Using Stereo Matching with General Epipolar Geometry for 2D Face Recognition across Pose

Carlos D. Castillo, *Student Member*, *IEEE*, and David W. Jacobs, *Member*, *IEEE*

Abstract—Face recognition across pose 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 seen from different poses. 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 pose 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 data set and demonstrates superior performance compared to existing methods in the presence of pose variation. It also shows robustness to lighting variation.

Index Terms—Face recognition, pose, stereo matching, epipolar geometry.

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 constant pose [30]. There are also several approaches to handling pose variation [24], [15], [17], [8]. However, there is still a lot of room for improvement. Progress would be important in many applications, for example, surveillance, security, and the analysis of personal photos and other domains in which we cannot control the position of subjects relative to the camera.

Correspondence seems crucial to producing 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, Fig. 1.

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.

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 show that the stereo matching cost is robust to pose variations. Consequently, we can use the stereo matching cost as a measure of similarity between two face images.

Note that we are not interested in performing 3D 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 3D 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).

We can therefore use 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. 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. We also tried obtaining the epipolar geometry from each pair of images using the method of Domke and Aloimonos [11], [12]. In this case, our method requires no hand-clicked points. We verified that there is no decrease in recognition performance in a fully automatic system.

Putting these steps together, we have the following 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 4. 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 data set and the Labeled Faces in the Wild (LFW) data set. Our results show that with pose variation at constant illumination our method is more accurate than previous methods of Gross et al. [17], Chai et al. [8], and Romdhani et al. [24]. 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. [15], which is designed to handle lighting variation, though it is not as accurate as the method of Romdhani et al. [24]. The experiments on the LFW data set show reasonable performance in an unconstrained setting (where there is simultaneous variation in pose, illumination, and expression).

This is an extended version of our conference paper [7]. The original conference version does not include our method with four feature points or experiments using four feature points, includes limited experiments with lighting change, and does not include the results on the LFW data set. Additionally, in the conference paper we did not develop a fully automatic system. However, the conference version of our paper includes an analysis of stereo matching for face recognition that has been eliminated from this version due to space constraints.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 discusses issues related to image alignment

- *The authors are with the Department of Computer Science, University of Maryland, 4420 A.V. Williams Bldg., College Park, MD 20742. E-mail: {carlos, djacobs}@cs.umd.edu.*

Manuscript received 28 Aug. 2008; revised 1 Apr. 2009; accepted 29 Apr. 2009; published online 21 May 2009.
Recommended for acceptance by M.-H. Yang.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-08-0575.
Digital Object Identifier no. 10.1109/TPAMI.2009.123.

Fig. 1. Example images from the CMU PIE data set. Observe that no linear transformation can make corresponding boxes have equal size because a linear transformation can only linearly scale their size.

and epipolar geometry, Section 4 presents the details of our face recognition method, and Section 5 presents and analyzes all experiments. Section 6 concludes.

2 Related Work

Zhao et al. [30] 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. Table 1 presents a summary of existing methods of face recognition across pose.

Some early approaches compensate for some 2D deformations in matching, which may partially account for the effects of pose. A notable example is the work of Wiskott et al. [29]. 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 [4]. In their work, they generated 2D 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 (see also [19] and [23]). Once the virtual view had been generated, the images were compared. Our method is similar to theirs in the sense that both are decidedly 2D 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 3D 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. [24] also focuses on 3D morphable models. In this work, shape and texture parameters of a 3D morphable model are recovered from a single image.

Basri and Jacobs [3] use a 3D 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 3D rigid transformation. Georghiades et al. [13] computed a 3D model 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 2D model for each one.

Gross et al. [15] presented two appearance-based algorithms for face recognition across pose and illumination. 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. The other method presented by Gross et al. [15] is called Bayesian Face Subregions (BFS). The algorithm models the appearance changes of the different face regions in a probabilistic framework.

TABLE 1
Existing Methods for Face Recognition across Pose

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

There have been several recent approaches to face recognition across pose that are based on patches. Chai et al. [8] presented a learning, patch-based rectification method based on locally linear regression. Lucey and Chen [22] present a patch-based algorithm for face recognition across pose of sparsely registered images (four manually selected points). Closely related, the work of Ashraf et al. [2] presents a new method to discover viewpoint-induced spatial deformations for general patch-based methods of face recognition across pose.

All of the methods previously mentioned in this section use intensity images of the face. This type of face recognition, based on 2D images constitutes the vast majority of face recognition research. There is, however, a significant amount of work done acquiring, matching, and performing recognition using 3D 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.

