Watch List Face Surveillance Using Transductive Inference

Fayin Li and Harry Wechsler

Department of Computer Science
George Mason University, Fairfax, VA, 22030, USA
{fli, Wechsler}@cs.gmu.edu

Abstract. The *open set* recognition task, most challenging among the biometric tasks, operates under the assumption that not all the probes have mates in the gallery. It requires the availability of a *reject* option. For face recognition open set corresponds to the *watch list face surveillance* task, where the face recognition engine must *detect* or *reject* the probe. The above challenges are addressed successfully in this paper using *transduction*, which is a novel form of inductive learning. Towards that end, we introduce the *Open Set TCM-kNN* algorithm, which is based upon algorithmic randomness and transductive inference. It is successfully validated on the (small) watch list task (80% or more of the probes lack mates) using FERET datasets. In particular, Open Set TCM-kNN provides on the average 96% correct detection / rejection and identification using the PCA and/or Fisherfaces components for face representation.

1 Introduction

The *open set* recognition task, most challenging among the biometric tasks, operates under the assumption that not all the probes have mates in the gallery. The above task, which has to detect the presence or absence of some biometric within the gallery while correctly finding the identity for the same biometric, requires the availability of a *reject* option. For face recognition open set corresponds to the *watch list face surveillance* task, where the face recognition engine must first find out if an individual probe is, or is not, on the watch list. If the probe is on the watch list, the face recognition engine must then *identify / recognize* the individual.

The performance index for correctly detecting and identifying an individual on the watch list is called the *detection and identification/ recognition rate*. "Typically, the watch list task is more difficult than the identification [recognition] or verification task alone. For the best system [FRVT2002 face recognition engine] using a watch list of 25 people, the detection and identification rate is 77%. Systems achieve better performance for a smaller watch list. If the impetus of the watch list application is to detect and identify [recognize] the "most wanted" individuals, the watch list should be kept as small as possible" (Phillips et al., 2003). The watch list face surveillance task is addressed successfully in this paper using *transduction*, which is a novel form of inductive learning (Vapnik, 1998). Towards that end, we introduce a novel approach,

the *Open Set TCM-kNN* (Transduction Confidence Machine – *k* Nearest Neighbor), which is based upon algorithmic randomness and transductive inference. Open Set TCM-kNN is successfully validated on the (small) watch list task (80% or more of the probes lack mates) using FERET datasets characterized by challenging (varying illumination and changing facial expression) face images. In particular, Open Set TCM-kNN provides on the average 96% correct detection / rejection and identification / recognition using the PCA and/or Fisherfaces components for face representation.

2 Transduction Confidence Machine (TCM) – kNN

Confidence measures can be based upon universal tests for randomness, or their approximation. 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 Gammerman et al. (1998) and Proedrou et al. (2001) to define information quality. Given a sequence of proximities (distances) between the given training (gallery) set and a probe *i*, the *strangeness* of *i* with putative label *y* in relation to the rest of the training set exemplars is defined as:

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

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)*. Based on the *strangeness*, a valid randomness test (Nouretdinov et al., 2001) defines the *p*-value measure of a probe with a possible classification assigned to it as

$$p = \frac{f(\alpha_1) + f(\alpha_2) + \cdots + f(\alpha_m) + f(\alpha_{new})}{(m+1) f(\alpha_{new})} \quad (2)$$

where *f* is some monotonic non-decreasing function with $f(0) = 0$, e.g., $f(\alpha) = \alpha$, *m* is the number of training examples, and α_{new} is the strangeness measure of a new test probe exemplar c_{new} with a possible classification. An alternative definition available for the *p*-value is $p(c_{new}) = \#\{i: \alpha_i \geq \alpha_{new}\}/(m+1)$.

