Utilizing Template Diversity for Fusion Of Face Recognizers

Sergey Tulyakov Nishant Sankaran Srirangaraj Setlur
Venu Govindaraju
Center for Unified Biometrics and Sensors
University at Buffalo, NY, USA
tulyakov,ns6,setlur,govind@buffalo.edu

Abstract

If multiple face images are available for the creation of person's biometric template, some averaging method could be used to combine the feature vectors extracted from each image into a single template feature vector. Resulting average feature vector does not retain the information about image feature vector distribution. In this paper we consider the augmentation of such templates by the information about diversity of constituent face images, e.g. sample standard deviation of image feature vectors. We consider the theoretical model describing the conditions of the usefulness of template diversity measure, and see if such conditions hold in real life templates. We perform our experiments using IARPA face image datasets and deep CNN face recognizers.

1. Introduction

Traditionally, biometric template construction methods consider a single image or a sensor scan for feature extraction. But for some biometric modalities a series of images or scans could be readily available instead of a single image or scan. This is the case for face recognition considered in the current paper - both single images and video sequences of a person can be utilized for the construction of biometric template. Recent facial biometrics databases, e.g. IJB-A [9] or YTF [19], include the video sequences of the facial images and stimulate future research into construction of biometric templates from scan sequences.

It appears that there are two main directions of creating templates from the sets of images. The first direction is the feature vector averaging or aggregation [2, 21], where the feature vectors extracted from single images are averaged by some algorithm to produce a single feature vector representing the template load; such feature vector is typically of the same dimension as source image feature vectors. The second direction is to keep separate feature vectors for each image in the template, and during the matching perform set-to-set feature vector comparisons; the final score is the aggregate of the comparison scores between individual image feature vectors in two templates [23]. Since the second approach was delivering worse performance than the first one on considered data set in our preliminary experiments, we concentrate on the first approach in current paper.

When the feature vectors of separate images are aggregated into a single feature vector of the template, useful matching information might be lost. In this paper we are interested in the spread of original face feature vectors. For example, the template might originate from a single face image or from a video sequence of face images containing essentially the same face with the same pose and light conditions. Alternatively, the template might be constructed from a variety of face images in different poses, illuminations and facial expressions. Intuitively, the second case contains more information than the first one, and we want our *template diversity* to reflect this. Note that during template construction and feature vector aggregation we end up with a single feature vector in any case; we can assume that in the second case our feature vector is more reliable and the template diversity measure should reflect this information.

In addition to deriving the formulas for template diversity (section 3) and determining their usefulness in the facial recognition system (section 6), we conduct a series of experiments on simulated model to find out the conditions under which the diversity measures perform best (section 4), and check if such conditions hold for real life facial templates (section 5).

2. Prior Work

The template diversity measure can be viewed as one in the category of biometric template quality measures, and a significant number of works have been presented over

2019 IEEE 5^{th} International Conference on Identity, Security, and Behavior Analysis (ISBA)
978-1-7281-0532-1/19/\$31.00 ©2019 IEEE

the years trying to improve the performance of biometric matchers by incorporating some template quality information. Typically, some measure of quality is extracted from image or biometric scanner data, and it is subsequently fused with original biometric comparison scores [18]. Instead of fusing quality measures with matching scores, more detailed matching algorithm modifications could be used, such as the modification of algorithm steps based on quality measures [3] or selection of good quality images for enrollment and matching [1].

The examples of more complex approaches to incorporating quality measures into biometric algorithms also exist. Kryszczuk and Drygajlo [10] present a framework for building a classifier ensemble which incorporates the quality measures along with matching scores. Poh and Kittler [15] perform clustering of biometric samples based on the vector of quality measures, and each cluster gets its own decision or classification parameters.

Some research also exists into deriving automatic measures of quality. For example, Grother and Tabassi [6] tried to use the distributions of matching scores to derive quality measures. Yang et al. [21] train neural network to perform the aggregation of image-wise feature vectors into a single template; it appears that the network is trained to implicitly give bigger weights for better quality images.

One common feature of most presented approaches is that the quality measure relates to a particular image or a biometric scan; in many cases, especially in fingerprint matching research, a separate algorithm is developed to estimate the quality of the image. If template is composed of a number of images, then the image qualities could be averaged, or some images could be given more weight according to quality measure. In contrast, in our work we are looking at the templates composed of many images, and try to measure how well such sets reflect the variability of biometric observations. Our current work derives diversity measures directly from the feature vectors of images included into template, but it could be possible to extend this work by deriving diversity measures directly from images by separate algorithms.

