide extensive experimental analysis to show that the proposed approaches perform superior to the state-of-the-art methods on popular datasets. Face fiducial detection approach manifests as two algorithms, one based on optimizing an objective function with quadratic terms (refer to Section 3.4.5) and the other based on simple KNN(refer to Section 3.4.4). Proposed face frontalization approach can be used either as a pre-processing step in face recognition, gender identification algorithms or also in rendering face video for face reenactment.

Proposed face fiducial is *initialization-insensitive, pose/occlusion and expression-robust* approach 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.

We compare our approach with five of *state-of-the-art* methods on three popular datasets such as LFPW, COFW and AFLW. In some cases, we report as much as 17% improvement in the accuracy.
For face frontalization, we employ an exemplar based approach to find the transformation that relates the profile view to the frontal view, and use it to generate realistic frontalizations. In specific,

- Our method does not involve estimating 3D model of the face, which is a common approach in previous work in this area. This leads to an efficient solution, since we avoid the complexity of adding one more dimension to the problem
- Our method also retains the structural information of the individual as compared to that of a recent method, which assumes a generic 3D model for synthesis

We compare our approach with a recent *state-of-the-art* method. We provide qualitative comparison on various faces extracted from the videos available online. We also provide quantitative result on a face recognition dataset by frontalizing all the faces before the recognition task and show that our method performs significantly better and efficient.

1.5 Thesis Overview

In this chapter, we introduced the problem and also the motivation in choosing the aforementioned problems along with the contributions. The rest of the thesis is divided into four more chapters. Chapter 2 briefly introduces the fundamental concepts which are used in the thesis. Chapter 3 describes in detail about the face fiducial detection which includes defining the problem, related work, approaches proposed and quantitative comparative experiments. Similarly Chapter 4 deals with face frontalization in detail pertaining to the problem definition, previous methods proposed, our approach, qualitative and quantitative results. And finally we end with conclusion which lead to the future direction of the work in Chapter 5.

2

Related Concepts

2.1 K-Nearest Neighbour (KNN) method

KNN is one of the simplest and effective classification algorithm used in the community. If one does not know the distribution of the data at hand, it is generally preferred to go with KNN as wrong assumption of distribution in other algorithms leads to bad performance. KNN is an instance based learning or lazy learning as it depends on the samples from a small neighborhood. All the training data is carried over till testing phase, where the label of the unknown test data is classified to a label represented by the majority label of its k-nearest neighbors. Figure 2.1 shows the example for labeling the test sample when K is equal to one and four.

The performance of KNN is dependent on the chosen value of K and also the distance metric used. The neighborhood distance of the sample depends on the Kth nearest neighbor. Different K results in different distances and different conditional probabilities. If value of K is very small, sample ends up with a small neighborhood and could result in poor performance because of data sparseness, noise, ambiguous or poorly labeled data. If we try to smoothen the effects by increasing the value of K, it results in introduction of outliers from other class and result in over smoothing.

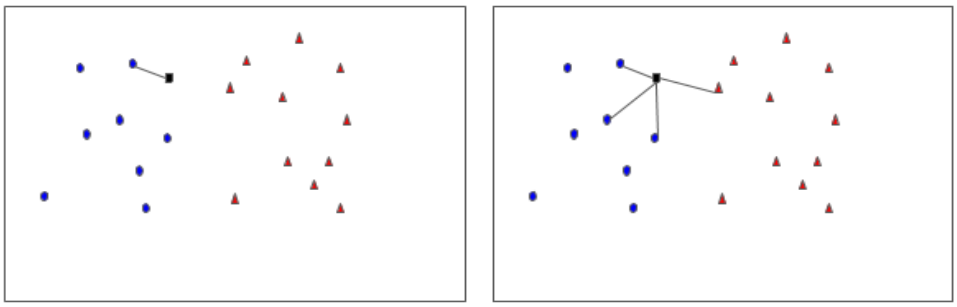


Figure 2.1: Left hand figure shows an example for 1-NN decision rule and the right hand figure shows the example for 4-NN

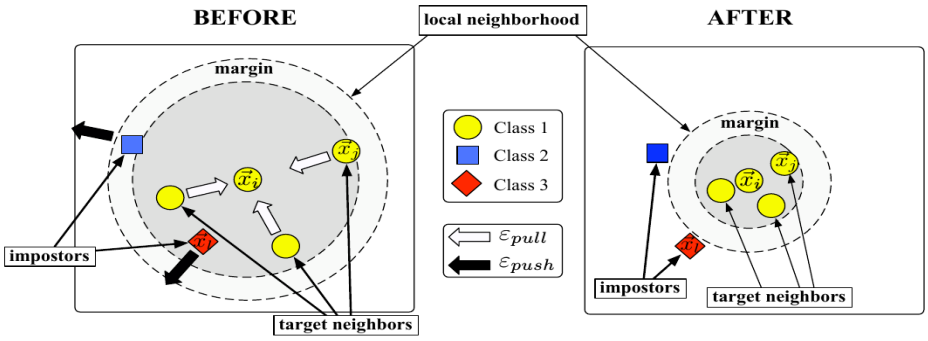


Figure 2.2: Pictorial representation of metric learning and transformed space. Left image shows the representation of samples belonging to different classes. Right image shows the transformed space from the learnt parameters. Notice that the samples from same class are moved closer to each other and are pushed away from other class samples. Figure reference: Weinberger *et al*. [49].

In our experiments, we use KNN in classification setting. Assuming face fiducial detection is a function given an image, we want to select one best performing function from among the functions in the appearance space where we have the training samples. The distance metric would give the degree of dissimilarity between the points.

2.2 Metric Learning

Type of metric used to measure distance between two points in KNN algorithm plays an important role in determining its performance. If no prior knowledge of the data is available, KNN algorithm is generally used with Euclidean distance measure. Since Euclidean metric does not hold any statistical measure of data with respect to labels, it would lead to sub-optimal solutions. Methods such as [15], [19], [37], [39] show that the performance of KNN improves by learning an appropriate distance metric. For example, distance metric learnt for face recognition task and the gender detection task would be significantly different.

Consider $x = [x_1, x_2, \cdots, x_n]$ and $y = [y_1, y_2, \cdots, y_n]$ as two points in a n dimensional space. The Euclidean distance, d between these two points is defined as

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} \tag{2.1}$$

Even simple linear transformation learnt in supervised manner can lead to better performance in KNN classification as shown by [37] and [19]. Consider a linear transformation matrix, L which satisfies pseudo-metric properties such as triangular inequality, non-negativity, symmetry and uniqueness. The distance measured in the transformed space can be represented by,

$$d_L = ||L(x-y)||^2 \tag{2.2}$$

The generalization of Euclidean metric is the Mahalanobis metric. The Mahalanobis distance between two vectors is defined as,

$$d_M = \sqrt{(x-y)^T M(x-y)} \tag{2.3}$$

Where M is a positive semi-definite matrix. Euclidean metric is a special case with $M = I$.

In this thesis (refer 3.5.2), we use metric to define degree of dissimilarity and also as probability. The vector under consideration for our approaches are to deal with the appearance of key feature on the face or structural representation of the key features on the face. Clearly there is some statistical information embedded across various points derived from these features. For example, we could see significant correlation within the points representing an eye corner or a nose tip. Also correlation is different when comparing between the eye corner and nose tip points. We would like to use these statistical information residing in the labeled data to derive a suitable distance metric to improve the accuracy of our KNN based algorithm.

More specifically we use *large margin nearest neighbor* (LMNN) classification algorithm [50] to learn a Mahalanobis metric specifically designed for KNN classification. LMNN works on the intuitive idea that the test sample would be classified correctly if it lies near to samples of same label. LMNN learns a linear transformation by minimizing a loss function that consists of two terms, where first term ensures the samples matching the labels are pulled together and second term pushes the samples with non-matching labels with large margin. Pictorially, it can be represented as in Figure 2.2

2.3 Bayes' Theorem

Bayesian methods help in providing coherent reasons in the face of uncertainty. It is based on mathematically handling uncertainty proposed by Bayes and Laplace in 18^{th} century and developed further by statisticians and philosophers in the 20^{th} century. Bayesian methods have emerged as popular models in the field of multisensory integration, motor learning, neural computation and as the base of machine learning.

Bayes rule states that,

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \tag{2.4}$$

It can be derived from basic probability theory. Here x can be considered as the data point and the θ as the model parameters. $P(\theta)$ is the probability of θ and is referred as the prior. Prior is obtained before observing any information on x. $P(x|\theta)$ is considered as *likelihood* and is the probability of x

conditioned on θ. $P(\theta|x)$ is considered as *posterior* probability of θ after observing x. $P(x)$ is the normalizing factor.

For a dataset of N points, $D = x_1, x_2, ..., x_N$, and model m with model parameters θ:

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)} \tag{2.5}$$

We compute the above quantity for many different models m and select the one with highest posterior probability as the best model for our data.

$$P(D|m) = \sum_i P(D|\theta_i, m)P(\theta_i|m) \tag{2.6}$$

and is called the *marginal likelihood*.

To predict the probability of new data points, x^*, which have not been observed yet,

$$P(x^*|D, m) = \sum_i P(x|\theta_i)P(\theta_i|D, m) \tag{2.7}$$

where

$$P(\theta|D, m) = \frac{P(D|\theta, m)P(\theta|m)}{P(D|m)} \tag{2.8}$$

is the posterior probability of model parameters θ conditioned on the data D and is based on Bayes rule.

2.4 Convex Optimization

Optimization in mathematical sense is to select a particular sample out of all possible samples which yields best solution in some sense. Based on the sample spaces and the type of function which defines the outcome, we can categorize the optimization problem in various types. We describe few types of optimization and its solutions which are relevant to this thesis in the following section.

Linear optimization is special case of mathematical optimization in which the solution space is defined by the linear equality and inequality constraints with a linear objective function. The solution space is a convex polytope. Linear optimization problem can be canonically represented as

$$O(x) = \arg\max_x (c^T x) \tag{2.9}$$

subjected to $Ax \leq b$ and $x \geq 0$. Where x is the variable vector, A is a matrix and b is vector which defines the solution space. There are various algorithms like *Simplex algorithm, Criss-Cross algorithm, Interior point method*, which can solve for global optimum.

Quadratic programming is another special case of mathematical optimization with a quadratic objective function subjected to linear constraints. It can be formulated as follows,

$$O(x) = \arg\max_x (\frac{1}{2} x^T Q x + c^T x) \tag{2.10}$$

subjected to $Ax \leq b$. Q is real symmetric matrix. Similar to linear programming, there are many algorithms like *Interior point, gradient projection, conjugate gradient* to solve general problems.

Qudratically constrained quadratic program is another case similar to Quadratic programming, but with both objective and constraints are quadratic functions.

In this thesis, we end up with a linear objective function with quadratic constraints and solutions need to be integers. Since it can not be solved in polynomial time, we aim to get approximate solution with integer relaxation and solve it as a linear optimization problem.

2.5 Affine Transformation

In this thesis (Chapter 4), we model face as a combination of planes in 3D. For example, consider the region formed by the polygon with the end points as nose tip, point exactly between the eyes and bottom right end of the nose. This region can be approximated to be a plane in 3D. We use a fiducial detector to find the key points on the face and triangulate to form various planes. The defined planes would undergo transformations to achieve certain goal. The transformation of the defined planes would involve translation, rotations and scaling. This section provides mathematical formulation to address the various aspects of handling transformations of planes.

