Face Mis-alignment Analysis by Multiple-Instance Subspace

Zhiguo Li[1], Qingshan Liu[12], and Dimitris Metaxas[1]

[1] Department of Computer Science, Rutgers University
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{zhli,qsliu,dnm}@cs.rutgers.edu

Abstract. In this paper, we systematically study the effect of poorly registered faces on the training and inferring stages of traditional face recognition algorithms. We then propose a novel multiple-instance based subspace learning scheme for face recognition. In this approach, we iteratively update the subspace training instances according to diverse densities, using class-balanced supervised clustering. We test our multiple instance subspace learning algorithm with Fisherface for the application of face recognition. Experimental results show that the proposed learning algorithm can improve the robustness of current methods with poorly aligned training and testing data.

1 Introduction

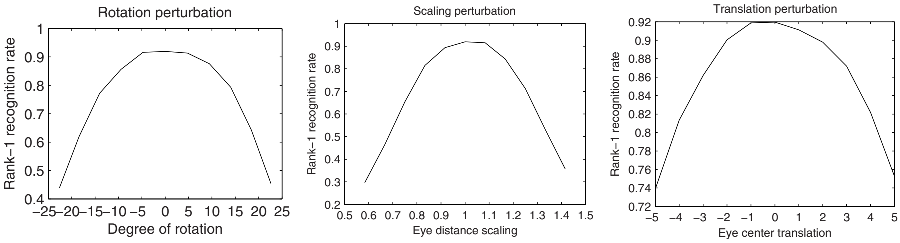
Face recognition has been one of the most successful applications of image analysis due to its wide range of potential commercial, security and entertainment applications. Depending on the type of features used, face recognition algorithms can be classified into two categories: shape based approaches, such as elastic bunch graph matching [1], and appearance based approaches, such as eigenfaces [2,3] and Fisher-faces [4,5] etc.

Accurate face alignment is critical to the performance of both appearance-based and shape-based approaches. However, current feature extraction techniques are still not reliable or accurate enough. It is unrealistic to expect localization algorithms to always get very accurate results under very different lighting, pose and expression conditions. To get better recognition rate, we need to improve the robustness of existing recognition algorithms.

To illustrate the effect of the face alignment error on face recognition performance, we use the FERET face database [6] with ground truth alignment information available. We intentionally add some perturbations to the ground truth. Perturbations are added by moving the left center and right eye center ground truth with some random pixels.

Figure 1 shows that the rotation perturbation affects the recognition performance most, and the translation perturbation has the smallest effects. Overall, we can see that even small perturbations could reduce the recognition rate significantly.

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 901–910, 2007.



(a) (b) (c)

Fig. 1. Recognition Rate Change for FisherFace w.r.t (a) Rotation, (b) Scale and (c) Translation Perturbations

One intuitive way to make classifiers robust to image alignment errors is to augment the training sets by adding random perturbations to the training images. By adding noisy but identifiable versions of given examples, we can expand our training data and improve the robustness of the feature extraction against a small amount of noise in the input. The augmented training set can model the small image alignment errors. The other way is to add perturbations to the probe images during the testing stage. Adding perturbations to the training set requires that we know the ground truth before hand.

In multiple-instance learning algorithms, the task is to learn a classifier given positive and negative bags of instances. Each bag may contain many instances. A bag is labeled positive if at least one of the instance in it is positive. A bag is labeled negative only if all instances in it are negative. The face alignment problem can be explicitly formulated as a multiple-instance learning problem: we take the whole image as a bag, and all possible sub-windows within it as instances. If an image contains a face, then we label this image as a positive bag, since we know that there is at least a sub-window containing the face, but we don't know where exactly that sub-window is.