Based on strangeness and *p*-value definitions above, Proedrou et al. (2001) have proposed that TCM-kNN (Transduction Confidence Machine-*k* Nearest Neighbor) serves as a formal transduction inference algorithm. If there are *n* classes in the training data, there are *n p*-values for each probe exemplar. Using the *p*-value one can now predict the class membership as the one that yields the largest *p*-value. This is defined as the *credibility* of the assignment made. The associated *confidence* measure, which is either the 1st largest *p*-value or one minus the 2nd largest *p*-value, indicates how close the first two assignments are. 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 engine to adapt to existing conditions and act accordingly.

We have used several well-known similarity measures to evaluate their effect on different face representation (PCA and Fisherfaces). In those similarity

measures, L_1 defines the city-block distance, L_2 defines the Euclidean distance. Cosine, Dice, Overlap and Jaccard measure the relative overlay between two vectors. L_1, L_2 and cosine can also be weighted by the covariance matrix of training data, which leads to the Mahalanobis related distances. Our findings indicate that Mahalanobis related similarity distances are superior to others when expressive features (driven by PCA) are used; while overlay related similarity distances are superior when discriminating features are used. *Open Set TCM-kNN*, which is an instantiation of TCM-kNN appropriate for the watch list face surveillance problem, is described in the next section.

3 Open Set TCM-kNN

The *watch list face surveillance* problem operates under the *open set* assumption that not all the probes have mates in the gallery and it requires a *reject* option. The standard *Open Set* PCA and Fisherfaces classifiers learn the operational threshold from the intra- and inter-distance (similarity) distribution of each training exemplar. The statistics of intra-distance distribution set the lower bound of the threshold and the statistics of inter-distance distribution set the upper bound. As the minimum distance of the new exemplar to the prototypes of each subject becomes closer to or larger than the upper bound, the more likely the new testing exemplar will be rejected. Our experiments have shown that face recognition performance varies according to the threshold chosen. It is, however, not easy to determine the threshold needed for rejection even if the lower and upper bound are known. In addition, one needs to make a strong assumption that the distribution of the similarity distances is similar across different training and test sets. Those problems are overcome by the *Open Set TCM-kNN* classifier and the rationale is explained next.

Given a new exemplar, the *p*-value vector output from Open Set TCM-kNN gives the likelihoods that the new exemplar belongs to each subject in the training data. If some *p*-value significantly outscores the others, the new exemplar can be mated to the corresponding subject ID; otherwise, the probe is equally likely to be mated to each class and it should be rejected due to its ambiguity. If the top ranked (highest *p*-values) choices are very close to each other and outscore the other choices, the top choice should be accepted, i.e., should not be rejected, but its recognition be questionable due to the ambiguity involved. The ambiguity is measured by the *PSR* (peak-side-ratio) value and it determines if the test exemplar should be rejected or not. The Open Set TCM-kNN algorithm is:

```
Open Set TCM-kNN Algorithm

Calculate the alpha values for all 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 values for all the training
    exemplars classified as non-j;
  Calculate alpha value for t if it is classified as j
  Calculate p-value for t if it is classified as j
  end for
```

```
    Calculate the P_max, P_mean and P_stdev(standard deviation)of
     exemplar t's p-values
    Calculate the PSR value of t: PSR=(P_max- P_mean)/P_stdev
  end for
end for
Calculate the mean, stdev (standard deviation) for all the
PSR values
Calculate the rejection threshold = mean + 3*stdev
Calculate the distance of the probe s from all training
exemplars
for i = 1 to c do
 Calculate alpha value for s if it is classified as i
 Calculate p-value for s if it is classified as i
end for
Calculate the largest p-value max for s
Calculate the mean and stdev of probe p-values without max
Calculate the PSR value for the probe exemplar:
  PSR = (max-mean)/stdev
Reject the probe s if its PSR is less than or equal to the
threshold
Otherwise predict the class with the largest p-value
Output as prediction confidence 1st largest p-value or one
minus the 2nd largest p-value
Output as prediction credibility the largest p-value
```

In the algorithm, the threshold for rejection is learned from the training data set based on transductive inference. For each training exemplar X its mates are removed from the training data set, X is now treated as a probe, and Open Set TCM-kNN outputs the p-values for X. The corresponding PSR value for X is $PSR = (P_{max} - P_{mean}) / P_{stdev}$. Very low PSR values result for X since it should be rejected because it lacks any mates. The training exemplars' PSR distribution provides thus a robust method for deriving the operating threshold.

