Open World Face Recognition with Credibility and Confidence Measures

Fayin Li and Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030
{fli, wechsler}@cs.gmu.edu

Abstract. This paper describes a novel framework for the Open World face recognition problem, where one has to provide for the Reject option. Based upon algorithmic randomness and transduction, a particular form of induction, we describe the TCM-kNN (Transduction Confidence Machine – kNearest Neighbor) algorithm for Open World face recognition. The algorithm proposed performs much better than PCA and is comparable with Fisherfaces. In addition to recognition and rejection, the algorithm can assign credibility ("likelihood") and confidence ("lack of ambiguity") measures with the identification decisions taken.

1. Introduction

The choices facing face recognition systems should include: ACCEPT, REJECT ("is not here"), and AMBIGUITY ("need more information"). The inclusion of the REJECT option, which corresponds to an **open world** of (**face recognition**) hypotheses, adds complexity to the whole process and makes face recognition much harder compared to the more traditional closed world biometric systems available today. In addition to seeking how similar or close some probe face image is to each subject in the face gallery set, one needs some measure of confidence when making any identification decision.

This paper describes a novel methodology for handling an open world of hypotheses, including the REJECT option, and provides the means to associate credibility and confidence measures with each of the decisions made regarding HumanID. The proposed methodology, based upon randomness concepts and transductive learning, is formally validated on challenging (varying illumination) and large overlapping FERET data sets.

2. Randomness and *p*-Values

Confidence measures can be based upon universal tests for randomness, or their approximation. A Martin-Lof randomness deficiency (Li and Vitanyi, 1997) based on such tests is a universal version of the standard statistical notion of ***p***-values. Universal tests for randomness are not computable and hence one has to approximate the ***p***-values using non-universal tests.

We use the ***p***-value construction in Proedrou et al. (2001) to define the quality of information. The assumption used is that data items are independent and are produced by the same stochastic mechanism. Given a sequence of proximities (distances) between the given training (gallery) set and an unknown sample (test) probe, one quantifies to what extent the (classification) decision taken is reliable, i.e., non-random. Towards that end one defines the *strangeness* of the unknown sample probe *i* with putative label *y* in relation to the rest of the training set exemplars as:

$$\alpha = \left(\sum_{j=1}^{k} D_{ij}^{y} \right) \left(\sum_{j=1}^{k} D_{ij}^{-y} \right)^{-1}$$

The strangeness measure is the ratio of the sum of the *k* nearest distances *D* from the same class (***y***) to the sum of the *k* nearest distances from all other classes **(–*y*)**. The strangeness of an exemplar increases when the distance from the exemplars of the same class becomes larger and when the distance from the other classes becomes smaller. A valid randomness test (Nouretdinov et al., 2001) defines then the ***p***-value measure of a test exemplar with a possible classification assigned to it as

$$p = \frac{f(\alpha_1) + f(\alpha_2) + \cdots + f(\alpha_m) + f(\alpha_{new})}{(m+1) f(\alpha_{new})}$$

where *f* is some monotonic non-decreasing function with *f*(0) = 0, e.g., *f*(α) = α, *m* is the number of training examples, and α_{new} is the strangeness measure of a new potential test probe exemplar c_{new}. An alternative definition available for the ***p***-value is $p(c_{new}) = \#\{i : \alpha_i \geq \alpha_{new}\}/(m+1)$ **.** Using the ***p***-value one can now predict the class membership as the one that yields the largest ***p***-value, which is defined as the *credibility* of the assignment made. The associated *confidence* measure, which is one minus the 2nd largest ***p***-value, indicates how close the first two assignments are. The confidence value indicates how improbable the classifications other than the predicted classification are and the credibility value shows how suitable the training set is for the classification of that testing example. One can compare the top ranked assignments, rather than only the first two assignments, and define additional confidence criteria. Both the credibility and confidence measures allow the face recognition module to adapt to existing conditions and act accordingly.

