On accuracy estimation and comparison of results in biometric research

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

Abstract

The estimated accuracy of an algorithm is the most important element of the typical biometrics research publication. Comparisons between algorithms are commonly made based on estimated accuracies reported in different publications. However, even when the same dataset is used in two publications, there is a very low frequency of the publications using the same protocol for estimating algorithm accuracy. Using the example problems of face recognition, expression recognition and gender classification, we show that the variation in estimated performance on the same dataset across different protocols can be enormous. Based on these results, we make recommendations for how to obtain performance estimates that allow reliable comparison between algorithms.

1. Introduction

The estimated accuracy of an algorithm plays a central and essential role in every 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%. If we read in a paper on face expression recognition that *the accuracy obtained by the proposed method was 97.3%*, we could believe that 97.3% of the expressions that a person makes –whatever that means– will be correctly recognized. However, how confident is this value for the dataset used in the paper? Moreover, how generalizable is the proposed method for a wider variety of conditions? And can this 97.3% be compared with the 98.5% reported in another paper for expression recognition experiments on the same dataset?

When we attempt to answer such questions, we typically focus on the 'what' elements of the dataset: What is the number of images in the dataset? (Larger is better.) What kinds of expressions were taken into account? (A greater variety is generally 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? (Greater balance on these dimensions is generally 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?) 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 (unmixed)?) Is an estimate of the variance in the estimated accuracy reported? (How is it computed?) All of these considerations directly impact whether or not it is possible to make a valid comparison between performance figures published in different papers.

Detailed analyses of methods of statistical testing for machine learning algorithms can be found in the literature; for example, in Demsar's well-known work [7]. But the common practice in biometrics research is not at this level. There is recognition of the need for improved practice in the biometrics research community, as evidenced by papers on this topic in recent conferences [18], [31]. In this paper, we want to show how the experimental methods that the biometrics research community has been following in practice can lead to wrong conclusions. Thus we focus on a narrower issue than [18], [31], and attempt to outline how two different publications can follow a similar enough methodology that their accuracy estimates can reliably be compared.

In this paper, we explore the problems that a researcher can have when experimenting on image databases (*e.g.* faces, iris, fingerprints, etc.) in terms of 'how' the images were used. We review the literature on three typical face image analysis challenges, and we discover that in each one there are so many experimental protocols that it is nearly impossible to make fair comparisons across the published literature. Moreover, many times a protocol is so intricate and so insufficiently detailed that is not possible to be confident in repeating it. Our paper 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. Based on these results, we make recommendations for how to obtain performance estimates that allow reliable comparison between algorithms. Thus, we propose the "EPD" methodology: [E] Experiments: Wherever a subject-disjoint train-and-test split would be possible, it should be used. [P] Protocol: The protocol should ideally be to report the mean and standard deviation of some number of randomized 10-fold cross-validation (10f CV) trials. Reviewers should accept accuracy reported on a single hold-out (HO) trial only if there is a clear justification made. [D] Data: Any downsampling from the collected dataset should be described and justified. Wherever possible, results should be presented with and without the downsampling, so that reviewers can judge its effect.

2. Protocols

This section presents the definition of various protocols that are commonly used in biometrics research papers for evaluating the accuracy of a classifier. An important element of this is the definition of the 'training data' and the 'testing data'[1]. In general, there is a set $\mathbb{D}$ that contains all collected data. The original raw data is typically a face image, iris image, fingerprint, or other biometric sample. The data used in estimating accuracy of a classifier is typically not the raw data, but rather the set of feature vectors and labels derived from the raw data. For each element of raw data, there is a feature vector and one or more corresponding labels. The labels are "meta-data" that specifies the identity of the person, the facial expression being made, the gender of the person, or other properties of the raw data. Often in biometric studies, from the set $\mathbb{D}$ of all data collected, a subset $\mathbb{X} \subset \mathbb{D}$ is selected out for some reason. For example, data may be dropped from the study because the person's eyes are closed in the image, or the person wore glasses, or some other reason. It is important that any selection out of the original dataset be described and justified in the experimental context. We will call subset $\mathbb{X}$ the 'used data' because it is used to evaluate the accuracy of a classifier. Set $\mathbb{X}$ is $\{\mathbf{X}, \mathbf{d}\}$, *i.e.*, it consists of *i)* a matrix $\mathbf{X}$ of size $N \times p$, for N samples and p features for each sample; and *ii)* a vector $\mathbf{d}$ of N elements with the labels (one label per sample). The general protocol to estimate the accuracy of a classifier is as follows:

