

# **Stereo Matching for Unconstrained Face Recognition**

Ph.D. Proposal

Carlos D. Castillo  
University of Maryland  
Department of Computer Science  
College Park, MD 20742  
carlos@cs.umd.edu

May 10, 2009

## Abstract

Unconstrained face recognition is a problem of fundamental importance in computer vision. We propose to address this problem by using stereo matching to judge the similarity of two, 2D images of faces. Stereo matching allows for arbitrary, physically valid, continuous correspondences. We show that the stereo matching cost provides a very robust measure of similarity of faces that is insensitive to a wide range of variations. To enable this, we show that for conditions common in face recognition, the epipolar geometry of face images can be computed using either four or three feature points. We also provide a straightforward adaptation of a stereo matching algorithm to compute the similarity between faces. The proposed approach has been tested on the CMU PIE dataset and demonstrates superior performance compared to existing methods in the presence of pose variation.

We have also studied the problem of face recognition with weight variation. Using this dataset, we empirically study how weight gain and loss affects face recognition performance. We present baseline experiments using a wide variety of existing methods. These show that weight change can significantly degrade the performance of recognition algorithms. Our results also show that correspondence-based methods exhibit the most robust performance as the weight difference increases.

The research plans include: building a stereo method for matching in the presence of deformation and illumination change and learning from pose invariant descriptors built using cost and correspondence information.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Face Recognition Across Pose . . . . .	3
<b>2</b>	<b>Background</b>	<b>6</b>
2.1	Face Recognition Across Pose . . . . .	6
<b>3</b>	<b>Preliminary Work</b>	<b>10</b>
3.1	Analysis of Stereo Matching for Face Recognition . . . . .	10
3.2	Alignment . . . . .	14
3.2.1	Epipolar Geometry under Scaled Orthographic Projection . . . . .	15
3.2.2	Epipolar Geometry and Horizontal Movement . . . . .	17
3.3	Stereo Matching and Face Recognition . . . . .	18
3.3.1	Rectification and Matching Costs . . . . .	20
3.4	Experiments . . . . .	21
3.4.1	PIE Pose Variation: 34 Faces . . . . .	22
3.4.2	PIE Pose Variation: 68 Faces . . . . .	26
3.4.3	PIE Pose and Illumination Variation . . . . .	26
3.5	Conclusion . . . . .	30
<b>4</b>	<b>Research Plan</b>	<b>31</b>
4.1	Stereo Matching with Illumination Variation . . . . .	31
4.1.1	Plan . . . . .	34
4.2	Image Representation, Correspondences and Learning for Unconstrained Face Recognition . . . . .	34
4.2.1	Image Representation . . . . .	36
4.2.2	Descriptor Generation . . . . .	36
4.2.3	Plan . . . . .	38
4.3	Face Recognition with Weight Variation . . . . .	38
4.3.1	Experimental Evaluation . . . . .	39
4.3.2	Discussion . . . . .	42
4.3.3	Plan . . . . .	44

4.4 Closing Remarks ..... 45

# Chapter 1

## Introduction

Face recognition is a fundamental problem in computer vision. There has been a lot of progress in the case of images taken under controlled conditions [46]. There are many approaches for handling, variation of illumination and expression. There are also several approaches to handling pose variation [37, 18, 20, 8]. However, there is still a lot of room for improvement. When multiple confounding factors occur simultaneously the problem is often termed unconstrained face recognition. Progress in unconstrained face recognition would be important in many applications, for example: surveillance, security, the analysis of personal photos and other domains in which we cannot control the conditions under which the images are taken.

Existing systems achieve excellent results when images are taken under controlled conditions, so that there is no variation in viewing conditions. Recently, there has been a good deal of work on recognition in the case of variations in viewing conditions that occur over a short period of time, such as variations in pose or lighting. Variations that occur over longer periods of time (such as aging and weight gain) have proven harder to study.

### 1.1 Face Recognition Across Pose

Correspondence seems crucial to produce meaningful image comparisons. The importance of good correspondences is even greater in the case of face recognition across pose. Standard systems often align the eyes or a few other features, using translation, similarity transformations, or perhaps affine transformations. However, when the pose varies these can still result in fairly significant misalignments in other parts of the face. Observe, for example, that in Figure 1.1 no linear transformation can make corresponding boxes have equal size, because a linear transformation can only linearly scale their size.

To handle this situation, we use stereo matching. This allows for arbitrary, one-to-one continuous transformations between images, along with possible occlusions, while maintaining an epipolar constraint. We show that the greater generality provided by stereo matching, which efficiently computes dense correspondences, may be necessary for effective face recognition across pose.

The purpose of stereo matching is to compute correspondences between scan lines of pixels in



Figure 1.1: Example images from the CMU PIE dataset. Observe that no linear transformation can make corresponding boxes have equal size.

images. Correct correspondences can be many-to-one and can involve occlusions. This means that situations like the one presented in Figure 1.1 can be handled by stereo matching.

In the process of computing the correspondences between scan lines in two images a *stereo matching* cost is optimized, which reflects how well the two images match. We can use the stereo matching cost as a measure of similarity between two face images.

Note that we are not interested in performing 3-D reconstruction, which is the most common purpose of stereo matching. In reconstruction the stereo matching costs are discarded and the correspondences are used along with geometric information about the camera layout to compute a 3-D model of the world. We have no use for the correspondences except to compute the stereo matching costs. We are therefore unaffected by some of the difficulties that make it hard to avoid artifacts in stereo reconstruction. For example, ambiguities frequently arise when different correspondences produce similar costs; in this case selecting the correct correspondence is essential for reconstruction, but not very important for judging the similarity of two images.

Prior to stereo matching, we need to estimate the epipolar geometry. In almost all applications of face recognition, the size of the face is small relative to its distance to the camera. Therefore we can approximate the projection of the face to the camera using scaled-orthographic projection (weak perspective). Under scaled-orthographic projection all epipolar lines are parallel to each other (the epipole is at infinity). This simplifies the problem of determining the epipolar geometry.

We propose two methods. One method uses four feature points to estimate the epipolar geometry of the two faces. The images are then rectified, and the similarity score is computed by adding the stereo matching cost of every row of the rectified images. The method works with general camera movement under the (very reasonable) assumption of scaled orthographic projection. We also study a specific case in which the camera is at the same height as the eyes of an upright subject. In this case, the epipolar lines are parallel to the lines that connect the two eyes. In this case we can determine epipolar geometry using only three points.

Putting these steps together, we have the following, remarkably simple algorithm:

- Prior to recognition, build a gallery of 2D images of faces, each with three to four landmark

points specified.

- Given a 2D probe image, find three to four corresponding landmark points.
- Compare the probe to each gallery image as follows:
  - Using landmark points, rectify the probe and gallery image.
  - Run a stereo algorithm on the image pair, using the enhancements described in Section 3.3. Discard the correspondences and use the matching cost as a measure of image similarity.
- Identify the probe with the gallery image that produces the lowest matching cost.

We will show that this method works very well even for large viewpoint changes. We evaluate our method using the CMU PIE dataset. Our results show that with pose variation at constant illumination our method is more accurate than previous methods due to Gross et al. [20], Chai, et al. [8] and Romdhani et al. [37]. While our method is designed to only handle pose variation, we also test it with pose and illumination variation to verify that our method does not fall apart in such a setup. Surprisingly, our method is more accurate than the method of Gross et al. [18], which is designed to handle lighting variation, though it is not as accurate as the method of Romdhani, et al. [37].

## Chapter 2

# Background

Face recognition is a fundamental problem in computer vision. It has been widely studied for the past 30 years. There has been a lot of progress in this research area, see [46] for an excellent survey.

In the past few years, there has been great interest in face recognition in unconstrained settings. By unconstrained it is understood, simultaneous variation in illumination, pose, expression, age and weight. To systematically study these variations, they have been separated into tractable groups, such as variation of pose and illumination, expression and pose, aging, etc. There are several variations in which a lot of progress has been made and there are other variations in which there is a lot of room for improvement.

### 2.1 Face Recognition Across Pose

Zhao et al. [46] review the vast literature on face recognition. Although the bulk of this work assumes fixed pose, there have been a number of approaches that do address the problem of pose variations.

Correspondences are fundamental for face recognition across pose. Many of these methods use some 3-D knowledge of faces to compensate for pose. In this case, obtaining correspondences becomes an operation of aligning the 3-D model to 2-D images: morphable model fitting, 3-D rigid transformations, sampling images from a 3-D model have been proposed. Other methods only use image information (i.e., they don't use 3-D knowledge). In this case obtaining correspondences becomes a 2-D matching problem: optical flow, estimating the light-field of the object and a wide variety of patch-based methods have been proposed. Typically, these approaches all rely on some initial manual correspondences. It is expected that if a method obtains good correspondences, it should obtain effective performance at face recognition across pose. Table 2.1 presents a summary of existing methods of face recognition across pose.

Historically, many approaches compensate for some 2-D deformations in matching, which may partially compensate for the effects of pose. A notable example is the work of Wiskott et al. [43]. This work was among the first to present a face recognition method that was robust to alignment issues. They developed a method called Elastic Bunch Graph Matching (EBGM). The comparison

function used *Gabor jets* at manually clicked feature points, and geometric information of distances between the feature points. Correspondences were obtained for the feature points only.

One of the first methods to study face recognition across pose was proposed by Beymer and Poggio [3]. In their work they generated 2-D virtual views from a single image per person using prior knowledge of the object class (in particular symmetry and prototypical objects of the same class) using optical flow. Once the virtual view had been generated the images were compared. Our method is similar to theirs in the sense that both are decidedly 2-D and stress the importance of finding good correspondences. In this approach the correspondences are obtained using optical flow between the two facial images.

Blanz and Vetter [5] use laser scans of 200 subjects to build a general 3-D morphable model of three dimensional faces. Then, with the aid of manually selected features, they fit this model to images. The parameters of the fit to two different images can be compared to perform recognition. In their experiments they show strong results for a subset of the poses in the PIE database. The work of Romdhani et al. [37] also focuses on 3-D morphable models. In this work shape and texture parameters of a 3-D morphable model are recovered from a single image. They present exhaustive results of experiments with pose variations for the PIE dataset and show strong results (the best prior results of which we are aware with pose variation). In these methods the correspondences are obtained by fitting the 3-D morphable model to the 2-D images. These type of methods solve a very difficult intermediate problem (fitting or inverse rendering) which is useful for graphics, but may not be needed for recognition.

Basri and Jacobs [2] use a 3-D model to generate a low dimensional subspace containing all the images that an object can produce under lighting variation. Pose is determined using manually selected point features. Correspondences are obtained by computing a 3-D rigid transformation that aligns the features of a 3-D model with the corresponding features of the 2-D images.

In Georghiades et al. [16] a 3-D model is computed for each person using a gallery containing a number of images per subject taken with controlled illumination at a constant pose. Pose variation is handled by sampling the set of possible poses, and building a 2-D model for each one. They evaluate their method using the Yale Face Database B. Correspondences with 2-D images are obtained by sampling the individual 3-D head model.

In Gross et al. [18] two appearance-based algorithms for face recognition across pose and illumination are presented. One of them is called eigen light-fields. At the core of the method is the *plenoptic function* or light field. To use this concept, all of the pixels of the various images are used to estimate the (eigen) light-field of the object. Correspondences are obtained by computing the light-field angles using the camera intrinsics and the relative orientation of the camera to the object (which are assumed to be known). They evaluate their results using the CMU PIE dataset [39]. In its assumptions, (recognizing faces across general unknown poses), this method is the most similar to ours. However our approach is simpler and our results are better.

The other method presented in Gross et al. [18] is called Bayesian Face Subregions (BFS). The algorithm models the appearance changes of the different face regions in a probabilistic framework.

Using probability distributions for similarity values of face subregions, the method computes the likelihood of probe and gallery images coming from the same subject. The method is designed to handle the case of simultaneous variation in pose and illumination. In this patch-based method, correspondences are computed trivially on a quadrilateral grid that includes the two eyes and the mouth as edges.

There have been several recent approaches to face recognition across pose that are based on patches. In Chai et al. [8], the authors present a learning, patch-based rectification method based on locally linear regression. Given a non-frontal facial image, the method provides a prediction strategy to generate the frontal view. In their experiments, the method compares well to other recent methods on the PIE dataset. Lucey and Chen [27] present a patch-based algorithm for face recognition across pose of sparsely registered images (4 manually selected points). Closely related, the work of Ashraf et al. [1] presents a new method to discover viewpoint-induced spatial deformations for general patch based methods of face recognition across pose.

All the methods previously mentioned in this section use intensity images of the face. This type of face recognition, based on 2-D images constitutes the vast majority of face recognition research. There is, however, a significant amount of work done acquiring, matching and performing recognition using 3-D reconstructions of faces (see [6] for a survey).

While progress has been made in handling pose variations, significant challenges remain. For this problem, current methods have substantially worse performance than when pose is fixed between the probe and gallery. In addition, many methods for handling pose variation require substantially more computation than other methods, and can be very slow. This is in part because the process of finding a correspondence between the probe and gallery requires expensive optimization processes.

Table 2.1: Key aspects of existing methods for face recognition across pose.

Method	Type	Correspondences	# of manually specified points
Wiskott et al.	2-D	Jets only at points with manually specified correspondences	4-7
Beymer and Poggio	2-D	Optical flow	4-6
Blanz and Vetter	3-D general model	3-D model fitting	10-20
Romdhani et al.	3-D general model	3-D model fitting plus extensions	10-15
Basri and Jacobs	3-D person-specific model	3-D rigid transformation	5
Gheorgiades et al.	3-D person-specific model	Sampling from a built 3-D model	Requires training and test images in the same pose
Gross et al. (ELF)	2-D	Computing the eigenlightfield, known camera geometry	3/40+
Gross et al. (BFS)	2-D	Patches, sampled uniformly on the central region of the face	3
Chai et al.	2-D	Rectification through locally-linear regression	5
Lucey and Chen	2-D	Patches, learning patch dependency	4
Ashraf et al.	2-D	Patches, learning the spatial deformation of the patches	4

## Chapter 3

# Preliminary Work

Our preliminary work is presented in this section. In this section we will present an effective method face recognition across pose using stereo matching.

### 3.1 Analysis of Stereo Matching for Face Recognition

Most work in image-based recognition aligns regions to be matched with a low-dimensional transformation, such as translation, or a similarity or affine transformation. Instead, we use stereo matching. When we enforce the ordering constraint, this allows for arbitrary, one-to-one continuous transformations between images, along with possible occlusions, while maintaining an epipolar constraint. In this section we show that the greater generality afforded by stereo matching may be necessary for face recognition, and that stereo matching will not be too sensitive to noise in determining the epipolar lines.

We illustrate this using a *very* simplified model of faces, in which we calculate the disparity maps that will correctly match two images. We do not attempt to accurately capture face shape in this example. Rather, we just provide a coarse demonstration of the disparity variation that can occur under viewing conditions similar to those that typically occur in face recognition.