In this paper, we systematically investigate the effect of mis-aligned face images on face recognition systems. To make classifiers robust to the unavoidable face registration error, we formulate the face alignment problem within the multiple-instance learning framework. We then propose a novel multiple-instance based subspace learning scheme for face recognition tasks. In this algorithm, noisy training image bags are modeled as the mixture of Gaussians, and we introduce a supervised clustering method to iteratively select better subspace learning samples. Compared with previous methods, our algorithm does not require accurately aligned *training and testing* images, and can achieve the same or better performance as manually aligned face recognition systems. In this paper, we used the term *"**noisy images**"* to denote poorly aligned images.

1.1 Related Work

Researchers have been trying to overcome the sensitivity of subspace based face recognition algorithms to image alignment errors. Martinez [7] proposed a method to learn the subspace that represents the error for each of training images. Shan *et al.* [8] studied the effect of the mis-alignment problem, and for each training image they generated several perturbed images to augment the training set and thus modeling the mis-alignment errors. Compared with the aforementioned work, our algorithm requires the ground truth for neither the ***training*** set nor the ***testing*** set. Multiple-instance learning approach, on the other hand, such as MILBoost, was used in [9] for face detection problems. In their work, Viola *et al.* formulated the face detection problem as a multiple-instance learning approach, and AnyBoost was modified to adapt to multiple-instance learning condition. Several multiple-instance learning methods have been proposed, such as diverse density [10] and MI-SVMs [11]. Diverse density algorithm tries to find the area which is both of high density positive points and of low density negative points. kNN is adopted for multiple-instance learning by using Hausdorff distance in the work of Wang *et al.* [12].

2 Multiple-Instance Subspace Learning

2.1 Motivation

Given a limited set of noisy training images, we augment the training set by perturbing the training images. The augmented larger training set will normally cover more variations for each subject and thus model the alignment error, however, it could also introduce some very poorly registered faces into the training set, which will have negative effect for the learning process.

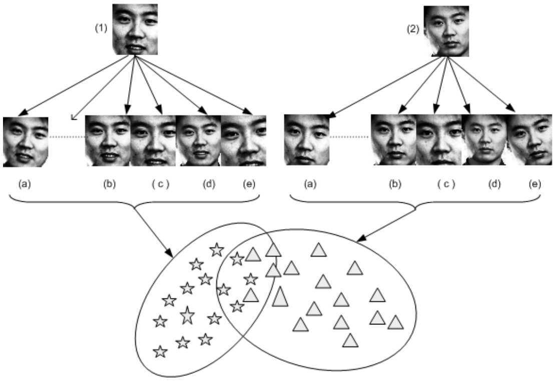


Fig. 2. Bags of Instances

Figure 2 shows two noisy training images (1) and (2). From each of the noisy image, we generate two bags each with multiple instances, denoted by $(a), (b), ...(e)$ in the figure. While image (b) and (d) will certainly benefit the training process, image (e) will most likely cause confusion for the classifier, since it could be more similar to other subject. As will be shown later, those very poorly registered images will indeed increase the recognition error. Thus given noisy training images, we must build algorithm that can automatically select those "good" perturbed images from training bags, and exclude those very poorly registered images from being selected.

2.2 Approximating the Constrained k-Minimum Spanning Tree

Excluding very poorly registered images from the noisy bags can be formulated within the multiple-instance learning framework. One assumption is that the good perturbed images from the same subject tend to be near to each other. The high density areas correspond to the good perturbed images, while the low density areas correspond to poorly perturbed images, and those are the bad images we want to exclude from the training set. As shown in figure 2, the good perturbed images will lie in the intersection area of the two bags. The idea is very similar to the diverse density approach used by Maron [10] for multiple-instance learning. Since the the perturbed noisy images have irregular distribution, we use non-parametric method to find out the high density area. Our non-parametric method is based on k-minimum spanning tree [13]: given an edge-weighted graph $G = (V, E)$, it consists of finding a tree in G with exactly $k < |V| - 1$ edges, such that the sum of the weight is minimal. In our face recognition application, the nodes will be the face image instances, and the edges represent the Euclidean distance between face image instances. The problem is known as NP-complete problem, and we don't need to get the exact solution. We used heuristic method to find out the approximate k-minimum spanning tree. Firstly, for each instance, we build its k-nearest neighbor graph. Among all the instances, we find the one with minimum k-nearest neighbor graph. Since the size of the neighbors is fixed by k, the one with minimum sum of k-nearest neighbor graph will have the highest density, and thus corresponds to the good perturbed image area. Although in this high density area, there will still exist some noisy images, those noisy images are identifiable and useful to our learning algorithm.