3. Transduction Confidence Machine (TCM)- kNN

Another form of learning, beyond induction, is transduction. Given an unlabeled validation test, in addition to the training set, the task now is to estimate the class for each unlabeled pattern in order to construct the best classifier rule for both the training and validation sets.

TCM-*k*NN Algorithm

```
for $i = 1$ to $m$
    Find and store $D_i^y$ and $D_i^{-y}$
end for
Calculate the alpha strangeness values for all the training exemplars
Calculate the similarity dist vector as the distances of the new exemplar from all the training exemplars
for $j = 1$ to $c$ do
    for every training exemplar $t$ classified as $j$ do
        if $D_{ti}^{j} > dist(t)$ , $i = 1\ldots k$, recalculate the alpha value of exemplar $t$
    end for
    for every training exemplar $t$ classified as non-$j$ do
        if $D_{ti}^{-j} > dist(t)$, $i = 1\ldots k$, recalculate the alpha value of exemplar $t$
    end for
    Calculate alpha value for the new exemplar classified as $j$
    Calculate $p$-value for the new exemplar classified as $j$
end for
Predict the class with the largest $p$-value
Output as confidence one minus the 2nd largest $p$-value
Output as credibility the largest $p$-value
```

The constraints on the (geometric) layout of the learning space and the search for improved classification margins are addressed in this paper using algorithmic randomness (Vovk et al., 1999), universal measures of confidence randomness (Vovk et al., 1999), and transductive confidence (learning) machines (TCM) (Proedrou et al., 2001). The experimental data presented later on that validates our approach, is based on TCM-kNN which is an augmented TCM using locality-based evidence, e.g., the k-Nearest Neighbors (kNN) concept.

The similarity distances *dist* (in script) used are shown next. Given two *n*-dimensional vectors $X, Y \in \Re^n$, the distance measures used are defined as follows:

$$d_{L1}(X,Y) = |X - Y| = \sum_{i=1}^{n} |X_i - Y_i| \qquad d_{L2}(X,Y) = \|X - Y\|^2 = (X - Y)^T (X - Y)$$

$$d_{\cos}(X,Y) = -\frac{X^T Y}{\|X\|\|Y\|} \qquad d_{Dice}(X,Y) = -\frac{2X^T Y}{\|X\|^2 + \|Y\|^2} = -\frac{2X^T Y}{X^T X + Y^T Y}$$

$$d_{Jaccard}(X,Y) = -\frac{X^TY}{\|X\|^2 + \|Y\|^2 - X^TY} = -\frac{X^TY}{X^TX + Y^TY - X^TY}$$

$$d_{Mah+L2}(X,Y) = (X-Y)^T \Sigma^{-1}(X-Y) \qquad d_{Mah+\cos}(X,Y) = -\frac{X^T\Sigma^{-1}Y}{\|X\|\|Y\|}$$

where Σ is the scatter matrix of the training data. For PCA, Σ is diagonal and the diagonal elements are the (eigenvalues) variances of the corresponding components. The Mahalanobis + L1 distance defined only for PCA is

$$d_{Mah+L1}(X,Y) = \sum_{i=1}^{n} \left(\frac{|X_i \quad Y_i|}{\sqrt{\lambda_i}} \right)$$