The importance of diversity in images used for template creation is frequently emphasized in another face recognition technique consisting in augmentation of the training sets or original imagery by differently altered images [20]. Our approach does not change the diversity of the template or training set, but only adds its measure to the matching process.

Our work is also connected to the work on score fusion methods incorporating user specific fusion functions [14] or decision thresholds [8]. The methods presented in such works either assume that there is a sufficient number of templates for a single user to estimate the user specific distribution of genuine comparison scores, or use some parametric models to estimate the user specific changes in score distributions from the set of impostor scores. In contrast to these works, we derive diversity measures internally for a given template using its images, and not performing comparisons with other templates.

3. Diversity Measure Construction

Suppose the facial template is constructed using N facial images, and each facial image produces a feature vector $\boldsymbol{f}_n$. We will assume that the template will be represented by the mean of these feature vectors:

$$\bar{\boldsymbol{f}} = \frac{1}{N} \sum_{n=1\ldots N} \boldsymbol{f}_n \tag{1}$$

We want the diversity measure to reflect the spread of individual image facial feature vectors around the mean, and thus we will define the template diversity measure as

$$d = \sqrt{\frac{1}{N-1} \sum_{n=1\ldots N} ||\bar{\boldsymbol{f}} - \boldsymbol{f}_n||^2} \tag{2}$$

Note, that if $\boldsymbol{f}_n$ were one-dimensional vectors, then this definition would have coincided with the traditional definition of sample standard deviation. Although scatter matrices represent a more straightforward extension of standard deviation concept to multidimensional space, their use in current problem might pose difficulty since number of images in the templates is usually less than the dimension of feature vectors. Instead, we utilize eq. 2, which have been used before in different applications and it is termed as radial standard deviation [7].

We hypothesize that the usefulness of particular diversity measure will greatly depend on the statistical distribution of face image feature vectors. Thus, for different data sets and feature vector extraction methods other diversity measures could perform better and should be used instead. In this paper we also performed experiments using mean absolute measure or Gini's mean difference [22]:

$$d = \frac{1}{N(N-1)} \sum_{k,n=1\ldots N, k\neq n} ||\boldsymbol{f}_k - \boldsymbol{f}_n|| \tag{3}$$

Although such measure of diversity, along with related Gini index, is more frequently used in different applications, it performed a little worse than the diversity of eq. 2 in our experiments.

4. Diversity Measure Simulations

In order to understand the impact of utilizing the diversity measures on the performance of face matchers, we performed a series of simulation experiments. Although, the performance improvement on real life data sets is most important factor in utilizing diversity measures, the simulations allow us to change the parameters of the model and understand the conditions when improvements occur.

We employ the following model of face feature vector distributions. Suppose that the feature vectors are located in L dimensional space. Let $\boldsymbol{m}_i$ be the master feature vector of person i, and suppose that all feature vectors of person i are generated using normal distribution $\boldsymbol{f}_{i,n} \sim \mathcal{N}(\boldsymbol{m}_i, \sigma_i \boldsymbol{I})$. Next, suppose that the master feature vectors of all persons are normally distributed around origin: $\boldsymbol{m}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma \boldsymbol{I})$. Finally, we define the common standard deviation $\sigma = 1$ and we make individual person's standard deviations randomly uniformly distributed on interval $\sigma_i \sim U(.5 - \alpha, .5 + \alpha)$.

Note that depending on the parameter α we would have less or more variation in the individual person's standard deviation of feature vectors. In turn, this will result in less or more dispersion of the generated person's feature vectors, and thus in smaller or greater template diversity measure. Therefore, this model would allow us to investigate the connection between assumed person's individual degree of face variation, resulting changes in diversity measures, and the benefits of utilizing diversity measures.

During the simulations we randomly generate the person's N feature vectors $\boldsymbol{f}_{i,n}$ according to above formulas and take the mean of these feature vectors as person's template (eq. 1), $\bar{\boldsymbol{f}}_i$. To calculate genuine matching scores, we randomly generate one more feature vector of the same person, $\boldsymbol{f}_{i,probe}$, and calculate a matching score as a distance: $s_{gen} = ||\bar{\boldsymbol{f}}_i - \boldsymbol{f}_{i,probe}||$. To calculate impostor score, we randomly generate a feature vector of some other person, $\boldsymbol{f}_{j,probe}$, and get corresponding distance as score : $s_{imp} = ||\bar{\boldsymbol{f}}_i - \boldsymbol{f}_{j,probe}||$. The diversity measure d for a particular template is generated according to formula 2.

