Face Recognition and Alignment using Support Vector Machines

Antony Lam
University of California, Riverside
antonylam@cs.ucr.edu

Christian R. Shelton
University of California, Riverside
cshelton@cs.ucr.edu

Abstract

Face recognition in the presence of pose changes remains a largely unsolved problem. Severe pose changes, resulting in dramatically different appearances, is one of the main difficulties. We present a support vector machine (SVM) based system that learns the relations between corresponding local regions of the face in different poses as well as a simple SVM based system for automatic alignment of faces in differing poses. We then present experimental results from multiple random splits of the CMU PIE Database to verify the strength of our approach.

1. Introduction

Automatic face recognition has numerous applications in areas as diverse as security, human computer interaction, and image search engines. As such, there has been much work on face recognition in the past decades and tremendous progress has been made. There already exist systems that can perform in excess of 90% accuracy under controlled conditions [6]. However, changes in illumination or pose remain largely unsolved problems [17]. In this paper we focus on the issue of recognition across pose. The general problem we wish to address is as follows: given two images of faces in arbitrary poses, indicate how likely they are to be the same person. This is similar to the "one sample per person" problem mentioned in [14]. This is an important problem because it is not always possible to have multiple images of the same person. This is especially true if one were attempting to determine the identity of a stranger. In this problem, the stranger may have pictures of himself scattered all over the Internet with no clear organization of those images but there still must be a way to query a search engine with just one facial image and get ranked results of other similar faces.

In recent years, there has been much work related to the problem of face recognition with pose changes. For example, Blanz and Vetter [1] built a system that uses a 3D morphable model to perform face recognition. In their work they built explicit 3D models of the head and face which has the advantage that their models can be very accurate. However, there have been other works using simpler models that have proven effective [3, 4]. (Although it should be noted that [4] did not perform as well on full profile views. Our system performs competitively on such views.) Eigen Light-Fields [5] addresses the problem by computing an eigenspace from the light-fields of the head and recognizing based on Eigen Light-Fields in a manner analogous to Eigenfaces [15]. It has the advantage of being able to use as many images as are available to improve its accuracy and does not require the gallery images to be in a canonical pose. However, it may be fruitful to investigate how a more explicit image-based representation of relations between pose can improve accuracy. The Eigen Light-Fields approach also does not make use of a component based decomposition of the face which has been shown in some cases to be more beneficial than a global approach to recognition [8].

There have also been other works that attempt to solve this problem using more explicit learning of pose relations through patch decompositions. Kanade and Yamada [10] presented a multi-subregion based approach which decomposes faces across pose into patches and learns the relations between corresponding patches from one pose to another under a Gaussian model. Their work showed that recognition between poses separated by as much as 45 degrees can still be done with accuracy in excess of 80%. However, a drawback to the multi-subregion system is its reliance on similarity in appearance between corresponding patches of the same people. Not surprisingly, at extreme differences in pose, accuracy drops. Liu and Chen [11] extended the work of [10] by introducing a texture map representation of the face. They assume the head to be an ellipsoid and determine what the texture map of such an ellipsoid head would be. The basic idea is that their transformation allows for facial features to maintain greater similarity over a wider range of pose changes. While this does improve results, their alignment procedure has to optimize over a total of eight parameters, for each image. There is also the drawback that at extreme poses, there is a limit to how much the texture map can transform facial features to appear similar.

978-1-4244-2154-1/08/$25.00

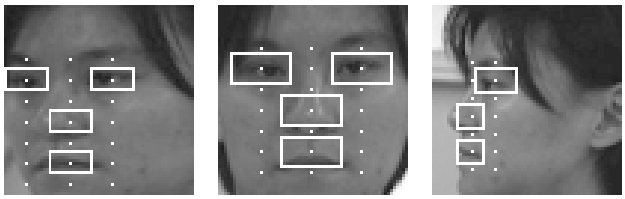

Figure 1. Centers of the Salient Regions and some of the Bounding Boxes

In this work, we offer our extension to the work of Kanade and Yamada [10] by modeling the relations between patches not based on their similarity but on their joint appearances. This is achieved through the application of support vector machines (SVMs) [2] for capturing patch relations between poses. We first present our algorithm in the next section followed by experimental results on manually aligned faces and preliminary results on recognition performance with automatically cropped and aligned faces.

