On accuracy estimation in face biometric problems

Domingo Mery
Pontificia Universidad Católica de Chile
dmery@ing.puc.cl

Yuning Zhao
University of Notre Dame
Yuning.Zhao.37@nd.edu

Kevin Bowyer
University of Notre Dame
kwb@nd.edu

The estimated accuracy of an algorithm plays a central and essential role in every face biometric problem. The accuracy goal is simple – the higher, the better. We know that the accuracy of an accepted biometric recognition system should be a number close to 100%. However, how confident is an estimated accuracy for a given dataset? Moreover, how generalizable is the proposed method for a wider variety of conditions? When we attempt to answer such questions, we typically focus on the 'what' elements of the dataset. What is number of images in the dataset? (Larger is better.) What kinds of expressions were taken into account? (More is better.) What are the illumination conditions in the images? (A broader range is generally better.) What is the gender, age and racial sampling of the data? (Broader is better.) Such questions are good and important, although many papers are published without such properties of the dataset being detailed. Nevertheless, the generalizability issue should also raise questions about 'how' the images are used to estimate accuracy, as well as 'what' is represented in the images. How is the accuracy estimated? (Mean, weighted mean, median?) How is the experimental protocol defined? (Leave-one-out? Half-Half? 10-fold cross-validation?) How are the images divided into train and test portions? (Randomly? Every N-th image? According to time of acquisition?) How is the data sampled from the underlying original data collection? (Is any data that was originally collected not used? If so, is this documented?) How is the person-specific nature of the data captured? (Are train and test splits person-disjoint?) How is the variance in the estimated accuracy estimated?

In order to illustrate the problematic nature of accuracy estimation, let us review one representative example. We found in paper [A][1], that the reported accuracy on face expression recognition on database X was 96.3% using 10-fold cross-validation. In paper [B][1], the reported accuracy on the same database was 70.0% using 10-fold cross-validation. Finally, in paper [C][1], the reported accuracy on the same database was 95.0%, however, the used experimental protocol was similar to this one: *we divide 10 facial expression sequences of every person into training and testing sets. Firstly, we use one expression image for testing, others for training. Then 14 images are used for training and 7 images left for testing. At last 7 images are used for training and 14 images for testing.* At first, we may think that method [A] is better than [B] and [C] because it has the highest reported accuracy. Nevertheless, method [C] uses such an uncommon way to evaluate the accuracy is not comparable. In addition, we might be tempted to think that the 96.3% in [A] and the 70.0% in [B] could be used in a fair comparison because both protocols use cross-validation with 10 folds. However, paper [A] is silent about how the folds are defined. Thus it is very difficult to establish if, for this accuracy estimate, the subjects whose images appear in one of the ten subsets also have images in the other nine subsets. Now, we can understand why method [B] has an accuracy estimate of 'only' 70.0%. The protocol used is more realistic because subjects that appear in the training set do not appear in the testing set. The accuracy estimate in this case has a much better chance of holding up in the face of new data. In conclusion, we can say that in this example that accuracies of [A], [B] and [C] are not comparable because the experimental protocols are too radically different.

In this work, we explore the problems that a researcher can have when experimenting on face image databases in terms of 'how' the images were used. We review the literature on three typical face image analysis challenges: expression recognition on JAFFE database (see Table 1), gender recognition on FERET database (see Table 2) and face recognition on AR database (see Table 3). We discover that in each one there are so many experimental protocols that it is nearly impossible to make fair comparisons. Moreover, many times a protocol is so intricate and so insufficiently

[1]To avoid hurt feelings, the reference is not given in this part, however, it is cited in our references.

detailed that is not possible to be confident in repeating it. Our work is focused on face databases, but we believe that the same issues arise for all biometric modalities. We claim that these two problems –no standard protocol, and ill-defined protocols– undermine the research on biometrics because they lead to confusing differences in strength of protocol with differences in estimated accuracy of algorithms.