The performance of the simulated system without diversity measure can be simply evaluated by generating the sets of genuine and impostor matching scores and by constructing ROC curves using these sets of scores. To observe the effect of utilizing diversity measure, we generate genuine and impostor samples as pairs $\{s_{gen}, d\}$ and $\{s_{imp}, d\}$, and approximate likelihood ratios $LR(s,d) = \frac{p_{gen}(s,d)}{p_{imp}(s,d)}$ in a grid $s \times d$ by accumulating samples in corresponding grid bins.

Tables 1 and 2 contain the results of simulation experiments; each table cell contain EER value of original system performance (without diversity measure) and performance of the system fusing matching score with the diversity measure. The number of simulation samples for each run (10^9) is chosen so that the EER is precise to approximately 10^{-5}. We performed the simulations using two most frequently used in face recognition distance measures - Euclidean and Cosine, and considered different values of N (number of images in the template) and α. Also we set the dimension of feature space $L = 5$.

α	$N=2$	$N=3$	$N=4$	$N=5$
$\alpha = 0$	8.83 8.83	7.76 7.76	7.21 7.21	6.88 6.88
$\alpha = 0.1$	9.35 9.31	8.27 8.19	7.72 7.61	7.38 7.25
$\alpha = 0.2$	10.74 10.30	9.62 8.96	9.04 8.26	8.69 7.83
$\alpha = 0.3$	12.49 11.25	11.31 9.70	10.69 8.94	10.31 8.48
$\alpha = 0.4$	14.30 12.08	13.07 10.49	12.42 9.73	12.02 9.28

Table 1. Improvements from utilizing diversity measures in simulated systems with Euclidean distance based matching scores. Each cell presents original matcher performance (top) and performance of original score fused with diversity measures (bottom) (% EER).

α	$N=2$	$N=3$	$N=4$	$N=5$
$\alpha = 0$	10.49 9.98	9.54 8.80	9.03 8.17	8.71 7.78
$\alpha = 0.1$	10.62 10.05	9.66 8.85	9.15 8.21	8.84 7.82
$\alpha = 0.2$	11.00 10.24	10.03 8.99	9.52 8.32	9.20 7.91
$\alpha = 0.3$	11.59 10.52	10.61 9.18	10.08 8.49	9.75 8.07
$\alpha = 0.4$	12.31 10.81	11.32 9.42	10.78 8.72	10.43 8.30

Table 2. Improvements from utilizing diversity measures in simulated systems with Cosine distance based matching scores. Each cell presents original matcher performance (top) and performance of original score fused with diversity measures (bottom) (% EER).

Some interesting properties could be observed from the experiments. First, the performance improvement can exist even if $\alpha = 0$, i.e. if there is no variation in each person's feature vectors and each person's face feature vectors are distributed the same way around person's master feature vector. This is the case with cosine distance scores. At the same time, Euclidean distance score simulations show that diversity might have no effect if there is no variation in each person's feature vector distributions. Second, the effect of utilizing template diversity measure seems to be bigger if there is a variation in the distribution of feature vectors of each person. For example, for $N = 5$ and Cosine distance calculation we get improvement around 1% for $\alpha = .1$, and around 2% for $\alpha = .4$. Finally, it appears that the diversity measure mitigates the effect of increasing variation between person's face feature vectors distributions. Thus if

α increases, then the error rate for system utilizing diversity measure increases much slower than the error rate of the original system.

5. Testing for Template Diversity

The simulation tests performed in section 4 show that the impact of using template diversity measures is more pronounced if there is a variation in the distribution of individual face image feature vectors around person's master, or mean, feature vector. Moreover, there could be no benefit of using template diversity if there is no such variation. Therefore, it would be interesting to verify the presence of such variation in real face data before utilizing diversity measure for score fusion.

The statistical tests on homogeneity of variance seek to verify if the samples in different groups of population are generated using same variance values, and we will use one of such tests, Levene's test [11], for our purpose. In our case we define the groups as the face feature vectors of individual persons. The test's null hypothesis is that all groups have samples generated with the same variance; for simulation model of the previous section it would mean that all σ_i are equal, and $\alpha = 0$.