2. SVM Based Recognition

SVMs have been shown to be a powerful tool for the task of face recognition [9, 8]. Jonsson et al. [9] showed empirical evidence that SVMs could effectively extract relevant discriminatory information from training images of frontal images. Heisele et al. [8] even applied SVMs to recognition in the presence of pose change. However, they only tested their system on a database of 10 subjects as their goal was to establish the strength of SVMs for face recognition and to compare the accuracies of global versus local approaches to analyzing faces. In our work, we continue this line of investigation by testing SVMs trained on local image patches of the face using the CMU PIE database as it has more individuals (68 subjects) and a greater set of ranges in face pose. As Kanade and Yamada [10] presented promising initial work on this problem, we will adopt their testing methodology so our work can be seen in context.

2.1. Local Patch Representation of Faces

In our work, we first chose to manually decompose frontal faces into 21 salient rectangular regions. We then manually located the corresponding salient regions to the frontal regions for all other poses. In the case where only a part of the face was visible due to self occlusion in more extreme poses, we only located the visible corresponding regions (see Figure 1). This is essentially the same as the decomposition performed in [10] except we did not define all regions to be of the same size. This is because corresponding regions across pose do not maintain the same size due to foreshortening and self-occlusion as the pose change deviates from one view to the next. A good example of this effect is in the eye. If we imagine a head turning from a frontal view towards the right, the bounding box around the person's right eye would appear increasingly smaller.

Note that this step is performed once per pose, not once for each image of a pose. We assume that the images are aligned, either manually or automatically (see Section 2.4). Thus, this step represents the input of knowledge about the position (although not appearance) of facial features under rotation. Currently this is manual. In future work we would like to automatically define the salient regions of all poses but we do this manually at present to test the effectiveness of such regions for recognition.

2.2. Discretely Separated Poses for Training Face Recognition

Kanade and Yamada [10] showed that recognition accuracies beyond 90% between images of faces in poses differing by as much as 30 degrees of out of plane rotation can be done by just measuring the sum squared error between corresponding patches (as part of a Gaussian model). This suggests that computers can tolerate some degree of pose change without very precise modeling of facial geometry. Intuitively, this makes sense because minor changes in facial pose do not change facial appearance dramatically. As long as we know the general locations of corresponding salient regions in the face, direct comparisons between the pixel values of the regions suffices when pose change is minor. Recognition accuracy mainly drops when the poses of two given facial images are too different. What this suggests is that there may be no need to model pose change as a continuous process for face recognition purposes. Instead, modeling how a discrete set of specific poses covering a wide range of rotations relate to each other may be all that is needed for recognition across pose. Like the work of [10] and [11], we adopted this same approach.

Thus, we only define the 21 regions above for a set of discrete poses. We feel this is enough to gain good accuracy on any intermediate poses. For this work, we chose the 13 poses from the CMU PIE Database (see Figure 3).

2.3. SVMs for Learning Pose Relations

We now define the main contribution of our work: a pose relation SVM approach to face recognition. The basic idea behind our application of SVMs to learning pose relations is to train SVMs to answer the question of whether two faces in pose m and pose n are of the same person or not. As our analysis of faces is based on the patch decomposition of [10], learning a relation from two given poses m and n requires training an SVM for each of the corresponding regions between the two poses to be able to decide if the corresponding regions belong to the same person or not.

To apply SVMs to the problem of learning pose relations, our algorithm first takes in images of multiple known subjects in different poses as training data. These people are not part of the set to be recognized (nor the query set). This

set is only used to learn a general relationship between the appearance of a region in one pose and the appearance of the same region in a different pose. The relations between every pair of poses, m and n, are independently learned.

The training procedure will be presented shortly but we shall first introduce some notation. Let $p_m(i)$ be an image of subject i in pose m and $p_m^k(i)$ be the kth region of the image $p_m(i)$. Similarly, let $p_n^k(j)$ be the corresponding kth region of $p_m^k(i)$ in image $p_n(j)$. Let $v_m^k(i)$ be region $p_m^k(i)$ represented as a vector and $v_{m,n}^k(i,j)$ be the concatenation of $v_m^k(i)$ and $v_n^k(j)$. The training procedure between poses is as follows.

1. For each $v_{m,n}^k(i,j)$ if $i = j$, consider it to be a positive case for region k; otherwise, consider it to be a negative case for region k.

2. For each region k, train an SVM $R_{m,n}^k$ (with a radial basis function kernel) using the corresponding dataset.

This procedure aims to build independent pose relation SVMs for each of the corresponding regions k between two given poses. The learned functions are used to determine a score for how likely two novel images of faces are to be the same person. To employ the learned functions, we take two novel facial images q and t in poses a and b respectively and subdivide each of them into the regions associated with their pose. For each pair of the corresponding regions l in the two images, we concatenate the vectors describing each of the regions l into a single vector $v_{a,b}^l(q,t)$ and feed it to the function $R_{a,b}^l$ which returns whether the two regions match.