References

[1] L. A. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern recognition letters*, 2010. 4

[2] S. Baluja and H. A. Rowley. Boosting sex identification performance. *International Journal of Computer Vision*, 2007. 4

[3] I. Buciu, C. Kotropoulos, and I. Pitas. ICA and Gabor representation for facial expression recognition. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, pages 855–858. IEEE, 2003. 4

[4] J. Chen and Z. Yi. Sparse representation for face recognition by discriminative low-rank matrix recovery. *Journal of Visual Communication and Image Representation*, 2014. 4

[5] F. Cheng, J. Yu, and H. Xiong. Facial expression recognition in JAFFE dataset based on Gaussian process classification. *IEEE Transactions on Neural Networks*, 21(10):1685–1690, Oct. 2010. 4

[6] H. B. Deng, L. W. Jin, L. X. Zhen, and J. C. Huang. A new facial expression recognition method based on local gabor filter bank and pca plus lda. *International Journal of ...*, 2005. 4

[7] W. Deng, J. Hu, and J. Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012. 4

[8] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, pages 399–406, 2013. 4

[9] R.-X. Ding, H. Huang, and J. Shang. Patch-based locality-enhanced collaborative representation for face recognition. *IET Image Processing*, 9(3):211–217, March 2015. 4

[10] X. Feng. Facial expression recognition based on local binary patterns and coarse-to-fine classification. In *Computer and Information Technology, 2004. CIT '04. The Fourth International Conference on*, pages 178–183. IEEE, 2004. 4

[11] X. Feng, M. Pietikainen, and A. Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4):592–598, Dec. 2007. 4

[12] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang. Single Sample Face Recognition via Learning Deep Supervised Auto-Encoders. *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, 10(10):2108–2118, OCTOBER 2015. 4

[13] G. Guo and C. R. Dyer. Learning from examples in the small sample case: face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(3):477–488, June 2005. 4

[14] J.-M. Guo, C.-C. Lin, and H.-S. Nguyen. Face Gender Recognition Using Improved Appearance-Based Average Face Difference and Support Vector Machine. In *2010 International Coriference on System Science and Engineering*, 2010. 4

[15] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic Estimation from Face Images: Human vs. Machine Performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1148, 2015. 4

[16] S. L. Happy and A. Routray. Automatic Facial Expression Recognition Using Features of Salient Facial Patches. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 6(1):1–12, January-March 2015. 4

[17] T. Jabid, M. H. Kabir, and O. Chae. Gender Classification using Local Directional Pattern (LDP). In *2010 International Conference on Pattern Recognition*, 2010. 4

[18] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *European Conference on Computer Vision (ECCV 2012)*, pages 331–344. Springer, 2012. 4

[19] Z. Jiang, Z. Lin, and L. S. Davis. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013. 4

[20] A. Koutlas and D. I. Fotiadis. An automatic region based methodology for facial expression recognition. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 662–666, 2008. 4

[21] H.-W. Kung, Y.-H. Tu, and C.-T. Hsu. Dual Subspace Nonnegative Graph Embedding for Identity-Independent Expression Recognition. *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, 10(3):626–639, March 2015. 4

[22] M. Kyperountas, A. Tefas, and I. Pitas. Salient feature and reliable classifier selection for facial expression classification. *Pattern Recognition*, 2010. 4

[23] P. Li, S. L. Phung, A. Bouzerdoum, and F. H. C. Tivive. Feature selection for facial expression recognition. In *IEEE 2nd. European Workshop on Visual Information Processing*, pages 35–40. IEEE, 2010. 4

[24] D. Liang, J. Yang, Z. Zheng, and Y. Chang. A facial expression recognition system based on supervised locally linear embedding. *Pattern recognition letters*, 26(15):2374–2389, Nov. 2005. 4

[25] S. Liao, W. Fan, A. C. S. Chung, and D.-Y. Yeung. Facial Expression Recognition using Advanced Local Binary Patterns, Tsallis Entropies and Global Appearance Features. In *Image Processing, 2006 IEEE International Conference on*, pages 665–668. IEEE, 2006. 4