We also need to add the constraint to include at least one instance from each bag during the base selection phase. The idea is similar to that of MI-SVM. In MI-SVM, for every positive bag, we initialize it with the average of the bag, and compute the QP solution. With this solution, we compute the responses for all the examples within each positive bag and take the instance with maximum response as the selected training example in that bag. In our k-nearest neighbor graph algorithm, if some bag is far from other bags, using only the k-nearest neighbor graph to select training images may not include any instance from this isolated bag. We force the algorithm to accept at least one instance from every bag. If all the instances in a bag fall outside the most compact k-nearest neighbor graph, we select the instance with the minimum distance to the k-nearest neighbor graph.

The iterative multiple-instance based FisherFace [4] [5] learning procedure is shown in the following algorithm 1.

The learning procedure normally takes 2-3 iterations to converge. In our experiments, we use bag size of 25, i.e., each original training image is perturbed

Algorithm 1. Multiple-Instance Subspace Learning Algorithm

```
Input:
$S$: number of subjects
$N_s, s = 1...S$: number of noisy image for subject $s$
$R$: number of instances per bag
$K$: target number of nearest neighbors

Initialize $x_b^{(0)} = 1/R \sum_{i=1}^{R} x_{bi}^{(0)}$;
while Base is still changing do
    Compute Sufficient Statistics
    $x_b^{(t)} = \frac{1}{\mathcal{Z}} \sum_{g \in \mathcal{G}_s^{(t)}} x_{bg}^{(t)}$;
    Compute Multiple-Instance Eigenbase
    $m_s^{(t)} = \frac{1}{N_s} \sum_{b=1}^{N_s} x_b^{(t)}$
    $m^{(t)} = \frac{1}{\sum_{s=1}^{S} N_s} \sum_{s=1}^{S} N_s m_s^{(t)}$
    $S_W^{(t)} = \sum_{s=1}^{S} \sum_{b=1}^{N_s} \sum_{g \in \mathcal{G}_s^{(t)}} (x_{bg}^{(t)} - m_s^{(t)})(x_{bg}^{(t)} - m_s^{(t)})^T$
    $S_B^{(t)} = \sum_{i=s}^{S} N_s (m_i^{(t)} - m^{(t)})(m_i^{(t)} - m^{(t)})^T$
    $W^* = arg \max_{W} \frac{W^T S_B^{(t)} W}{W^T S_W^{(t)} W}$
    Base Selection
    $y = W^{*T} * x$
    Select good perturbed training samples $\mathcal{G}_s$ for each subject by finding the most
    compact $k$-nearest neighbor graph from projected subspace $y$.
end while
```

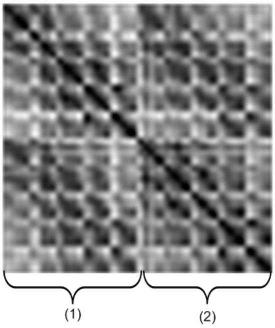


Fig. 3. Bag Distances Map

to generate 25 images. Each subject has 1-4 training images, and we take k as 60% of each subject's total number of perturbed noisy images.

To show that good perturbed images are similar to each other, figure 3 shows an example distance map for two bags (1) and (2). Each bag has 25 instances, which are generated by adding 25 random perturbations to a well-aligned image. The instances around the middle of the two bags have smaller perturbations, i.e., they are good perturbed images. In the distance map, the darker the color, the similar the two instances. From the graph we can see that an instance from bag (1) is not necessarily always nearer to instances in bag (1) than in bag (2), which means that two aligned different face images from one subject could be more similar than the same image to itself perturbed by noises. Also we can see that instances around the middle of bag (1) are more similar to those instances around the middle of bag (2), which means good perturbed images from the same subject are similar to each other, and thus confirmed our assumption.