We use homogeneous notation for representing points as it can address the most generic projective transformations. Homogeneous representation of a 2D point (x, y) in Euclidean space is represented as as a 3 dimensional vector by adding a final coordinate of 1, $(x, y, 1)^T$. A planar projective transformation of 3 dimensional vector by a non-singular 3×3 matrix is represented by,

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Based on the in-variances in properties of transformations, there are 4 sub groups under projective transformations. They are isometrics, similarity transformations, affine transformations and projective transformations.

Isometric transformation is composition of just translation and rotation. Here the length (the distance between two points), angle (between two lines) and area is preserved. Similarity transformations preserves the form and is composed of isometric scaling along with translation and rotation. In-variances include angle, parallelity, ratio of lengths. Both isometrics and similarity transformations require two point correspondences to compute transformation matrix. Affine transformation is a non-singular linear

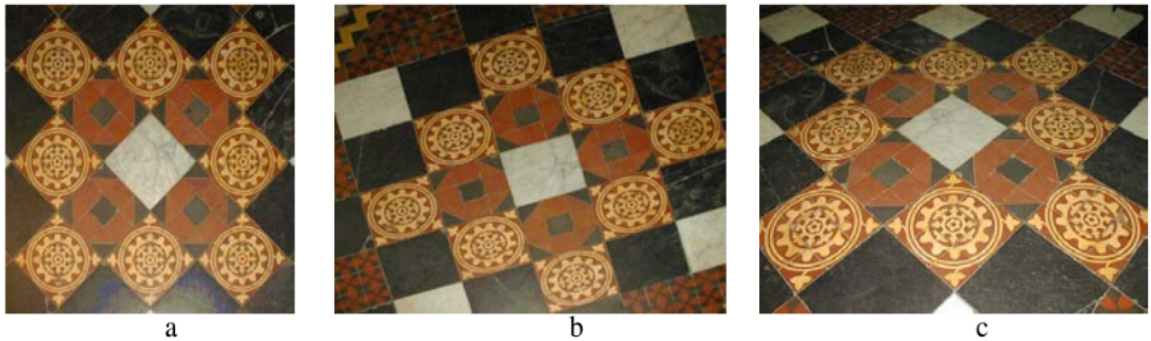


Figure 2.3: Pictorial representation of various transformations. (*a*) **Similarity:** Observe the patterns preserved such as circular pattern, square shaped tiles, parallel and perpendicular lines. (*b*) **Affine:** In this case, circles are imaged as ellipses, orthogonal lines are no more orthogonal, but the parallel lines are preserved. (*c*) **Projective:** Area of tiles closer to camera is larger than the ones away and parallel world lines are converging. Image courtesy: Multiple View Geometry in Computer Vision [21]

transformation followed by a translation which has 6 degrees of freedom and can be thought of as a combination of rotations and non-isotropic scalings along with translation. It requires 3 point correspondences to compute the transformation matrix. In this case, parallel lines are preserved along with the ratio of lengths of parallel line segments and ratio of areas. And finally the most generalized projective transformation has 8 degree of freedom and would require 4 point correspondences to compute the transformation matrix.

In this thesis, we consider affine transformation to model the planes on the face images. Since we triangulate regions on the face image using the face fiducial detector, we have 3 point correspondences to model the corresponding planes between any two faces.

3

Fiducial detection

3.1 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 3.1, second row, eyes), drastic appearance and illumination variation (Figure 3.1, third row, pose) or expression variation (Figure 3.1, first row, mouth). Though there is no consensus yet on even the number of fiducial points assigned to a face [42], 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 [8, 11, 26, 42, 55, 59], since it automatically lends to better performance of systems that rely on fiducial detection.

While a number of different approaches like active shape models [30], regression based methods [55], cascaded neural networks [56], tree based methods [60] and exemplar based approaches [8] 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 [11] , LFPW [8] (Figure 3.4) and AFLW [26] offer images varying widely in appearance, pose, expression, illumination and occlusion, each of these algorithms demonstrate their strengths in specific areas like occlusion handling [11], or robust performance in the case of profile views [60]. Indeed, while regression based approaches are better suited to perform well on metrics that measure pixel-wise accuracy of detection [30, 55], exemplar or mixture-of-trees based approaches [8, 60] 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 thesis, 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,

Figure 3.1: 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.

our algorithms use several state-of-the-art candidate algorithms [5, 11, 46, 54, 60] 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 preprocessing 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 this chapter is as follows. In section 3.2, we review related work with a perspective to distill out complimentary advantages of different approaches to fiducial detection. This is followed in section 3.4 by the formulation in section 3.4.1 and outline of our approach with focus on exemplar selection (section 3.4.3), output selection (section 3.4.4 for the kNN algorithm, section 3.4.5 for the optimization algorithm) and implementation details (section 3.4.7). We then follow up with an extensive experimental section 3.5, 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.

3.2 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 [7, 18] and the traditional AAM based methods have not been suitable for fiducial detection *in the wild* [20, 33]. However, some recent methods that deviate from the traditional pixel-value based texture model have shown new promise [3, 6].

Constrained Local Models (CLMs): The CLM framework has existed for a decade [16, 34] 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 [34] for facial shape estimation. A regression based strategy for CLM optimization has also been proposed recently [4].

Exemplar Methods: Exemplar based approaches have been popular since Belhumeur *et al.*'s work [8]. Zhao *et al*. [58] 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*. [41] and Shen *et al*. [38] 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*. [59] 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 3.4.4) or *smoothen* candidate outputs using optimization (Figure 3.7) 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 [5, 12, 32, 46, 54]. 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 regression based strategy is adopted to learn this mapping and has been shown to converge in 4-5 iterations [5, 54].

A recent work of Smith *et al*. [42] addresses the problem of analyzing the quality of facial fiducial results using an exemplars based approach. However, several difference exist between our approaches. 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* [38], while the goal of the work presented 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 [59], while we either side-step such issues (section 3.4.4) or handle them using optimization (section 3.4.5).

Recently, some promising attempts have also been made to approach the problem of facial fiducial detection in the deep-learning framework [56]. However, most of the proposed deep-learning based models work on low resolution images [56, 57]. This prevents us from getting accurate fiducials on actual data. In this thesis, we present a fully-automatic and principled approach for selecting the best fiducial location by combining results from multiple candidate algorithms for every image.

3.3 Complementarity Analysis

In this section, we discuss the reason and degree of complementarity between *state-of-the-art* methods. Methods initially proposed worked on datasets which consists of mostly frontal view faces [29] in a lab environment. Since the practical scenarios leads to much more complicated settings, *in-the-wild* sort of datasets [26], [11], [8] were released. Deformable part based methods [60] model each part of the face separately. Also, since same part can look significantly different in different poses, separate model are considered. This leads to a better performance in profile view faces. Most of the regression based model does not consider this scenarios. Regression based methods tend to do fairly well in getting the accurate estimation in frontal view faces as compared to that of part based methods. Since Artizzu *et al*. [11] explicitly model occlusion during training, they do well in the case of occluded faces. Table3.1 shows the complementarity performance of various methods. Best performing experiment is done by selecting candidate algorithm output nearest to ground truth by simple Euclidean metric. And average experiment was performed by averaging the estimates of all candidate algorithms.

Chehra	Zhu	Intraface	RCPR	Ours (kNN)	Ours (Opt)	Avg	Best
7.21	7.60	7.79	9.28	4.31	4.83	12.43	2.56

Table 3.1: Table shows the failure rates of various *state-of-the-art* methods and also the failure rate considering the average of all the estimates and selecting the best performing fiducial among the methods for COFW dataset.

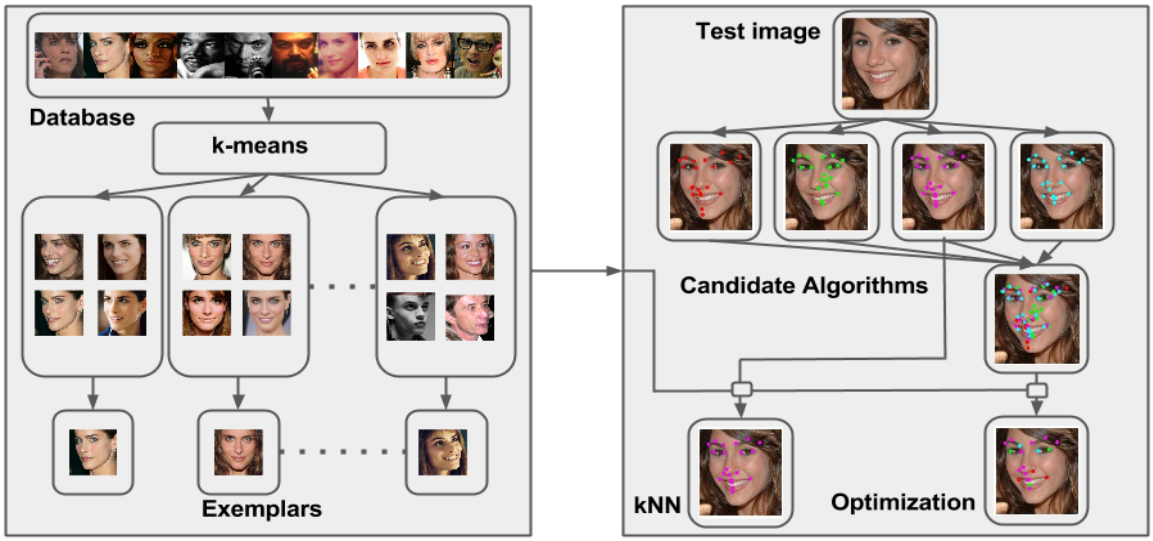


Figure 3.2: Left box pictorially represents exemplars selection. Right box represents our two algorithms for output selection. One by using kNN approach and other using optimization. Best viewed in color.

3.4 Algorithm

In this section, we first outline our formulation in section 3.4.1, 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 3.4.3) and *output selection* (section 3.4.4, section 3.4.5). A flowchart of our approach is illustrated in Figure 3.2.