It would be natural to sum the resulting number of positive classifications (across the region l) from the SMVs to determine how likely it is that the two images are of the same person. However, we discovered that due to the small training data sizes and the relatively large portion of negative examples in the training sets, all of the R outputs would be -1 (*i.e.* not a match) for all our test data.

Instead, we discovered that the raw distances to the hyperplane for each SVM provided indications of which subjects were likely to be the same person. Thus if we let $R_{a,b}^l$ be not the thresholded output of the support vector machine, but rather the distance to the hyperplane (*i.e.* the value prior to thresholding), although all of the outputs would be negative, the total sum would still be a good measure. Thus we use

$$s_r(q,t) = \sum_{k=1}^{K} R_{a,b}^k(v_{a,b}^k(q,t)) \tag{1}$$

to score whether two images q and t with poses a and b respectively are of the same person. The higher the score, the more likely it is that the two observed faces belong to the same person.

Figure 2. Examples of the Alignment Grids

2.4. Automated Face Cropping and Alignment

We performed experiments using manually cropped and aligned images of the faces but we also developed an automation of that procedure. To automate the process, we first used the Viola-Jones face detector [16] to automatically crop faces from images. We found the detector to be robust enough to crop faces in a wide range of poses (including profile views) but the faces would typically have some variation in alignment. This is partially due to the detector not always cropping faces consistently and also from differences in the way the subjects positioned their heads in the images. As such, an alignment procedure was found to be necessary to position the faces in canonical positions so that their salient regions could be better compared.

We now present work on a simple SVM based alignment procedure that aligns given images reasonably well for a large number of poses. The procedure learns canonical alignments for a set of discretely separated poses based on provided manual alignments as training data. The alignment algorithm uses a set of "alignment SVMs" for each pose p_n which are trained as follows.

1. Gather a training set of manually aligned faces in pose n.

2. For each image, divide the entire image into K evenly sized regions $r_n^1, r_n^2, ..., r_n^K$ over the entire image.

3. For each region r_n^k, train an SVM A_n^k (using a radial basis function kernel) to classify all examples of region r_n^k as positive and any other region r_n^l where $l \neq k$ as negative.

In our experiments, we set $K = 25$ (see Figure 2). Once the training procedure is complete, the 25 SVMs can be used to score alignments of facial images $p_n(i)$ through observation of its 25 evenly sized regions. The scoring function is defined as

$$s_a(p) = \sum_{k=1}^{K} A_n^k(r_n^k(p)) \tag{2}$$

where $r_n^k(p)$ is the region k in image p (assuming pose n). We note that the SVMs here are separate from the SVMs

for recognition. Each pose n has a set of alignment SVMs specifically trained for it. These determine what type of appearance each local region r_n^k should have and the SVMs vote using their raw distances to the hyperplane on whether their own region is consistent with a good alignment.

Using the alignment scoring procedure, a search for a good alignment can be done simply by a brute force search over the alignment parameters of translation, scale, and in-plane rotation. However, a brute force search is very slow so we adopted a heuristic. We observed that facial crops found by the Viola-Jones detector are restricted to a certain range of differences in translation, scale, and rotation. We decided to perform a local gradient ascent by iteratively optimizing over each parameter independently and observed good results. However, the automatic face crops were less consistent in profile views and those images did not automatically align as well. To address the problem with the profile views, we introduced random restarts into our search procedure and applied the same procedure (with random restarts) to all the poses (including frontal views). We outline the specific details of our alignment procedure below.

1. Maximize Equation 2 over each alignment parameter (scale, rotation, x-translation, and y-translation) in turn, while keeping the others fixed. (All parameter searches on the variables are within some bounded region from the variable's current value.)

2. Repeat the above step until convergence or a maximum number of iterations has been reached.

3. If the score of the found alignment exceeds a minimum alignment score threshold, accept the found alignment.

4. Otherwise, perform a random restart to some other point in the parameter space within a bounded region from the alignment that was just found and repeat from step 1. (We do a maximum of 10 random restarts.)

5. If no alignment with a score exceeding the minimum score threshold was found, select the alignment with the highest score.

In our outline, we introduced a minimum alignment score threshold. This threshold is determined based on the training data used for the particular pose being aligned. We set the threshold to be the mean of the training scores minus half of the standard deviation of the same scores. (Higher scores indicate better alignment. This threshold serves to trade-off accuracy in alignment for speed.)