2.3 Testing Procedure

During testing stage, we used the nearest neighbor algorithm as our classification algorithm. The distance metric we used is the modified Hausdorff distance. The Hausdorff distance provides a distance measurement between subsets of a metric space. By definition, two sets $\mathcal{A}$ and $\mathcal{B}$ are within Hausdorff distance of d of each other iff every point of $\mathcal{A}$ is within distance of d of at least one point of $\mathcal{B}$, and every point of $\mathcal{B}$ is within distance d of at least one point of $\mathcal{A}$. Formally, given two sets of points $\mathcal{A} = \{A_1, ..., A_m\}$ and $\mathcal{B} = \{B_1, ..., B_n\}$, the Hausdorff distance is defined as: $H(\mathcal{A}, \mathcal{B}) = \max\{h(\mathcal{A}, \mathcal{B}), h(\mathcal{B}, \mathcal{A})\}$, where $h(\mathcal{A}, \mathcal{B}) = \max_{A_i \in \mathcal{A}} \min_{B_j \in \mathcal{B}} \|A_i - B_j|$. This definition is very sensitive to outliers, so we used a modified version of the Hausdorff distance. In this paper, we take the distance of bag A and bag B as $H(\mathcal{A}, \mathcal{B}) = \min_{\mathcal{A}_i \in \mathcal{A}} \min_{\mathcal{B}_j \in \mathcal{B}} \|\mathcal{A}_i - \mathcal{B}_j\|$. For single instance probe and gallery testing case, we use the nearest neighbor method based on Euclidian distance in the subspace.

3 Experimental Results and Discussions

We used the well known FERET database [6] in our experiments. One reason to use this data set is that it's relatively a large database available, and the testing results will have more statistical significance. The training set, which is used to find the optimal FisherFace subspace, consists of 1002 images of 429 subjects, with all subjects at near-frontal pose. The testing set consists of the gallery set and the probe set. The gallery set has 1196 subjects, each subject has one near-frontal image with under normal lighting condition. The probe set has 1195 subjects, each subject has one image with the same condition as the probe set, but with different expressions. For comparison purposes, we have the ground truth positions of the two eye centers for training, probe and gallery images.

In this paper, we denote ***noisy bag*** as a bag generated from a noisy image, and ***aligned bag*** as a bag generated from a well-aligned image. We use "single" in comparison to bag.

Since we have many possible experimental setup combinations (training data, gallery data, probe data, noisy image, well-aligned image, single image and bag of images etc), we use table 1 and table 3 to explain our experimental setup.

3.1 Testing with Well-Aligned Training Data

To see how the introduction of the augmented training bags will affect the recognition performance, we first test on the well-aligned training data.

Table 1. Testing combinations for aligned training data

Base 1	single aligned training
Base 2	aligned bag training
Testing 1	single aligned gallery, single aligned probe
Testing 2	single aligned gallery, single noisy probe
Testing 3	aligned bag gallery, noisy bag probe
Testing 4	single aligned gallery, noisy bag probe

Table 2. Results comparison

	base 1	base 2
testing 1	0.9247	0.9749
testing 2	0.8795	0.9665
testing 3	0.9674	0.9849
testing 4	0.9431	0.9774

From table 2 we have the following notable observations:

- The recognition rate is always higher if we use aligned bag instead of single image as training data, which motivates the aforementioned perturbation based robust algorithms. However, it's not true anymore if we don't have well-aligned training data, i.e., we only have some noisy training images, and we add perturbations to generate noisy bags. Using the noisy bags as training data may not necessarily improve recognition performance, since the very poorly aligned images will confuse the classifier.
- If we take the baseline algorithm as the case of single aligned training, single aligned gallery and single aligned probe, then the rank-1 recognition rate for the baseline algorithm is 92.47%.
- If we use aligned bag probe and noisy bag probe, the rank-1 recognition rate is 96.74%, which is better than the baseline algorithm. It means adding perturbations to the gallery and probe set can make the algorithm robust to alignment errors.