3.4.1 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 thesis, we consider $K = 20, \ \& \ n = 20$ since that is the set of common fiducials detected by algorithms presented in recent literature [5, 11, 46, 54, 60]. Note that recent approaches [42] 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

$$X^* = \arg\max_{\tilde{X}} P(X \mid R) \tag{3.1}$$

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 3.1 as

$$X^* = \arg\max_{\tilde{X}} P(X \mid R, \mathcal{X}) = \arg\max_{i} P(\bar{X}_i | R) \tag{3.2}$$

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 [8], equation 3.2 can now be elaborated as

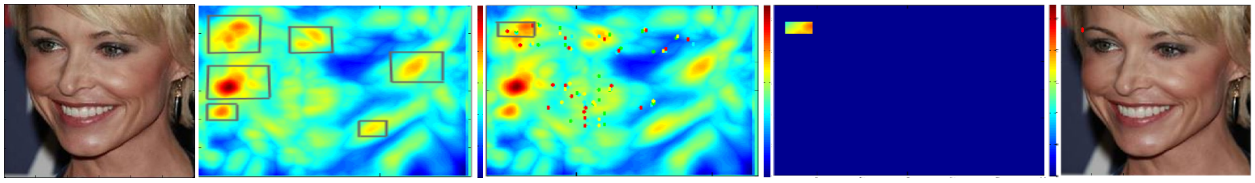
$$\begin{aligned} P(\bar{X}_i \mid R) &\propto P(R \mid \bar{X}_i) && (3.3) \\ &\propto \sum_{k \in K} P(R \mid X_k, \bar{X}_i) P(X_k \mid \bar{X}_i) && (3.4) \end{aligned}$$

where we marginalize over all exemplars X_k. Note that equation 3.4 *splits* the probability into comparison between *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

$$P(\bar{X}_i \mid R) \propto \sum_{k \in K} \prod_j P(R \mid \mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) P(\mathbf{x}_k^j \mid \bar{\mathbf{x}}_i^j) \tag{3.5}$$

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

$$\begin{aligned} P(R \mid \mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) &= (1/\alpha) \exp(-\|F_k^j - F_i^j\|^2) && (3.6) \\ P(\mathbf{x}_k^j \mid \bar{\mathbf{x}}_i^j) &= (1/\beta) \ dist(\mathbf{x}_k^j, \bar{\mathbf{x}}_i^j) && (3.7) \end{aligned}$$



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

Figure 3.3: An example of fiducial detection of eye corner in a test image. Best viewed in color.

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 3.5 entails summing over SIFT and HOG distances between candidate and exemplar fiducials. Finally, one could alternatively choose to optimize equation 3.2 using an optimization function as outlined in section 3.4.5. In this work, candidates are generated using algorithms of Zhu *et al*. [60], Xiong *et al*. [54], Asthana *et al*. [5], Artizzu *et al*. [11], and Tzimiropouluos *et al*. [46].

Example: In equation 3.5, 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 3.6. This is better illustrated with an example. In Figure 3.3, we show an input image for which the *minimum distance* in SIFT+HOG space from a set of exemplars is shown in Figure 3.3b, 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 3.3c, with bounding box), which is then selected and isolated using equation 3.7 (Figure 3.3d), leading to a correct location of the eye fiducial in the final output (Figure 3.3e).

3.4.2 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 3.5.2.

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.

Algorithm 1 Algorithm for Exemplar Selection (`ComputeDatasetExemplars`)

input Training image data $\mathcal{D}$, fiducials $\mathcal{F}_d$.
  $\mathcal{E} = \emptyset, \mathcal{R} = \emptyset, \mathcal{S} = \emptyset, \mathcal{F} = \mathcal{F}_d$
  $Cntrs$ = `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)
    $\mathcal{S} = \mathcal{S} \cup \{I_i, F_i, Feat_i\}$
    $\mathcal{R} = \mathcal{R} \setminus Feat_i$, $\mathcal{F} = \mathcal{F} \setminus F_i$
  end for
output $\mathcal{E}, \mathcal{S}$

3.4.3 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 to 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 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.

Figure 3.4: Examplars automatically selected by our clustering approach in Section 3.4.3 for LFPW dataset. Best viewed in color.

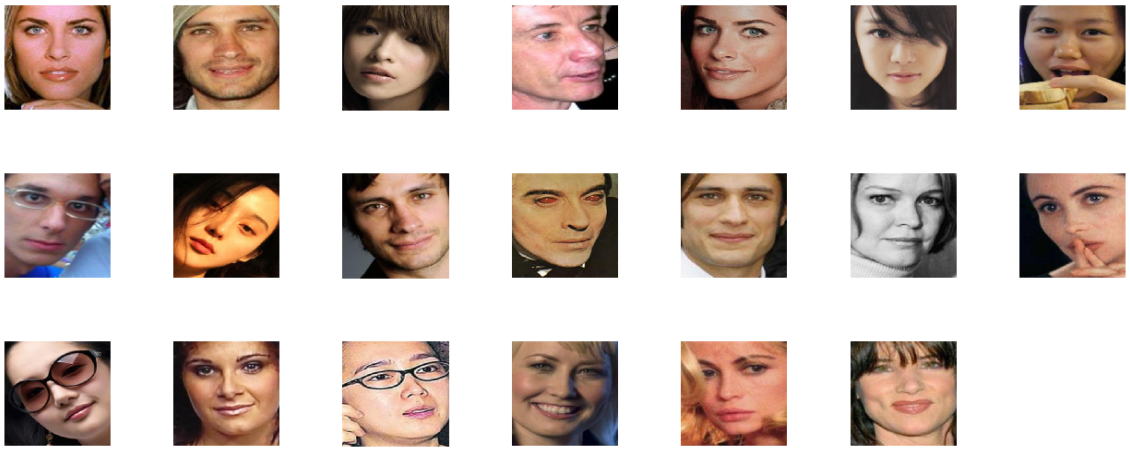


Figure 3.5: Examplars automatically selected by our clustering approach in Section 3.4.3 for COFW dataset. Best viewed in color.

3.4.4 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.

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 3.5) (Algorithm2). Pictorial representation of this method with the selected exemplars for each of the candidate algorithm is shown in figure 3.6 Note that this is a simple kNN based approach, where k=1. Alternatively, we also consider the idea of

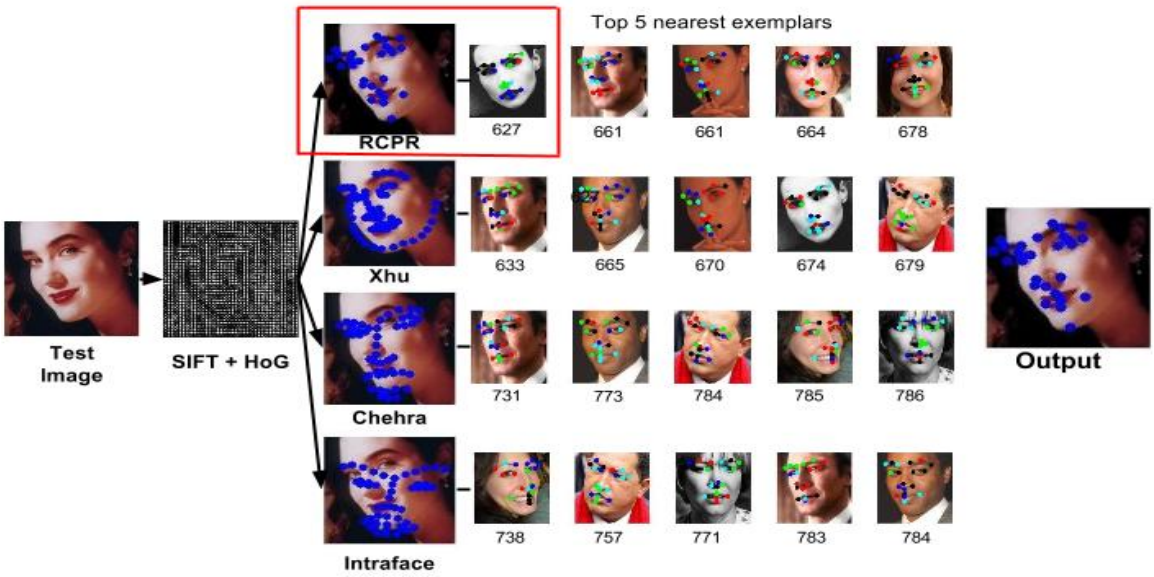


Figure 3.6: Figure shows how the best solution is selected by using kNN. It shows the top 5 nearest exemplars for each candidate algorithm. Observe if the fiducials are off, the nearest exemplar tend to be dissimilar to the input image.

selecting individual fiducials from various candidate algorithm outputs which minimizes an objective function. This is explained in the following section.

3.4.5 Output selection by Optimization

Instead of selecting fiducials 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 3.2, 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.

Algorithm 2 Algorithm for Output Selection by KNN

input Training image data $\mathcal{D}$, testing image data $\mathcal{T}$, training fiducials $\mathcal{F}_d$, testing fiducials $\mathcal{F}_t$.
  $\mathcal{E}$ = `ComputeDatasetExemplars(` $\mathcal{D}$, $\mathcal{F}_d$ `)`
  $F_{out} = \varnothing$
  for Each image-fiducial pair (I, F) in $\mathcal{T}$, $\mathcal{F}_t$ **do**
    I_{face} = `CropAndResize(` I, F `)`
    $\mathcal{O}_{all}$ = `AllFidDetectAlgos(` I_{face} `)`
    for All results F_i in $\mathcal{O}_{all}$ **do**
      $Feat_{test}$ = `ComputeFeatures(` I_{face}, F_i `)`
      for Each pair $(I_e, Feat_e)$ in $\mathcal{E}$ **do**
        $dist_{i,e}$ = `DistFunc(` $Feat_e$, $Feat_{test}$ `)`
      end for
      $dist_i$ = $\arg\min_e dist_{i,e}$
    end for
    $\{dist_{min}, \quad i\}$ = $\arg\min_i dist_i$
    $F_{out} = F_{out} \cup F_i$
  end for
output F_{out}

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

$$u_i^j = \arg\min_k \|f(x_i^j) - f(\mathcal{E}_k^j)\|^2 \tag{3.8}$$

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 3.7. 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

$$p_{cd}^{ab} = \|s(\mathcal{E}_{m(a,c)}) - s(\mathcal{E}_{m(b,d)})\|^2, \qquad a \neq b \tag{3.9}$$

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

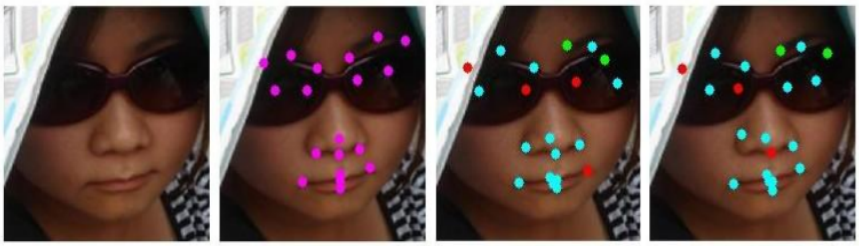


Figure 3.7: From left to right, we observe input test image, output selection by kNN, output selection by optimization without structural costs and output selection by optimization with structural costs. Observe that the left eye prediction suffers in third image because of not considering structural costs for optimization.

methods should not be simultaneously selected. This is easily enforced by the constraint

$$\sum_{i} x_i^j = 1 \tag{3.10}$$

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

$$O(X,Y) = \sum_{i=1}^{5}\sum_{j=1}^{20}(x_i^j \times u_i^j) + \sum_{c=1}^{20}\sum_{d=c+1}^{20}\sum_{a=1}^{5}\sum_{b=1}^{5}(y_{cd}^{ab} \times p_{cd}^{ab}) \tag{3.11}$$

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 [13] 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

$$\begin{aligned} O(X,Y) \quad = \quad & \sum_{i=1}^{5}\sum_{j=1}^{20}(x_i^j \times u_i^j) + \\ & \sum_{c=1}^{20}\sum_{d=c+1}^{20}\sum_{a=1}^{5}\sum_{b=1}^{5}(y_{cd}^{ab} \times p_{cd}^{ab}) \\ & 0 \leq x_i^j, y_{cd}^{ab} \leq 1, x_c^a \geq y_{cd}^{ab}, x_d^b \geq y_{cd}^{ab} \\ & x_c^a + x_d^b \leq y_{cd}^{ab} + 1 \end{aligned} \tag{3.12}$$

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.