1) Select training ($\mathbf{X}_{\text{train}}, \mathbf{d}_{\text{train}}$) and testing data ($\mathbf{X}_{\text{test}}$, $\mathbf{d}_{\text{test}}$) from $\mathbb{X}$. In some biometric contexts, *e.g.* expression recognition, an algorithm may be trained on images of some set of subjects and then tested on unseen subjects, *i.e.* a subject-disjoint train-and-test data selection[2] is used in order to estimate the accuracy in a more realistic way.

2) Train a classifier using training data ($\mathbf{X}_{\text{train}}, \mathbf{d}_{\text{train}}$):

$$\boldsymbol{\Theta} = \text{ClassifierTrain}(\mathbf{X}_{\text{train}}, \mathbf{d}_{\text{train}}) \quad (1)$$

where $\boldsymbol{\Theta}$ is a vector that contains all parameters of the classifier that was trained. For instance, in a simple classifier like Euclidean minimal distance we store in $\boldsymbol{\Theta}$ only the centers of mass of each class in the training set.

3) Predict the labels of each testing sample using the features of the testing data $\mathbf{X}_{\text{test}}$, the learned classifier and its parameters $\boldsymbol{\Theta}$. Store the prediction in vector $\mathbf{d}_s$ of N_{test} elements: $\mathbf{d}_s = \text{Classify}(\mathbf{X}_{\text{test}}, \boldsymbol{\Theta})$. In this step it is not allowed to use the labels of the testing data $\mathbf{d}_{\text{test}}$.

4) Compute the accuracy of the testing data defined as

$$\eta_i = \frac{\text{\# test samples correctly predicted}}{N_{\text{test}}}, \quad (2)$$

where the numerator corresponds to the number of elements of $\mathbf{d}_{\text{test}}$ and $\mathbf{d}_s$ that are equal.

5) Repeat steps 1–4 n times (using index i and some data selection criterium) and compute the final estimation by averaging over n times: $\eta = \frac{1}{n}\sum_i \eta_i$

In the following, we summarize details of the typical protocols used in the literature[3].

- **Hold-out:** We take a percentage S of $\mathbb{X}$ for training and the rest for testing. In our general methodology, this protocol corresponds to $n = 1$ in (**??**). This is the simplest way to evaluate the accuracy. In the tables and text in the remainder of the paper, we will refer to instances of this protocol as '(S)-(100-S) HO', *e.g.* for $S = 80\%$ it is '80-20 HO'.
- **Cross-validation:** The data is divided into v folds. A portion $S = (v-1)/v$ of the whole data is used for training and the rest ($1/v$) for testing. This experiment is repeated

[1] A good practice to split the data into three groups, namely training, validation (for parameter tuning), and testing, but only a few numbers of the reported papers in this area use them.

[2] This selection of samples is also known as 'unmixed'.

[3] Bootstrap has been used in some biometric problems, see for example [40], however, it is not included in our work because in the cases we studied bootstrap is not common.

v times rotating train and test data to evaluate the stability of the classifier. The estimated accuracy, η, is calculated as the mean of the v individual accuracies of the true classifications that are tabulated in each case, *i.e.* $n = v$ in Step 5). We call this protocol 'vf CV', *e.g.* for $v = 10$ it is '10f CV'. An extended cross-validation can be computed by repeating a v-fold cross-validation k times, with the samples randomly selected each time. The estimated accuracy is the average over the k estimations. We call this protocol '$k \times v$f CV', *e.g.* for $k = 20$ and $v = 10$ it is '20×10f CV'.