The test's results are presented in table 3. In this table, the test values are computed as [13]:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i.})^2} \tag{4}$$

where N_i is the number of images in the template of person i, k is the number of considered templates, $Z_{ij} = ||\bar{\boldsymbol{f}}_i - \boldsymbol{f}_{i,j}||$, $\boldsymbol{f}_{i,j}$ is the j-th feature vector of the template of person i, $\bar{\boldsymbol{f}}_i$ is the mean of $\boldsymbol{f}_{i,j}$, $\bar{Z}_{i.}$ is the mean of Z_{ij} taken over feature vectors of template i, $\bar{Z}_{..}$ is the mean of Z_{ij} taken over all feature vectors of all persons. The critical values of F-distribution are calculated using the program on https://www.waterlog.info/f-test.htm.

Our data set and face feature vector extractors are described in section 6. Since there is a guideline for a minimum number of samples in each group, we consider only those facial templates, which contain at least M ($M = 5, 10, 15$) images or feature vectors. Additionally, since used data set contains both single images and video sequences, and the frames from the same video will produce similar feature vectors spoiling the test's results, we left a single random frame from each video in the templates. The total number of resulting templates and feature vectors are given table 3.

Since the test values in all cases are bigger than the corresponding critical values (at .01 significance level) of the tests, the null hypothesis of equal variance is rejected in all cases. According to the simulation analysis of section 4, the use of diversity measures should be beneficial for our system.

Note that the results presented in this section seem to imply the existence of variation in the distributions, and in particular, variance, of facial images for different persons. Thus, for one person with bigger variance more diverse face imagery could be generated, and for another person with smaller variance less diverse imagery could only be generated. This property of facial images could potentially be exploited not only for the definition of template diversity measures, but for other face recognition related tasks, e.g. creating facial models.

6. Experiments

Section 5 suggests that there is a variability in the distributions of feature vectors for different persons in the considered face recognizers, and section 4 suggests that we should see a performance increase in our systems when we try combine the diversity measure with matching scores. But, since the precise distribution of feature vector for real life data is not known, the performance improvement is not guaranteed and can only be verified by fusion experiments.

We conduct our experiments on the IARPA Janus Benchmark-A (IJB-A) dataset [9], and on the later IARPA Janus Benchmark-C (IJB-C) [12] dataset, which is a superset of the original IJB-A. The testing protocols specify gallery and probe templates with different numbers of constituent face images and video frames (from 1 to more than 100). Thus, these sets suits well for our task of investigating the proposed template diversity measures.

Three different deep CNN face recognizers are used to extract feature vectors from each database face image: CNN1 [2], CNN2 [17], and CNN3 [16]. All three recognizers construct templates by averaging the image feature vectors and calculate matching scores using cosine distance between averaged feature vectors, and in this paper we do it the same way. Note, that all three referenced method perform additional feature vector embedding (Joint Bayesian Embedding for CNN1 and Triplet Probabilistic Embedding for CNN2 and CNN3), which we omitted for the experiments in this paper.

Since cosine distance is used for matching these templates (it performs better than Euclidean distance), we also used cosine distance for calculating template diversity measures ($||.||$ of eqs. 2 and 3). Also, since cosine distance cd is not truly a distance, but rather a confidence value and has a range from -1 to 1, we replaced it by $1 - cd$ while calculating diversity measures. After calculating diversity measures and face recognition scores we employ a traditional backpropagation neural network to fuse them. Thus, our fusion is represented as a function $F(s_1, s_2, s_3, d_g, d_p)$, where s_i is the comparison score output by CNN i, d_g is the diversity measure computed for gallery template, and

Min # of images	Total Templates	Total Samples	Critical Value	Test Value CNN1	Test Value CNN2	Test Value CNN3
15	26	595	1.81	10.53	2.05	5.24
10	60	973	2.30	6.53	4.38	5.81
5	365	2754	1.20	3.95	3.23	4.20

Table 3. Result of Levene's test on homogeneity of variances.

d_p is the diversity measure computed for probe template. For baseline performance with no diversity measures we do not include any diversity measure into neural network inputs: $F(s_1, s_2, s_3)$. And in order to separately judge the benefits of gallery and probe diversity measures, we train fusion networks accepting corresponding sets of parameters: $F(s_1, s_2, s_3, d_g)$ and $F(s_1, s_2, s_3, d_p)$. In all cases, we used same training sets and same network architecture (3 layer fully connected perceptron) with the exception of the number of input parameters.