3 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, in which case the epipoles are at infinity and the epipolar lines are parallel in both images. 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. Even with scaled orthographic projection, there can be considerable variation in disparity between two images. See [7] for an analysis of this. For an excellent overview of epipolar geometry and scaled orthographic projection, see [18].

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 most of our experiments, we specify them by hand. Some applications involving offline recognition may use such hand-clicked points directly. At the same time, there is a lot of work on automatic detection of facial features [28], [14]. By reducing the number of points needed for recognition, we can make it easier to use these detectors to build fully automatic recognition systems.

3.1 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: (θ, γ, s, t). θ is the angle of the epipolar lines in the first image. γ 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. With four corresponding points, we get a nonlinear system of equations which we solve in a straightforward way.

3.2 Epipolar Geometry and Horizontal Movement

We will now consider a special 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 can determine the epipolar geometry using three feature points. The two eyes will define the direction of the epipolar lines. This tells us θ (an unknown γ still allows for in plane rotation of the images). Given a correspondence between three points we can 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.

4 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 frontoparallel planes. A number of methods might be suitable. We have decided to use a 1D dynamic-programming-based algorithm, which is quite fast. We have used the method given by Criminisi et al. [10][1] which has been developed for video conferencing applications and so seems to fit our needs. 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.

It is important to stress that, provided we get good correspondences, 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.

1. We also tried the method described by Cox et al. [9] and found it to be about twice as fast but less accurate (the accuracy was about 8 percent lower on average on several gallery-probe experiments with a gallery of 68 individuals).

4.1 Stereo Matching

The core of the stereo method calculates a matching between two scan lines (rows of each face). 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 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 [9]. See [10] for a complete presentation of the matching algorithm. In the rest of the section, we point out some of the details we use in our image comparison algorithm.

The cost of matching the two scan lines l_1 and l_2, denoted $\mathrm{cost}(l_1, l_2)$, 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 $\Sigma^\star$) that encodes the optimal matching to a given matching problem between scan lines $I_{1,i}$ 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 scan line 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.

4.2 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 the method presented by Criminisi et al. [10] to match rectified images in which the length of scan lines varies.

All solutions found by the method of Criminisi et al. have length equal to the sum of both scan lines being matched. However, since each cost is going to be compared to other costs matched over scan lines of potentially different lengths, we need some normalization strategy. The cost used weighs every match equally:

$$\mathrm{cost}(I_1, I_2) = \frac{\sum_{i=1}^{n} \mathrm{cost}(I_{1,i}, I_{2,i})}{\sum_{i=1}^{n} |I_{1,i}| + |I_{2,i}|}. \tag{1}$$

The cost expressed in (1) 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 over all scan lines.

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.

$$\mathrm{similarity}(I_1, I_2) = \min \begin{cases} \mathrm{cost}(\mathrm{rectify}(I_1, I_2)), \\ \mathrm{cost}(\mathrm{rectify}(I_2, I_1)), \\ \mathrm{cost}(\mathrm{rectify}(\mathrm{flip}(I_1), I_2)), \\ \mathrm{cost}(\mathrm{rectify}(I_2, \mathrm{flip}(I_1))). \end{cases} \tag{2}$$

TABLE 2
Results for Pose Variation with 3ptSMD

az.	-66	-47	-46	-32	-17	0	0	0	16	31	44	44	62	
alt.	3	13	2	2	2	15	2	1.9	2	2	2	13	3	
Probe	c34	c31	c14	c11	c29	c09	c27	c07	c05	c37	c25	c02	c22	avg
Gallery														
c34	-/-	79/91	85/82	74/74	59/62	29/32	37/35	15/18	47/50	51/56	49/65	60/71	54/71	53/58
c31	84/94	-/-	81/88	68/88	60/76	78/91	44/59	22/32	43/56	54/65	90/97	68/82	65/76	62/75
c14	96/94	85/91	-/-	100/100	100/100	76/82	93/91	60/76	79/88	82/82	56/56	82/88	50/56	80/83
c11	94/94	90/97	100/100	-/-	100/100	88/88	100/100	90/94	90/94	96/97	51/53	90/94	53/65	86/89
c29	88/88	79/88	100/100	100/100	-/-	99/100	100/100	97/100	100/100	96/97	54/62	90/94	50/53	87/90
c09	44/59	96/100	72/76	88/94	97/100	-/-	97/97	76/82	96/97	91/88	93/97	79/82	66/79	82/87
c27	60/74	62/76	93/97	100/100	100/100	100/100	-/-	97/97	100/100	100/100	62/65	97/100	46/53	84/88
c07	25/29	34/41	72/74	87/91	96/97	76/79	97/100	-/-	97/97	85/82	31/38	62/65	16/24	64/68
c05	60/68	60/76	90/100	91/97	100/100	97/97	100/100	96/94	-/-	100/100	76/85	100/100	63/79	86/91
c37	74/79	69/82	93/97	97/100	94/94	91/91	100/100	84/91	99/100	-/-	88/94	100/100	79/94	88/93
c25	44/47	93/94	35/44	40/59	41/44	85/91	40/47	16/18	66/68	79/85	-/-	72/79	88/97	58/64
c02	75/79	74/79	87/94	88/91	76/88	68/82	94/94	60/62	96/100	99/97	85/94	-/-	88/94	82/87
c22	56/68	62/65	47/53	43/53	37/44	53/65	32/38	13/21	41/53	56/68	87/88	69/79	-/-	49/57