3.2 Testing with Noisy Training Data

To show that if we don't have well-aligned training data, adding random perturbations to augment the training set may help much, we performed various experiments. More importantly, we also show that after selecting good perturbed images using our multiple-instance based scheme from the set of augmented data, the recognition performance improves a lot. Table 4 shows the testing results, and we have the following notable observations:

- When we use single noisy training image without adding perturbations (base 1), the recognition rate is very low for all the testing combinations. This indicates that the within-subject scatterness for poorly registered face in the training set is so high that they overlap with other subjects' clusters and

Table 3. Testing combination for noisy training data

Base 1	single noisy training
Base 2	Iteration 1, noisy bag training
Base 3	Iteration 3, noisy bag training
Testing 1	single aligned gallery, single aligned probe
Testing 2	single aligned gallery, single noisy probe
Testing 3	aligned bag gallery, noisy bag probe

Table 4. Results comparison

	testing 1	testing 2	testing 3
base 1	0.3213	0.1941	0.5431
base 2	0.9540	0.9364	0.9766
base 3	0.9690	0.9590	0.9833

lead to confusion for the classifiers. For Fisherface, it means the objective function it tries to minimize is ill-conditioned, which will lead to the failure of the algorithm.

- For base 2 case, we augment the noisy images by adding perturbations to generate noisy bags, then the recognition rate increases greatly compared to using noisy images directly.
- Base 3 shows that it's not good to treat all the instances from the noisy bags as the same. We used our multiple-instance based subspace learning method to remove those "bad" instances from the augmented noisy bags. The resulting training set increases the discriminative power of the classifier, but not to disperse the within subject cluster and cause confusion.
- Given only noisy training and probe set, we still achieved much higher recognition rate of 98.33% than the baseline algorithm of 92.47% as shown in table 2, and roughly the same as the optimal case of 98.49%, where all noisy bags are generated by perturbing the aligned images.

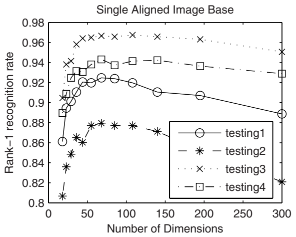


Fig. 4. Single Aligned Training Base

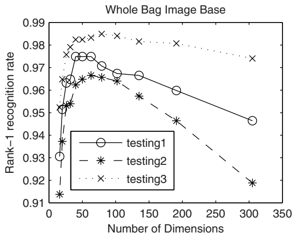


Fig. 5. Aligned Bag Training Base

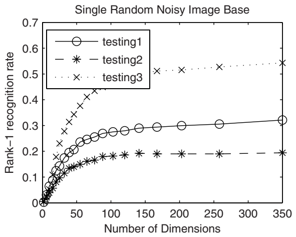


Fig. 6. Single Noisy Training Base

Figures 4 shows testing results with single aligned image as training data. Figure 5 shows testing results with aligned bag as the training data. Both figures show the change of recognition rate w.r.t. the change of the number of dimension used by FisherFace. In both cases, the recognition rate has the following order: testing 3 > testing 1 > testing 2, where all the testings have the same meaning as explained in table 1.

Figure 6 shows how noisy training images could affect the recognition rate. It's obvious that when the training set is not aligned very well, all the testing cases fail, including using probe bags and gallery bags. So it's very important to remove noisy training images from corrupting the training subspace.

Figure 7, 8 and 9 show recognition error rates on three different testing combinations. The testings have the same meaning as explained in table 3. Optimal 1 means training with aligned bags, and optimal 2 means training with aligned single images. Iter1 and Iter3 means the first iteration and the 3rd iteration of the base selection procedure. We can see that in all cases, the 3rd iteration results is better than the 1st iteration results. It supports our claim that extremely poorly registered images will not benefit the learning algorithm. We use our multiple-instance learning algorithm to exclude those bad training images from corrupting the training base. Also interestingly, in all tests, optimal 1 always performs worst, which indicates that by adding perturbations to the training base, even very noisy images, we can improve the robustness of learning algorithms. Note that in all cases, when the number of dimensions increases, the error rate will first decrease and then increase. Normally we get the best recognition rate using around the first 50 dimensions (account for 70% of total energy).