1. We model the face as a cylinder. Perturbations to this model, such as adding a nose, can be handled fairly easily.
2. We assume the face is viewed by two cameras with image planes that are rectified to be perpendicular to the  $z$  axis and that the cylinder axis is the  $y$  axis. This is roughly the situation when an upright person photographs another upright person. For simplicity, we will assume that the cylinder lies on the  $z$  axis, that the camera focal points lie on the  $x$  axis at points symmetric about the  $z$  axis (see Figure 3.1). We call the left and right focal points  $f_l$  and  $f_r$  respectively.
3. We assume that the distance from the camera to the person is much bigger than the radius of the cylinder that represents the person. Specifically, we assume that vectors from the camera

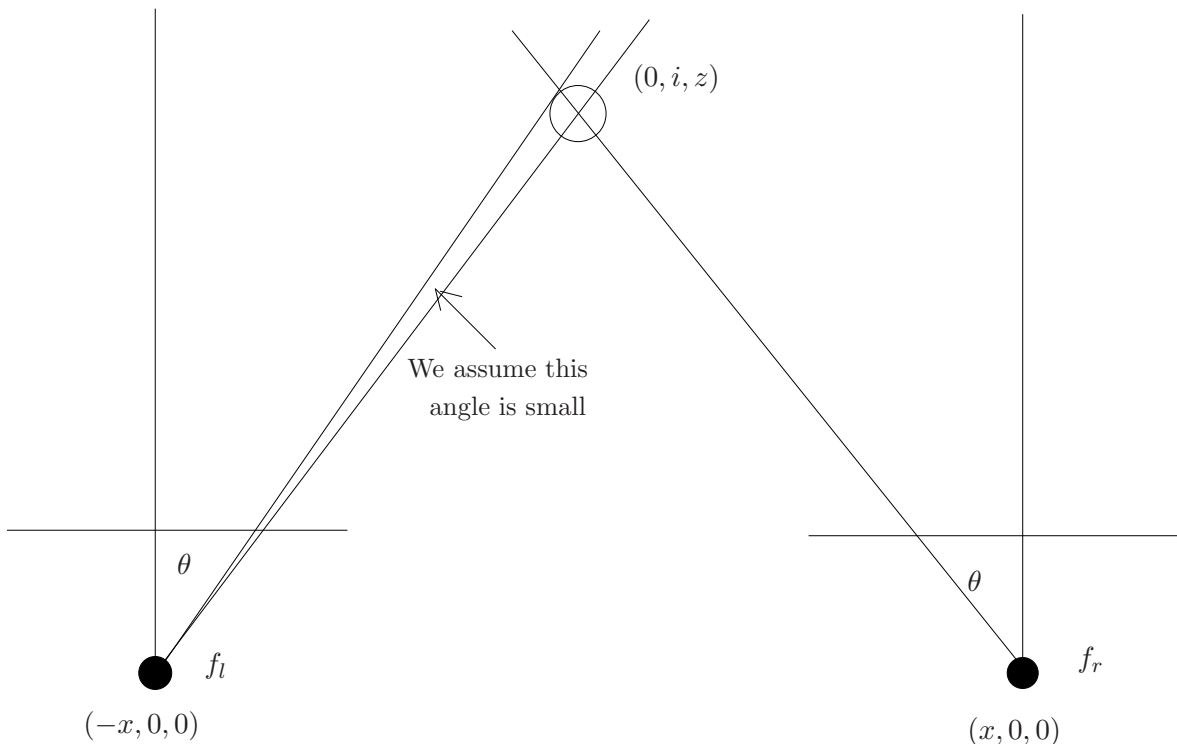


Figure 3.1: Our very simplified model of faces.

focal point to any location on a horizontal cross section of the cylinder have the same direction. If we imagine that the cylinder (face) has a radius of three inches, and the distance from the camera to the face is 8 feet, we can calculate that a vector from the focal point to the center of a cross-section of the cylinder will be within 5.5 degrees of a vector to any point on the cylinder cross section, so this approximation is not too bad.

These assumptions simplify our presentation, which could be readily extended to other settings.

We will analyze disparities on the  $y = 0$  plane. Given these assumptions, each camera will see half of a circular cross-section. They will not see exactly the same half-circle, however, as there will be some occlusion. Without loss of generality assume the radius of the circle is 1. We will denote the angle between the  $z$  axis and a vector from  $f_l$  to the cylinder by  $\theta$ . The corresponding angle for the right camera will then be  $-\theta$ . Define  $l_1$  and  $l_2$  to be two points on the circle, such that the tangent lines to the circle at  $l_1$  and  $l_2$  pass through  $f_l$ . That is,  $l_1$  and  $l_2$  are the first and last points on the circle that are visible in the left image. Define  $L$  to be the line connecting  $l_1$  and  $l_2$ . We can similarly define  $r_1$  and  $r_2$  for the right image. So, for example, the region of the circle between  $r_1$  and  $l_2$  is visible in both images.

Note that every line connecting  $f_l$  to  $L$  intersects the circle in a single point that will be visible in the left camera. So one way to determine the image of the circle in the left camera is to project the visible half-circle onto  $L$  using these lines, and then to consider how  $L$  is projected onto the left camera. Because we assume the cylinder is small relative to its distance to the camera, we can

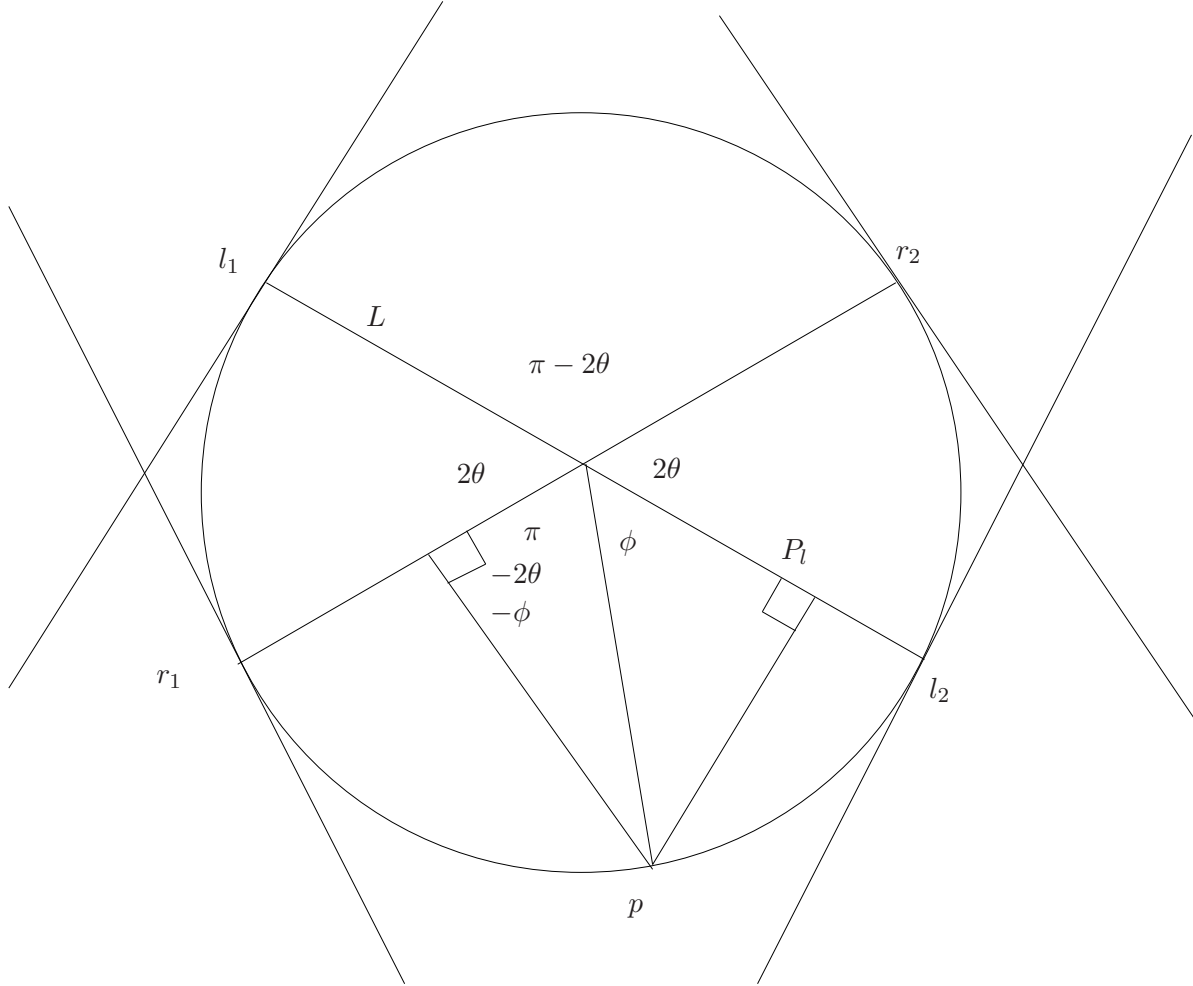


Figure 3.2: The circle parameterized by the angle  $\phi$ .

approximate the projection of  $L$  into the left camera using scaled-orthographic projection. Without loss of generality we can normalize the left image so that the width of the circle's projection is 1 (this is in image units, which may differ from 3D units), and the  $x$  coordinate of the image of  $l_1$  is 0. This is illustrated in Figure 3.2.

We can parameterize points on the circle by the angle  $\phi$ , which we take relative to  $l_2$  (see Figure 3.2). Consider some such point  $p$ . We can determine the location of  $p$  in the left image, by considering the line through  $p$  and  $f_l$ . The point where this line intersects  $L$ , call it  $P_l$ , will appear in the same image location as  $p$ . Define the distance from  $P_l$  to  $l_1$  to be  $d(l_1, P_l)$ . Then the  $x$  coordinate of  $p$  in the left image is  $d(l_1, P_l)/2 = (1 + \cos \phi)/2$ . Similarly, its position in the right image will be  $(1 - \cos(\pi - 2\theta - \phi))/2$ . If we define the disparity,  $d$ , in a matched point to be its  $x$  coordinate in the left image minus the  $x$  coordinate in the right; we get:

$$d = (\cos \phi + \cos(\pi - 2\theta - \phi))/2 \tag{3.1}$$

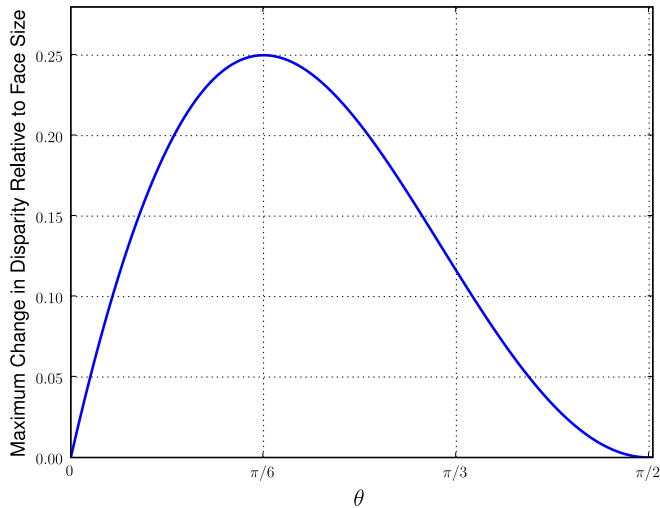


Figure 3.3: Change in disparity relative to the size of the face as a function of  $\theta$ .

It is straightforward to show that disparity is minimized by  $\phi = 0$  or  $\phi = \pi - 2\theta$ , which are the furthest points visible in both cameras, and maximized by  $\phi = (\pi - 2\theta)/2$ , which corresponds to the point closest to the cameras.

We are interested in the variation between the minimum and maximum disparity values,  $\Delta d$ . We have:

$$\Delta d = \cos\left(\frac{\pi}{2} - \theta\right) - \frac{1}{2} - \frac{\cos(\pi - 2\theta)}{2} \quad (3.2)$$

This is maximized for  $\theta = \pi/6$ , when  $\Delta d = 1/4$ . Figure 3.3 shows how the maximum change in disparity varies with  $\theta$ . In Figure 3.3, we can see that for a large range of  $\theta$ , disparity changes quite a bit within the image.

From this analysis, we can see that for a cylinder, disparity in an image can vary by as much as 1/4 of the apparent width of the cylinder, and frequently varies substantially. These variations in disparity cannot be accounted for by aligning the images with a linear transformation, since linear transformations can only create linear disparity maps. In contrast, the disparity map for this cylinder is highly non-linear, since the smallest disparity is at the two ends of the image, and the greatest disparity occurs in the middle of the image. In fact, in scenarios such as the one described here, because of the symmetry of the viewing conditions, we can demonstrate that the optimal linear transformation to align the two images will simply be the identity transformation, which does not account for any of these variations in disparity. Note that the amount of disparity is independent of the distance from the cameras to the face, because we measure disparity relative to the apparent size of the face.

Ideally, one should determine the epipolar geometry prior to matching two faces. However, in

many cases, images result from an upright photographer taking a picture of an upright subject. This results in epipolar lines that are approximately horizontal. If we align the eyes in two photographs, this will align corresponding horizontal epipolar lines. However, error will result when epipolar lines are not purely horizontal. To get a sense of the possible magnitude of this error, we analyze a simple example.

Consider the case in which we take two pictures of a face that is five feet high, at a distance of eight feet. But suppose that the disparity is vertical instead of horizontal, because one photograph is taken from a height of five feet, and the second is taken from a height of six feet. Vertical disparity will be zero at the eyes, which are aligned, and will be maximized at the point that is closest to the cameras, the tip of the nose. If we assume that the nose is about one inch long, then using similar triangles we can determine that it appears at the same image location as a point  $1/8$  of an inch below the nose, in the second image. For a face that is six inches long, the vertical disparity will therefore be about 2% of the height of the face in the image. This error is small compared to the variations of up to 25% in horizontal disparity that can arise in the situation we analyze above. Of course, this is just an illustrative example; the error introduced by mis-estimation of the epipolar lines will depend in practice on the viewing conditions typical in a specific application. Our example simply makes the point that in some common settings, this error will be quite small, while stereo matching can compensate for correspondence errors that will be large.

## 3.2 Alignment

In order to perform stereo matching we first need to know the epipolar geometry. In the most general case this requires eight corresponding points. We can reduce this by assuming that images are generated by scaled orthographic projection. This model is valid when the average variation of the depth of the object along the line of sight is small compared to the distance of the camera to the object and the field of view is small as is generally the case with facial images. Note that, as shown in Section 3.1, even with scaled orthographic projection there can be considerable variation in disparity between two images.

To begin, consider the case of two images generated with orthographic projection. Orthographic projection occurs with a perspective camera model when the focal point is at infinity. The *baseline*, which connects the two focal points, is therefore a line at infinity. The *epipole* of each image, then, is a point where this line at infinity intersects the image plane. This means that the epipoles are points at infinity in each image plane. The *epipolar lines* in each image therefore intersect at a point at infinity, meaning that they are parallel. If we also allow for scaling in each image, this may alter the distance between corresponding epipolar lines, but will not affect the fact that they are parallel.