The cell format is accuracy with 68 faces/accuracy with 34 faces. The diagonals are not included in any average. The global averages are 74.5 percent for 68 faces and 79.8 percent for 34 faces.

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 four-point case or, in the three-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 of the images in the gallery should be in the same pose. Our C implementation of the method can compare two 60×72 facial images in 1 second. The number of operations done to compare two $w \times h$ images is $O(hw^2)$.

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 and 2) similarity computation using stereo matching. In the next section, we will see that this very straightforward approach demonstrates excellent performance.

5 Experiments

We have tested our algorithm using the CMU PIE database [25]. 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 have a significantly higher camera altitude (poses: c31, c09, and c02) and one last pose 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. 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 three-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 four-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 those of Gross et al. [15], [17] and Chai et al. [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 the method of Romdhani et al. [24], 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 the method of Gross et al. [15], [17] and one to compare with the method of Romdhani et al. [24].

5.1 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 [15], eigenfaces [26] and FaceIt as described in [15], [17]. FaceIt[2] 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 [15], [17]. Table 4 summarizes the average recognition rates. Table 2 presents detailed results for this experiment using 3ptSMD, and Table 3 presents detailed results for this experiment using 4ptSMD. Fig. 2 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 percent) and when they don't (with an average of 75.4 percent) and overall (with an average of 78.5 percent, as reported in Table 6) 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.

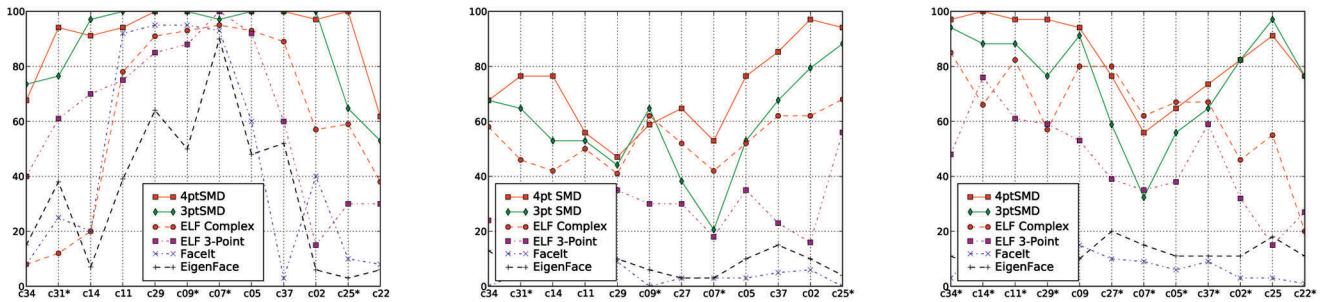
Fig. 2 shows a comparison with the results presented in the paper of Gross et al. [15], [17]. In this experiment, we observe that