[26] P. Liu, S. Han, Z. Men, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014. 4

[27] L. Lu and P. Shi. Fusion of multiple facial regions for expression-invariant gender classification. *IEICE Electronics Express*, 2009. 4

[28] E. Makinen and R. Raisamo. Evaluation of Gender Classification Methods with Automatically Detected and Aligned

Faces. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 30(3):514–547, March 2008. 4

[29] D. Mery and K. Bowyer. Face Recognition via Adaptive Sparse Representations of Random Patches. In *IEEE Workshop on Information Forensics and Security (WIFS 2014)*. IEEE, 2014. 4

[30] D. Mery and K. Bowyer. Recognition of facial attributes using adaptive sparse representations of random patches. In *Workshop on Soft Biometrics in conjunction with European Conference on Computer Vision (ECCV 2014)*, 2014. 4

[31] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):707–711, 2002. 4

[32] I. Naseem, R. Togneri, and M. Bennamoun. Linear Regression for Face Recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):2106–2112, Nov 2010. 4

[33] H.-T. Nguyen and A. Caplier. Local Patterns of Gradients for Face Recognition. *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, 10(8):1739–1751, AUGUST 2015. 4

[34] J.-S. Pan, Q. Feng, L. Yan, and J.-F. Yang. Neighborhood Feature Line Segment for Image Classification. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, 25(3):387–398, MATCH 2015. 4

[35] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 4

[36] Y. Shinohara and N. Otsu. Facial Expression Recognition Using Fisher Weight Maps. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2004. 4

[37] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee. Human Facial Expression Recognition Using Stepwise Linear Discriminant Analysis and Hidden Conditional Random Fields. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 24(4):1386–1398, April 2015. 4

[38] G. Srinivas, H. Jeffrey R. J., J. P, and W. Harry. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE TRANSACTIONS ON Neural Networks*, 11(4):948–960, July 2000. 4

[39] J. E. Tapia and C. A. Perez. Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information From Histogram of LBP, Intensity, and Shape. *Information Forensics and Security, IEEE Transactions on*, 8(3):488–499, 2013. 4

[40] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu. Robust Face Recognition via Adaptive Sparse Representation. *IEEE Transactions on Cybernetics*, (99):1, 2014. 4

[41] Y. Wang, H. Ai, B. Wu, and C. Huang. Real time facial expression recognition with AdaBoost. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, pages 926–929. IEEE, 2004. 4

[42] X. Wei, C. T. Li, and Y. Hu. Face recognition with occlusion using Dynamic Image-to-Class Warping (DICW). In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013)*, pages 1–6, 2013. 4

[43] R. Weng, J. Lu, J. Hu, G. Yang, and Y.-P. Tan. Robust Feature Set Matching for Partial Face Recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 601–608, Dec 2013. 4

[44] J. Wu, W. A. Smith, and E. R. Hancock. Facial gender classification using shape-from-shading. *Image and Vision Computing*, 28(6):1039–1048, June 2010. 4

[45] Z. Yang, M. Li, and H. Ai. An Experimental Study on Automatic Face Gender Classification. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, pages 1099–1102. IEEE, 2006. 4

[46] T. Zavaschi, T. H. H. Zavaschi, A. S. Britto Jr, A. S. Britto, Jr, les Oliveira, L. E. S. Oliveira, and A. L. Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications: An International Journal*, 40(2), Feb. 2013. 4

[47] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2691–2698. IEEE, 2010. 4

[48] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (KCCA). *Neural Networks, IEEE Transactions on*, 17(1):233–238, 2006. 4

[49] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn. Graph-Preserving Sparse Nonnegative Matrix Factorization With Application to Facial Expression Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(1):38–52, Feb. 2011. 4

[50] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using markov random fields. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1050–1057, Sept 2009. 4