As we will demonstrate, we can calculate the epipolar geometry under the scaled orthographic model using four feature points. We will not focus our attention on how these points can be obtained; in our experiments we specify them by hand. Some applications involving off-line recognition may use such hand clicked points directly. At the same time there is a lot of work on automatic detection

of facial features [17, 22, 10, 38]. By reducing the number of points needed for recognition, we can make it easier to use these detectors to build fully automatic recognition systems.

### 3.2.1 Epipolar Geometry under Scaled Orthographic Projection

We now want to consider arbitrary viewpoint changes, still using scaled orthographic projection. Under scaled orthographic projection the epipolar geometry can be characterized as a tuple:  $(\theta, \gamma, s, t)$ .  $\theta$  is the angle of the epipolar lines in the first image.  $\gamma$  is the angle of the epipolar lines on the second image.  $s$  is the relative scale; that is, scaling the second image by  $s$  will cause the distance between two epipolar lines in the second image to match the distance between corresponding lines in the first image. Finally,  $t$  is the translation perpendicular to the epipolar lines needed to align corresponding lines.

Solving for this type of epipolar geometry requires four corresponding points. We formulate this by encoding the three variables relating to the second image,  $(\gamma, s, t)$ , as a similarity transformation, with the added constraint that the translation must be perpendicular to the epipolar lines. Given corresponding points in the two images, this similarity transformation must transform each point in the second image onto a line in the first image that passes through the corresponding point, at an angle  $\theta$ . This yields a constraint for each point of the form:

$$(P_{2x}^i, P_{2y}^i, 1) \begin{pmatrix} a & -b & 0 \\ b & a & 0 \\ T_x & T_y & 1 \end{pmatrix} \begin{pmatrix} l_{1x}^i \\ l_{1y}^i \\ l_{1c}^i \end{pmatrix} = 0 \quad (3.3)$$

$(P_{2x}^i, P_{2y}^i)$  are the coordinates of the  $i$ 'th point in the second image, while  $(l_{1x}^i, l_{1y}^i, l_{1c}^i)$  represents the line of slope  $\tan(\theta)$  that passes through each of the points in the first image. Note that it is convenient to represent the lines in parametric form (in terms of  $\sin(\theta)$  and  $\cos(\theta)$ ), so that after multiplying the first two components of Eqn. 3.3 each restriction becomes:

$$(aP_{2x}^i + bP_{2y}^i + T_x, -bP_{2x}^i + aP_{2y}^i + T_y, 1) \begin{pmatrix} -\sin(\theta) \\ -\cos(\theta) \\ \sin(\theta)P_{1x}^i - \cos(\theta)P_{1y}^i \end{pmatrix} = 0 \quad (3.4)$$

The final constraint comes from the fact that  $T_x$  and  $T_y$  are not independent, they are constrained to be translations perpendicular to the angle of the epipolar lines in the first image  $\theta$ :

$$\cos(\theta)T_x + \sin(\theta)T_y = 0 \quad (3.5)$$

We can think of  $\sin(\theta)$  and  $\cos(\theta)$  as separate variables, with the constraint  $\sin(\theta)^2 + \cos(\theta)^2 = 1$ . Then, with Eqns. (3.4) and (3.5), we have a system of bilinear and a quadratic equation. This has six unknowns,  $a, b, T_x, T_y, \sin(\theta)$  and  $\cos(\theta)$ , and  $n + 2$  equations, given  $n$  point correspondences. We solve this in a very simple way. Noting that the equations become linear when  $\theta$  is known, we simply consider a brute-force sampling of  $\theta$ , and check which value produces a consistent set of

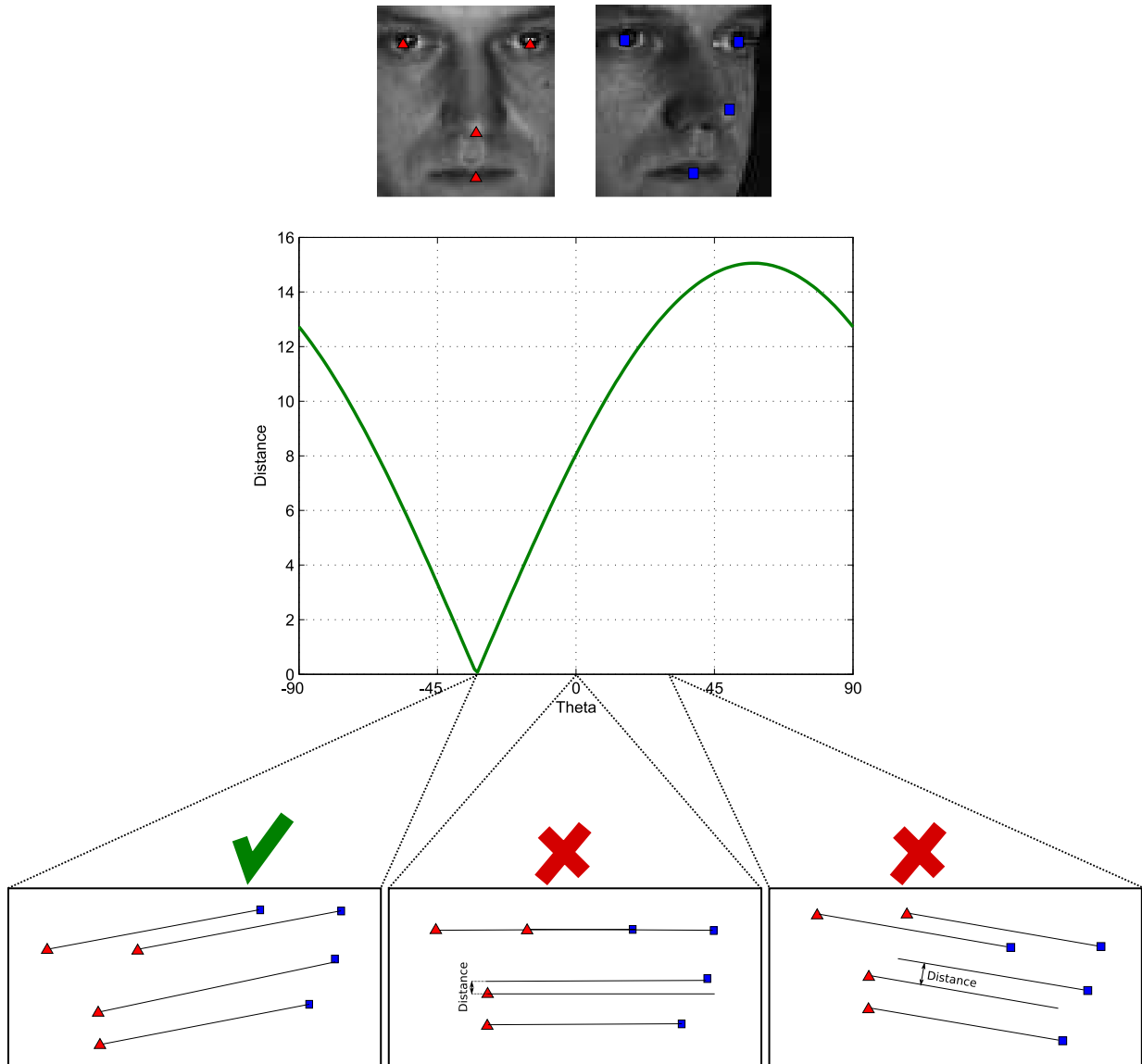


Figure 3.4: Example of our method to compute the epipolar under scaled orthographic projection. For each angle  $\theta$  we compute the distance perpendicular to it of a fourth point in the two images. We choose the epipolar geometry that has the smallest distance.

linear equations. For each  $\theta$  we compute the alignment (a candidate epipolar geometry) given 3 points. When this has been done, we use the fourth point to compute, the quality of the alignment.

1. Use 3 points to solve for  $(a, b, T_x, T_y)$  using Eqns. (3.4) and (3.5).

2. Apply the similarity transform  $M = \begin{pmatrix} a & b & T_x \\ -b & a & T_y \\ 0 & 0 & 1 \end{pmatrix}$  to the second image.

3. Use the distance of the 4th point in the direction perpendicular to  $\theta$  to determine how good the match is. The best transformation  $M$  is the one that minimizes this distance.

The rectification is procedure is, therefore, applying the best  $M$  to the second image and then rotating both images by  $\theta$  in such a way that the epipolar lines become horizontal.

After this is done, we are ready to compute the stereo matching cost to determine the image similarity.

### 3.2.2 Epipolar Geometry and Horizontal Movement

We will now study a particular case of the general setup: an upright person with both images taken with the camera located at the same height as the person’s head (in fact, our reasoning applies to any situation in which the eyes and both camera focal points are coplanar). In that case we know that the epipolar lines are parallel to the lines connecting the eyes. For this case we only determine the epipolar geometry using three feature points. The two eyes will define the direction of the epipolar lines. This tells us  $\theta$ . Given a correspondence between three points, Eqns. (3.4) and (3.5) then provide four linear constraints on four unknowns, allowing us to solve for the epipolar geometry linearly. Moreover, our experiments show that in many practical situations, even when the cameras are not perfectly at eye level these alignments work reasonably well.

Since this is the simplest alignment method we study, this procedure is, additionally, the base procedure we use to generate the thumbnails for the four-point alignment procedure explained in Section 3.2.1. The method presented in this section is equivalent to the case presented in Section 3.2.1 when  $\theta = 0$ .

We now describe a simple method of rectifying the two images so that horizontal rows of each image contain corresponding epipolar lines. Note that this rectification does not require that the three matched landmark points in the two images must coincide, just that corresponding points should lie on corresponding horizontal lines in the rectified images.

1. Rotate the image so the eyes are horizontal.
2. Scale the image so that the vertical distance between the eyes and the mouth is an arbitrary but fixed  $d$ .
3. Translate the images up/down in such a way that eyes are on an arbitrary but fixed line  $y_e$ .

4. Translate in the x direction so the center of mass of the x coordinates is 0. This step is not needed to align corresponding epipolar lines, but is convenient.
5. Cut a thumbnail in such a way that the height is arbitrary but fixed and the thumbnail includes the three feature points.

Note that this procedure will produce thumbnails that will have different widths but a fixed height. This is appropriate, since given our assumptions the apparent height of a face will be the same for all images, but its apparent width may vary with the viewing direction.

### 3.3 Stereo Matching and Face Recognition

There exist a wide variety of stereo algorithms. We require an efficient stereo algorithm appropriate for wide baseline matching of faces. Since faces are very slanted objects, we require the algorithm to have excellent support for surfaces that are not fronto-parallel planes. A number of methods might be suitable. We have decided to use a 1-D dynamic-programming based algorithm, which is quite fast. We have used Criminisi et al. [13]<sup>1</sup> which has been developed for video conferencing applications and so seems to fit our needs. This algorithm handles slanted surfaces in a very elegant way. It is not obvious that it will work for the large changes in viewpoint that can occur in face recognition, but we will show that it does.

In this section we will review the stereo matching method of Criminisi et al. [13] as it is presented by its authors. In the following section we will describe how we adapt the algorithm for the purpose at hand.

It is important to stress that we are relatively unaffected by some of the difficulties that make it hard to avoid artifacts in stereo reconstruction. For example, when many matches have similar costs, matching is ambiguous. One weakness of dynamic programming stereo algorithms is that when matching is ambiguous, it can be difficult to produce correspondences that are consistent across scan lines. Selecting the right match is difficult, but important for good reconstructions. Since we only use the cost of a matching, selecting the right matching is unimportant to us in this case. Also, errors in small regions, such as at occluding boundaries, can produce bad artifacts in reconstructions, but that is not a problem for our method as long as they don't affect the cost too much.

The core of the stereo method calculates a matching between two scanlines (rows of each face). The algorithm is a dynamic programming stereo matching algorithm that is fast and performs well when compared to other methods.

The algorithm accounts for exactly one pixel in one image with each step taken. Each step involves a transition from one point to another in four planes (or cost matrices) called  $C_{Lo}$ ,  $C_{Lm}$ ,  $C_{Ro}$  and  $C_{Rm}$ . Each point in a matrix represents the last point in each image that has been accounted

---

<sup>1</sup>We also tried the method described in Cox et al. [12] and found the method to be about twice as fast but less accurate (about 8% on average on several gallery-probe experiments with a gallery of 68 individuals) than the method described in Criminisi et al. [13].

for, along with the nature of the last step used to account for a point. Points are accounted for by matching (m) and occlusions (o) in the left (L) and right (R) images. The planes naturally define the persistence of states. By setting the state transition costs adequately many state transitions can be favored or biased against. For example long runs of occlusions can be favored over many short runs by setting a high cost for entering or leaving an occluded state. This formulation handles slanted surfaces well (because it allows many-to-one matches) and offers better control over the occlusion costs than traditional one plane models [12].

The elements of the cost matrix are initialized to  $+\infty$  everywhere except in the right occluded plane where:

$$C_{Ro}[i, 0] = i\alpha \quad \forall i = 0 \dots W - 1 \quad (3.6)$$

$\alpha$  is the cost of a persistent occlusion.

The forward step of the 4-state DP computes the four cumulative cost matrices according to the following recurrence relation, in which  $\beta$  is the cost of beginning an occlusion, and  $\beta'$  is the cost of ending one:

$$C_{Lo}[l, r] = \min \begin{cases} C_{Lo}[l, r - 1] + \alpha \\ C_{Lm}[l, r - 1] + \beta \\ C_{Rm}[l, r - 1] + \beta \end{cases} \quad (3.7)$$

$$C_{Lm}[l, r] = M(l, r) + \min \begin{cases} C_{Lo}[l, r - 1] + \beta' \\ C_{Lm}[l, r - 1] + \gamma \\ C_{Rm}[l, r - 1] \\ C_{Ro}[l, r - 1] + \beta' \end{cases} \quad (3.8)$$

where  $M(l, r)$  is the cost of matching the  $l$ th pixel in the left scanline with the  $r$ th pixel in the right scanline.  $\alpha$ ,  $\beta$ ,  $\beta'$  and  $\gamma$  are parameters that can be set experimentally.  $C_{Ro}$  and  $C_{Rm}$  are symmetric. Our experiments show that the method is rather insensitive to these parameters and all experiments shown here are run with  $\alpha = 0.5$ ,  $\beta = \beta' = 1.0$  and  $\gamma = 0.10$  as recommended in [13].  $M(l, r)$  is a fast approximation to the normalized cross correlation of a  $3 \times 7$  window around the points  $(l, s)$  and  $(r, s)$  of the images, where  $s$  is the current scanline.