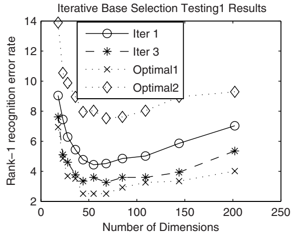


Fig. 7. Single aligned gallery, single aligned probe

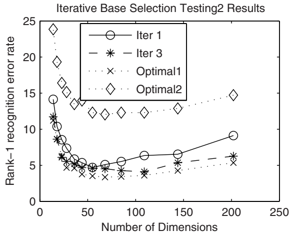


Fig. 8. Single aligned gallery, single noisy probe

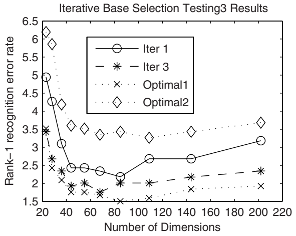


Fig. 9. Aligned bag gallery, noisy bag probe

4 Conclusions

In this paper, we systematically studied the influence of image mis-alignment on face recognition performance, including mis-alignment in training sets, probe sets and gallery sets. We then formulated the image alignment problem in the multiple-instance learning framework. We proposed a novel supervised clustering based multiple-instance learning scheme for subspace training. The algorithm proceeds by iteratively updating the training set. Simple subspace method, such as FisherFace, when augmented with the proposed multiple-instance learning scheme, achieved very high recognition rate. Experimental results show that even with the noisy training and testing set, the Fisherface learned by our multiple-instance learning scheme achieves much higher recognition rate than the baseline algorithm where the training and testing images are aligned accurately. Our algorithm is a meta-algorithm which can be easily used with other methods. The same framework could also be deployed to deal with illumination and occlusion problems, with different definition of training bags and training instances.

Acknowledgments

The research in this paper was partially supported by NSF CNS-0428231.

References

1. Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. In: Sommer, G., Daniilidis, K., Pauli, J. (eds.) CAIP 1997. LNCS, vol. 1296, pp. 456–463. Springer, Heidelberg (1997)
2. Kirby, M., Sirovich, L.: Application of the karhunen-loeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(1), 103–108 (1990)
3. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 586–591. IEEE Computer Society Press, Los Alamitos (1991)
4. Belhumeur, P.N., Hespanha, J., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
5. Etemad, K., Chellappa, R.: Discriminant analysis for recognition of human face images. In: Bigün, J., Borgefors, G., Chollet, G. (eds.) AVBPA 1997. LNCS, vol. 1206, pp. 127–142. Springer, Heidelberg (1997)
6. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
7. Martinez, A.: Recognizing imprecisely localized, partially occuded and expression variant faces from a single sample per class. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(6), 748–763 (2002)
8. Shan, S., Chang, Y., Gao, W., Cao, B.: Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, pp. 314–320 (2004)
9. Viola, P., Platt, J.C, Zhang, C.: Multiple instance boosting for object dection. In: Proceedings of Neural Information Processing Systems (2005)
10. Maron, O., Lozano-Perez, T.: A framework for multiple-instance learning. In: Proceedings of Neural Information Processing Systems, pp. 570–576 (1998)
11. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vecctor machines for multiple-instance learning. In: Proceedings of Neural Information Processing Systems, pp. 561–568 (2002)
12. Wang, J., Zucker, J.D.: Solving multiple-instance problem: A lazy learning approach. In: Proceedings of International Conference on Machine Learning, pp. 1119–1125 (2000)
13. Blum, C., Blesa, M.J.: New metaheuristic approaches for the edge-weighted k-cardinality tree problem. Computers and Operations Research 32(6), 1355–1377 (2005)