While there exist 3D facial alignment methods such as [7], these methods rely on 3D laser scans to use as training data. Our alignment algorithm only requires a set of 2D images. (Although we should note that the work here on alignment is still somewhat preliminary.) While our paper's emphasis is mainly on exploring the performance of SVMs in face recognition across pose, we include this alignment procedure and corresponding results to show more general use of region-based SVMs in face recognition.

3. Experiments

We tested our system on the CMU PIE database and adopted the general protocol used by Kanade and Yamada [10]. The protocol is to choose half of the subjects for training the recognition system and the other half for testing. However, when researchers develop such systems and are using the same test data to verify their algorithms (during development), they may unknowingly overfit their test data. This is due to the fact that algorithms under development can be adjusted and modified in many ways. We first developed our system using the first half of the subjects as the training set and the last half as the testing set (in which the frontal pose was used as the gallery database and the non-frontal poses used as queries) and only coarsely tuned the parameters of our algorithm. After we were satisfied with our system, we selected five random splits of the subjects. Each split would have half the subjects randomly selected for training and the other half for testing. We then tested our system without adjusting any parameters on these five random splits and determined the mean and standard deviations of our accuracies for recognition. The same splits were used to test our implementations of the systems described in [10] and [11] for comparison purposes.

3.1. The CMU PIE Database

The CMU PIE Database [13] contains images of 68 subjects taken under 13 different poses, 21 different illuminations, and 2 occasions resulting in over 37,000 images of people. In our experiments, we use only the frontal illumination, neutral expression, no glasses subset of the database. This means we work with the 68 subjects where each one has 13 poses. The poses are denoted by their camera labels (*e.g.* c27 for the frontal view and c11 for one of the 45 degree views).

3.2. Multi-Subregions

Kanade and Yamada [10] developed a system for recognition across pose based on the similarity of local regions on the face between different poses of the same people and different people. They first manually determined the locations of the eyes and mouth for all facial images and used those locations to define a 7-by-3 lattice of points on the face starting from the eyebrow and extending down to the chin. These points were then used to define 9-by-15 pixel subregions on the face and the similarities between corresponding regions between poses were then modeled using two Gaussians, one for the similarities of image patches be-

Figure 3. Examples of the Poses in the CMU PIE DB (From the CMU PIE DB Website)

tween same identities and the other for different identities. In our implementation of their system, we needed to make one modification because they assume all regions are 9-by-15 pixels in size. Our regions are variably sized (as discussed in Section 2.1) so we had to normalize the sizes between corresponding regions. We do this normalization by resizing the smaller patch to be the same size as the larger patch. In a way, this illustrates a strength of our approach. Our system can learn relations between subregions of different poses without the need to fix patch sizes to be the same between all poses.

3.3. Texture Maps

Liu and Chen [11], developed a transformation in which the head is assumed to be an ellipsoid and a texture map based on this assumption is computed for each face. The idea is that if the head were an ellipsoid, the out of plane rotation of the head would be easier to recognize after this transformation. The hope is that the texture maps would help to preserve similarities of facial features between the same person across a wider range of poses than in [10]. In their work, Liu and Chen first manually cropped out faces. They then applied texture map transformations to the faces based on a best fit to their Universal Mosaic Model. As we did not have such a model available to us, we used our manually aligned faces and determined (manually) the texture mapping parameters for fitting to a canonical mosaic model. We spent a long time optimizing these parameters to achieve good results and visually inspected the texture map results to verify that they were reasonable fits.

3.4. Results

In our experiments, we tested each split independently with all facial crop sizes normalized to 64-by-64 pixels. We kept galleries of frontal images from each of the five splits and used all the other poses as probe images. At the moment, we assume that the pose of each image is specified so the recognition accuracies determined were done independently for each pose for all three systems. There also already exist systems in the literature such as [12] where face detection and pose estimation are done simultaneously

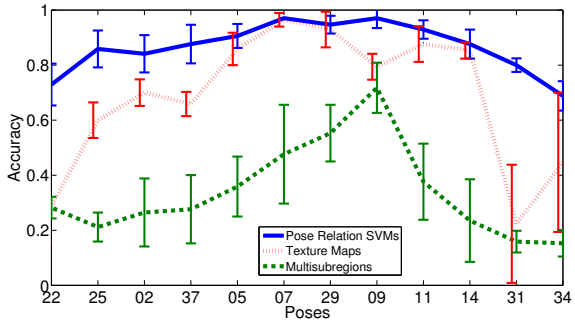

Figure 4. Rank 1 Recognition Results on Manually Aligned Faces (means with standard deviations)