2. Version 2.5.0.17 of the FaceIt recognition engine was used.

TABLE 3
Results for Pose Variation with 4ptSMD

az.	-66	-47	-46	-32	-17	0	0	0	16	31	44	44	62	
alt.	3	13	2	2	2	15	2	1.9	2	2	2	13	3	
Probe	c34	c31	c14	c11	c29	c09	c27	c07	c05	c37	c25	c02	c22	avg
Gallery														
c34	-/-	79/91	91/91	78/82	65/68	38/44	44/44	26/35	50/50	50/53	60/65	71/74	56/65	59/63
c31	91/97	-/-	99/100	96/97	94/97	78/94	65/76	50/56	62/65	65/74	84/91	72/82	60/76	76/83
c14	97/100	100/100	-/-	97/100	91/97	87/85	79/91	71/71	79/91	76/82	59/68	76/91	78/85	82/88
c11	94/97	97/97	99/100	-/-	100/100	97/94	94/94	94/97	88/100	94/97	79/82	87/94	65/74	90/93
c29	87/85	97/97	96/97	100/100	-/-	100/100	99/97	100/100	96/97	94/97	82/85	81/94	53/53	90/91
c09	54/62	91/97	84/91	99/100	100/100	-/-	100/100	97/97	94/100	94/97	85/91	90/91	65/76	87/91
c27	60/68	93/94	91/91	97/94	99/100	99/100	-/-	100/100	97/100	99/100	97/100	97/97	62/62	90/92
c07	40/41	62/71	79/79	97/97	100/100	96/100	100/100	-/-	100/100	99/97	88/85	97/94	32/35	82/83
c05	71/79	79/85	90/100	93/100	97/97	97/94	99/100	100/100	-/-	100/100	100/100	99/100	78/91	91/95
c37	66/74	74/79	85/94	94/97	90/91	91/88	97/97	99/100	100/100	-/-	100/100	100/100	91/97	90/93
c25	65/76	79/88	56/68	66/76	71/82	85/88	91/97	79/82	97/100	100/100	-/-	99/97	94/97	81/87
c02	81/88	71/85	74/88	81/85	69/85	93/94	90/91	85/94	93/97	100/100	99/97	-/-	99/100	86/92
c22	57/68	62/76	66/76	56/56	44/47	49/59	47/65	35/53	66/76	76/85	88/94	91/97	-/-	61/71

The cell format is accuracy with 68 faces/accuracy with 34 faces. The diagonals are not included in any average. The global average is 82.4 percent for 68 faces and 86.8 for 34 faces.



(a) (b) (c)

Fig. 2. Cross sections with fixed gallery pose for the results presented in Table 4. Probe poses marked with * have a vertical misalignment of about 10 degrees with the corresponding gallery pose. (a) Gallery Pose c27. (b) Gallery Pose c22. (c) Gallery Pose c31.

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 3ptSMD at the poses where there is considerable misalignment (the poses marked with $*$).

Table 5 shows a comparison with the method of 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, c11, c07, and c09. Note that this is a slice of data from Table 2. Our 3ptSMD method produces nearly perfect results in these conditions, results that are much better than those reported by Chai et al.

5.2 PIE Pose Variation: 68 Faces

We also compared our results with the ones presented by Romdhani et al. [24]. 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 4 summarizes the results of this experiment.

The global average for the method of Romdhani et al. [24] is 74.3 percent, the global average for our 3ptSMD method is almost the same, at 74.5 percent. For the subset of poses in which the epipolar lines fit perfectly, our average performance is 80.8 percent, while theirs is 71.6 percent. 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 percent. Our 4ptSMD achieves overall accuracy of 82.4 percent, which is considerably higher than the performance reported by Romdhani et al. Our method runs about 40 times faster than the method presented in [24], requires fewer manually specified points, and is much simpler. Detailed results are presented in Tables 2 and 3.

We also tried the fully automatic method (using probabilistic egomotion [11], [12] to compute the epipolar geometry) on the PIE data set on a small subset of the pose combinations obtaining the same results as with four hand-clicked points. Because the egomotion computation is a bit slow and must be computed for each pair of images, it is not practical to use it on the entire PIE data set.

5.3 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

TABLE 4
A Comparison of Our Stereo Matching Distance with Other Methods across Pose

34 Faces

Method	Accuracy
Eigenfaces [15], [17]	16.6%
FaceIt [15], [17]	24.3%
Eigen light-fields (3-point norm.) [15], [17]	52.5%
Eigen light-fields (Multi-point norm.) [15], [17]	66.3%
3-point Stereo Matching Distance	**79.8%**
4-point Stereo Matching Distance	**86.8%**

68 Faces

Method	Accuracy
LiST (Romdhani et al. [24])	74.3%
3-point Stereo Matching Distance	**74.5%**
4-point Stereo Matching Distance	**82.4%**

TABLE 5
Comparisons over a Slice of the Data with the Method of Chai et al. [8] and Gross et al. [17]

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

The gallery pose is c27 and contains 34 faces. The table layout is the same as the one presented in [8].

lighting variation, the use of normalized correlation in matching 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.