Table 1. Literature review on expression recognition using JAFFE

No.	Method	Accuracy	Evaluation	unmix
1	LP-LBP [11]	93.8	20 $\times$ 10-fold CV	no
2	SLLE [24]	91.5	10-fold CV, 14 images/class for training	no
3	SLLE [24]	92.7	10-fold CV, 21 images/class for training	no
4	Boosted-LBP [35]	81.0	10-fold CV	no
5	Ensamble [46]	96.2	10-fold CV	no
6	L-SVM [13]	92.4	10-fold CV	no
7	PDM-Gabor [20]	90.2	10-fold CV	no
8	SH-FER [37]	96.3	10-fold CV	no
9	Salient Facial Patches [16]	91.8	10-fold CV	no
10	Hybrid Filter [23]	96.7	10-fold CV	no
11	SLLE [24]	86.8	Leave one subject out	yes
12	SFRCS [22]	85.9	Leave one subject out	yes
13	Ensamble [46]	70.0	Leave one subject out	yes
14	DSNGE [21]	65.6	Leave one subject out	yes
15	GP [5]	55.2	Leave one subject out	yes
16	HLAC [36]	69.4	Leave one subject out (only nine women instead of ten)	yes
17	Coarse to Fine [10]	77.0	Leave one subject out	yes
18	BDBNJ [26]	91.8	Leave one subject out	yes
19	KCCA [48]	77.1	Leave one subject out	yes
20	BDBNJ+C [26]	93.0	Leave one subject out using CK+ in training too.	yes
21	ASR+ [30]	94.3	Leave on sample out. Training 203 samples. Repetitions 350.	no
22	SFRCS [22]	96.7	Leave one sample out	no
23	GWs+SVM [3]	90.3	Leave one sample out	no
24	KCCA [48]	98.4	Leave one sample out	no
25	GP [5]	93.4	Leave one sample out	no
26	ALBP [25]	88.3	Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
27	Tsallis [25]	85.4	Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
28	ALBP+Tsallis [25]	91.9	Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
29	ALBP+Tsallis+NLDAI [25]	94.6	Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
30	GSNMF [49]	91.0	Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
31	Gabor+PCA+LDA [6]	97.3	3 $\times$ Hold out. Training: 2 samples of each facial expression for each person. Testing: remaining images.	no
32	Adaboost [41]	98.9	Reclassification. The goal was to use JAFFE for training and another DB for testing	no
33	Boosted-LBP [35]	41.3	Training: CK+ Testing: JAFFE	yes
34	BDBN [26]	68.0	Training: CK+ Testing: JAFFE	yes

Table 2. Literature review on gender recognition using FERET

No.	Method	Accuracy	Images	M/F	Evaluation	unmix
1	SVM-RBF [31]	96.6	1755	1044/711	5-fold CV	?
2	Read AdaBoost [45]	93.8	3529	?	5-fold CV	no
3	AdaBoost [2]	94.4	2409	1495/914	5-fold CV	yes
4	AdaBoost [2]	97.1	2409	1495/914	5-fold CV	no
5	Fusion (L6) [1]	99.1	411	212/119	5-fold CV	yes
6	Fusion [39]	99.1	411	212/119	5-fold CV	yes
7	Fusion (L6) [39]	97.8	411	211/119	5-fold CV	yes
8	2DPCA-SVM [27]	94.8	800	400/400	5-fold CV	?
9	DIF [15]	96.8	2729	1722/1007	5-fold CV (unclear)	no
10	ASR+ [30]	95.0	1051	602/448	Leave on sample out. Training 880 samples. Repetitions 400.	yes
11	manual alignment [28]	87.1	411	212/119	74-26 HO	yes
12	AAFD [14]	88.9	2722	1713/1009	80-20 HO	yes
13	recovered needle-map[44]	84.3	200	100/100	70-30 HO	yes
14	ERBF2 - C4.5 [38]	96.0	3006	1906/1100	30 male and 30 female for Training, others for testing, 20 repetitions.	no
15	Read AdaBoost [45]	92.0	3529	?	Training with Chinese Database, Testing on FERET	yes
16	LDP [17]	95.1	2000	1100/900	not mentioned	not mentioned