3.4.6 Improvement of Zhu *et al.*

We modify Zhu *et al*. [60] approach by replacing feature pyramid constructed using the HoG filters for each part which represents the likelihood of part location, by response based on the feature distance

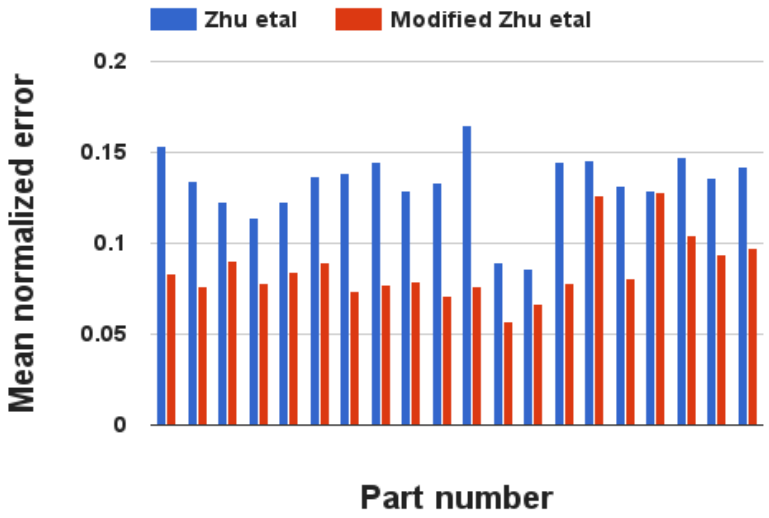


Figure 3.8: Comparison between Zhu *et al*. [60] (blue bars), and its modification using our exemplar approach (red bars). For each part, the y-axis plots the mean pixel error normalized by interocular distance over the entire COFW dataset.

at each pixel in euclidean space with respect to corresponding feature of the part in the exemplars. Here we restrict the likelihood area for each part in the response within the boundary which encloses the prediction of corresponding part in candidate methods that we use. The likelihood score is inversely proportional to the distance of the feature at each pixel with respect to the corresponding part in the exemplar. Within that boundary, we compute the feature based on SIFT and HoG and compute the distance with respect to the feature computed at corresponding part in 20 exemplars. We choose the smallest distance among them to score the likelihood for that pixel. With this modification we see significant reduction in the failure rate and mean error of Zhu *et al*. [60] in COFW dataset which has lot of occluded faces.

3.4.7 Implementation Details

In this section, we present some implementation details 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 [5, 11, 46, 54, 60] along with the trained models. Experiments were conducted on the same test split for candidate and our algorithms for all the datasets.

3.5 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 Labeled Face Parts in the Wild (LFPW), Caltech Occluded Faces in the Wild (COFW) and Annotated Facial Landmarks in the Wild (AFLW). Before we present the quantitative result (produced in Table 3.2) 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 [8], 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*. [11], 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 [26], and contains several annotated face images in extreme settings. It is considered one of the toughest datasets in fiducial detection literature [42, 56], as it has larger pose variations, partial occlusions and illumination variation compared to other datasets. Like [56], we sample 1000 training images and 3000 testing images randomly from the dataset, while ensuring no overlap between the two sets.

3.5.1 Quantitative Results

In this section, we outline the basis for future experiments detailed in the next sections. Table 3.2 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 3.4 illustrates our results of exemplar selection on the LFPW dataset. SIFT features for each fiducial are calculated at

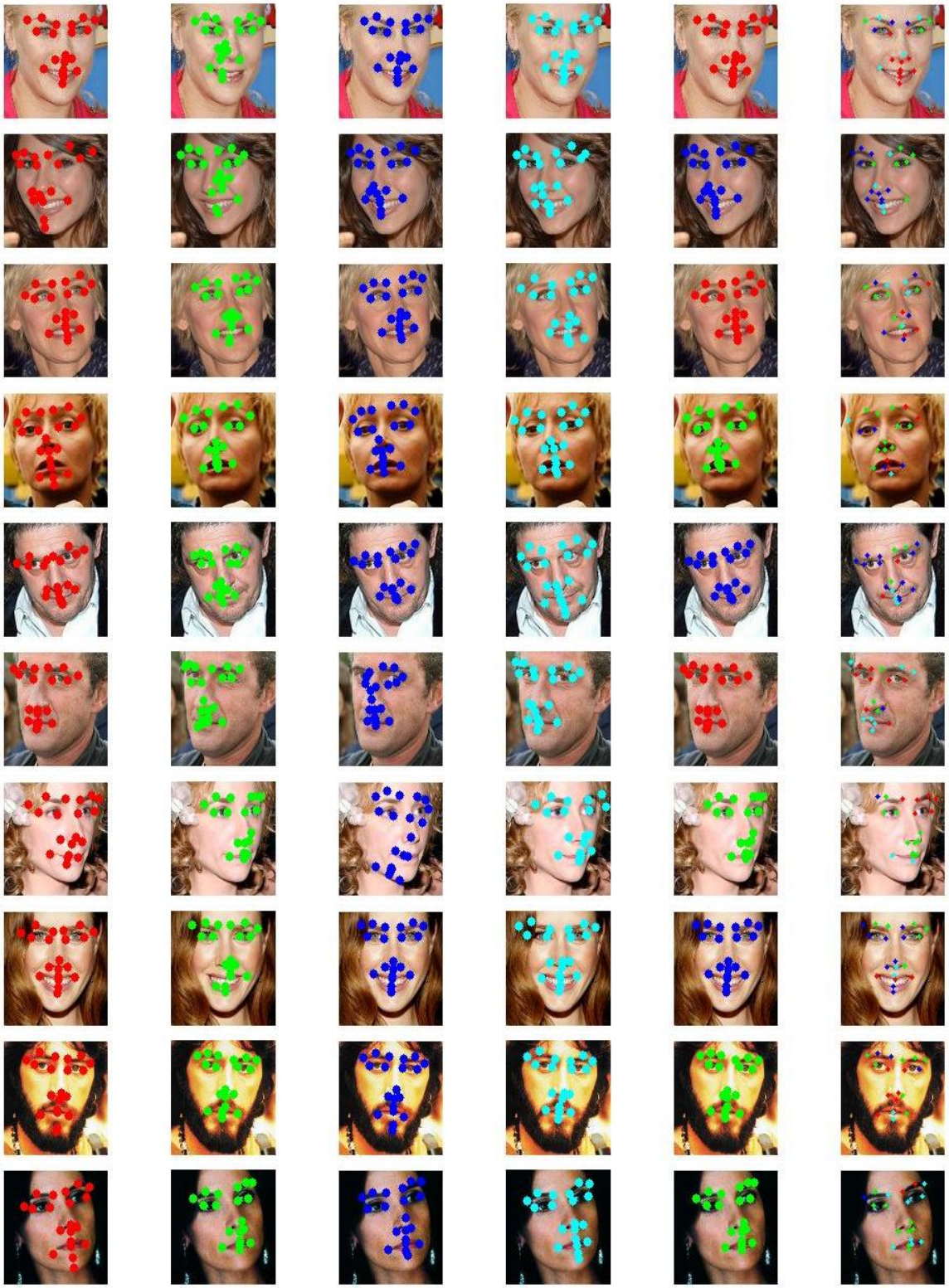


Figure 3.9: Figure shows the qualitative results of candidate and our algorithms. Fiducials by Chehra [5] (red points), Zhu et al. [60](green points), Intraface [54](magenta points), RCPR [11](cyan points), Output selection by KNN and Output selection by Optimization can be observed in column 1, 2, 3, 4, 5 and 6 respectively

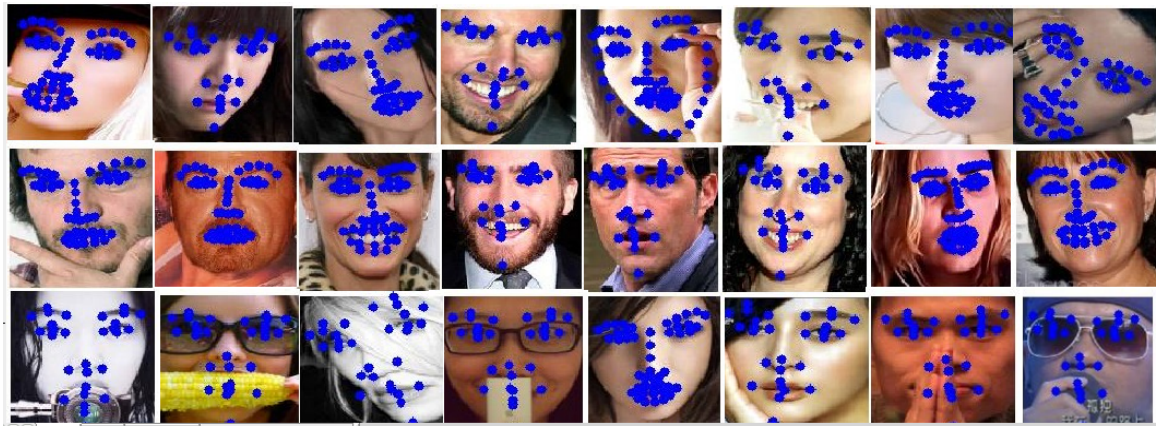


Figure 3.10: Results with varying pose (Row 1), expression (Row 2) and occlusion (Row 3). Best viewed in color.

Dataset	Chehra	Zhu	Intraface	RCPR	PO	Ours (kNN)	Ours (Opt)
LFPW	7.21	7.60	7.79	9.28	**4.82**	**4.31**	4.83
COFW	7.95	15.76	7.22	7.30	6.73	**5.98**	**6.28**
AFLW	40.44	**25.88**	47.98	39.78	46.67	**19.93**	32.08

Table 3.2: Table shows the mean error 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.

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 3.2, 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 3.13. 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

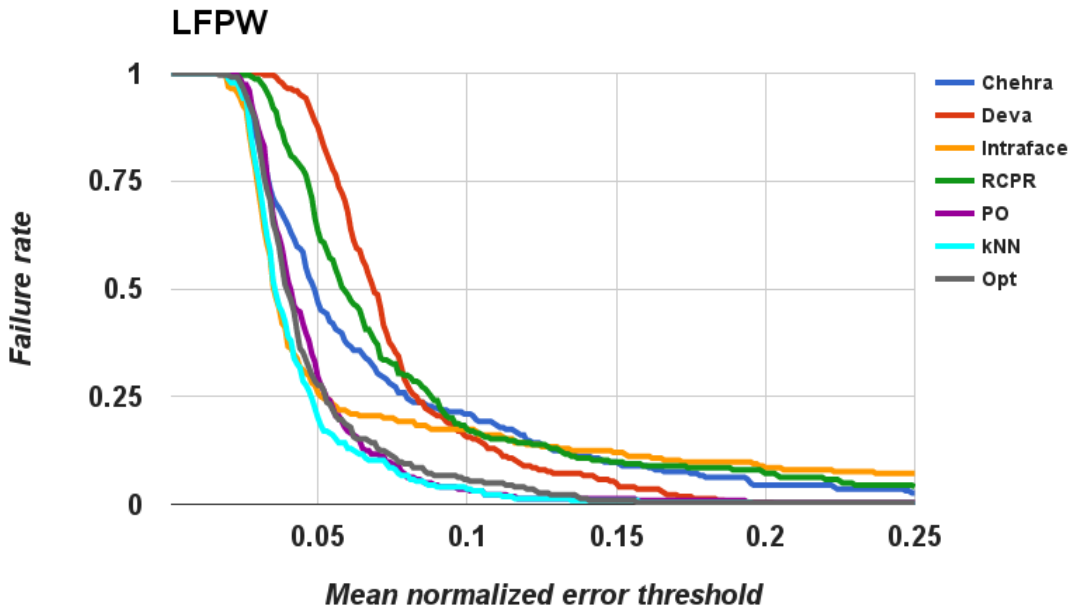


Figure 3.11: Results of our approach on LFPW dataset. Drop in failure rate with the change in cut-off threshold of mean error normalized with interocular distance. Lower curve means more accurate results. Best viewed in color.