First, we compare our method to BFS [15] 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 Fig. 3. We observe that our algorithm strictly dominates BFS over all probe poses. For lighting invariance, they use [16] 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 Blanz and Vetter [5] and Romdhani et al. [24]. For this experiment, we used images of the faces of 68 individuals viewed from three poses (front: c27, side: c5, and profile: c22) and illuminated from 21 different directions. We used light number 12 for the gallery illumination to be able to compare our results with those of Romdhani et al. [24]. They select that lighting because "... the fitting is generally fair at that condition." Our results are presented in Table 7. We do not expect our results to be as good as those of Romdhani et al. [24] because our algorithm only accounts for lighting variation by using a fast approximation to normalized cross correlation as described by Criminisi et al. [10], while Romdhani et al. [24] have a 3D model and perform an optimization to solve for the lighting that best matches the model to the image. We also tested on the part of the PIE data set without ambient lights which has harsh shadows. Other works on pose 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 that of Romdhani et al. [24] when there is no pose change (gallery-probe combinations:

TABLE 6
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 Misalign-ment**	**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%

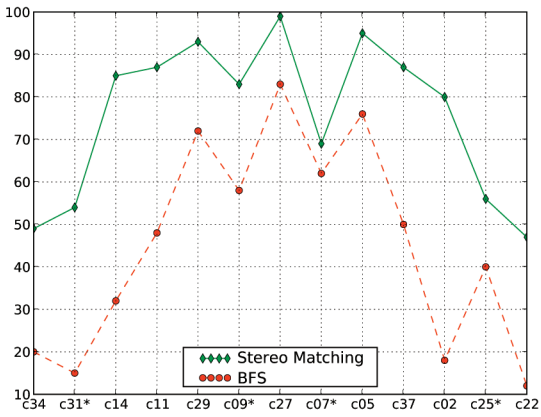


Fig. 3. A comparison of our method with BFS. Gallery pose is frontal (c27) probe. Poses are as indicated in the x axis; we report the average over the 21 illuminations.

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 percent, while the global average of Romdhani et al. [24] is 81 percent, which is considerably better.

5.4 Unconstrained Face Recognition: Labeled Faces in the Wild

This section of the experimental evaluation is done with the LFW data set [20]. The data set is a collection of images from the news in which the Viola and Jones [28] face detector has detected a face. The data set therefore falls into the category of unconstrained face recognition. The LFW data set is designed with a specific protocol to test learning-based recognition methods.

With these experiments, we want to show how our method works in an unconstrained setting: expression, illumination, and pose change that occur at the same time. In this section, the epipolar geometry was obtained from each pair of images using the probabilistic egomotion method of Domke and Aloimonos [11], [12]. Note that in this configuration our method requires no hand-clicked points. We evaluate two methods based on SMD, one using NSSD as described before and one using the SIFT-like [21] DHOG [27] descriptor (dense histogram of gradient orientations).

TABLE 7
Accuracy Percentage with Pose and Illumination Variation

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

The cell format is: (with ambient)/(without ambient). Mean w/ambient lights: 74.6 percent and mean wo/ambient lights: 67.4 percent.

TABLE 8
A Comparison of Different Methods as They Perform on the LFW Data Set

Method	Accuracy
Eigenfaces	0.6002 ± 0.0079
MERL	0.7052 ± 0.0060
SMD + NSSD+ Probabilistic Egomotion, funneled	$\mathbf{0.7092 \pm 0.0055}$
Nowak, original	0.7245 ± 0.0040
SMD + DHOG + Probabilistic Egomotion, funneled	$\mathbf{0.7251 \pm 0.0050}$
3x3 Multi-region histograms	0.7295 ± 0.0055
Nowak, funneled	0.7393 ± 0.0049
MERL+Nowak, funneled	0.7618 ± 0.0058
Hybrid descriptor-based, funneled	0.7847 ± 0.0051

All of the results were taken from [1].

The results are presented in Table 8. SMD performs as well as any prior method based on image comparison. However, our current approach does not incorporate learning, and works considerably less well than learning-based approaches. We believe that learning is critical for this type of uncontrolled variation. It remains a future direction to determine how best to incorporate learning into SMD.

6 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 2D 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 straightforward 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 data set 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. [17], Chai et al. [8], and Romdhani et al. [24]. 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.

Acknowledgments

The authors gratefully acknowledge financial support from a fellowship from P. Horvitz (Apptis, Inc.), from the Honda Research Initiation Grant and from the US Office of Naval Research Under MURI Grant N00014-08-10638. They would like to thank Olga Veksler for her advice about stereo. They would like to thank Peter Belhumeur for his insights on the interactions between pose and illumination in face recognition.