4 Data Collection and Experimental Results

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 normalized (zero mean and unit variance) images of size 128 x 128 with 256

gray scale levels. Each column in Fig. 1 corresponds to one subject. The normalized face images are processed to yield 400 PCA coefficients and 200 Fisherfaces using FLD (Fisher Linear Discriminant) on the reduced 200 dimensional PCA space.

TCM-kNN yields similar performance with kNN on the closed set face recognition problem except that its output includes confidence and credibility for each decision made. The best similarity distances for PCA and Fisherfaces representations were found to be {Mahalanobis + (L_1, L_2 or cosine)} and {cosine, Dice, Jaccard, (Mahalonobis + cos)}, respectively. Only Mahalanobis + L_1 and cosine distances, which yield slightly better performance than the other distances, are used by Open Set TCM-kNN for the watch list face surveillance experiments reported next.

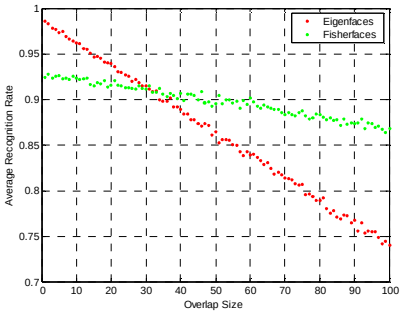

Figure 2. Mean Detection and Identification Rates vs. Overlap Size

Fig. 2 shows the mean detection and recognition rates for PCA and Fisherfaces components for different overlap percentages (between the gallery [watch list] and the probe sets) using Mahalanobis + L_1 and cosine distance, respectively. There are 200 subjects (see Sect. 4), two images are used for training while the third image is available for testing, the size for both the watch list [gallery] and the probe sets is 100 subjects, and the overlap between the watch list and the probe set varies from 0 to 100 subjects. The detection and recognition rate is the percentage of subjects whose probe is correctly detected (or rejected) and identified. We report the average results obtained over 100 randomized runs. The performance goes down, almost linearly, as the overlap between the watch list and the probe sets increases. One gets overall better performance when using Fisherfaces components compared to PCA components; for small overlap, i.e., the small watch list size, which is most relevant, PCA yields slightly better performance than Fisherfaces. The explanation for the observed performance is that as both the size of watch list increases it becomes more difficult to detect and identify individuals on the watch list. This confirms the FRVT2002 findings (Phillips et al., 2003).

The next experiment uses the Open Set and Open Set TCM-kNN classifiers on small watch lists, whose size varies from 10 to 40 subjects, and reports the mean performance (detection and identification) rates obtained over 100 randomized runs. Let the watch list size be k subjects, each of them having 2 (two) images in the gallery. Then there are $600 - 2k$ face images in the probe set, k stands for the number of subjects on the watch list and $3 \times (200 - k)$ stands for the number of face images that come from subjects that are not on the watch list.

Table 1 and 2 shows the mean performance of Open Set and Open Set TCM-kNN algorithms for different watch list sizes. For watch list size k, the accuracy [detection and identification rate] is (*average correct rejection + average correct recognition*) / (*600 - 2k*). The numerical results when the number of subjects on the watch list is k should be interpreted as follows. The average results are better the closer the *correct rejection* number is to $3 \times (200 - k)$, the closer the *correct recognition* number is to the watch list size and the higher the *accuracy* is.