The cost of matching the two scan lines  $l_1$  and  $l_2$ , denoted  $\text{cost}(l_1, l_2)$ , is:  $C_{Ro}[l - 1, r - 1]$ . The optimal matching solution will be a sequence of symbols in the alphabet:  $\Sigma = \{C_{Lo}, C_{Lm}, C_{Ro}, C_{Rm}\}$  which can be obtained by following a backward step. A solution (a word in  $\Sigma^*$ ) that encodes the optimal matching to a given matching problem between scanlines  $I_{1,i}$  and  $I_{2,i}$  has length equal to  $|I_{1,i}| + |I_{2,i}|$ . We have no use for the optimal matching itself, we only use its cost and its length to normalize it.

One of the key ingredients to the flexibility of this method is the ability to match multiple pixels in one scanline to one pixel in the other. This is done by concatenating several consecutive  $C_{Lm}$

(or  $C_{Rm}$ ) in the word that encodes the solution.

### 3.3.1 Rectification and Matching Costs

When we match a probe image to different gallery images, we obtain different rectifications. While the original thumbnails are axial rectangles, the rectified thumbnails will be arbitrarily rotated rectangles that will contain varying numbers of rows with valid pixels, and different numbers of valid pixels in each row. It is therefore important to avoid any bias in our image comparisons which favor some thumbnail orientations over others. In this section we explain how to adapt Criminisi et al. [13] to match rectified images in which the length of scanlines varies.

The equations presented in Eqns. 3.7 and 3.8 are an effective measure of similarity when the two images are square and of identical size. When these assumptions are broken, Eqns. 3.7 and 3.8 stop being an effective measure of similarity because now in image comparisons there will be a different number of pixels in each image. We will focus this section in adapting this metric for the purpose at hand.

As previously mentioned, all solutions found by the method of Criminisi et al. have length equal to the sum of both scan lines being matched. This is due to the fact that the algorithm at every step accounts for exactly one pixel. However, since each cost is going to be compared to other costs matched over scanlines of potentially different lengths, we need some normalization strategy.

There are two sensible normalizations that can be used, one that weighs every row equally:

$$\text{cost}(I_1, I_2) = \frac{1}{n} \sum_{i=1}^n \frac{\text{cost}(I_{1,i}, I_{2,i})}{|I_{1,i}| + |I_{2,i}|} \quad (3.9)$$

and one that weighs every match (arc in the graph) equally. This is the one we actually use:

$$\text{cost}(I_1, I_2) = \frac{\sum_{i=1}^n \text{cost}(I_{1,i}, I_{2,i})}{\sum_{i=1}^n |I_{1,i}| + |I_{2,i}|} \quad (3.10)$$

The cost expressed in Eqn. 3.10 is a sensible measure of similarity since it is not dependent on the relative scale of the images, it just calculates the average cost per match made (that is per arc in the graph) over all scan lines. However, the costs in Eqns. 3.9 and 3.10 are built on top of the structure of the match found. This property is useful because it makes the cost not depend on the shape of the non-data that is present in the image, and therefore there will be no biases towards matches with scan lines in both images at the same angles.

We identify two special cases to Eqns. (3.9) and (3.10):

1. When two scan lines of non-data are being matched
2. When a scan line of non-data is being matched to scan line of data

We could pay a constant penalty for each of these special situations but doing so would artificially add noise to the similarity cost. We decide to not include these special cases in the average described

in Eqn. (3.10) and let the other pixels, for which there is actual data to match, decide what the average cost per match should be.

Let  $\text{cost}(I_1, I_2)$  be defined as either of the two cases studied above. Since we do not know which image is left and which image is right we have to try both options. One of them will be the true cost, the other cost will be noise and should be ignored.

$$\text{similarity}(I_1, I_2) = \min \begin{cases} \text{cost}(\text{rectify}(I_1, I_2)) \\ \text{cost}(\text{rectify}(I_2, I_1)) \\ \text{cost}(\text{rectify}(\text{flip}(I_1), I_2)) \\ \text{cost}(\text{rectify}(I_2, \text{flip}(I_1))) \end{cases} \quad (3.11)$$

Additionally, *flip* produces a left-right reflection of the image and adjusts the hand clicked positions of the four points accordingly. *flip* is helpful when two views see mainly different sides of the face. In this case, a truly correct correspondence would mark most of the face as occluded. However, since faces are approximately vertically symmetric, *flip* approximates a rotation about the  $y$  axis that creates a virtual view so that the same side of the face is visible in both images. For example, if we viewed a face in left and right profile, there would be no points on the face visible in both images, but flipping one image would still allow us to produce a good match. *rectify* performs the rectification described in the 4-point case, or in the 3-point case does nothing at all, since all images are already partially rectified to handle this case.

Finally, we perform recognition simply by matching a probe image to the most similar image in the gallery. For the method to work well all the images in the gallery should be in the same pose.

Before closing this section it is important to note how simple the proposed approach is. It is a two step process: (1) alignment according to assumptions regarding the viewing conditions, (2) similarity computation using stereo matching. In the next section we will see that this very straight-forward approach demonstrates excellent performance.

## 3.4 Experiments

We have tested our algorithm using the CMU PIE database [39]. This database consists of 13 poses of which 9 have approximately the same camera altitude (poses: c34, c14, c11, c29, c27, c05, c37, c25 and c22). Three other poses that have a significantly higher camera altitude (poses: c31, c09 and c02) and one last pose that has a significantly lower camera altitude (pose c07). We say that two poses have aligned epipolar lines if they are both from the set: {c34, c14, c11, c29, c27, c05, c37, c25, c22}. If not, we say that two poses have misaligned epipolar lines.

The thumbnails used were generated as described in Section 3.2.2. All images have a height of 72, a pose-dependent width and a distance between the eyes and the mouth of  $d = 50$  and the eyes are horizontally located in  $y_e = 13$ . For the 3-point Stereo Matching Distance (3ptSMD) this is all

the image processing performed, the stereo matching cost was then computed and normalized and this cost is the image similarity between the two faces. For the 4-point Stereo Matching Distance (4ptSMD) the epipolar rectification was then performed on the thumbnail. After rectification, the stereo matching cost was computed and this cost is the image similarity between the two faces.

A number of prior experiments have been done with pose variation using the CMU PIE database, but somewhat different experimental conditions. We have run our own algorithm under a variety of conditions so that we may compare to these. For example, to compare results with [18, 20, 8] we need to use a subset of 34 people because they use 34 people for training and the remaining 34 for testing. We do not require training, but we are interested in comparing the methods in equal conditions so we tested on individuals 35-68 from the PIE database. To compare with [37] we used 68 people as a test set. Then to illustrate that our method works in more realistic situations we evaluated simultaneous variation in pose and illumination. This too is done in two separate experiments, one to compare with [18, 20] and one to compare with [37].

### 3.4.1 PIE Pose Variation: 34 Faces

We conducted an experiment to compare our method with four others. We compared with two variants of eigen light-fields[18], eigenfaces[41] and FaceIt as described in [18, 20]. FaceIt<sup>2</sup> is a commercial face recognition system from Identix which finished top overall in the Face Recognition Vendor Test 2000. Eigenfaces is a common benchmark algorithm for face recognition. Finally, eigen light-fields is a state of the art method for face recognition across pose.

In this experiment we selected each gallery pose as one of the 13 PIE poses and the probe pose as one of the remaining 12 poses, for a total of 156 gallery-probe pairs. We evaluated the accuracy of our method in this setting and compared to the results in [18, 20]. Table 3.3 summarizes the average recognition rates. Table 3.1 presents detailed results for this experiment using 3ptSMD and Table 3.2 presents detailed results for this experiment using 4ptSMD. Figure 3.5 shows several cross-sections of the results with different fixed gallery poses.

The fact that 3ptSMD performs solidly both when the epipolar lines fit (with an average of 81.4%) and when they don't (with an average of 75.4% ) and overall (with an average of 78.5% as reported in Table 3.7) shows that assuming horizontal epipolar geometry is not a bad approximation for real applications of face recognition across pose, even when this assumption does not hold perfectly.

Figure 3.5 shows a comparison with the results presented in the paper of Gross et al. [18, 20]. In this experiment we observe that in all gallery poses our method outperforms all the other methods for the extreme probe poses (c34, c31, c14, c02, c25 and c22). Observe that the 4ptSMD method is considerably better than than 3ptSMD at the poses where there is considerable misalignment (the poses marked with \*).

Table 3.4 shows a comparison with Chai et al. [8], using the experimental conditions described in their paper. The gallery pose is c27 and contains 34 faces, the probe poses are: c05, c29, c37,

---

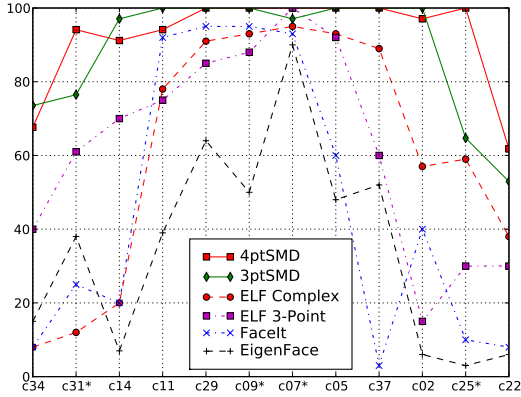
<sup>2</sup>Version 2.5.0.17 of the FaceIt recognition engine was used.

Table 3.1: Results for pose variation for 34 faces with 3ptSMD. The diagonals are not included in any average. The table layout is the same as [37] and [21]. The global average for this table is 79.8%.

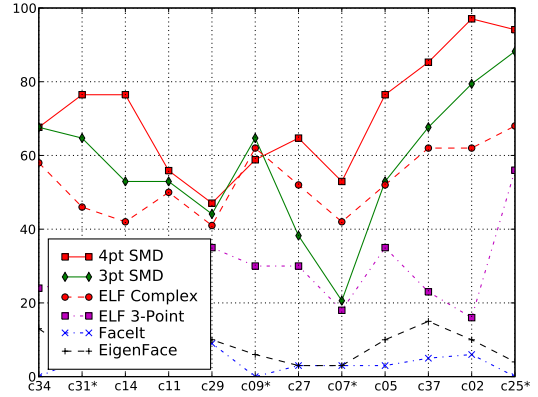
azimuth	-66	-47	-46	-32	-17	0	0	0	16	31	44	44	62	
altitude	3	13	2	2	2	15	2	1.9	2	2	2	13	3	
Probe Pose	c34	c31	c14	c11	c29	c09	c27	c07	c05	c37	c25	c02	c22	avg
Gallery Pose														
c34	-	91	82	74	62	32	35	18	50	56	65	71	71	58
c31	94	-	88	88	76	91	59	32	56	65	97	82	76	75
c14	94	91	-	100	100	82	91	76	88	82	56	88	56	83
c11	94	97	100	-	100	88	100	94	94	97	53	94	65	89
c29	88	88	100	100	-	100	100	100	100	97	62	94	53	90
c09	59	100	76	94	100	-	97	82	97	88	97	82	79	87
c27	74	76	97	100	100	100	-	97	100	100	65	100	53	88
c07	29	41	74	91	97	79	100	-	97	82	38	65	24	68
c05	68	76	100	97	100	97	100	94	-	100	85	100	79	91
c37	79	82	97	100	94	91	100	91	100	-	94	100	94	93
c25	47	94	44	59	44	91	47	18	68	85	-	79	97	64
c02	79	79	94	91	88	82	94	62	100	97	94	-	94	87
c22	68	65	53	53	44	65	38	21	53	68	88	79	-	57

Table 3.2: Results for pose variation for 34 faces with 4ptSMD. The diagonals are not included in any average. The table layout is the same as [37] and [21]. The global average is 86.82%.

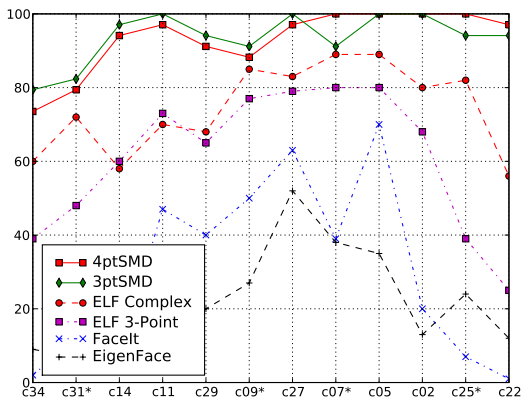
azimuth	-66	-47	-46	-32	-17	0	0	0	16	31	44	44	62	
altitude	3	13	2	2	2	15	2	1.9	2	2	2	13	3	
Probe Pose	c34	c31	c14	c11	c29	c09	c27	c07	c05	c37	c25	c02	c22	avg
Gallery Pose														
c34	-	91	91	82	68	44	44	35	50	53	65	74	65	63
c31	97	-	100	97	97	94	76	56	65	74	91	82	76	83
c14	100	100	-	100	97	85	91	71	91	82	68	91	85	88
c11	97	97	100	-	100	94	94	97	100	97	82	94	74	93
c29	85	97	97	100	-	100	97	100	97	97	85	94	53	91
c09	62	97	91	100	100	-	100	97	100	97	91	91	76	91
c27	68	94	91	94	100	100	-	100	100	100	100	97	62	92
c07	41	71	79	97	100	100	100	-	100	97	85	94	35	83
c05	79	85	100	100	97	94	100	100	-	100	100	100	91	95
c37	74	79	94	97	91	88	97	100	100	-	100	100	97	93
c25	76	88	68	76	82	88	97	82	100	100	-	97	97	87
c02	88	85	88	85	85	94	91	94	97	100	97	-	100	92
c22	68	76	76	56	47	59	65	53	76	85	94	97	-	71



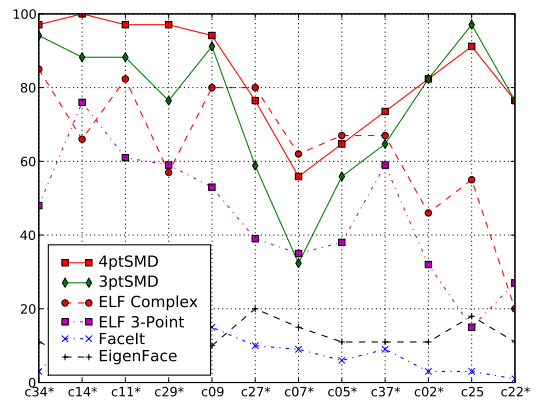
(a) Gallery Pose c27



(b) Gallery Pose c22



(c) Gallery Pose c37



(d) Gallery Pose c31

Figure 3.5: Cross-sections with fixed gallery pose for the results presented in Table 3.3. Probe poses marked with \* have a vertical misalignment of about 10 degrees with the corresponding gallery pose.

Table 3.3: A comparison of our stereo matching distance with other methods across pose.