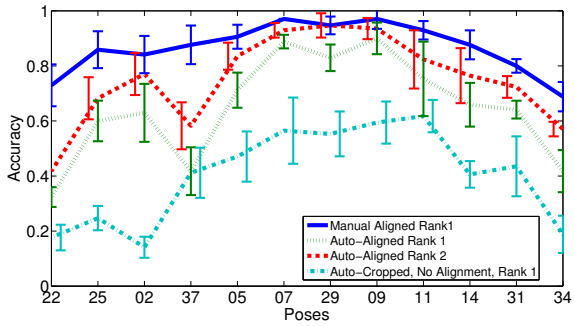

Figure 5. Recognition Results on Automatically Cropped and Aligned Faces (means with standard deviations)

and such systems could be integrated into ours.

Figure 4 compares our system to the other two systems with manually aligned images. It can be seen that our method outperforms the other two methods as the pose becomes increasingly distant from the frontal view. The reason we do not have as much advantage over the texture map method in the less extreme poses is because the appearance of subregions does not vary greatly when pose change is minor. In this case, the other systems may actually be more generally applicable than our SVM-based system since they are based on direct measurements of similarity between image patches. (The multi-subregion method would have been expected to perform well in the closer to frontal cases but did not. This is likely due to the differently sized subre-

gions that we selected.) SVMs on the other hand require sufficient training data to choose support vectors that would cover a wide enough range of cases in facial appearance.

However in the case of extreme pose change, the use of similarity between images patches fails because the same facial feature can appear dramatically different (*e.g.* the nose). In this case, use of SVMs provides better accuracy. It should be noted that although we made a best effort to produce a fair comparison of our work to the work of [11], many factors such as the way images were cropped or our choice of texture parameters can have an effect on their accuracy. We note that [11] reported accuracies of 60% and 70% for their most extreme poses so it may be that our implementation of their system is not optimally tuned. However, they use more facial features. For example, they note the forehead as being a strong indicator of identity between poses. We chose to only limit ourselves to the regions defined in [10] which focuses primarily on the face only. It can be seen that even with fewer features and a more rigorous 5-split testing procedure we still achieve about 70% accuracy for both extreme poses.

Figure 5 compares our automatic alignment system's performance to that of manually aligned images and automatic crops without automatic alignment. For these results, we also retrained the alignment system based on each of the random splits. It is not surprising that there is a degrading of performance. The performance is especially degraded in the extreme poses. Although our automatic alignments were actually quite close to the manual alignments, accuracy was likely affected by slight inconsistencies in scale and translation. Our current use of SVMs examines the pixel values directly and is thus more sensitive to misalignments. However, it is encouraging that the rank 2 accuracies of the automatically aligned images can be 15% greater than the rank 1 accuracies. If a person were using such a face recognition system in a search engine, a set of top ranked images would still provide useful results to the user.

4. Conclusions

We present a method for using region-based pose relation support vector machines to learn aligning and recognition scoring functions. The results are good and do not require any manual intervention except the definition of 21 regions for each pose. We are especially encouraged by the quality of the results given the small training set size (only 34 images per pose). In most applications, a larger database of images would be available.

References

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. fifth workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[3] S. Du and R. Ward. Face recognition under pose variations. *Journal of the Franklin Institute*, 343(6):596 – 613, 2006.

[4] D. Gonzlez-Jimnez and J. L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *IEEE Transactions on Information Forensics and Security*, 2:413–429, 2007.

[5] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):449 – 465, 2004.

[6] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? Technical Report CMU-RI-TR-01-17, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2001.

[7] L. Gu and T. Kanade. 3D alignment of face in a single image. In *Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1305–1312. IEEE Computer Society, 2006.

[8] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91(1-2):6–21, 2003.

[9] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Learning support vectors for face verification and recognition. In *Proc. the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pages 208–213. IEEE Computer Society, 2000.

[10] T. Kanade and A. Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *Proc. 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 2, pages 954 – 959, 2003.

[11] X. Liu and T. Chen. Pose-robust face recognition using geometry assisted probabilistic modeling. In *Proc. the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 502–509. IEEE Computer Society, 2005.

[12] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *Joural of Machine Learning Reserach*, 8:1197–1215, 2007.

[13] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, 2003.

[14] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006.

[15] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 511–518. IEEE Computer Society, 2001.

[17] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.