- **Leave-one-out:** It is essentially an extreme version of the cross-validation technique, with N folds, where N is the number of samples of $\mathbb{X}$. In each experiment, one sample is used for testing and the rest ($N - 1$ samples) are used for training. After repeating the experiment N times, every sample is selected as test data once. We call this protocol 'LOO'. A variation of LOO protocol that reduces the computational time can be computed as follows. $M < N$ random samples are taken for training, and one random sample (that was not included in training stage) is used for testing. This process is repeated $n = K < N$ times, and the estimated accuracy is the average over the K estimations. We call this protocol 'LOO(M, K)', *e.g.* for $M = 100$ and $K = 500$ it is 'LOO(100,500)'. In some face biometric problems, when we are interested in recognizing a facial attribute (gender, expression, etc.), an *unmixed* version of leave-one-out can be used, that is subjects that appear in the training set do not appear in the testing set. In this case, the protocol is called 'leave-one-subject-out' or 'LOSO'.

3. Literature review

In this Section, we review the literature on some relevant databases: (a) face recognition on AR database [30], (b) face expression recognition using the JAFFE database [28], and (c) gender recognition using the FERET database [39][4] (see a summary of each review in Tables 1, 2 and 3 respectively). These particular datasets are of course not the only ones that can be used in these problems[5], but they are ones that have been widely used, and the issues illustrated are not dependent on the particular dataset.

- **AR – Face recognition:** The AR database [30] consists of

Table 1: Review on face recognition (AR)

Method	Year	η	Subjects	Images/sub.	Illum.	Sunglass	Scarf	Evaluation
01) LPOG [37]	2015	99.1	134	13	yes	yes	yes	1-12 HO*, single sample per person
02) NFLS-I [38]	2015	99.0	120	14	yes	no	no	LOO
03) LC-KSVD [20]	2013	97.8	100	26	yes	yes	yes	20-6 HO*
04) PLECR [10]	2015	98.2	100	26	yes	yes	yes	10×13-13 HO*
05) DKSVD [54]	2010	95.0	100	26	yes	yes	yes	3×20-6 HO*
06) LC-KSVD [20]	2013	97.8	100	26	yes	yes	yes	20-6 HO*
07) SSRC [9]	2013	98.0	100	26	yes	yes	yes	10×13-13 HO*
08) DLRR [5]	2014	89.7	100	26	yes	yes	yes	3×9-17 HO*, training: no disguise, sunglass, scarf
09) SSRC [9]	2013	90.0	100	26	yes	yes	yes	3×9-17 HO*, training: no disguise, sunglass, scarf
10) ASR+ [32]	2014	100.0	100	20	yes	yes	yes	LOO(200,10000)
11) MLERPM [48]	2013	98.0	100	20	yes	yes	no	14-6 HO*, training: no disguise, testing: disguise
12) MLERPM [48]	2013	97.0	100	20	yes	no	yes	14-6 HO*, training: no disguise, testing: disguise
13) SSRC [9]	2013	90.9	100	20	yes	yes	no	3×8-12 HO*, training: no disguise, sunglass
14) SSRC [9]	2013	90.9	100	20	yes	no	yes	3×8-12 HO*, training: no disguise, scarf
15) DLRR [5]	2014	91.4	100	20	yes	yes	no	3×8-12 HO*, training: no disguise, scarf, sunglass
16) DLRR [5]	2014	90.2	100	20	yes	no	yes	3×8-12 HO*, training: no disguise, scarf, sunglass
17) ASRC [46]	2014	75.5	100	14	yes	no	no	2-12 HO*
18) ASRC [46]	2014	94.7	100	14	yes	no	no	7-7 HO*
19) DLRR [5]	2014	93.7	100	14	yes	no	no	7-7 HO*, training: session 1, testing: session 2
20) DICW [47]	2013	99.5	100	14	no	yes	no	8-6 HO*, training: no disguise, testing: disguise
21) DICW [47]	2013	98.0	100	14	no	no	yes	8-6 HO*, training: no disguise, testing: disguise
22) ASR+ [32]	2014	100.0	100	13	yes	yes	yes	LOO(1300,10000)
23) Mod LRC [36]	2010	95.5	100	10	no	no	yes	8-2 HO*, training: no disguise, testing: disguise
24) LRC [36]	2010	96.0	100	10	no	yes	no	8-2 HO*, training: no disguise, testing: disguise
25) ℓ_{struct} [19]	2012	92.5	100	10	?	yes	no	799-200 HO**, training: no disguise, testing: disguise
26) ℓ_{struct} [19]	2012	69.0	100	10	?	no	yes	799-200 HO**, training: no disguise, testing: disguise
27) ASR+ [32]	2014	97.0	100	9	yes	yes	yes	LOO(900,10000)
28) ASR+ [32]	2014	99.0	100	8	yes	yes	yes	LOO(800,10000), training: no disguise, testing: disguise
29) ASR+ [32]	2014	95.0	100	5	yes	yes	yes	LOO(500,10000)
30) ASR+ [32]	2014	98.0	100	7	yes	yes	yes	LOO(700,10000)
31) SSAE [12]	2015	85.2	80	13	yes	yes	yes	1-79 HO*, single sample per person
32) ASR+ [32]	2014	100.0	80	13	yes	yes	yes	LOO(1040,8000)
33) ESRC [8]	2012	95.0	80	13	yes	yes	yes	1-12 HO*, single sample per person

