Dependence Characteristics of Face Recognition Algorithms

Andrew Rukhin
University of Maryland Baltimore County
Baltimore, MD
rukhin@math.umbc.edu

Patrick Grother, P. Jonathon Phillips, Stefan Leigh, Alan Heckert
NIST, Gaithersburg, MD
{pgrother, jphillips, leigh, heckert}@nist.gov

Elaine Newton
Rand Corporation
Pittsburgh, PA
enewton@rand.org

Abstract

Nonparametric statistics for quantifying dependence between the output rankings of face recognition algorithms are described. Analysis of the archived results of a large face recognition study shows that even the better algorithms exhibit significantly different behaviors. It is found that there is significant dependence in the rankings given by two algorithms to similar and dissimilar faces but that other samples are ranked independently. A class of functions known as copulas is used; it is shown that the correlations arise from a mixture of two copulas.

1 Introduction

The subset of images misrecognized by face recognition algorithms is large and the causes of failure not well understood. Although it is known that image specific properties like low contrast or shadow or viewing angle cause difficulties it is also widely accepted that certain faces are difficult to recognize [5]. It is also known that certain algorithms are vulnerable to specific properties of the images. However, this paper shows that even algorithms that are known to perform well do not succeed and fail on the same images. We introduce statistical measures of this heterogeneity.

The FERET [4] protocol defined a biometric non-specific framework for evaluation of recognition systems. An algorithm compares all images of a target set, T, and a query set, Q, and reports a similarity matrix, S, whose elements, s_{ij}, state some measure of the sameness of the identity of the individual in the i-th image of T and the j-th image of Q. Because subsets of T and Q have different subject- or acquisition-specific properties certain elements of S, corresponding to gallery, G \subseteq T, and probe sets, P \subseteq Q, may be used to make post-hoc performance estimates on targeted recognition tasks. This study draws on S matrices archived under the FERET program. Those tests were notable for the large number of subjects, 1196.

Here a standard gallery and probe set known as dupII is considered. Each of the 234 probe images was taken between 540 and 1031 days after its gallery match. The gallery contains 1196 images. The best recognition rate of 52% was achieved by the USC algorithm. Figure 1 shows that only around 2% of images are matched correctly by all 15 algorithms, and about 27% are classified correctly by just one. This suggests that fusion of two or more algorithms may yield superior performance. This study considers measures of the correlation between recognition systems appropriate to determining complementary behavior.

2 Algorithm Comparison

Algorithms may be compared by considering their similarity matrices; the probes in P specify the columns and the galleries G select the rows. Because the origin of the s_{ij} is generally unknown (they are propriety and/or undoc-
2.1 Rank Correlation

The Spearman rho, \( \rho_S \), and Kendall tau, \( \tau_T \), coefficients are the most widely used measures of rank correlation. Their respective definitions are:

\[
\rho_S = \frac{\sum_{i=1}^{N} (R(i) - \bar{R})(S(i) - \bar{S})}{\sqrt{\sum_{i=1}^{N} (R(i) - \bar{R})^2 \sum_{i=1}^{N} (S(i) - \bar{S})^2}},
\]

\[
\rho_K = \frac{\sum_{i,j} \text{sign}((R(i) - R(j))(S(i) - S(j)))}{N(N-1)/2},
\]

where \( R(i) \) and \( S(i), i = 1, \ldots, N, \) are the ranks given to the \( N \) gallery images by two algorithms using one probe image, and \( \bar{S} = \bar{R} = (N + 1)/2 \) is the mean of the ranks \( 1, \ldots, N \).

The advantage of these definitions is that by using the Central Limit Theorem one can establish asymptotic normality of both coefficients as \( N \to \infty \).

The \( \rho_S \) and \( \rho_K \) coefficients for seven algorithms are shown in Table 1. The values are averages over the 254 probes. Note that all values are positive, the smallest being 0.13, and the largest values are found in the lower right 3x3 sub-matrix corresponding to three algorithms that share common image normalization implementations. Algorithms five to seven differ only in the distance metric applied after PCA feature extraction. Note also that the two MIT algorithms are correlated.

Empirically \( \rho_S \) and \( \rho_K \) are linearly related for small \( \rho \). As discussed below, the Kendall tau statistic is smaller than the Spearman rho values.

2.2 Partial Rank Correlation

We considered the weak correlations of Table 1 to be the result of including essentially independent rankings for large \( R \) and \( S \). That is, beyond some small fraction of \( N \) the algorithms order images that are dissimilar to the probe image arbitrarily. To study this hypothesis we restricted the correlation to the most similar gallery images, i.e. those with rank \( \leq K \). The corresponding characteristics are based on the so-called partial rankings [1] [2].