4. Data Collection

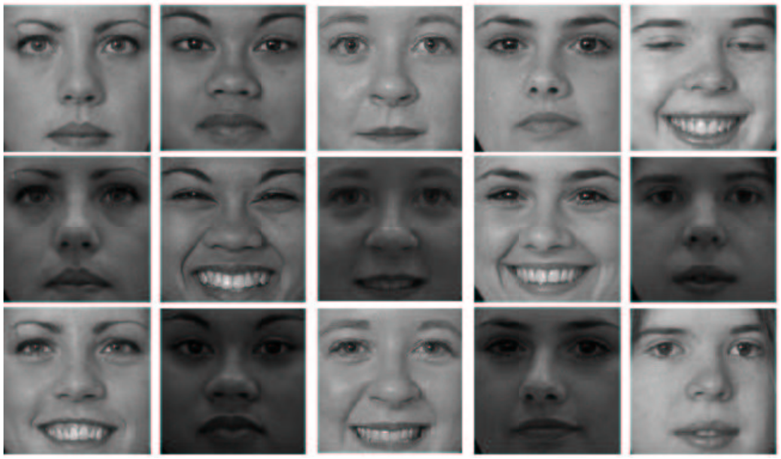

Figure 1. Face Images

Our data set is drawn from the FERET database, which has become a de facto standard for evaluating face recognition technologies (Phillips et al., 1998). The data set consists of 600 FERET frontal face images corresponding to 200 subjects, which were acquired under variable illumination and facial expressions. Each subject has three images of size 256x384 with 256 gray scale levels. Face image normalization is carried out as follows: first, the centers of the eyes of an image are manually detected, then rotation and scaling transformations align the centers of the eyes to predefined locations, and finally, the face image is cropped to the size of 128x128 to extract the facial region. The extracted facial region is further normalized to zero mean and unit variance. Fig. 1 shows some exemplar images used in our experiments that are already cropped to the size of 128x128. Each column in Fig. 1 corresponds to one subject. Note that for each subject, two images are randomly chosen for training, while the remaining image (unseen during training) is used for testing.

The normalized face images are processed to yield 400 PCA coefficients, according to eqs. 7 – 9 from Liu and Wechsler (2002), and 200 Fisherfaces using FLD (Fisher Linear Discriminant), according to eqs. 10 – 12 from Liu and Wechsler (2002) on a reduced 200 dimensional space PCA space.

5. Open World Face Recognition Algorithms

Open World TCM-kNN Algorithm

```
Calculate the alpha values for all the training exemplars
for $i$ = 1 to $c$ do
    for every training exemplar $t$ classified as $i$ do
        for $j$ = 1 to $c$ and $j$ != $i$ do
            Assume $t$ is classified as $j$, which should be rejected
            Recalculate the alpha value for all the training exemplars classified as non-$j$
            Calculate alpha value for the exemplar $t$ classified as $j$
            Calculate $p$-value for the exemplar $t$ classified as $j$
        end for
        Calculate the $P_{\max}$, $P_{\text{mean}}$ and $P_{\text{stdev}}$ (standard deviation) for the $p$-value of exemplar $t$
        Calculate the $PSR$ value for exemplar $t$: $PSR = (P_{\max} - P_{\text{mean}})/P_{\text{stdev}}$
    end for
end for
Calculate the $mean$, $stdev$ (standard deviation) for all the $PSR$ values
Calculate the $mean$ + 3*$stdev$ as $threshold$ for rejection
Calculate the distances of the probe exemplar from all the training exemplars
for $i$ = 1 to $c$ do
    Calculate alpha value for the probe exemplar classified as $i$
    Calculate $p$-value for the probe exemplar classified as $i$
end for
Calculate the largest $p$-value $max$ for the probe exemplar
Calculate the $mean$ and $stdev$ for the probe $p$-value without $max$
Calculate the $PSR$ value for the probe exemplar: $PSR = (max - mean)/ stdev$
Reject the probe exemplar if its $PSR$ is less than or equal to the $threshold$.
Otherwise predict the class with the largest $p$-value
```

Open World {PCA, Fisherfaces} Algorithm

```
for  $i = 1$ to $m$
    Find the maximum intra-**within**-distance and minimum inter-**between**-distance
end for
Calculate the mean and standard deviation for all maximum intra-distances and minimum inter-distances: $mean_{\text{intra}}$, $mean_{\text{inter}}$, $stdev_{\text{intra}}$ and $stdev_{\text{inter}}$
Calculate $mean_{\text{intra}} + 3* stdev_{\text{intra}}$ as the lower bound of the threshold
Calculate $mean_{\text{inter}} - 3* stdev_{\text{inter}}$ as the upper bound of the threshold
Choose the *threshold* based on the lower and upper bound
Calculate the distances of the probe exemplar from all the training exemplars
Find the minimum distance $dist_{\text{min}}$ of the probe exemplar
If $dist_{\text{min}} >= threshold$, then reject the probe exemplar
Else predict the class with the minimum distance $dist_{\text{min}}$
```