**34 Faces**

Method	Accuracy
Eigenfaces [18, 20]	16.6%
FaceIt [18, 20]	24.3%
Eigen light-fields (3-point norm.) [18, 20]	52.5%
Eigen light-fields (Multi-point norm.) [18, 20]	66.3%
<b>3-point Stereo Matching Distance</b>	<b>79.8%</b>
<b>4-point Stereo Matching Distance</b>	<b>86.8%</b>

**68 Faces**

Method	Accuracy
LiST (Romdhani et al. [37])	74.3%
<b>3-point Stereo Matching Distance</b>	<b>74.5%</b>
<b>4-point Stereo Matching Distance</b>	<b>82.4%</b>

Table 3.4: Comparisons over a slice of the data with the method of Chai et al. [8] and Gross et al. [20]. The gallery pose is c27 and contains 34 faces. The table layout is the same as the one presented in [8].

Probe Pose	Methods			
	<b>3ptSMD</b>	LLR-step5 with PCA+LDA	ELF (3-P Normalization)	ELF (Complex)
c05	<b>100%</b>	98.5%	88%	93%
c29	<b>100%</b>	100%	86%	91%
c37	<b>100%</b>	82.4%	74%	89%
c11	<b>97%</b>	89.7%	76%	78%
c07	<b>100%</b>	98.5%	100%	95%
c09	<b>100%</b>	98.5%	87%	93%
Mean	<b>99.5%</b>	94%	85.1%	89.8%

Table 3.5: Confusion matrix for pose variation for 68 faces with 3ptSMD. The diagonals are not included in any average. The table layout is the same as [37] and [21]. The global average for this table is 74.5%.

azimuth altitude Probe Pose	-66 3 c34	-47 13 c31	-46 2 c14	-32 2 c11	-17 2 c29	0 15 c09	0 2 c27	0 1.9 c07	16 2 c05	31 2 c37	44 2 c25	44 13 c02	62 3 c22	avg
Gallery Pose														
c34	-	79	85	74	59	29	37	15	47	51	49	60	54	53
c31	84	-	81	68	60	78	44	22	43	54	90	68	65	62
c14	96	85	-	100	100	76	93	60	79	82	56	82	50	80
c11	94	90	100	-	100	88	100	90	90	96	51	90	53	86
c29	88	79	100	100	-	99	100	97	100	96	54	90	50	87
c09	44	96	72	88	97	-	97	76	96	91	93	79	66	82
c27	60	62	93	100	100	100	-	97	100	100	62	97	46	84
c07	25	34	72	87	96	76	97	-	97	85	31	62	16	64
c05	60	60	90	91	100	97	100	96	-	100	76	100	63	86
c37	74	69	93	97	94	91	100	84	99	-	88	100	79	88
c25	44	93	35	40	41	85	40	16	66	79	-	72	88	58
c02	75	74	87	88	76	68	94	60	96	99	85	-	88	82
c22	56	62	47	43	37	53	32	13	41	56	87	69	-	49

c11, c07 and c09. Note that this is a slice of data from Table 3.1. Our 3ptSMD method produces nearly perfect results in these conditions, results that are much better than those reported in Chai et al.

### 3.4.2 PIE Pose Variation: 68 Faces

We also compared our results with the ones presented in Romdhani et al. [37]. These results are, to our knowledge, the best reported on the whole PIE database for pose variation. In this work all 68 images were used, so for this part we report our results using all 68 faces. Table 3.3 summarizes the results of this experiment.

The global average for the method of Romdhani et al. [37] is 74.3%, the global average for our 3ptSMD method is about the same, at 74.5%. For the subset of poses in which the epipolar lines fit perfectly our average performance is 80.8%, while theirs is 71.6%. We consider the case where all epipolar lines fit to be the best possible scenario for the 3ptSMD. When the epipolar lines are misaligned the average for 3ptSMD is 69.2%. Our 4ptSMD achieves overall accuracy of 82.4%, which is considerably higher than the performance of Romdhani et al. Our method runs about 40 times faster than the method presented in [37], requires fewer manually specified points, and is much simpler. Detailed results are presented in Tables 3.5 and 3.6.

### 3.4.3 PIE Pose and Illumination Variation

We also evaluated the performance of the method across pose and illumination. Although our method is not designed to handle lighting variation, the use of normalized correlation in matching

Table 3.6: Confusion matrix for pose variation for 68 faces with 4ptSMD. The diagonals are not included in any average. The table layout is the same as [37] and [21]. The global average for this table is 82.4%

azimuth altitude Probe Pose	-66 3 c34	-47 13 c31	-46 2 c14	-32 2 c11	-17 2 c29	0 15 c09	0 2 c27	0 1.9 c07	16 2 c05	31 2 c37	44 2 c25	44 13 c02	62 3 c22	avg
Gallery Pose														
c34	-	79	91	78	65	38	44	26	50	50	60	71	56	59
c31	91	-	99	96	94	78	65	50	62	65	84	72	60	76
c14	97	100	-	97	91	87	79	71	79	76	59	76	78	82
c11	94	97	99	-	100	97	94	94	88	94	79	87	65	90
c29	87	97	96	100	-	100	99	100	96	94	82	81	53	90
c09	54	91	84	99	100	-	100	97	94	94	85	90	65	87
c27	60	93	91	97	99	99	-	100	97	99	97	97	62	90
c07	40	62	79	97	100	96	100	-	100	99	88	97	32	82
c05	71	79	90	93	97	97	99	100	-	100	100	99	78	91
c37	66	74	85	94	90	91	97	99	100	-	100	100	91	90
c25	65	79	56	66	71	85	91	79	97	100	-	99	94	81
c02	81	71	74	81	69	93	90	85	93	100	99	-	99	86
c22	57	62	66	56	44	49	47	35	66	76	88	91	-	61

Table 3.7: Summary of the cases where the camera movement is horizontal and when it is not over the experiments with 3ptSMD and 4ptSMD.

Method	# Faces	Epipolar Alignment	Epipolar Misalignment	Average
3ptSMD	34	84.8%	75.6%	79.8%
3ptSMD	68	80.8%	69.2%	74.5%
4ptSMD	34	87.2%	86.5%	86.8%
4ptSMD	68	82.6%	82.3%	82.4%

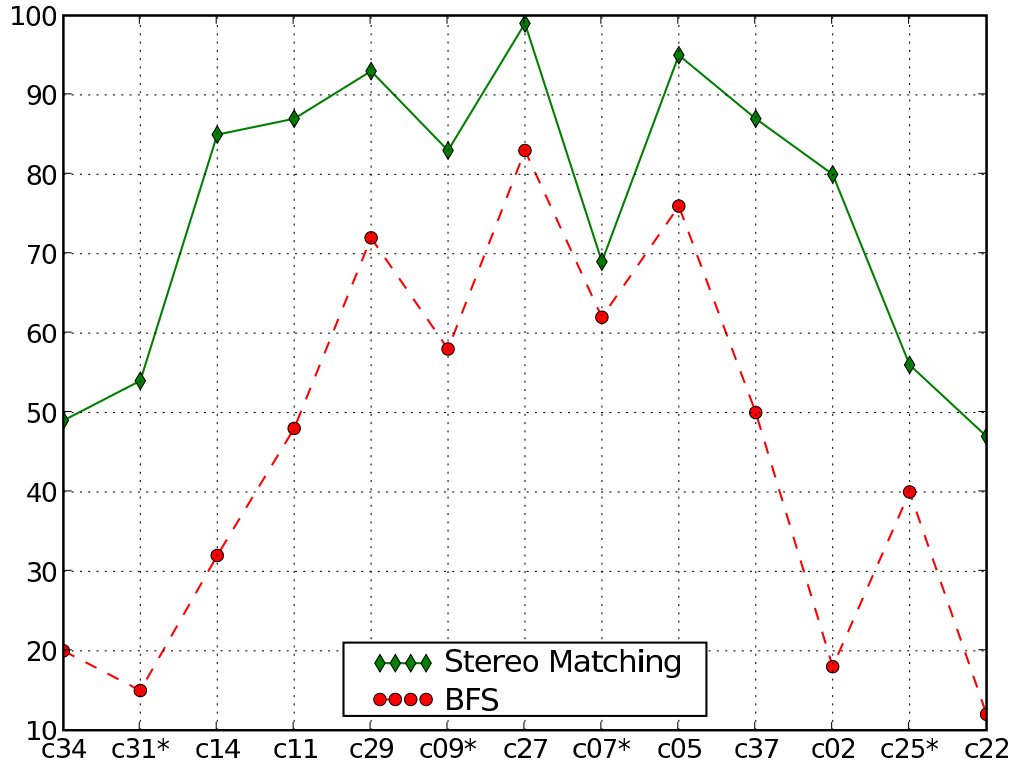


Figure 3.6: (A comparison of our method with BFS. Gallery pose is frontal (c27) probe poses are as indicated in the x axis, we report the average over the 21 illuminations.

may provide some robustness to lighting changes. The objective of this experiment is to verify that the good performance obtained when there is variation in pose (the previous experiments) are not an artifact of the (constant) illumination condition, and that the system degrades gracefully with lighting changes.

In this section we compare our method to Bayesian Face Subregions (BFS) [18] in the case of simultaneous variation of pose and illumination. For this experiment, the gallery is frontal pose and illumination. For each probe pose, the accuracy is determined by averaging the results for all 21 different illumination conditions. The results of this comparison are presented in Figure 3.6. We observe that our algorithm strictly dominates BFS over all probe poses.

For lighting invariance they use [19] which computes the reflectance and illumination fields from real images using some simplifications, while we simply use an approximation to normalized correlation.

We also performed experiments in such a way that we can compare with [5] and [37]. For this experiment we used images of the faces of 68 individuals viewed from 3 poses (front: c27, side: c5 and profile: c22) and illuminated from 21 different directions. We used light number 12 for the

Table 3.8: Accuracy percentage with pose and illumination variation. The cell format is: (with ambient lights)/(without ambient lights). Three galleries and three probes were used. F: Frontal, S: Side, P: Profile. The table layout is similar to [37]

light	F Gallery			S Gallery			P Gallery		
	F	S	P	F	S	P	F	S	P
2	94/38	93/44	32/4	85/41	100/53	41/6	26/21	18/25	100/47
3	96/68	96/76	35/13	93/65	100/85	41/9	29/25	16/25	100/51
4	97/82	94/87	37/24	96/82	100/94	35/12	34/25	24/31	100/66
5	99/100	99/97	35/34	99/97	100/100	47/32	38/35	25/29	100/94
6	100/99	100/99	41/35	100/99	100/100	57/56	38/29	43/24	100/100
7	100/99	99/97	37/34	99/87	100/100	53/49	29/21	35/16	100/100
8	100/100	100/100	44/37	100/100	100/100	56/60	35/19	43/25	100/100
9	100/100	100/100	44/44	100/100	100/100	65/62	40/35	47/46	100/100
10	99/90	99/93	29/34	99/88	100/99	49/35	32/28	28/21	100/87
11	100/100	100/100	46/44	100/100	100/100	60/56	47/32	49/35	100/100
12	-/-	100/100	53/44	100/100	-/-	71/62	49/46	56/53	-/-
13	100/100	100/100	46/41	100/100	100/100	63/49	44/43	49/49	100/100
14	100/100	100/100	47/43	100/100	100/100	66/49	44/46	59/53	100/100
15	100/100	99/94	46/31	100/100	100/100	54/40	37/46	60/54	100/100
16	100/100	97/74	40/21	100/97	100/99	51/32	40/41	53/47	100/100
17	100/90	96/49	35/19	99/75	100/84	49/26	32/41	44/47	100/100
18	99/91	99/97	37/28	99/90	100/97	38/25	35/37	22/32	100/79
19	100/100	100/99	38/29	100/99	100/100	54/38	43/35	44/32	100/99
20	100/100	100/100	44/38	100/100	100/100	63/51	49/41	51/40	100/100
21	100/100	100/100	50/40	100/100	100/100	65/54	47/47	57/53	100/100
22	100/100	100/99	50/37	100/100	100/100	57/40	38/46	60/54	100/100
avg	99/92	98/90	41/32	98/91	100/95	54/40	38/35	42/37	100/91

gallery illumination to be able to compare our results with [37]. They select that lighting because “...the fitting is generally fair at that condition”. Our results are presented in Table 3.8. We do not expect our results to be as good as those of [37], because our algorithm only accounts for lighting variation by using a fast approximation to normalized cross correlation as described in Criminisi et al. [13], while [37] has a 3-D model and performs an optimization to solve for the lighting that best matches the model to the image. We also tested on without ambient lights part of the PIE dataset which has harsh shadows. Other works have not reported results without ambient lights.

Our stereo matching method degenerates into an approximation to normalized correlation over small windows when there is no change in pose. Our method performs better than Romdhani et al. [37] when there is no pose change (gallery probe combinations: F-F, S-S and P-P). It is surprising that our method works better than theirs in this case because we are using a simple illumination insensitive image comparison technique and they perform an optimization to solve for lighting. Overall, for this experiment our global average is 74.6% while the global average of Romdhani et al. [37] is 81%, which is considerably better.

## 3.5 Conclusion

We have presented a simple, general method for face recognition with pose variation that is based on stereo matching. Our approach is motivated by the observation that correspondence is critical for face recognition across pose. Finding correspondences in 2-D is exactly the problem that stereo matching solves. We use stereo matching for face recognition across pose and show that this method exhibits excellent performance when compared to existing methods.

Our method is very simple. The formulation itself is straight-forward yet it is based on a very well-understood problem (stereo matching). The implementation can be done in C in a couple hundred lines of code.

The method we presented also degrades gracefully in the case of simultaneous variation of pose and illumination. Although our method is not really meant to handle lighting variation, since it uses normalized correlation it is somewhat robust to changes in illumination.

We evaluated our method using the CMU PIE dataset under a wide variety of conditions. Our results show that with pose variation and constant illumination our method is much more accurate than the methods of Gross, et al. [20], Chai et al. [8] and Romdhani et al. [37]. Additionally, our method is robust to some variation in lighting.

We feel that the main difference between our method and prior approaches is the use of stereo matching to find correspondences. Our method compares corresponding pixels very simply, using normalized correlation; this is a much more naive comparison than in many prior approaches. Therefore, we feel that the main reason for the superior experimental performance of our system lies in our emphasis on comparing images based on these correspondences.

## Chapter 4

# Research Plan

In this section I present my research plan. I will present a description of the problems I plan to study in the rest of my dissertation work. The general plan is the following:

- Construct a method for stereo matching in the presence of illumination change. In particular a method that is robust to changes in viewpoint and illumination when matching very slanted objects.
- Integrate large-scale spatial learning into our face recognition method. This involves solving two distinct alignment issues for pose invariant description of image differences. This will allow us to learn how to compensate for variations in the images that are not being explicitly accounted for by the (pose+illumination) model.
- Evaluate the use of our methods both in real-world unconstrained settings and in the context of facial variations that have not been previously studied like face recognition with weight change.

In the following two sections I will describe the general ideas behind each research direction and my detailed plans for each.

### 4.1 Stereo Matching with Illumination Variation

Finding reliable correspondences in two images of a scene taken from arbitrary viewpoints with possibly different cameras and in different illumination conditions is a difficult and key problem in computer vision.