Analogues of the Spearman rho and Kendall tau coefficients for partially ranked data can be obtained by using this modification of the ranks [1]:

\[
\tilde{R}(i) = \begin{cases} 
R(i) & \text{if } R(i) \leq K \\
\frac{N+K+1}{2} & \text{otherwise,}
\end{cases}
\]

which retains the ranks of the \( K \) closest images and substitutes the average of the largest \( N - K \) ranks for the others, so that \( \sum_{i=1}^{N} \tilde{R}(i) = \sum_{i=1}^{N} R(i) = N(N+1)/2 \). Using eq. 3 in the formulae for \( \rho_S \) and \( \rho_K \) gives expressions for partial rank correlations. Their denominators reduce to expressions involving only \( K \) and \( N \). The appendices give formulae for P-values.

Figure 2 plots both \( \rho_S \) and \( \tau_K \) as functions of \( K \). The rightmost values are in Table 1. The results agree with those from the three other standard FERET recognition tasks. All values are positive; the absence of negative values makes sense for algorithms that are intelligently designed. The P-values are all small, and are therefore not graphed indicating significant correlation. The relative ordering of algorithms is preserved for both statistics at all \( K \); this seems to be a desirable property given our intention to identify correlated algorithms. Finally, the general form of the curves is an initial brief decline followed by an almost monotonic rise for \( \rho_S \) and a plateau for \( \tau_K \). This difference is explained in reference 3. For small \( K \), the majority of the values \( \tilde{R} \)
and $\tilde{S}$ are equal to the replacement value in eq. (3). Thus the behavior of the statistics is determined by only a small number of the ranks; after that, as $K$ increases the curves become flatter and smoother. As we will see below, this effect is consistent with uncorrelated $R(i)$ and $\tilde{S}(i)$ for $i \gg 1$. It seems that the aggregative nature of the partial rank correlation statistics with increasing $K$ limits their usefulness. Instead local correlation measures must be considered.

2.3 Rank Co-occurrence

When two algorithms rank a set of gallery images one can consider the number of gallery images for which both algorithms mutually rank images that are similar to the probe image. Over the 105 pairs of algorithms mutually rank images that are similar to the probe image. Figure 3 shows that co-occurrence exhibits a characteristic "bathtub" form. At the left hand end, the algorithms behave independently. Specifically, for a fixed window length $K$, define the co-occurrence statistic $T(u)$ as

$$T(u) = \text{card } \{i: u \leq R(i) < u + K, u \leq S(i) < u + K\}. \quad (4)$$

If $R$ and $S$ are independent and uniformly distributed permutations, then the random variable $T$ has a hypergeometric distribution with moments

$$ET = \frac{K^2}{N}, \quad \text{Var}(T) = \frac{K^2(N-K)^2}{N^2(N-1)}. \quad (5)$$

$T(u)$ values are obtained, as before, by averaging over all probes. Figure 3 shows that co-occurrence exhibits a characteristic "bathtub" form. At the left hand end, the algorithms mutually rank images that are similar to the probe image. Over the 105 pairs of 15 algorithms with $K = 50$, $T(1)$ has median 14 and maximum 42. This behavior is expected and desirable. At the other end, the algorithms mutually reject gallery images that are dissimilar to the probe. This is a novel result. Here $T(N-K)$ has median 7 and maximum 41. The P-values are very small, indicating $T(u)$ significantly above the expected value; in the central region the curves are flat over a wide range. For pairs of algorithms with large $\rho_K$, for example, the plateau is significantly above the expected value; for low correlation the algorithms behave independently.

3 Discussion

For an explanation of the behavior of the co-occurrence statistic we assume that the cumulative distribution functions $F_X$ and $F_Y$ of the random similarity scores $X$ and $Y$ from two algorithms are continuous. Then both $F_X(X)$ and $F_Y(Y)$ have the uniform distribution on interval $(0,1)$. Moreover with fractional ranks $(u/N \rightarrow z$ and $K/N \rightarrow a$) co-occurrence becomes

$$G(z) = \lim_{N \rightarrow \infty} N^{-1} ET(u) = \int \frac{a}{z} \int \frac{a}{z} f(u,v) \, du \, dv$$

where $f(u,v)$ is the joint distribution function of $(U = F_X(X), V = F_Y(Y))$, and $F_X(X)$ and $F_Y(Y)$ is the corresponding density. A function $C(u,v)$ is called copula [3]. It is known that for such functions $C(u,1) = C(1,u) = u, C(u,0) = C(0,u) = 0$. Copulas satisfy certain inequalities like

$$\max(u + v - 1, 0) \leq C(u,v) \leq \min(u,v).$$

Their importance for studying rank is their invariance under monotone transformations: if $\alpha$ and $\beta$ are strictly increasing functions, then $C_{\alpha(X),\beta(Y)}(u,v) = C_{X,Y}(u,v)$.

So, in terms of copula $C$, the approximate value of the co-occurrence statistic's mean as a function of $z$ is

$$G(z) = C(a + z, a + z) + C(z, z) - C(z, a + z) - C(a + z, z) \quad (6)$$

with $0 < z < 1 - a$ and $G(z) \geq 0$. For independent variables $U$ and $V$, $C(u,v) = uv$, so that $G(z) = a^2$.