It might be possible to further increase the impact of utilizing template diversity by deploying more complicated fusion architectures, as it is done in other works utilizing template quality measures [5]. But in our work we deployed such rather straightforward fusion approach possibly providing a more objective comparison on the benefits of utilizing different diversity measures.

Table 4 contains the results of experiments on IJB-C dataset fusing all three CNN recognizer scores with the diversity measures calculated from either gallery or probe templates, or both of them. Since IJB-C dataset does not have separate training subset, we employ bootstrap training and testing procedure. Thus, in each bootstrap iteration we randomly select about a half gallery and a half probe templates for training and other halves for testing. After making sure that there is no intersection between training and testing sets (no same person), we obtain sets of approximately 5000 probe and 900 gallery templates. The bootstrap procedure is repeated 100 times and the mean and 95% confidence interval results are reported in the table.

Table 5 contains the results of experiments on IJB-A dataset. We follow the IJB-A testing protocol defining 10 splits of dataset into training, testing gallery and testing probe subsets. Since we need the samples of genuine and impostor training scores to train the fusion networks, and IJB-A protocol does not provide training gallery and training probe subsets, we perform random selection of such subsets from training set, and perform training/testing bootstrap experiments 10 times for each split. The means and 95% confidence interval of all 100 (10 splits * 10 bootstraps) experimental performance measures are reported in the table.

Our diversity measures for the experiments are calculated using eqs. 2 and 3 with the following modifications. First, we average the feature vectors of images or frames having the same media identifier (the database metadata contains this information) and obtain media ID based clusters. At the second step, we use these clusters as feature vectors $\boldsymbol{f}_n$ to calculate diversity measures by eqs. 2 and 3. This approach is consistent with the reference methods of template averaging in [2, 17, 16]. Otherwise, if we treat the video frames and separate images of the template with equal weights, then diversity measures give only very small improvements over baseline methods. We hypothesized that in this case the video frames produce large numbers of close feature vectors, which influences the calculation of proper template diversity measures. Also note, that the diversity measures derived using different CNNs are highly correlated, and give approximately the same benefits during fusion. Thus, the fusion experiments presented in this section utilized only the diversity measures derived a single CNN3 [16].

Table 5 contains also the results of augmenting the template adaptation method of comparison score calculation [4] with our template diversity measures. For a given template consisting of a set of person's images, the template adaptation method constructs a linear SVM separating this set from a large set of negative samples, or a reference set of face images disjoint from that person's images. The comparison score between that template and unknown template is calculated as a margin of unknown template's averaged feature vector calculated with the trained template's SVM. Separate SVMs are trained for both gallery and probe templates, and the final comparison score is an average of gallery's SVM evaluated at probe's feature vector, and probe's SVM evaluated at gallery's feature vector. Consistent with the results of [4], the template adaptation method did provide the improvement over our baseline cosine distance based score calculation, with the biggest improvements observed at relatively big FAR values. Since the template adaptation method utilizes the sets of images of particular template to build template specific linear SVMs, and incorporates SVM margins into score calculations, we speculated that template adaptation method implicitly incorporates the characteristics of template's image distributions, i.e. template diversity. But in our experiments, we saw that the addition of diversity measures to the template adaptation comparison scores still has significant benefits to the system performance.

Overall, the experiments of this section suggest that per-

Diversity Method	FAR	No Diversity	Gallery d	Probe d	Probe& Gallery d
Mean Ave Dist	.1%	$95.33 \pm .09$	$95.52 \pm .08$	$95.50 \pm .09$	$95.60 \pm .08$
	.01%	$90.42 \pm .15$	$90.70 \pm .13$	$90.54 \pm .15$	$90.96 \pm .12$
Radial SD	.1%	$95.33 \pm .09$	$95.58 \pm .08$	$95.49 \pm .09$	$\mathbf{95.76 \pm .09}$
	.01%	$90.42 \pm .15$	$90.95 \pm .14$	$90.88 \pm .14$	$\mathbf{91.18 \pm .14}$

Table 4. Performance (% TAR at FAR=.1% and at FAR=.01%) of systems merging three considered CNN face recognizers and different template diversity measures on IJB-C dataset.