In image matching methods, lighting variation is typically handled by methods invariant to additive and multiplicative changes in image intensity [35]. This is commonly referred to as contrast invariance. There are several well understood methods to compare images in a way that is robust to additive and multiplicative changes in image intensity: comparing the direction of gradient, computing the normalized correlation over small windows, and jets of oriented filters such as Gabor jets or Gaussians. See [35] for an excellent review.

Historically the main use of stereo matching has been for 3-D reconstruction. We have shown [7] that stereo matching also provides an effective method for image comparison for face recognition across pose. That is, we have shown that stereo matching provides a sensible way to compare facial images in a way that is very robust to viewpoint changes. But our work so far is limited to pose variation under constant illumination conditions.

The objective of this part of the work is to build a stereo matching method that is both locally contrast invariant and robust to viewpoint changes. Image comparison methods that are very robust to illumination change (such as direction of gradient [11] and normalized correlation [35]) are very fragile with respect to viewpoint changes. On the other hand pixel comparison methods that are robust to viewpoint change (such as SSD [12] or the method of Birchfield and Tomassi [4]) are very fragile with respect to illumination variation.

Most of the work done on stereo matching assumes an image of the same scene taken at the same instant of time. We would like to study the problem of stereo matching in the presence of illumination change, these conditions imply that the images weren't taken at the same instant of time. Many methods (see Ogale and Aloimonos [34], for example) have provisions to handle small variations in illumination to compensate for photometric issues. On the other hand, we're interested in dense stereo with major changes illumination and viewpoint and the interaction of changes in illumination and viewpoint when matching very slanted surfaces.

There is a body of work that tries to integrate photometric stereo with binocular stereo (for example Moses and Shimshoni [31], Lim et al. [24], Simakov et al. [40]). These works concern themselves primarily with reconstruction and their baselines are small. There is also some work done on wide baseline stereo that is somewhat robust to lighting and deformation but the algorithms don't provide dense correspondences and therefore are not fit for our purposes, see Matas et al. [28].

We are interested in the problem from a recognition point of view – that is to compute useful discriminative costs between classes of objects. But this is not an issue in terms of the stereo setup: we're interested in computing a dense matching in different illumination conditions. We feel that the task of comparing two images is much simpler than the task of reconstructing a 3-D scene, however a dense stereo matching method that is robust to illumination variation is useful for both tasks. The problem we are interested in has several important characteristics:

1. We require dense correspondences. Every pixel in each image should be accounted for. We assume that the problem of motion estimation has been solved for our purposes.
2. We require effective handling for wide baselines. Wide baselines have important consequences on the treatment of slant.
3. The images were not taken at the same instant of time. There is illumination variation and perhaps even some deformation.

There is a critical implication of the presence of slanted surfaces in stereo matching [33]. To illustrate this concept consider Figure 4.1. A detail of this figure is presented in Figure 4.2. Let point A

have coordinates  $(X_A, Z_A)$  with respect to camera 1. Similarly, let the point  $B$  have coordinates  $(X_B, Z_B)$ . Let the cameras be separated by a translation  $t$ . Therefore:

$$L_1 = X_B/Z_B - X_A/Z_A \quad (4.1)$$

$$L_2 = (X_B - t)/Z_B - (X_A - t)/Z_A \quad (4.2)$$

Therefore except in the case of no movement ( $t = 0$ ) or frontoparallel surface ( $Z_A = Z_B$ ) slanted planes will always project onto segments of different length. Similarly the appropriate window width should change (points  $a$  and  $b$  in Figure 4.2):

$$w_1 = X_b/Z_b - X_a/Z_a \quad (4.3)$$

$$w_2 = (X_a - t)/Z_a - (X_a - t)/Z_a \quad (4.4)$$

However we will assume that the width of the window is small compared to the distance from the camera to the matching window. Even when using an algorithm that can match  $N$  pixels in one scanline to  $M$  pixels in another scanline algorithms similar in style to Criminisi et al. [13] when matching surfaces of different slant the size of the matching window has to be adjusted. An illustration of this situation is presented in Figure 4.2. We've defined the slant as the angle normal to the window in the direction of the principal point of the camera (see Figure 4.2) these angles are called  $\alpha_1$  and  $\alpha_2$ . Elementary trigonometry gives us that:

$$\cos(\alpha_1) = \frac{w_1}{w_f} \quad \text{and} \quad \cos(\alpha_2) = \frac{w_2}{w_f} \quad (4.5)$$

where  $w_f$  is a fixed width, the one used if the surface was fronto-parallel and additionally, we know that relative slant is continuously the derivative of the disparity or discretely its first difference:

$$\tan(\alpha_1) - \tan(\alpha_2) = \Delta d \quad (4.6)$$

which makes for a fairly constrained search problem.

The importance of correctly handling slanted surfaces becomes evident in wide baseline stereo. The problem is exacerbated when there is variation in illumination, in this case, the property of local contrast invariance is of fundamental importance. It is very difficult to provide contrast invariance in stereo matching without estimating the slant. These two conditions commonly occur in face recognition across pose and illumination. But face recognition across pose and illumination is not the only problem that would benefit from a deeper understanding of the fundamental issues behind performing stereo matching when there is illumination change.

We still need to gain a lot of insight into this theoretical problem but it is undeniable that it falls at the core of several important applications.

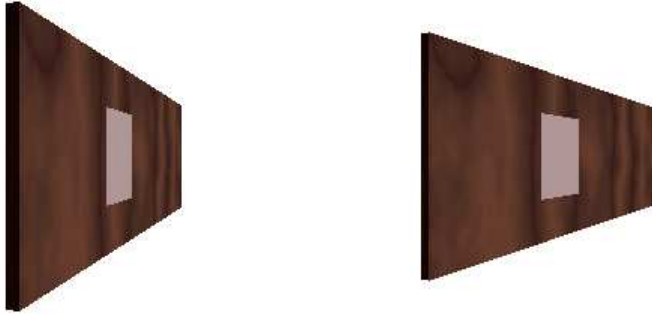


Figure 4.1: A wooden wall with a seen from two viewpoints. This is our example to illustrate the critical importance of handling slant correctly.

#### 4.1.1 Plan

- Gain a theoretical understanding on the fundamental issues present when there is simultaneous variation in viewpoint and illumination in stereo matching.
- Construct a dense stereo matching method that is both locally contrast invariant and robust to viewpoint changes. This entails being able to estimate the slant from the disparity information on the fly (in one pass) and being able to correct the (local) image comparison for the interaction of illumination change and scene structure (slant).
- Test on ray traced images, then on a dataset where there is control over the illumination conditions such as CMU PIE and Multipie, and then on datasets like Labeled Faces in the Wild [23].

## 4.2 Image Representation, Correspondences and Learning for Unconstrained Face Recognition

Image representation is one of the central topics in computer vision. An ideal image representation would be robust to a wide range of variations making image comparisons straight forward and making the recognition of objects of a class of interest trivial. Many image representations have been proposed [29].

Several methods have been proposed, that based on a representation learn the spatial weights of different regions of the face:

$$\text{similarity}(x_1, x_2) = \vec{w}^T(\text{difference}(x_1, x_2)) \quad (4.7)$$

where  $x_1$  and  $x_2$  are images,  $\vec{w}$  is called a weight vector and  $\text{difference}(x_1, x_2)$  is called a difference descriptor. For example if  $\vec{w}$  is a vector of ones and difference is the component-wise squared difference, then the similarity measure becomes SSD. Several representations for  $x_1$  and  $x_2$ , several

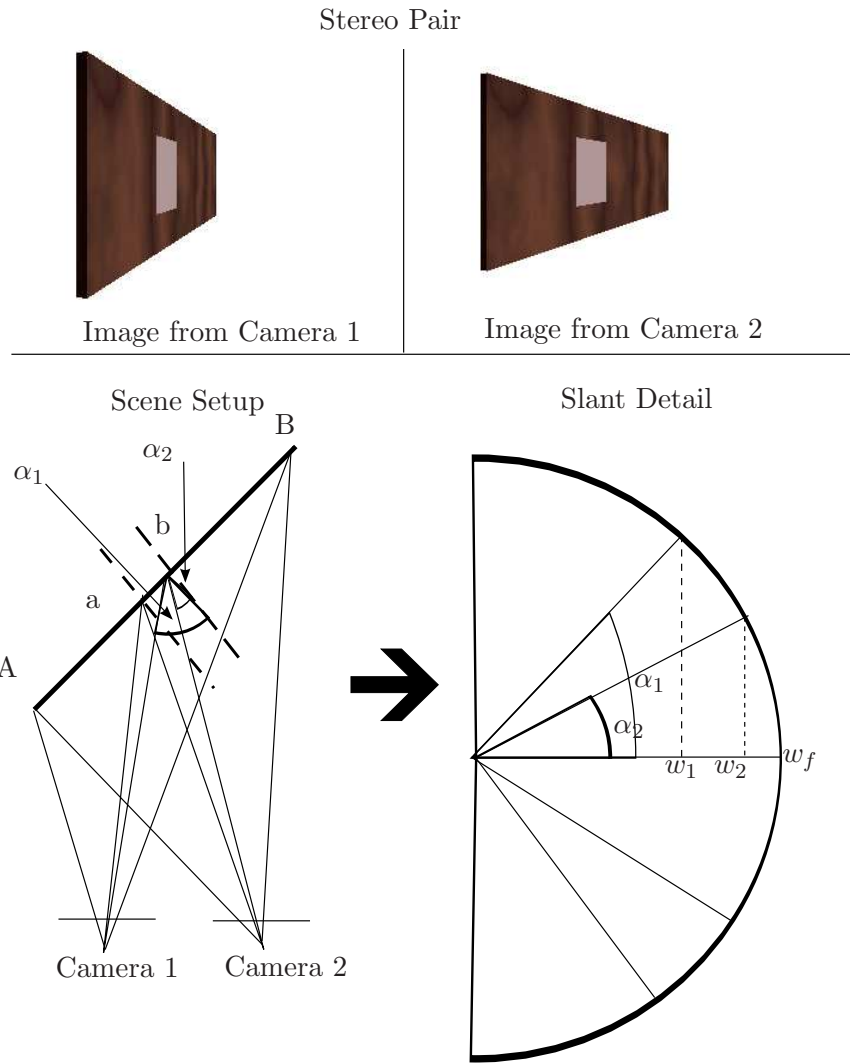


Figure 4.2: Relation of slant with window size as seen from above. It is clear that longer window is needed on the image captured by camera 2 (right) than the image captured by camera 1 (left). In general as the slant increases the matching window has to be made smaller.

ways of learning  $\vec{w}$  and several ways for computing the difference given  $x_1$  and  $x_2$  have been proposed. These methods have proven useful in face recognition for compensating for several types of variations such as: lighting change and aging, see for example Ling et al. [25].

Methods based on learning the spatial weights of features for face recognition are a useful technique [30], [45], [36], [25]. These type of methods have not been explored when there is variations in pose. We have also shown that using correspondences is useful for face recognition across pose [7]. The main objective of this part of the work is to understand how to compare images based on labeled evidence when there is pose variation. To effectively do so, we need a deeper understanding of the interaction of image representation, the representation of correspondences/deformation and spatial learning.

We would like to explore how to represent  $x_1, x_2$  (image representation) and how to compute the descriptor difference( $x_1, x_2$ ) (descriptor generation) when there is pose variation. The key is not the learning machinery itself, the key is, we believe, the representation of the correspondences, the occlusions, the epipolar geometry, the similarity function and the image description.

### 4.2.1 Image Representation

We're interested in studying the image representation for spatial learning purposes. Multiple methods have been proposed to describe images in a way robust to deformation [29]. Several advances have been made in studying image representations for spatial learning purposes [30], [36], [25].

It is expected that a method that is globally viewpoint invariant, locally contrast invariant and robust to deformation should be based on such an image description. The image description methods we're interested in evaluating are the SIFT-like [26] DHOG [42] descriptor (dense histogram of gradient orientations) and MSER (Maximally Stable Extremal Regions) [28].

The MSER image representation has shown promise in wide baseline stereo (though not dense) and SIFT and SIFT-like image descriptions have been shown to be very effective in describing image regions for correspondence.

We have evaluated describing facial images using the SIFT-like [26] DHOG descriptor from the VLFeat library. The scanlines are compared using our SMD method [7] using this DHOG description. Our preliminary results show that in unconstrained environments the DHOG descriptor is more accurate than the windowed NSSD image comparison metric, but NSSD and DHOG perform equally well in the more controlled PIE dataset.

### 4.2.2 Descriptor Generation

We are interested in studying how to describe the image differences when there is pose variation. Given two scanlines  $s_1$  and  $s_2$  of length  $l_1$  and  $l_2$  respectively the stereo method computes an optimal matching  $M$  which is a sequence (a word) of length  $l_1 + l_2$ . The optimal matching will be a sequence of symbols in the alphabet:  $\Sigma = \{C_{Lo}, C_{Lm}, C_{Ro}, C_{Rm}\}$ . Each symbol accounts for one pixel in either  $s_1$  or  $s_2$ . This means that we can associate a cost with each pixel in each image being compared. The costs are associated according to the Table 4.1.

Table 4.1: Decoding of a matching  $W = \langle c_1, \dots, c_n \rangle$  into two descriptors  $D_1$  and  $D_2$  of the same length of the scanlines matched.

$c_k$	$c_{k-1}$	$D_{1,i}$	$D_{2,j}$
$C_{Lo}$	$C_{Lo}$	$\alpha$	-
$C_{Lo}$	$C_{Lm}$	$\beta$	-
$C_{Lo}$	$C_{Rm}$	$\beta$	-
$C_{Ro}$	$C_{Lm}$	-	$\beta$
$C_{Ro}$	$C_{Ro}$	-	$\alpha$
$C_{Ro}$	$C_{Rm}$	-	$\beta$
$C_{Lm}$	$C_{Lo}$	$\beta' + M(i, j)$	-
$C_{Lm}$	$C_{Lm}$	$\gamma + M(i, j)$	-
$C_{Lm}$	$C_{Ro}$	$\beta' + M(i, j)$	-
$C_{Lm}$	$C_{Rm}$	$M(i, j)$	-
$C_{Rm}$	$C_{Lo}$	-	$\beta' + M(i, j)$
$C_{Rm}$	$C_{Lm}$	-	$M(i, j)$
$C_{Rm}$	$C_{Ro}$	-	$\beta' + M(i, j)$
$C_{Rm}$	$C_{Rm}$	-	$\gamma + M(i, j)$

Note that due to the effects of the rectification procedure the image is a quadrilateral, not necessarily a square. Once the matching has been decoded, the matching cost arrays (images) can be made an axis aligned image (of the same size as the input image) using the inverse of the rectification transformation. After performing the transformation, the cost array is a square of the same size as the original image being matched. We call this procedure back-projection. Back projection is performed on both input images.

When there is variation in pose, it becomes clear that there are two distinct usages of the alignment:

1. Alignment for image comparison purposes: for this purpose we use epipolar geometry and stereo matching. Here point features need to be aligned in such a way that they fall on corresponding epipolar lines.
2. Alignment for spatial learning purposes: this is needed because we need to generate feature vectors (descriptors) with features that are in correspondence. Here, the images need to be put in the same reference frame.