As copulas are known to capture the properties of the joint distribution which are invariant under monotone transformations, all traditional measures of dependence including Spearman's rho, and Kendall's tau can be expressed in terms of copulas [3].

This suggests that the co-occurrence corresponds to a mixture of two copulas: one at $(1,1)$ (large ranks) and another with more of the mass at $(0,0)$ (small ranks). Then the joint density $f(u,v)$ is the weighted sum of two densities, and the set $\{(u,v) : f(u,v) \geq c\}$ can be represented as a union of two sets: $C_0$ and $C_1$ which are star-shaped about $(0,0)$ and $(1,1)$.

The joint distribution with such a copula satisfies the definition of left tail monotonicity of one random variable $U$ in
another random variable V. Namely, \( P(U \leq u \mid V \leq v) \) is a nondecreasing function of \( v \) for any fixed \( u \). Each of these conditions implies positive quadrant dependence \( P(U \leq u, V \leq v) \geq P(U \leq u)P(V \leq v) \), i.e. \( C(u, v) \geq uv \), and, under these conditions, \( \rho_S \geq \rho_K > 0 \) as noted in section 2.1.

In practice, tail monotonicity means that the strongest correlation between algorithms occurs for both large and small rankings. Thus, all algorithms behave somewhat similarly not only by assigning the closest images in the gallery, but also by deciding which gallery object is most dissimilar to the given image. It seems that for most of the pairs the statistic. akin to co-occurrence 5 arising under the independence assumption.

To confirm tail monotonicity properties in our situation, we can estimate the density \( f(u, v) \), \( 0 < u, v < 1 \), by using the statistic, akin to co-occurrence

\[
f(u, v) = N^{-1} \text{card} \left\{ i : u < R(i)/N \leq u + a, \quad v < S(i)/N \leq v + a \right\}
\]

for a sufficiently small \( a \). The surface of \( f \), if plotted, has two peaks at \((0, 0)\) and \((1, 1)\).

4 Conclusion

We have shown that correlated behavior between algorithms exists at both small and large ranks, and not in between. We have defined co-occurrence as an informative measure of this and demonstrated, using copulas, that tail monotonicity properties explain the characteristic bathtub shape.

Appendix A P-values for \( \rho_S \) and \( \rho_K \)

P-values for the hypothesis that both permutations \( R \) and \( S \) are independent and uniformly distributed are obtained as follows.

**Theorem 4.1** Assume that both random permutations \( R \) and \( S \) are independent and uniformly distributed. When \( N \to \infty \) and \( K \to \infty \), the distribution of the coefficient \( s = \sqrt{N - 1} \rho_S \) converges to the standard normal distribution \( N(0, 1) \).

Thus for sufficiently large \( K \), \( P(|s| \geq z) \approx 2 - 2\Phi(z) \), from which an approximate P-value is: \( 2 - 2\Phi(|z|) = \text{erfc} \left( \frac{|z|}{\sqrt{2}} \right) \), where \( \Phi \) denotes the distribution function of standard normal distribution, and \( \text{erfc} \) is the complementary error function. Simulation shows the normal approximation to be satisfactory only when \( K \geq 0.1N \), otherwise the distribution of both partial correlation coefficients is very skewed to the right. It is then better to use \( T(1) \) (eq. 4) instead of \( \tilde{\rho} \) itself. Under the null hypothesis above, the random variable \( T \) has a hypergeometric distribution with parameters \((N, K, K)\) with \( K < N/2 \). The values of this statistic are recorded in \( n \) trials, and the frequencies \( \nu_k \) of each value \( k, k = 0, 1, \ldots, K \) are found. Then they are conjoined by the classical goodness-of-fit \( \chi^2 \)-statistic,

\[
\chi^2 = \sum_{k=0}^{K} \frac{[\nu_k - nP(T = k)]^2}{nP(T = k)}.
\]

Under the independence hypothesis \( T \) has the \( \chi^2 \)-distribution with \( K \) degrees of freedom. A difficulty here is that some of these probabilities are very small, and some pooling into larger classes is needed.

The P-values for this discrete distribution can be found from the formula

\[
1 - P \left( \frac{K \chi^2}{2} \right),
\]

where \( P(u, x) = \Gamma \left( \frac{x}{2} \right) e^{-u} u^{x-1} \) is the incomplete gamma function. This formula only works well for \( K \leq 100 \).

Appendix B P-values for \( T(u) \)

The classical goodness-of-fit \( \chi^2 \)-statistic, is evaluated on the basis of frequencies \( \nu_k \) of each value \( k, k = 0, 1, \ldots, K \) as above. An alternative way is to use normal approximation of the hypergeometric distribution, according to which the standardized values, \( z_T = (T - ET)/\sqrt{var(T)} \), are approximately standard normal. Then the approximate P-values can be found from the formula \( 2 - 2\Phi(\lvert z_T \rvert) \).

References