Matching method	FAR	No Diversity	Mean Ave Dist	Radial SD
Cosine	.1%	93.40 ± 0.19	$\mathbf{93.75 \pm 0.17}$	93.58 ± 0.15
Distance	.01%	88.82 ± 0.53	$\mathbf{89.71 \pm 0.45}$	89.60 ± 0.41
Template	.1%	94.37 ± 0.18	$\mathbf{94.94 \pm 0.11}$	94.78 ± 0.12
Adaptation	.01%	88.91 ± 0.92	91.23 ± 0.28	$\mathbf{91.67 \pm 0.33}$

Table 5. Performance (% TAR at FAR=.1% and at FAR=.01%) of systems merging three considered CNN face recognizers and different template diversity measures on IJB-A dataset.

formance improvements from using template diversity measures are somewhat limited, but consistent and agree with the theoretical analysis showing that improvements are very likely. Moreover, some crafting and accounting for data peculiarities might help to construct better performing diversity measures. Radial standard deviation method of eq. 2 seems to perform better than mean average distance method of eq. 3 for media id cluster based diversity measures.

7. Conclusion

In this paper we achieved the following goals:

- We presented two possible ways to define the diversity of facial templates.
- We conducted simulation tests showing the impact of variation in person's feature vector distributions on the effectiveness of diversity measure utilization.
- We showed that real life facial images have such variation (inhomogeneity of variances of feature vectors belonging to different persons).
- We presented the results of fusing the diversity measures with original deep CNN matching scores.

The performed experiments showed that the improvements are consistent, and some experimentation with the diversity calculation formula might be needed to achieve best benefits. The template diversity measure can also readily complement other traditional template quality measures to achieve superior system performance.

Future research directions could include the construction of template diversity measures by using some auxiliary information, either extracted by separate algorithms or given as a an image metadata, e.g. face pose information, or constructing a trainable algorithm for calculating diversity measures.

Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] A. J. Abboud and S. A. Jassim. Biometric templates selection and update using quality measures. In *SPIE Defense, Security, and Sensing*, volume 8406, pages 8406–09. SPIE, 2012.

[2] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep CNN features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[3] Y. Chen, S. C. Dass, and A. K. Jain. Fingerprint quality indices for predicting authentication performance. In *Audio- and Video-Based Biometric Person Authentication: 5th International Conference, AVBPA 2005, Hilton Rye Town, NY, USA, July 20-22, 2005. Proceedings*, pages 160–170. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[4] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. In *2017 12th IEEE International Conference*

on Automatic Face & Gesture Recognition (FG 2017), pages 1–8, 2017.

[5] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho. Multiple classifiers in biometrics. part 2: Trends and challenges. *Information Fusion*, 44:103–112.

[6] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[7] F. E. Grubbs. On the distribution of the radial standard deviation. *Ann. Math. Statist.*, 15(1):75–81, 1944.

[8] A. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *International Conference on Image Processing. 2002*, volume 1, pages I–57–I–60 vol.1, 2002.

[9] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[10] K. Kryszczuk and A. Drygajlo. Improving biometric verification with class-independent quality information. *IET Signal Processing*, 3(4):310 EP – 321, 2009.

[11] H. Levene. Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press, 1960.

[12] B. Maze, N. D. Kalka, J. Duncan, A. Jain, J. Adams, T. Niggel, P. Grother, T. Miller, J. Cheney, and C. Otto. Iarpa janus benchmark c: Face dataset and protocol. In *International Conference on Biometrics*, 2018.

[13] NIST/SEMATECH. e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm.

[14] N. Poh and J. Kittler. Incorporating model-specific score distribution in speaker verification systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):594–606, 2008.

[15] N. Poh and J. Kittler. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):3–18, 2011.

[16] R. Ranjan, S. Sankar, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24, 2017.

[17] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016.

[18] K.-A. Toh, W.-Y. Yau, E. Lim, L. Chen, and C.-H. Ng. Fusion of auxiliary information for multi-modal biometrics authentication. In D. Zhang and A. K. Jain, editors, *Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004. Proceedings*, pages 678–685. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[19] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011.

[20] Y. Xu, Z. Li, B. Zhang, J. Yang, and J. You. Sample diversity, representation effectiveness and robust dictionary learning for face recognition. *Information Sciences*, 375(Supplement C):171–182, 2017.

[21] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] S. Yitzhaki. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron - International Journal of Statistics*, pages 285–316, 2003. LXI.

[23] J. Zhao, J. Han, and L. Shao. Unconstrained face recognition using a set-to-set distance measure on deep learned features. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017.