The alignment for image comparison is obtained by computing the epipolar geometry. In our preliminary work we have done this by using hand clicked points for initial experimental purposes. But we're interested in having an end-to-end system for face recognition, we therefore have also tried using an off-the-shelf egomotion estimation method. We use the egomotion estimation method of Domke and Aloimonos [15, 14]. We use this method to estimate the epipolar geometry of a pair of face images. It is unclear that this method would work on a pair of images that are of different individuals and are taken under different illumination conditions and different expressions, but our initial experiments show that it does.

The alignment for spatial learning purposes has to be done to a reference image in order for the keypoints to be put in the same reference frame. This can be done in two different ways:

1. Solving for an affine transformation of keypoints: a dense correspondence between the two images (one query image and the reference image) is computed. Using the dense correspondence of a subset of points an affine correspondence can be solved for.
2. Transferring alignment: a dense correspondence between the two images (one query image and the reference image is computed). Using the dense correspondence the costs of matching the query image are transferred to the reference image.

This formulation using a stereo matcher, egomotion, and spatial learning has the inherent benefit of not requiring hand-clicked points. Hand-clicked points are pervasive in face recognition. However, the more unconstrained the imaging conditions are, the more unreasonable the assumption that an off-the-shelf fiducial point detector would be able to detect accurate points.

### 4.2.3 Plan

- Construct an image description based on the image comparison method that is simultaneously robust to pose and illumination changes and encodes correspondence (disparity) information and on top of which learning can be performed.
- Construct an image registration method that is robust to pose variations for spatial learning purposes that doesn't require hand clicked points.
- Evaluate how to describe the image in such a way that there is robustness to deformation. There are several image descriptions that are robust to deformations, for example the SIFT-like[26] DHOG descriptor and the MSER descriptor.
- Evaluate how to effectively perform large-scale, spatial learning on these descriptors.
- Test on face recognition datasets in which there is a wide range of variations such as Labeled Faces in the Wild [23] and our own Weight Variation Dataset.

## 4.3 Face Recognition with Weight Variation

We are also interested in considering another such type of face variation, changes in weight. Many applications, such as passport photo verification, police investigations, or the sorting of personal photographs require that we recognize an individual in photos taken months or years apart, in which the subject's weight may change considerably. Yet there has been no study of the effect that weight change has on the accuracy of recognition algorithms. We are interested in addressing this problem for the first time.

We have collected our own dataset of images with weight variation. In order to minimize the amount of time between photos. Some images were obtained from weight loss forums and personal

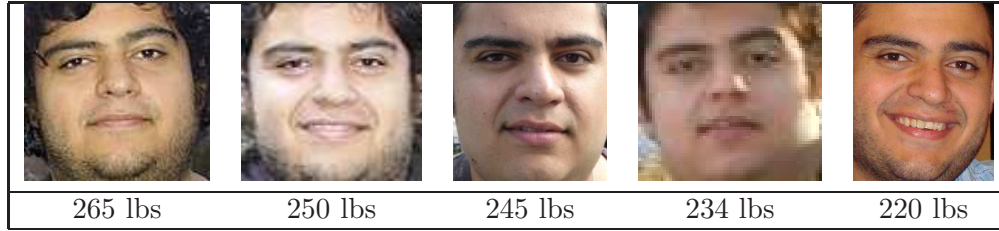


Figure 4.3: Facial changes as weight variation increases (images shown with permission of subject).

photo documentaries. Other images came from the TV show *The Biggest Loser*. Figure 4.3 shows an example of an individual’s weight variation.

We performed some preliminary experiments to get a sense of the data set. In our experiments, we have found that performance of existing algorithms degrades markedly as the amount of weight change increases. This suggests that weight change alone can have a large effect on recognition performance.

Third, we find that there are large differences in the relative performance of different algorithms as the amount of weight change varies. In particular, we find that of the recognition algorithms tested, the most robust performance is obtained by algorithms that stress finding correspondences.

### 4.3.1 Experimental Evaluation

We used the following algorithms:

- **NC:** Normalized correlation.
- **Window-NSSD:** Window NSSD with clipping.
- **FPLBP:** Four patch local binary pattern as described in Wolf et al. [44] using code provided by its authors and default parameters and SSD of the descriptors of the two images. The images are filtered using a non-linear noise removal method (`wiener2` in MATLAB).
- **SMD-0d:** Stereo Matching Distance at zero-disparity, to evaluate the benefit of having a structured occlusion cost but no non-trivial correspondences.
- **SMD:** Stereo Matching Distance by Castillo and Jacobs [7].
- **SVM-diff:** SVM trained on “differences” of face images normalized to zero mean and unit variance [36]. The  $\gamma$  and  $C$  parameter are evaluated by 5-fold cross validation on the training set on a grid of options for  $(\gamma, C)$  [9].
- **SVM-GO:** SVM trained on gradient orientation “differences” [25]. The  $\gamma$  and  $C$  parameter are evaluated by 5-fold cross validation on the training set on a grid of options for  $(\gamma, C)$  [9].
- **LBP-SVM:** An SVM is trained to integrate several LBP-based distance measures from FPLBP and TPLBP. The LBP descriptors are computed using code publicly available from Wolf et al. [44]. The images are filtered using a non-linear noise removal method (`wiener2` in MATLAB).
- **ERCF:** The images are classified using ERCF. The costs are computed using the Linux

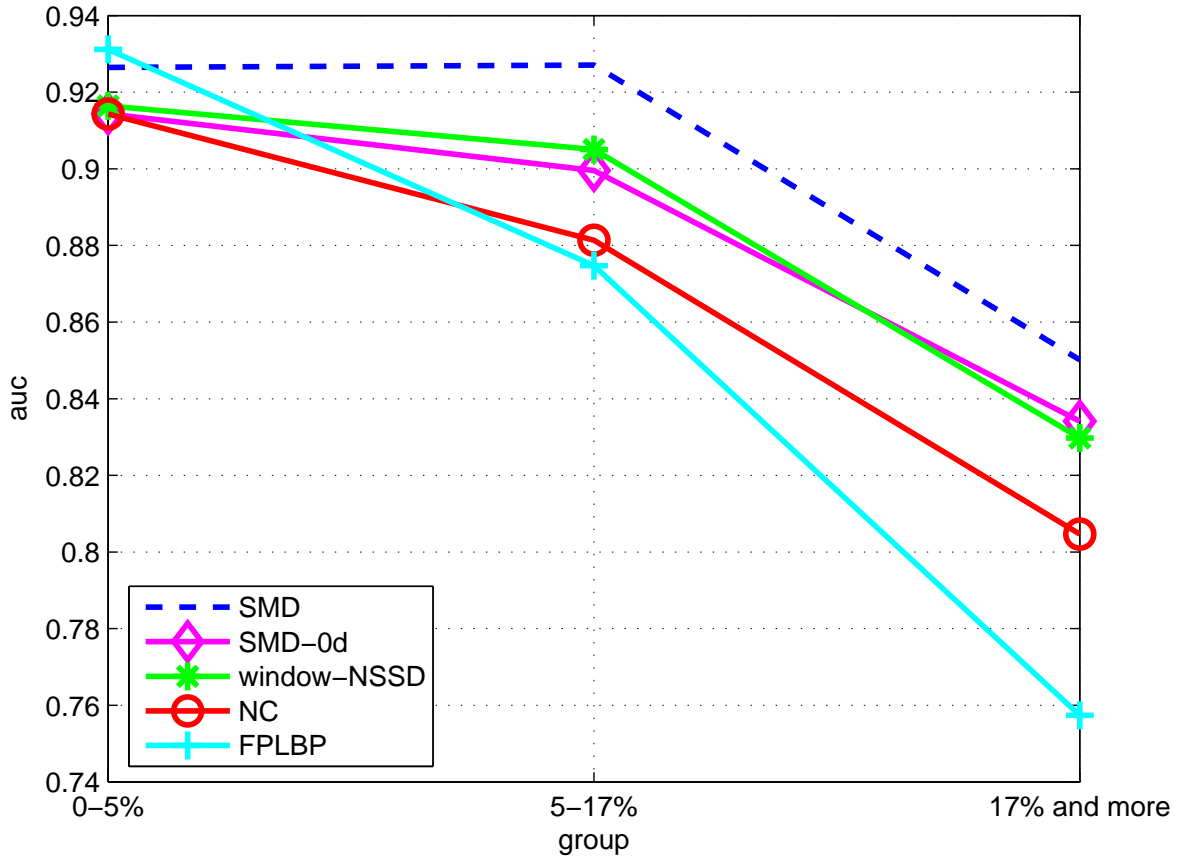


Figure 4.4: Performance of classifiers by groups of similar relative weight variation.

binaries publicly available from Nowak and Jurie [32].

Figure 4.4 shows how the performance of non-learning algorithms varies with weight change. First, we can see that performance of all algorithms drops as the amount of weight change increases. The magnitude of these changes suggest that weight change plays a very significant role in the difficulty of this task.

We can also see that different algorithms display different levels of robustness to weight change. SMD [7] is best when there is larger weight variation, but not when the weight change is small. FPLBP descriptors work very well when there is little weight variation but the performance decreases dramatically even in the presence of moderate weight gain.

Figure 4.4 also shows that there is a very slight difference between the performance of the two occlusion methods (window-nssd and SMD-0d). The performance of SMD-0d is slightly more robust to weight variation than window-nssd.

Figure 4.5 shows an ROC curve of all the non-learning methods compared on the entire Web Forum Dataset. This figure shows that SMD clearly and uniformly performs best.

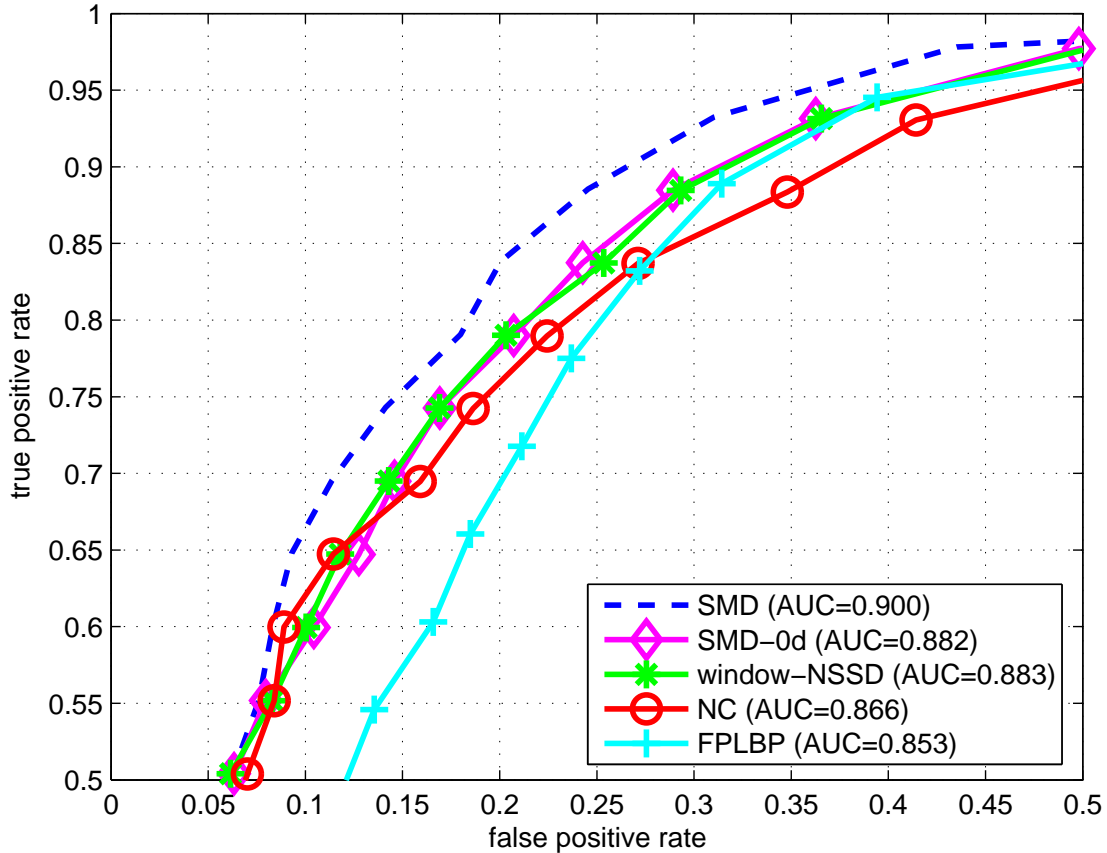


Figure 4.5: ROC curve comparing all the non-learning based methods.

From Figures 4.4 and 4.5 we observe that the two occlusion methods (window-nssd and SMD-0d) perform essentially equally well suggesting that it is not the treatment of occlusions but the ability to form correspondences with non-zero disparity that explains the difference between these two methods and SMD.

Figure 4.7 shows an ROC curve of all the methods that use learning. For this experiment we use half the dataset to train and half the dataset to test in a 2-fold cross-validation experiment. The curves presented are averages of each leg of the experiment. SMD (which was the best performing of the image matching methods) was also evaluated on the same testing set.

From Figure 4.7, we observe that ERCF and LBP-SVM perform best among the methods based on learning. The performance of SMD (which is not a learning based method) is better than the performance of the top two learning methods.