Dataset	Chehra	Zhu	Intraface	RCPR	PO	Ours (kNN)	Ours (Opt)
LFPW	20.98	15.62	17.41	17.41	**3.57**	**3.57**	5.8
COFW	21.89	49.70	18.15	14.20	9.27	**7.49**	**7.88**
AFLW	80.52	**71.28**	79.80	82.12	75.20	**59.03**	76.30

Table 3.3: Table shows the 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.

algorithms consistently perform better compared to other five algorithms at almost all cut-off ranges. Figure 3.10 and 3.9 illustrates some qualitative results using our approach.

3.5.2 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

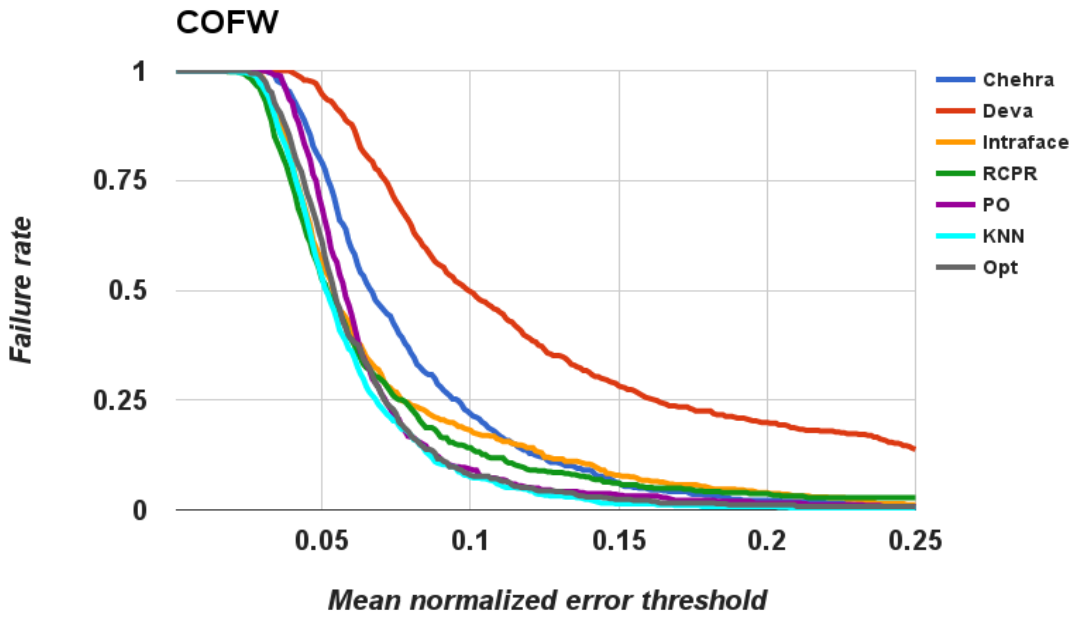


Figure 3.12: Results of our approach on COFW dataset. Drop in failure rate with the change in cut-off threshold of mean error normalized with interocular distance. Lower curve means more accurate results. Best viewed in color.

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 3.14b. 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 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 3.15.

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

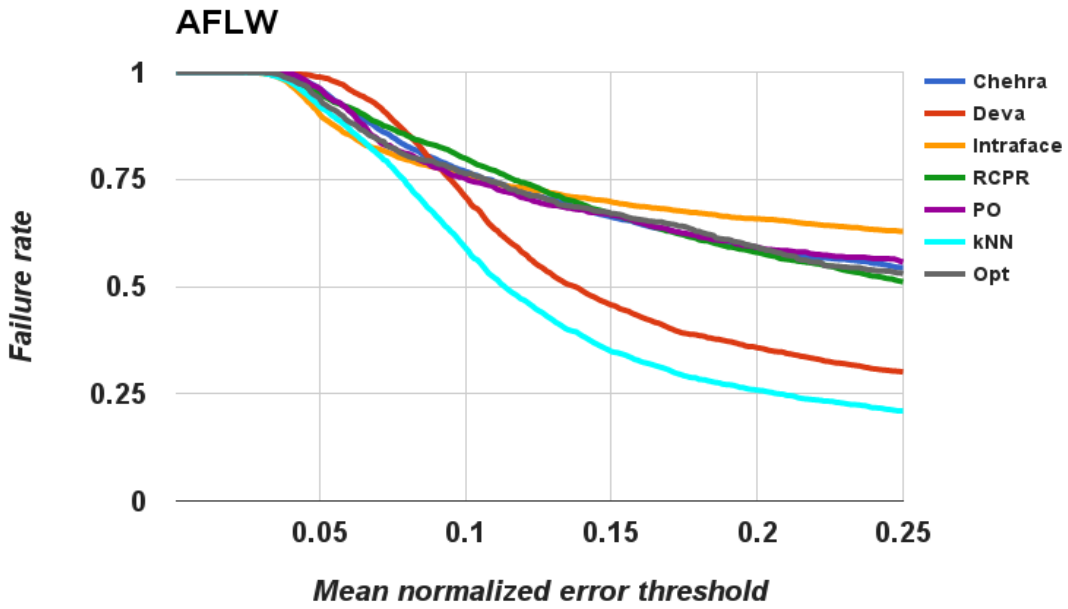


Figure 3.13: Results of our approach on AFLW dataset. Drop in failure rate with the change in cut-off threshold of mean error normalized with interocular distance. Lower curve means more accurate results. Best viewed in color.

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 3.7.

Shape vs Appearance Algorithm 1 outlines our approach of using both shape specific and appearance specific exemplars in output selection. In this experiment, we measure the relative importance of each type of exemplar. Figure shows results of our experiment, where we find that both have almost equal contributions to the superiority of output selection in comparison to competing methods.

Figure 3.17 shows our results when only one type of exemplars are used for output selection on the COFW dataset. Shape based and appearance based exemplars perform in a complimentary manner. Shape based exemplars provide robustness to partial occlusion, since they are better at identifying non-occluded fiducials, and generally result in nearest neighbors that are closer in pose to the test image. On the other hand, while appearance based exemplars falter in the presence of occlusion, they are better at identifying more accurate fiducials when all candidate algorithms give accurate outputs.

Clustering vs Eigenspace Analysis While `kmeans` has been the preferred choice of clustering method for Algorithm 1, we also experimented with using principal component analysis (PCA) instead. In order to select exemplars using PCA, we construct a shape matrix where each column represents an exemplar, and find its top 20 principal components. We then select one exemplar per component such that it maximizes its dot product with the corresponding principal component. Results, show negligible

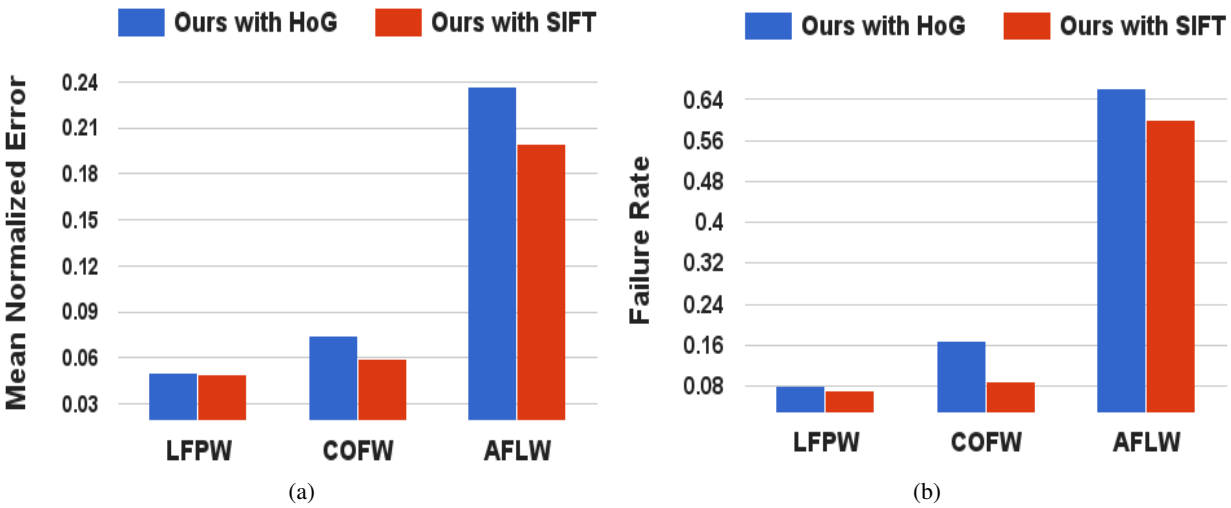


Figure 3.14: Comparison of mean error and failure rate for SIFT vs HOG experiment. Best viewed in color.