Table 3. Literature review on face recognition using AR

No.	Method	Accuracy	Subjects	Images/sub.	Illum.	Sunglass	Scarf	Evaluation
1	NFLS-I [34]	99	120	14	yes	no	no	Leave on sample out.
2	ASR+ [29]	97.0	100	9	yes	yes	yes	Leave on sample out. Training 900 samples. Repetitions 10.000.
3	ASR+ [29]	100.0	100	13	yes	yes	yes	Leave on sample out. Training 1.300 samples. Repetitions 10.000.
4	ASR+ [29]	99.0	100	8	yes	yes	yes	Leave on sample out. Training 800 samples (no disguise). Testing disguise. Repetitions 10.000.
5	ASR+ [29]	100.0	80	13	yes	yes	yes	Leave on sample out. Training 1.040 samples. Repetitions 8.000.
6	ASR+ [29]	95.0	100	5	yes	yes	yes	Leave on sample out. Training 500 samples. Repetitions 10.000.
7	ASR+ [29]	98.0	100	7	yes	yes	yes	Leave on sample out. Training 700 samples. Repetitions 10.000.
8	ASR+ [29]	100.0	100	20	yes	yes	yes	Leave on sample out. Training 200 samples. Repetitions 10.000.
9	ESRC [7]	95.0	80	13	yes	yes	yes	1-12 HO. Training: A single natural face.
10	Modular LRC [32]	95.5	100	10	no	no	yes	8-2 HO. Training: no disguise. Testing: disguise.
11	LRC [32]	96.0	100	10	no	yes	no	8-2 HO. Training: no disguise. Testing: disguise.
12	ASRC [40]	75.5	100	14	yes	no	no	2-12 HO
13	LC-KSVD [19]	97.8	100	26	yes	yes	yes	20-6 HO
14	ℓ_{struct} [18]	92.5	100	10	?	yes	no	799-200 HO. Training: no disguise. Testing: disguise.
15	ℓ_{struct} [18]	69.0	100	10	?	no	yes	799-200 HO. Training: no disguise. Testing: disguise.
16	SEC-MRF [50]	100.0	100	10	?	yes	no	799-200 HO. Training: no disguise. Testing: disguise.
17	SEC-MRF [50]	97.5	100	10	?	no	yes	799-200 HO. Training: no disguise. Testing: disguise.
18	MLERPM [43]	98.0	100	20	yes	yes	no	14-6 HO. Training: no disguise. Testing: disguise.
19	MLERPM [43]	97.0	100	20	yes	no	yes	14-6 HO. Training: no disguise. Testing: disguise.
20	LPOG [33]	99.1	134	13	yes	yes	yes	1-12 HO. Training: #1 neutral image, Testing: remaining 12 images.
21	PLECR [9]	98.2	100	26	yes	yes	yes	10 $\times$ 13-13 HO
22	DICW [42]	99.5	100	14	no	yes	no	8-6 HO. Training: no disguise. Testing: disguise.
23	DICW [42]	98.0	100	14	no	no	yes	8-6 HO. Training: no disguise. Testing: disguise.
24	DLRR [4]	91.4	100	20	yes	yes	no	3 $\times$ 8-12 HO. Training: 7 undisguised + 1 random sunglasses image.
25	DLRR [4]	90.2	100	20	yes	no	yes	3 $\times$ 8-12 HO. Training: 7 undisguised + 1 random scarf image.
26	ASRC [40]	94.7	100	14	yes	no	no	7-7 HO
27	DLRR [4]	93.7	100	14	yes	no	no	7-7 HO. Training: from session 1. Testing: from session 2.
28	SSAE [12]	85.2	100	13	yes	yes	yes	20-80 HO. Training: 20 subjects. Testing: 80 subjects.
29	DKSVD [47]	95.0	100	26	yes	yes	yes	3 $\times$ 20-6 HO
30	LC-KSVD [19]	97.8	100	26	yes	yes	yes	20-6 HO. Training: 20 random images per person.
31	SSRC [8]	98.0	100	26	yes	yes	yes	10 $\times$ 1300-1300 HO