From Figures 4.6 and 4.7 we observe that the performance of LBP-SVM is globally very good but note that of all the evaluated methods the performance of LBP-SVM degrades the most as weight change increases, therefore, the method is the least robust to weight variation of all the

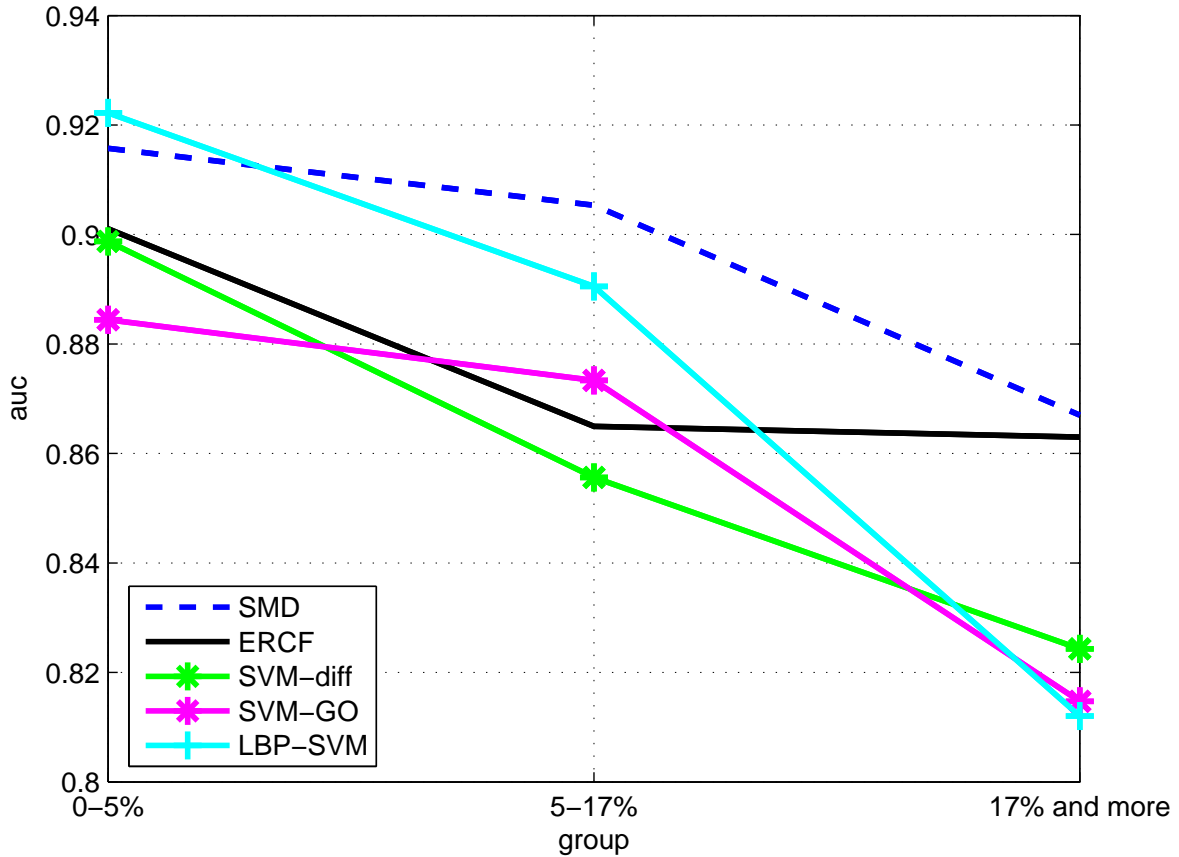


Figure 4.6: Performance of learning-based classifiers by groups of similar relative weight variation.

learning methods evaluated.

### 4.3.2 Discussion

First, results with all algorithms show that weight change can have a very significant effect on the accuracy of recognition algorithms. We consider group one to contain minor fluctuations in weight, zero to eight pounds for a 160 pound person. Group two contains weight changes that are commonly seen over a few years time, ranging from eight to twenty-seven pounds for a 160 pound person. Group three contains more extreme weight changes. Depending on the algorithm, the moderate weight changes in group two can account for an increase of between 10% and 50% in errors from group one to group two. The more extreme weight changes of group three create much more dramatic increases in error rates. This indicates that our dataset does indeed capture many of the special difficulties posed by weight change, and that weight change is an important challenge for face recognition algorithms.

Next we will discuss why some methods work better than others in the presence of weight

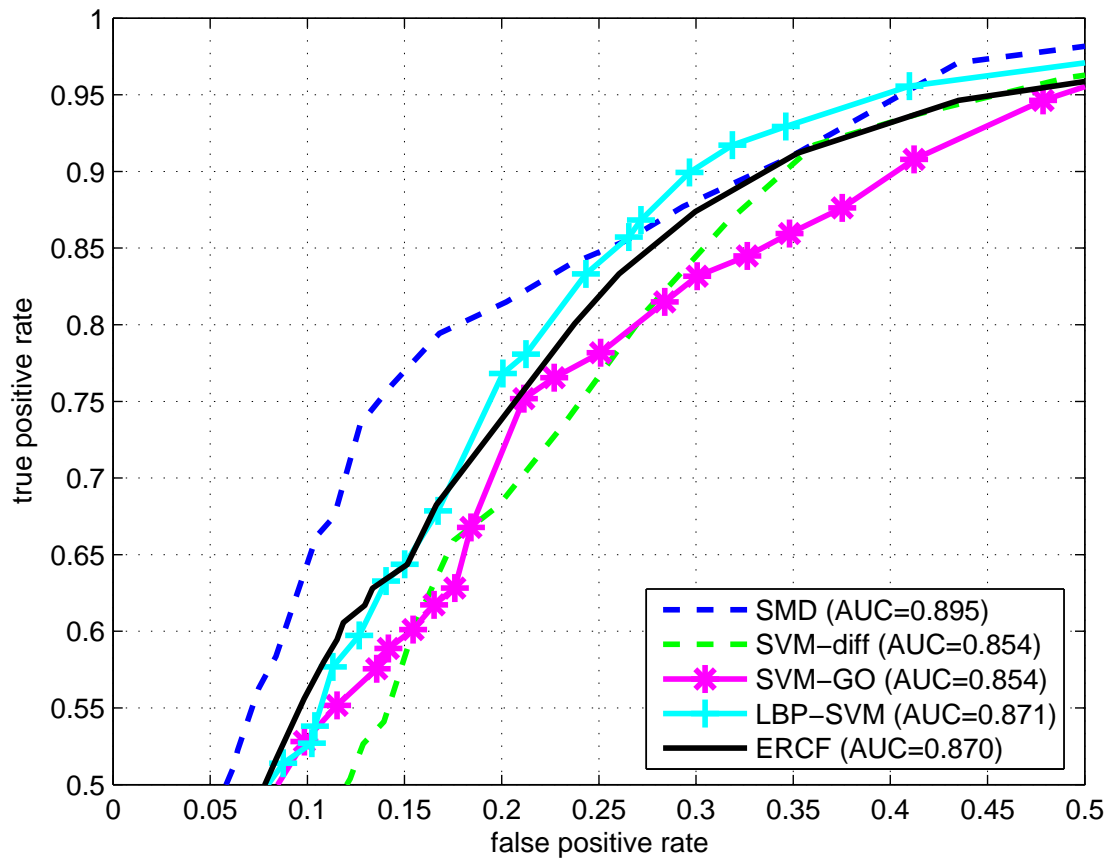


Figure 4.7: ROC curve comparing all the evaluated learning based methods. SMD was evaluated on the same testing set for comparison purposes.

variation. The top two methods (ERCF [32] and SMD [7]) have the common feature of finding non-trivial correspondences beyond those provided by alignment with a similarity transformation. While both methods do so taking very different approaches, experiments suggest that this results in better performance. These two methods do not explicitly account for weight variation but perform better in the presence of weight change, and are more robust to weight variation than other methods. For instance, methods based on local binary patterns (LBPs) perform remarkably well when there is little or no weight variation, but performance degrades rapidly when there is a large amount of weight variation.

The importance of correspondences is highlighted by the fact that SMD-0d is identical to SMD except that it only allows zero disparities. Therefore, the difference in performance between these two methods shows explicitly the importance of non-zero disparity correspondences.

Additionally the matching experiments show small yet significant difference in performance between normalized correlation and window-nssd with clipping and SMD-0d (the later two performing basically equally well). This illustrates the gain in accounting for occlusions. It is, however, unclear how such knowledge can be leveraged in a learning based method.

As there is limited data it is hard to draw definitive conclusions about learning algorithms. But we have verified that learning-based methods can perform well with this data. It is somewhat difficult to determine what constitutes a realistic scenario for learning algorithms when there is weight change. On one hand, it may be possible in the future to train learning systems with more data. On the other hand, our data mainly consists of image pairs with weight change. In many situations, learning based methods will be trained mostly using pairs that have limited weight change, which might hinder their ability to account for weight changes that do occur.

We have evaluated a variety of existing methods on our dataset. We have shown that weight variation is a significant confounding factor in face recognition and performance of all algorithms does, in fact, decrease as weight variations increase, suggesting that as face recognition methods move towards unconstrained settings weight variation needs to be accounted for.

Finally, our experiments also show that methods based on correspondences perform better as the weight variation increases. While not developed specifically for face recognition with weight variation the correspondence-based algorithms perform solidly and are quite robust to this variation.

### 4.3.3 Plan

- Construct a method for face recognition with weight gain. Apart from its practical importance, face recognition with weight variation is a problem that can be seen as a testbed for an algorithm that can compare images with deformation and illumination change. Having our own weight variation dataset and obtaining the self-consistent results we obtained in the preliminary evaluation is an important first step for this research direction.

## 4.4 Closing Remarks

The general direction for the rest of my dissertation work is going to be in the general area of methods for image comparison. I believe that direct image comparison is very powerful method for unconstrained face recognition. I believe fundamental issues in the area of image matching for face recognition require a deeper understanding. I consider that performing direct image comparison is a much simpler than reconstructing a scene (some type of scene reconstruction is present in many of the 3-D methods for face recognition across pose and illumination) while performing equally well.

I consider that as face recognition moves towards more unconstrained environments detecting high quality fiducial points is going to become increasingly infeasible, and therefore either algorithms for obtaining fiducial points need to advance significantly or, alternatively, good, useful face recognition methods will not require feature points and address the problems of registration and comparison end-to-end.

# Bibliography

- [1] Ahmed Bilal Ashraf, Simon Lucey, and Tsuhan Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008. 8
- [2] Ronen Basri and David Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233, 2003. 7
- [3] David Beymer and Tomaso Poggio. Face recognition from one example view. Technical Report AIM-1536, , 1995. 7
- [4] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(4):401–406, 1998. 32
- [5] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003. 7, 28
- [6] Kevin W. Bowyer, Kyong I. Chang, and Patrick J. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006. 8
- [7] Carlos D. Castillo and David W. Jacobs. Using stereo matching for 2-d face recognition across pose. In *CVPR*, 2007. 32, 36, 39, 40, 44
- [8] Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao. Locally linear regression for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 16(7):1716–1725, 2007. 3, 5, 8, 22, 25, 30
- [9] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 39
- [10] C.W. Chen and C.L. Huang. Human facial feature extraction for face interpretation and recognition. pages II:204–207, 1992. 15
- [11] Hansen F. Chen, Peter N. Belhumeur, and David W. Jacobs. In search of illumination invariants. In *CVPR*, pages 1254–1261, 2000. 32

- [12] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996. [18](#), [19](#), [32](#)
- [13] Antonio Criminisi, Andrew Blake, Carsten Rother, Jamie Shotton, and Philip H. S. Torr. Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming. *International Journal of Computer Vision*, 71(1):89–110, 2007. [18](#), [19](#), [20](#), [29](#), [33](#)
- [14] Justin Domke and Yiannis Aloimonos. A probabilistic framework for correspondence and egomotion. In *WDV*, pages 232–242, 2006. [37](#)
- [15] Justin Domke and Yiannis Aloimonos. A probabilistic notion of correspondence and the epipolar constraint. In *3DPVT*, pages 41–48, 2006. [37](#)
- [16] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001. [7](#)
- [17] Y. Gizatdinova and V. Surakka. Feature-based detection of facial landmarks from neutral and expressive facial images. 28(1):135–139, January 2006. [15](#)
- [18] Ralph Gross, Simon Baker, Iain Matthews, and Takeo Kanade. Face recognition across pose and illumination. In Stan Z. Li and Anil K. Jain, editors, *Handbook of Face Recognition*. Springer-Verlag, June 2004. [3](#), [5](#), [7](#), [22](#), [25](#), [28](#)
- [19] Ralph Gross and Vladimir Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Springer, June 2003. [28](#)
- [20] Ralph Gross, Iain Matthews, and Simon Baker. Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):449 – 465, April 2004. [3](#), [5](#), [22](#), [25](#), [30](#)
- [21] Ralph Gross, Jianbo Shi, and Jeffrey Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, December 2001. [23](#), [26](#), [27](#)
- [22] S.M. Hanif, L. Prevost, R. Belaroussi, and M. Milgram. Real-time facial feature localization by combining space displacement neural networks. 29(8):1094–1104, June 2008. [15](#)
- [23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV)*, 2008. [34](#), [38](#)
- [24] Jongwoo Lim, Jeffrey Ho, Ming-Hsuan Yang, and David J. Kriegman. Passive photometric stereo from motion. In *ICCV*, pages 1635–1642, 2005. [32](#)

- [25] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. A study of face recognition as people age. In *International Conference on Computer Vision (ICCV)*, 2007. [36](#), [39](#)
- [26] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. [36](#), [38](#)
- [27] Simon Lucey and Tsuhan Chen. A viewpoint invariant, sparsely registered, patch based, face verifier. *International Journal of Computer Vision (IJCV)*, December 2007. [8](#)
- [28] Jiri Matas, Ondrej Chum, Martin Urban, and Tomas Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002. [32](#), [36](#)
- [29] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. [34](#), [36](#)
- [30] Baback Moghaddam, Tony Jebara, and Alex Pentland. Bayesian modeling of facial similarity. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 910–916, Cambridge, MA, USA, 1999. MIT Press. [36](#)
- [31] Yael Moses and Ilan Shimshoni. 3d shape recovery of smooth surfaces: Dropping the fixed viewpoint assumption. In *ACCV (1)*, pages 429–438, 2006. [32](#)
- [32] Eric Nowak and Frederic Jurie. Learning visual similarity measures for comparing never seen objects. In *CVPR*, 2007. [40](#), [44](#)
- [33] Abhijit S. Ogale and Yiannis Aloimonos. Stereo correspondence with slanted surfaces: Critical implications of horizontal slant. In *CVPR (1)*, pages 568–573, 2004. [32](#)
- [34] Abhijit S. Ogale and Yiannis Aloimonos. Robust contrast invariant stereo correspondence. In *ICRA*, pages 819–824, 2005. [32](#)
- [35] Margarita Osadchy, David W. Jacobs, and Michael Lindenbaum. On the equivalence of common approaches to lighting insensitive recognition. In *ICCV*, pages 1721–1726, 2005. [31](#), [32](#)
- [36] P. Jonathon Phillips. Support vector machines applied to face recognition. In *Advances in Neural Information Processing Systems 11*, pages 803–809. MIT Press, 1998. [36](#), [39](#)
- [37] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *Computer Vision – ECCV’02*, volume 4, pages 3–19, Copenhagen, Denmark, 2002. [3](#), [5](#), [7](#), [22](#), [23](#), [25](#), [26](#), [27](#), [28](#), [29](#), [30](#)
- [38] P. Sankaran, S. Gundimada, R.C. Tompkins, and V.K. Asari. Pose angle determination by face, eyes and nose localization. pages III: 161–161, 2005. [15](#)
- [39] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003. [7](#), [21](#)

- [40] Denis Simakov, Darya Frolova, and Ronen Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *ICCV*, pages 1202–1209, 2003. 32
- [41] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. pages 586–591, 1991. 22
- [42] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/>, 2008. 36
- [43] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. In Gerald Sommer, Kostas Daniilidis, and Josef Pauli, editors, *Proc. 7th Intern. Conf. on Computer Analysis of Images and Patterns, CAIP'97, Kiel*, number 1296, pages 456–463, Heidelberg, 1997. Springer-Verlag. 6
- [44] Lior Wolf, Tal Hassner, and Yaniv Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV)*, 2008. 39
- [45] Yongbin Zhang and Aleix M. Martínez. Recognition of expression variant faces using weighted subspaces. In *ICPR (3)*, pages 149–152, 2004. 36
- [46] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, December 2003. 3, 6