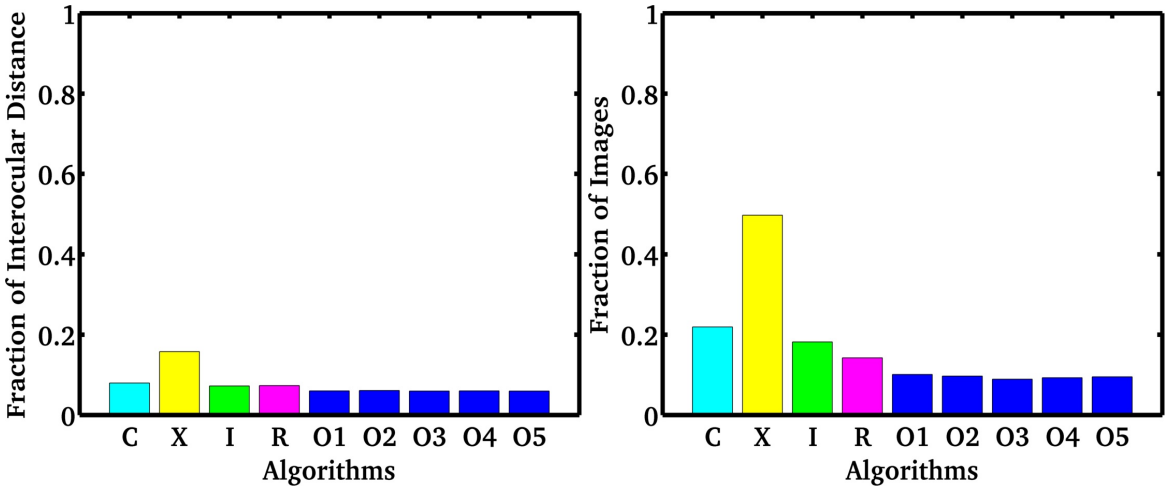
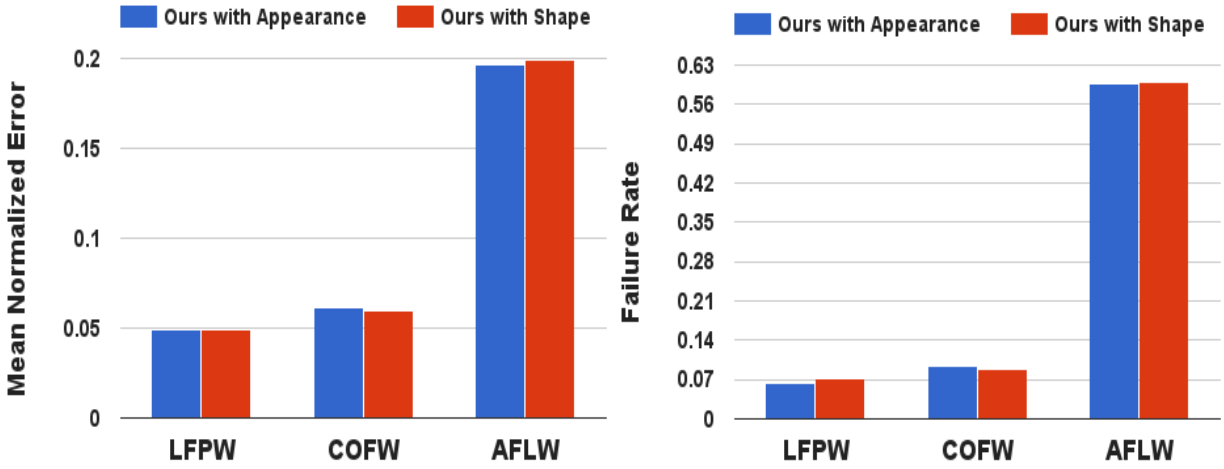


Figure 3.15: Comparison of mean error and failure rate when the number of exemplars is increased. Results O1-O5 correspond to our algorithm with number of exemplars (20, 30, 40, 50, 60) respectively. C, X, I and R corresponds to Chehra, Deva, Intraface and RCPR respectievely. Best viewed in color.



(a) (b)

Figure 3.16: Comparison of mean error and failure rates for the shape vs appearance experiment. Best viewed in color.

difference between the two approaches. We repeated the same experiment with appearance exemplars with similar results.

Euclidean distance vs Metric Learning We also performed an experiment to compare the performance of our algorithm when we use euclidean distance and Mahalanobis distance to find the similarity score between candidate results on a test image and exemplars. We learn the transformation matrix based on [50]. Metric is learnt with the objective of minimizing the distance of k nearest samples belonging to same class in the transformed domain and maximizing if the samples are of different class. Here we compute transformation matrix for each of the part in one vs rest fashion. Appearance feature vector (SIFT and HoG) computed given the ground truth location for a part is considered to be of one class and appearance feature vector of other parts are considered to be of another class. Transformation matrix is learnt. Learnt matrix is used to compute the similarity score for each of the part prediction in the test sample with the part in the exemplar. We use 200 samples for each part with the feature dimension of 535. We note that insignificant improvement is obtained using metric learning. Figure 3.18 shows the failure rate analysis of different datasets. We believe since the appearance representation of fiducials are too diverse for the metric to capture meaningful statistical information, metric learning failed to perform better than the simple Euclidean metric.

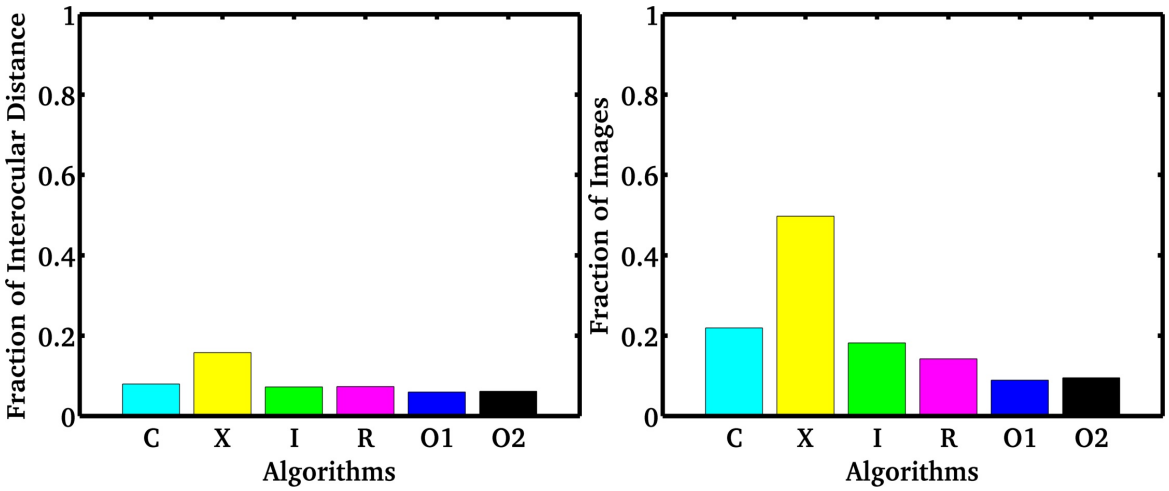


Figure 3.17: Comparison of exemplars for kmeans vs PCA. O1 represents our basic result and O2 represents PCA based results. Best viewed in color C, X, I and R corresponds to Chehra, Deva, Intraface and RCPR respectievely.

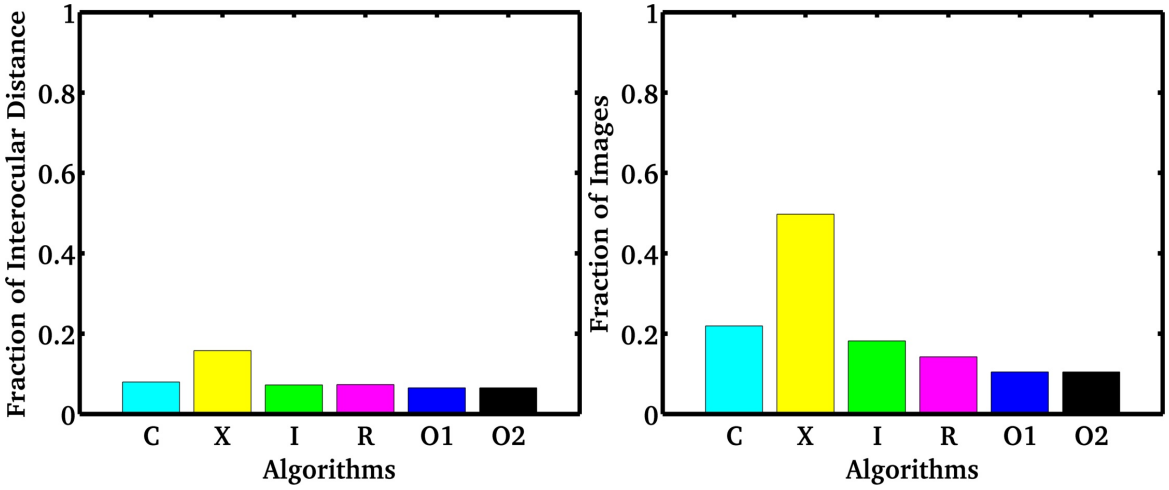


Figure 3.18: Comparison of Metric Learning Result. O1 represents our results with Euclidean metric and O2 with learnt metric. C, X, I and R corresponds to Chehra, Deva, Intraface and RCPR respectievely.

4

Frontalization

4.1 Introduction

Facial analysis in images for recognition/manipulation is a widely addressed and commercially important problem. Its applications range from surveillance to automatic tagging of photos on social websites. Recently, there are papers producing convincing results on *in-the-wild* datasets [22, 60]. These datasets differ from previous ones in their unconstrained nature of image capture. However such methods have two drawbacks. Firstly, a lot of these methods have degraded performance in profile view vs frontal view. Secondly, they require lot of training data [23]. One way to alleviate both problems is to be able to generate realistic frontal view faces for any person. This can be achieved, because faces have a definite structure. Eigen analysis [9], for example, has shown that faces exist in low dimensional sub spaces and can be represented as linear combinations of other faces. Also, it has been shown earlier that many face characteristics like expressions, hair *etc*., can be *transferred* from one person to another, in a very realistic manner [35].

In this thesis, we show that a pre-processing step of synthesizing frontal pose of the face significantly improves the accuracy of face recognition. Face frontalization is the process of synthesizing frontal pose

Figure 4.1: Left image shows the profile face. Second image is *face frontalized* by our method. Third image is of Hassner *et al.* [22] method. Right image is the natural frontal view of the individual. *Frontalization* helps in face recognition.

of the face, given a profile view of the face as shown in Figure 4.1. This step helps in simplifying the task of face recognition as recognition systems have more information and less occlusion to work with. Few methods counter this *frontalization* problem, by choosing to extract features only at the salient locations. Unfortunately, this leads to loss of structural relation between various parts of the face. However, as we will show that our method preserves this information as well.

Apart from aiding face recognition systems, frontalization techniques can also be used to generate a video out of a single image and can find applications in animation [35]. For example, if a family photograph has some people looking away from the camera, our approach can be used to correct this discrepancy [31].

Recent methods [44] [27] have proposed different ways of addressing the challenging problems of pose variations in images. Simonyan *et al.* [27] [40] choose to define features extracted out of large image regions to counter mis-alignments. Wolf *et al.* [24] [52] choose to align faces before extracting features. Sun *et al.* [23] use large datasets to create models robust to these challenges. In line with our approach, some recent works try to counter these challenging conditions of pose variation by synthesizing pose neutral faces from input images. Taigman *et al.* [45] try to estimate a 3D model of each input image. They then use this 3D information to synthesize the frontal view. On the other hand, [22] assumes a generic 3D model for all input images and produces convincing frontalization results. Even though the approach of [45] seems to be good, estimating 3D model from a single image is a hard problem. And assuming a generic 3D model in [22], leads to loss of structural information unique to an individual. Thus, in this work, we turn towards an exemplar based approach to fill the 3D information gap required by the previous approaches.

Lately, we have seen a surge of papers [28] [51] based on exemplar methods for solving computer vision problems. In these type of approaches, exemplars of the problem category are used instead of defining a generic model to solve the the problem at hand. For example, in the case of object detection, [28] trains a set of models using one positive exemplar each, instead of all the training set. And they show that the ensemble of such models give surprisingly good generalization. Similarly, our method is based on an exemplar based approach toward face frontalization. Consider a huge dataset of profile, frontal view face pairs of different. Chances of finding individuals with similar face structures to an input profile image is thus very high. Given such a match, the frontal view of the person in the database can then be used to frontalize the input image. Therefore, for our method, we collect a database of profile views and corresponding frontal view of a large number of individuals. We leverage the fact that faces lie in a low dimensional subspace and thus, many characteristics, like pose, expressions, *etc*., are transferable between people.