References

[1] Labeled Faces in the Wild Website, http://vis-www.cs.umass.edu/lfw/results.html, 2009.
[2] A.B. Ashraf, S. Lucey, and T. Chen, "Learning Patch Correspondences for Improved Viewpoint Invariant Face Recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2008.
[3] R. Basri and D. Jacobs, "Lambertian Reflectance and Linear Subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218-233, Feb. 2003.
[4] D. Beymer and T. Poggio, "Face Recognition from One Example View," Technical Report AIM-1536, 1995.
[5] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3d Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063-1074, Sept. 2003.
[6] K.W. Bowyer, K.I. Chang, and P.J. Flynn, "A Survey of Approaches and Challenges in 3d and Multi-Modal $3d + 2d$ Face Recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1-15, 2006.
[7] C.D. Castillo and D.W. Jacobs, "Using Stereo Matching for 2D Face Recognition across Pose," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.
[8] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally Linear Regression for Pose-Invariant Face Recognition," *IEEE Trans. Image Processing*, vol. 16, no. 7, pp. 1716-1725, July 2007.
[9] I.J. Cox, S.L. Hingorani, S.B. Rao, and B.M. Maggs, "A Maximum Likelihood Stereo Algorithm," *Computer Vision and Image Understanding*, vol. 63, no. 3, pp. 542-567, 1996.
[10] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P.H.S. Torr, "Efficient Dense Stereo with Occlusions for New View-Synthesis by Four-State Dynamic Programming," *Int'l J. Computer Vision*, vol. 71, no. 1, pp. 89-110, 2007.
[11] J. Domke and Y. Aloimonos, "A Probabilistic Framework for Correspondence and Egomotion," *Proc. Workshop Dynamical Vision*, R. Vidal, A. Heyden, and Y. Ma, eds., pp. 232-242, 2006.
[12] J. Domke and Y. Aloimonos, "A Probabilistic Notion of Correspondence and the Epipolar Constraint," *Proc. Third Int'l Symp. 3D Data Processing, Visualization, and Transmission*, pp. 41-48, 2006.
[13] 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 Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643-660, June 2001.
[14] Y. Gizatdinova and V. Surakka, "Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 135-139, Jan. 2006.
[15] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face Recognition across Pose and Illumination," *Handbook of Face Recognition*, S.Z. Li and A.K. Jain, eds., Springer-Verlag, June 2004.
[16] R. Gross and V. Brajovic, "An Image Preprocessing Algorithm for Illumination Invariant Face Recognition," *Proc. Fourth Int'l Conf. Audio- and Video-Based Biometric Person Authentication*, June 2003.
[17] R. Gross, I. Matthews, and S. Baker, "Appearance-Based Face Recognition and Light-Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449-465, Apr. 2004.
[18] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second ed. Cambridge Univ. Press, 2004.
[19] C.-K. Hsieh and Y.-C. Chen, "Kernel-Based Pose Invariant Face Recognition," *Proc. IEEE Int'l Conf. Multimedia and Expo*, pp. 987-990, 2007.
[20] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *Proc. Faces in Real-Life Images Workshop in European Conf. Computer Vision*, 2008.
[21] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[22] S. Lucey and T. Chen, "A Viewpoint Invariant, Sparsely Registered, Patch Based, Face Verifier," *Int'l J. Computer Vision*, vol. 80, pp. 58-71, Oct. 2008.
[23] A. Martinez, "Matching Expression Variant Faces," *Vision Research*, vol. 9, pp. 1047-1060, 2003.
[24] S. Romdhani, V. Blanz, and T. Vetter, "Face Identification by Fitting a 3d Morphable Model Using Linear Shape and Texture Error Functions," *Proc. European Conf. Computer Vision*, vol. 4, pp. 3-19, 2002.
[25] T. Sim, S. Baker, and M. Bsat, "The cmu Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615-1618, Dec. 2003.
[26] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 586-591, 1991.
[27] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," http://www.vlfeat.org/, 2008.
[28] P. Viola and M. Jones, "Robust Real-Time Object Detection," *Int'l J. Computer Vision*, 2002.
[29] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *Proc. Seventh Int'l Conf. Computer Analysis of Images and Patterns*, G. Sommer, K. Daniilidis, and J. Pauli, eds., pp. 456-463, 1997.
[30] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, Dec. 2003.