6. Experimental Results

We found that the best similarity distances for PCA and Fisherfaces are {Mahalanobis + (L_1, L_2 or cos)} and {cosine, Dice, Jaccard, (Mahalonobis + cos)}, respectively. Those distances are used in our experiments. The experiments reported were carried out on the data described in the previous section. Both the gallery and the probe sets consist of 100 subjects, and the overlap portion between the two sets on the average 50 subjects. The recognition rate is the percentage of the subjects whose probe is correctly recognized or rejected.

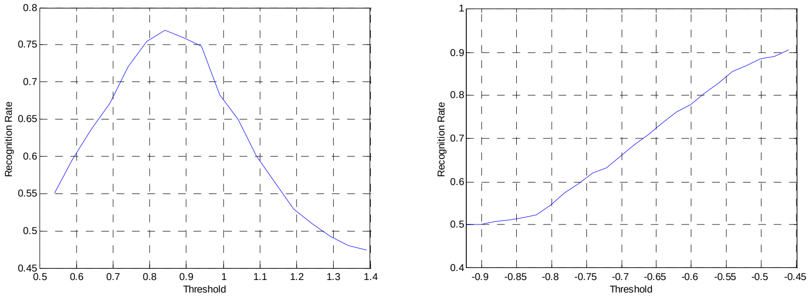

Figure 2. The Recognition Rate vs Threshold: PCA (Left) and Fisherfaces (Right)

Open World (PCA and Fisherfaces) Face Recognition

The data was randomly chosen, the same experiment was run 100 times, and Fig. 2 shows the mean recognition rates for different thresholds. The distance measurements for PCA and Fisherfaces, which yield the best results, are Mahalanobis + L_2 and cosine, respectively. Fig. 2 shows that the best recognition rate for PCA is 77% if the threshold can be chosen correctly, while for Fisherfaces is 91% if the threshold is chosen as its upper bound. The standard deviation for the best recognition rate for PCA and Fisherfaces are 3.7% and 2.4%, respectively.

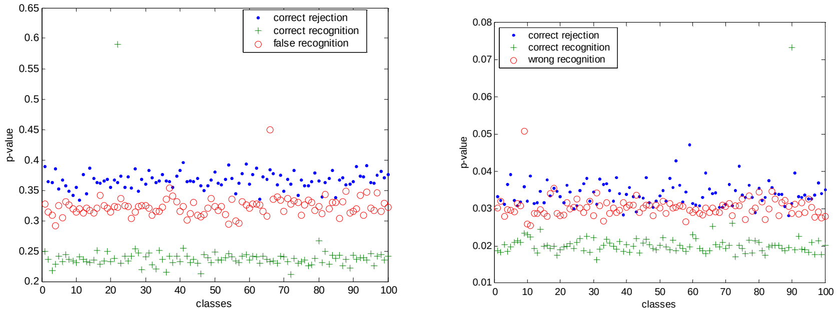

Figure 3. Test ***p***-value distribution of rejection, correct and false recognition using PCA with (Mahalanobis + L2) distance (Left) and Fisherfaces with cosine distance (Right)

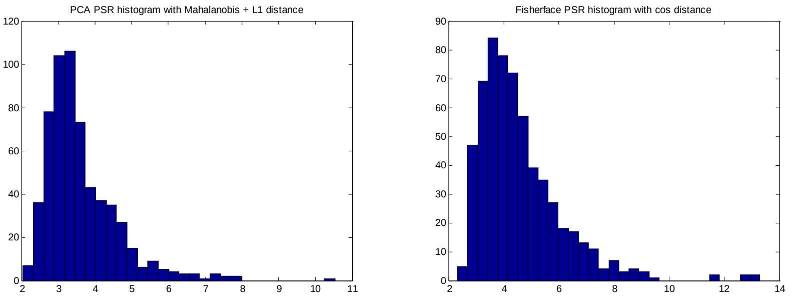

Figure 4. The *PSR* value histogram: PCA (Left) and Fisherfaces (Right)