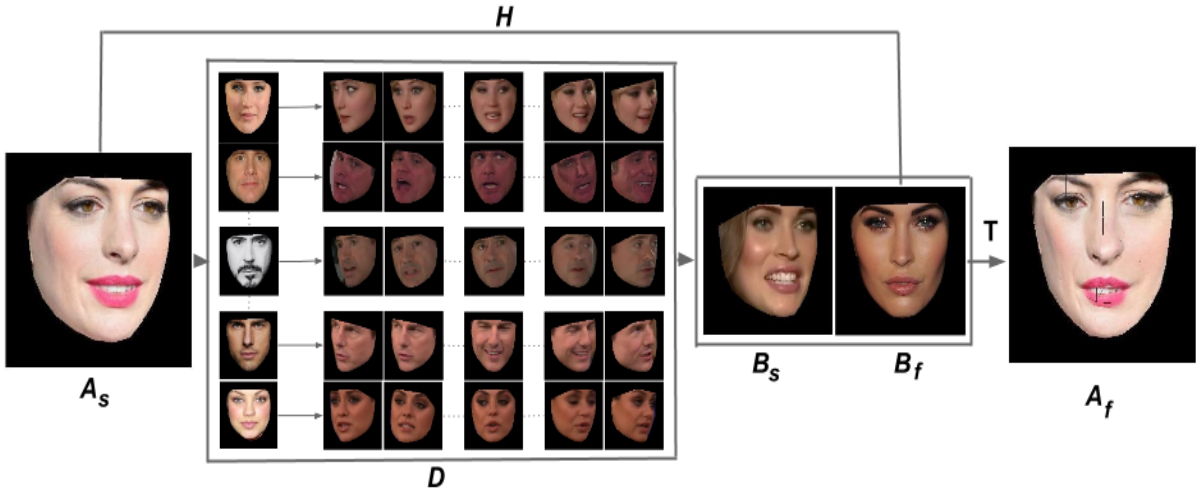


Figure 4.2: Figure shows the generic pipeline used in our approach. Given the input image (left most block) we use the exemplar database (second block) to compute the nearest profile view (third block, first image). We then use the correspondences between the profile and frontal views of the selected exemplar pair (third block) to compute the affine transformation H between the input image and the frontal exemplar, and use it to produce the *frontalized output* (right most block).

4.2 Face Frontalization

Our method takes as input, a profile view face, A_p, and an exemplar database, D, consisting of wide range of profile, frontal pose pairs for different persons. We then proceed to frontalize the face in two steps. First, we run facial landmark detection [25] on the input face, and using it we retrieve the most similarly posed face I_p^i and its corresponding frontal view face I_f^i, from database D. When profile views of two faces match, there is high likelihood that the two persons have similar facial structures. We exploit this property to get geometrical transformations required for frontalization of the input face. Simply put, we obtain the frontal view of A_p by using the affine transformations between A_p and I_f. One recently proposed state-of-the-art method [22] uses a generic 3D model for computing this transformation. This leads to loss of important discriminate structural information unique to an individual. Since we are finding a nearest profile exemplar and its corresponding frontal view face, structural information is still preserved for an individual face in our case.

Let $P = (X, Y)$, denote the landmark locations on the face, where $X = (x_1, x_2, \ldots, x_{68})$ and $Y = (y_1, y_2, \ldots, y_{68})$ are vectors of X and Y coordinates respectively. We consider 68 landmarks which includes feature points such as eye corners, nose tip, mouth line and jaw line. We use the dlib [25] implementation for landmark localization. Note that landmarks of images in D have been pre-computed. We also manually predefine 110 planes for a face using these landmarks. For example, the ends of two eyebrows and the beginning of the nose form a plane (see Figure 4.5). This has to be done only once

Figure 4.3: First row of images are the input profile images. Second row shows the retrieved faces from database.

since these planes connect the same fiducial points irrespective of the face. Each plane is defined by 3 landmark locations.

Let $T = \{t_1, t_2, \ldots, t_{110}\}$ represent the set of planes defined in 3D for a face image. Given planes of one profile-frontal image pair $T_p^m \& T_f^m$, in D, we define $H^m = \{H_1^m, H_2^m, \ldots, H_{110}^m\}$ as the affine transformations between corresponding planes, computed using point correspondences from the facial landmarks. That is,

$$t_f^{m,i} = t_p^{m,i} \times H_i^m \tag{4.1}$$

where the subscript p denotes profile, and the subscript f denotes frontal views.

Given the landmarks P for images in the database D, we now proceed to frontalize using the following steps.

4.2.1 Nearest exemplar selection

To retrieve the closest exemplar to the input face, we first need to define a similarity measure between faces. Let P^m and P^n represent landmarks of two faces. The similarity score between poses of two faces can then be defined as the Euclidean distance between P^m and P^n.

$$ds^{mn} = \sqrt[2]{\sum_{i=1}^{68}((x_i^m - x_i^n)^2 + (y_i^m - y_i^n)^2)} \tag{4.2}$$

However P^m and P^n are defined in different coordinate systems, separated by translation, rotation and scaling. We need to nullify the effect of translation and scaling and bring both sets of landmark positions to one coordinate system. Note that rotation is not considered as it is one of the parameters of pose and our exemplar database is exhaustive enough to take care of rotation variations in the input face. To remove the translational effect, we subtract the mean of X and Y coordinates from both the landmark

vectors, $X = (X - \mu_X), Y = (Y - \mu_Y)$. To remove the scaling effect, we multiply P_m by a factor of s, given by

$$s = \frac{\sum_{i=1}^{68}((x_i^m \times x_i^n) + (y_i^m \times y_i^n))}{\sum_{i=1}^{68}((x_i^m)^2 + (y_i^m)^2)} \tag{4.3}$$

which follows from a straightforward optimization procedure that minimizes *root mean squared error* (rmse) error between corresponding landmark positions. The derivation is omitted for brevity.

Pose is accurately defined by the position of landmarks on the face. We concatenate the landmark locations obtained on the input image into a single vector called the pose vector. We then use Euclidean distance as the metric of comparison to retrieve the most similarly posed face from the database D. To get an accurate measure, we convert the pose vector of input face to exemplar face coordinate system. Let P^t represent the landmarks of profile input face and P_p^i represent the landmarks of profile exemplars available in the database. The nearest exemplar is the one which has the least ds^{ti}.

$$i^* = \arg\min_i d^{ti} \tag{4.4}$$

Given the nearest profile image I_p^i, we retrieve its frontal image and pose I_f^i, P_f^i as shown in Figure 4.4. The first row of Figure 4.3 shows sample input faces and second row shows the nearest exemplars retrieved from D. Observe that men and women have slightly different facial structure, and this captured by our method, since women are retrieved as top exemplar candidates for input images of women.

4.2.2 Triangulation and Transformation

Once the frontal view of the nearest exemplar is obtained, we need to transform the input profile face to a frontal view. To do this, we first *transfer* correspondences between the exemplar pairs to the input image. This is done by replacing positions of the profile exemplar landmarks with those of the input image. Using the landmarks obtained, we define around 110 triangles on the face, each of which can be considered as a plane in the face coordinate system. Since the triangles are defined based on particular set of landmarks, we have correspondences between planes in the input image and corresponding frontal view exemplar. We obtain the affine transformations between the corresponding planes and then synthesize the frontal view of the input image using these transformations.

Figure 4.2 pictorially represents our method. For a given profile view input face, A_p, we retrieve most similar exemplar, I_p^i from the D along with its corresponding frontal view face, I_f^i. We then compute affine transformations, H^i, to transform planes of A_p to generate its frontal view.

4.2.3 Face Recognition

Our face recognition pipeline is based on the framework of Write *et al.* [53] who claim that faces of a particular individual lie in a low-dimensional subspace. In their method, training samples are repre-

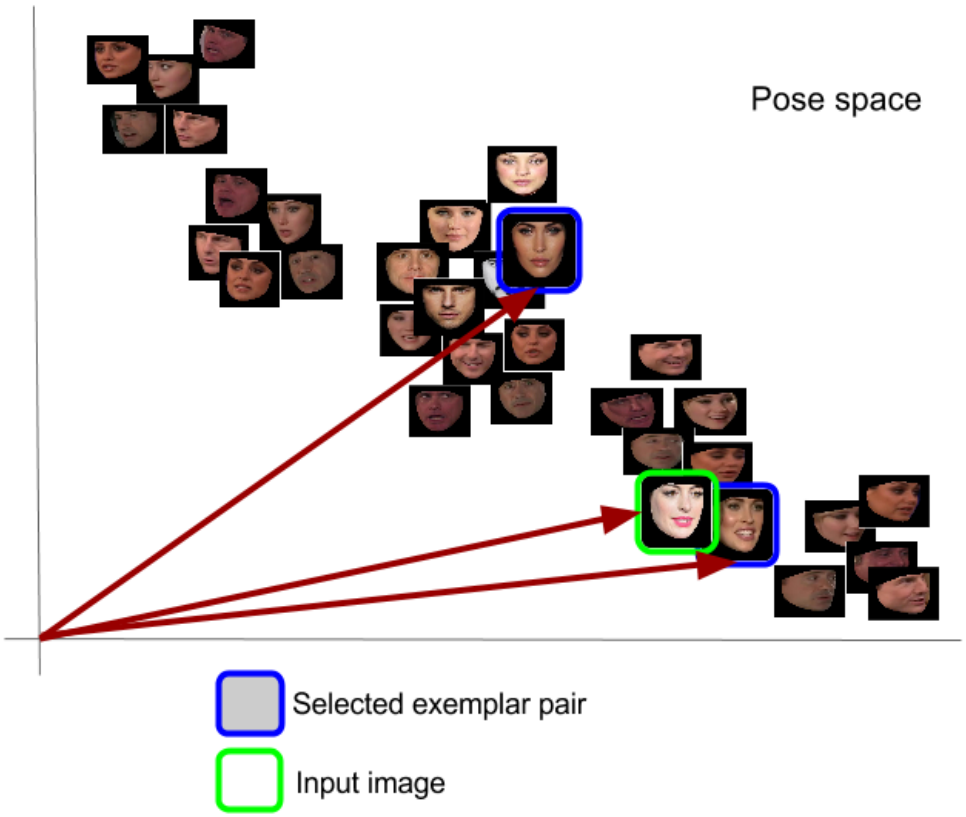


Figure 4.4: Figure shows the pictorial representation of faces in two dimensional pose space. Face with the green box is the input image. Faces in blue box are the exemplar pair selected.