Table 1. Mean Performance of Open Set Algorithm

Watch List Size	Eigenfaces (Mah+L_1 distance)			Fisherfaces (Cosine distance)		
	Average Correct Reject	Average Correct Recognition	Accuracy	Average Correct Reject	Average Correct Recognition	Accuracy
10	546.38	5.67	95.18%	550.00	7.98	96.20%
20	494.08	12.29	90.42%	502.36	15.93	92.55%
30	448.11	18.98	86.50%	457.83	24.41	89.30%
40	407.78	22.94	82.83%	416.05	32.24	86.21%

Table 2. Mean Performance of Open Set TCM-kNN

Watch List Size	Eigenfaces (Mah+L_1 distance)			Fisherfaces (Cosine distance)		
	Average Correct Reject	Average Correct Recognition	Accuracy	Average Correct Reject	Average Correct Recognition	Accuracy
10	536.62	8.67	94.02%	544.64	8.99	95.45%
20	522.38	16.97	96.31%	522.35	17.88	96.47%
30	495.65	25.16	96.44%	493.33	26.76	96.31%
40	467.39	33.42	96.31%	464.56	35.39	96.14%

The lower and upper bounds for the reject threshold for the Open Set algorithm are computed from the training set. The average correct rejection and identification numbers are recorded by varying the threshold from the lower bound to the upper bound. Since the watch list size is much smaller than the number of the subjects that should be rejected, the accuracy will be very high even if all the probes are rejected. Therefore, the average correct reject and recognition numbers as well as the detection and identification accuracy are needed to evaluate the performance of the algorithm. As an example, Table 1 shows that only about 60% (PCA) and 80% (Fisherfaces) of the subjects from the watch list (*average correct recognition /watch list size*) are recognized correctly even when the accuracy is high (about 96% for $k = 10$). Table 1 shows the results with the best threshold considering those three factors. As the watch list size increases, both the average rejection number, further away from $3 \times (200 - k)$, and the accuracy, drop dramatically.

Table 2 shows the average performance of Open Set TCM-kNN algorithm for different watch list sizes. PCA is very close to Fisherfaces in overall performance. If

the data in the table is examined carefully, PCA is a little better than Fisherfaces with the reject option while Fisherfaces is a little better than PCA when the decision of identification is made. Open Set TCM-kNN is much better than Open Set algorithm, when the correct rejection and correct recognition number as well as the accuracy are taken into account, especially when the watch list size is large. The overall performance for Open Set TCM-kNN, which keeps almost constant as the watch list size increases, is thus much more stable than the overall performance displayed by the Open Set algorithm.

The difference in performance shown by Fig. 2 and Table 2 indicates that the gallery size is also an important factor affecting algorithm performance. In Fig. 2 the gallery [watch list] size is always 100 subjects and only the overlap size between the gallery and probe sets changes, while in Table 2 the gallery size [watch list] varies according to *k*.

5 Conclusions

This paper introduced the *Open Set TCM-kNN* (Transduction Confidence Machine – *k* Nearest Neighbor) for the *watch list face surveillance* task. Open Set TCM-kNN, which is based upon algorithmic randomness and transductive inference, has been successfully validated on the (small) watch list task (80% or more of the probes lack mates) using FERET datasets characterized by challenging (varying illumination and changing facial expression) face images. In particular, Open Set TCM-kNN provides on the average 96% correct detection / rejection and identification accuracy using the PCA and/or Fisherfaces components for face representation.

References

1. A. Gammerman, V. Vovk, and V. Vapnik (1998), Learning by Transduction. In *Uncertainty in Artificial Intelligence*, 148 – 155.
2. I. Nouretdinov, T. Melluish, and V. Vovk (2001), Ridge Regression Confidence Machine, *Proc. 17th Int. Conf. on Machine Learning.*
3. S. Pankanti, R. M. Bolle and A. Jain (2000), Biometrics-the-Future of Identification, *Computer*, Vol. **33**, No. 2, 46 – 49.
4. 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.
5. P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi and M. Bone (2003), Face Recognition Vendor Test 2002 – Overview and Summary.
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. Vapnik (1998), *Statistical Learning Theory*, Wiley.