TCM-kNN

The data used are either the (400) PCA or (200) Fisherface components, and k = 1. The threshold is computed according to the algorithm described in Sect. 5 based on the training exemplars. The ***p***-value distributions shown in Fig.3 indicate that the test *PSR* values are useful for rejection and recognition. Recognition is driven by large *PSR* values. The best recognition rate using PCA components is 87.87% using the Mahalanobis + L_1 distance, and its standard deviation is 3.0%. The threshold is 6.57 computed from the PSR histogram shown in Fig. 4 (left). The best recognition rate using Fisherface components is 90% using the cosine distance, and its standard deviation is 2.7%. The threshold is 9.20 computed from the *PSR* histogram shown in Fig. 4 (right).

TCM-kNN provides additional information regarding the credibility and confidence in the recognition decision taken. The corresponding 2D distribution for correct and false recognition is shown in Fig. 5, where one can see that false recognition, for both the PCA and Fisherfaces components, shows up at low values.

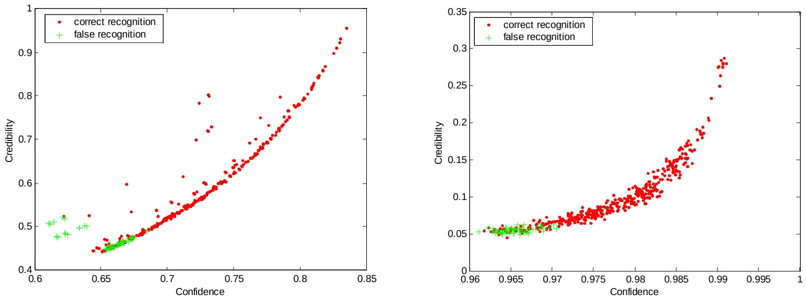

Figure 5: Distribution of confidence and Credibility: PCA (left) and Fisherfaces (right).

7. Conclusions

We introduced in this paper a new face recognition algorithm suitable for open world face recognition. The feasibility and usefulness of the algorithm has been shown on varying illumination and facial expression images drawn from FERET. Furthermore, both credibility and confidence measures are provided for both the recognition and rejection decisions. We plan to use those measures for optimal training of the face recognition system, such that the composition and size of the training set are determined using active rather than random selection

8. References

1. A. Gammerman, V. Vovk, and V. Vapnik (1998), Learning by Transduction. In *Uncertainty in Artificial Intelligence*, 148 – 155.
2. M. Li and P. Vitanyi (1997), *An Introduction to Kolmogorov Complexity and Its Applications*, 2ed. , Springer-Verlag.
3. C. Liu and H. Wechsler (2002), Gabor Feature Based Classification Using the Enhnaced Fisher Linera Discriminant Model for Face Recognition, *IEEE Trans. on Image Processing*, Vol. **11**, No. 4, 467 – 476.
4. I. Nouretdinov, T. Melluish, and V. Vovk (2001), Ridge Regression Confidence Machine, *Proc. 17th Int. Conf. on Machine Learning.*
5. P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss (1998), The FERET Database and Evaluation Procedure for Face Recognition Algorithms, *Image and Vision Computing*, Vol.**16**, No.5, 295-306.
6. K. Proedrou, I. Nouretdinov, V. Vovk and A. Gammerman (2001), Transductive Confidence Machines for Pattern Recognition, TR CLRC-TR-01-02, Royal Holloway University of London.
7. V. Vovk, A. Gammerman, and C. Saunders (1999), Machine Leraning Applications of Algorithmic Randomness, *Proc. 16th Int. Conf. on Machine Learning.*