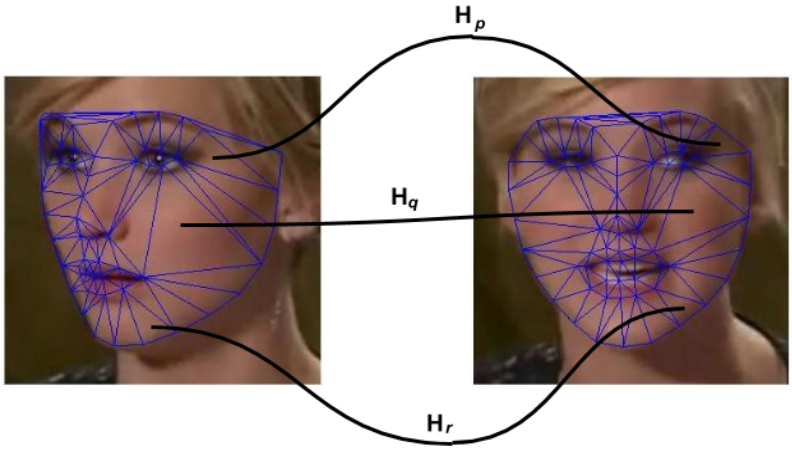


Figure 4.5: Image shows the planes represented as triangles and correspondences between two views of the same face. Note that each plane contains a fixed set of points irrespective of pose. For example, one plane contains two ends of the eyebrows and the top of the nose.

sented as a 2D matrix, where each column represents a feature extracted from one training image. The input test sample should then be represented as linear combination of samples from the corresponding training samples of the same class (person). This problem has to be posed as an $l0$ minimization problem as it selects a combination of samples from training set. Based on recent advancement in sparse representation and compressed sensing, the authors claim that when the solution is sparse enough, solving $l1$ minimization is equivalent to the $l0$ minimization problem. Using this insight, a solution can be obtained in polynomial time using linear programming models. We use their implementation as the basis for our experiments, while we train using our dataset.

4.3 Experiments and Results

In this section, we provide details pertaining to the exemplar database collection, qualitative and quantitative comparision with Hassner *et al*. [22].

4.3.1 Exemplar Database

For the exemplar database, we collected various face poses for 22 individuals (11 male and 11 female) from the talk shows available online. We selected sections which have complete swing of pose and expression changes. Approximately 15 exemplars and a frontal view were selected per individual. In total of around 400 exemplars and 22 frontal view faces were collected. For face recognition experiment, we collected approximately 50 training and 50 input faces of 6 celebrities online. We call this dataset the $PoseInTheWildFaceDataSet$ (PIWFDS). We consider a new dataset, as existing datasets do not contain profile-frontal image pairs and even state-of-the-art recognition systems perform poorly on it.

All our experiments were conducted using MATLAB. We used HoG [17] feature based face detector to find faces and its output is re-scaled to a 300×300 image for both the database images and our input. This is given to facial landmark detection code based on [25], which is publicly made available. This provides 68 landmarks on each face. Using the landmarks we divide the face surface into 110 planes (triangular in shape). Using the corresponding planes between input profile image and the exemplar frontal view image, we compute the homography transformations matrix using publicly available implementation of vgg_Haffine_from_x_MLE. Using this set of homographies, we synthesize the frontal view of the input image.

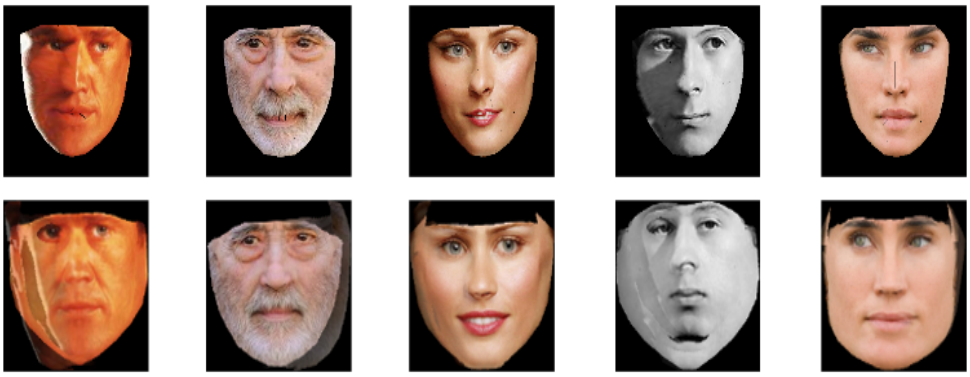


Figure 4.6: First row shows the output of our method and the second row is of Hassner *et al.* [22] for LFPW [10] dataset. Observe ghost like appearances, structure distortion, mirroring effects in Hassner *et al.* [22] output.

4.3.2 Comparison with Hassner et al

Figure 4.7 shows the comparative results between our method and that of Hassner *et al.* [22] for PIWFDS dataset. To show that our exemplar database is generic enough to extend to standard datasets, we provide qualitative results for LFPW [10] dataset in Figure 4.6. We use Hassner *et al.* publicly released code to obtain the result.

Observe ghost like appearances present in most of the cases from Hassner *et al.* [22] output. Also take into consideration, that the face structure of the actress in second row has been changed to a generic one. This is because they use a generic 3D model of the face to achieve the result.

4.3.3 Quantitative Results

For quantitative analysis, we used around 50 testing and 50 training samples of 6 classes for the face recognition task. Each sample is re-sized to a 300x300 image. After converting each sample from color to gray scale, we concatenated gray scale values to form a 90000 dimensional vector. We use Principal Component Analysis to reduce the dimensions to 40 using the training dataset. Each testing sample is also represented as 40 dimension vector as described above. We use publicly available implementation of Wright *et al.* [53] to recognize each input face.

Accuracy is calculated as fraction of testing samples classified correctly over the total number of samples. Our method achieved an accuracy of 31%, which is significantly better than 27% achieved by Hassner *et al.*.

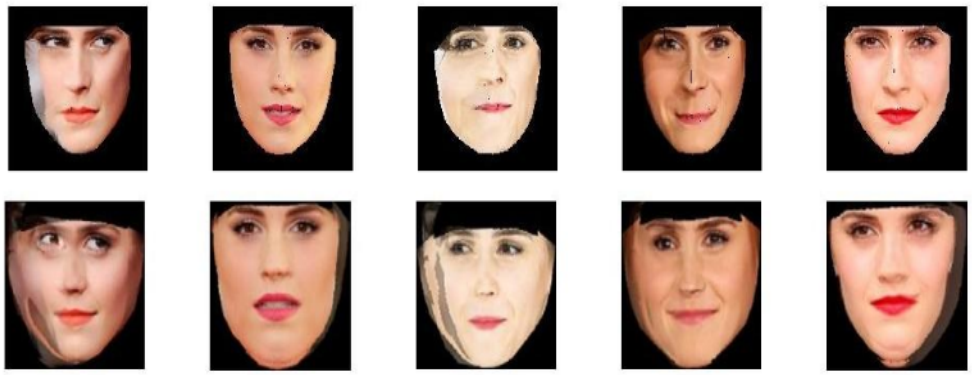


Figure 4.7: First row shows the output of our method and the second row is of Hassner *et al.* [22] for PIWFDS dataset.

4.4 Discussions

We proposed an efficient algorithm to synthesize a frontal view of the face without the problem of estimating 3D model of the face. Improving the quality of results would simply mean addition of distinct faces to the exemplar database. Automatic detection of distinct faces can be achieved by employing *state-of-the-art* face fiducial detectors to the newly seen face structures. k-d tree data structure can be used to make efficient search in the exemplar database. Also method proposed, can be easily extended to create face reenactment videos and other novel face image synthesis.

5

Conclusion

The human cortex is particularly large and therefore has a massive memory capacity. It is constantly predicting what you will see, hear, and feel, mostly in ways you are unconscious of. These predictions are our thoughts, and, when combined with sensory input, they are our perceptions. I call this view of the brain the memory-prediction framework of intelligence.

– Jeff Hawkins

If the machines ever to reach the capabilities of human brain, we ought to try things which are biologically inspired. Geared with the massive memory capability and growth of distributed computing, we hope that many of the computer vision problems can be formulated as memory association problem as compared to other traditional methods. Apart from conceptually simple, this sort of framework gives additional advantage of easy interpretability and parallelizability.

Following the above intuition, this thesis proposed exemplar based approaches for face fiducial detection and frontalization. Each of the method efficiently utilizes the information from the exemplars to achieve two different objectives. Both the approaches are easy to interpret for any outcome. While the face fiducial detection uses both the appearance information and structural information, face frontalization utilizes the strutural information of face from the exemplar database.

5.1 Summary

We showed that our approaches in general outperform *state-of-the-art* methods on popular datasets. We believe when the data at hand is diverse to a great extent and single model can not capture the diversity, it's better to go for exemplar based approaches. With the availability of large amount of data and very good distributed frameworks, we hope the community formulate various computer vision problems as memory association framework.

5.2 Future Direction

In this final section, we discuss how related research fields can also benifit from the methods proposed in this thesis, as well as directions in which the proposed methods can be extended for further improvement.

- Meta algorithm proposed in Chapter 3 can be extended to problems such as Human Pose estimation where there is enough hand annotated data to be used as exemplars and possible structures are fairly limited.
- Section 3.3 shows the quantitative analysis of degree of complementarity in the various candidate algorithms for face fiducial detection. Since there is huge scope for improvement, we hope this opens up new direction for further research.
- In the Chapter 3, we proposed two algorithms for fiducial detection, one which selects the entire output of one of the candidate algorithm and other which selects best performing parts from each of the candidate algorithm. Selection as whole performs better than that of by parts. Since the optimization framework allows only for approximate solution, there is scope to improve accuracy by other means.

Related Publications

- Mallikarjun B R, Visesh Chari, C V Jawahar, Akshay Asthana. Face Fiducial Detection by Consensus of Exemplars. *IEEE Winter Conference on Applications of Computer Vision (WACV), 2016*.
- Mallikarjun B R, Visesh Chari, C V Jawahar. Efficient Face Frontalization in Unconstrained Images. *National Conference on Computer Vision Pattern Recognition (NCVPRIPG), 2015*.

Bibliography

[1] KAIROS. http://www.kairos.com/face-recognition-api.

[2] Microsoft Emotion API. https://www.microsoft.com/cognitive-services/en-us/emotion-api.

[3] E. Antonakos, J. A. i medina, G. Tzimiropoulos, and S. Zafeiriou. Hog active appearance models. In *ICIP*, 2014.

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

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

[6] 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. *PAMI*, 2015.

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

[8] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *PAMI*, 2013.

[9] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI*, 1997.

[10] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

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

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

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

[14] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FG*, 2011.

[15] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

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

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

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

[19] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*. 2005.

[20] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *IMAVIS*, 2005.

[21] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[22] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. *CoRR*, 2014.

[23] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, 2014.

[24] G. B. Huang, V. Jain, and E. G. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[25] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[26] 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.

[27] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with gaussianface. In *AAAI*, 2015.

[28] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[29] A. Martínez and R. Benavente. The ar face database. Technical report, Computer Vision Center, 1998.

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

[31] B. M. Oh, M. Chen, J. Dorsey, and F. Durand. Image-based modeling and photo editing. In *Conference on Computer Graphics and Interactive Techniques*, 2001.

[32] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014.

[33] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *PR*, 2009.

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

[35] J. M. Saragih, S. Lucey, and J. F. Cohn. Real-time avatar animation from a single image. In *FG*, 2011.

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[37] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. ICML, 2004.

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

[39] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. ECCV, 2002.

[40] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013.

[41] 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.

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

[43] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. 2014.

[44] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *CoRR*, 2014.

[45] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. 2014.

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

[47] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPR Workshop*, 2006.

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

[49] K. Weinberger. *Metric Learning with Convex Optimization*. PhD thesis, University of Pennsylvania, 2007.

[50] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[51] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *CVPR*, 2008.

[52] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *ACCV*, 2010.

[53] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 2009.

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

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

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

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

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

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

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

[61] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *CoRR*, 2014.