x-y HO*: Training: x images per subject. Testing: y images per subject.

x-y HO**: Training: x images. Testing: y images per subject.

[4] A Google Scholar search shows that AR, FERET and JAFFE have on average 274, 364 and 71 citations per year, respectively, since 2010.

[5] Some databases such as LFW [16] do not have the mentioned problems because the majority of the reported results have followed the original protocol. There are however some exceptions, see for example the use of hold-out on LFW using 10 images per face for training and the rest for testing [51].

Table 2: Review on expression recognition (JAFFE)

Method	Year	η	Evaluation
01) LP-LBP [11]	2007	93.8	20 × 10-f CV (14 images/class for training, 21 images/class for training)
02) Boosted-LBP [41]	2009	81.0	10-f CV
03) Ensamble [53]	2013	96.2	10-f CV
04) PDM-Gabor [21]	2008	90.2	10-f CV
05) SH-FER [43]	2015	96.3	10-f CV
06) SFP [15]	2015	91.8	10-f CV
07) Hybrid Filter [24]	2010	96.7	10-f CV
08) ASR+ [34]	2015	96.7	10-f CV
09) SFRCS [23]	2010	85.9	LOSO^{+}
10) Ensamble [53]	2013	70.0	LOSO^{+}
11) DSNGE [22]	2015	65.6	LOSO^{+}
12) GP [6]	2010	55.2	LOSO^{+}
13) HLAC [42]	2004	69.4	LOSO^{+} (nine women)
14) BDBNJ [26]	2014	91.8	LOSO^{+}
15) KCCA [55]	2006	77.1	LOSO^{+}
16) BDBNJ+C [26]	2014	93.0	LOSO^{+} (CK+ & JAFFE)
17) ASR+ [33]	2014	94.3	LOO(203,350)
18) SFRCS [23]	2010	96.7	LOO
19) KCCA [55]	2006	98.4	LOO
20) GP [6]	2010	93.4	LOO
21) ALBP [25]	2006	88.3	HO*
22) Tsallis [25]	2006	85.4	HO*
23) ALBP+Tsallis [25]	2006	91.9	HO*
24) NLDAI [25]	2006	94.6	HO*
25) GSNMF [56]	2011	91.0	HO*
26) Boosted-LBP [41]	2009	41.3	$^{+}$Training: CK+ Testing: JAFFE
27) BDBN [26]	2014	68.0	$^{+}$Training: CK+ Testing: JAFFE

HO*: Training: 2 samples of each facial expression for each person. Testing: remaining images. $^{+}$ Unmixed evaluation: subject-disjoint train-and-test split.

Table 3: Review on gender recognition (FERET)

Method	Year	η	M/F*	Evaluation
01) SVM-RBF [35]	2002	96.6	1044/711	5-f CV
02) AdaBoost [52]	2006	93.8	?	5-f CV
03) AdaBoost [3]	2007	94.4	1495/914	5-f CV^{+}
04) AdaBoost [3]	2007	97.1	1495/914	5-f CV
05) ASR+ [34]	2015	94.1	600/440	5-f CV^{+}
06) Fusion (L6) [2]	2010	99.1	212/199	5-f CV^{+}
07) Fusion [45]	2013	99.1	212/199	5-f CV^{+}
08) Fusion (L6) [45]	2013	97.8	211/199	5-f CV^{+}
09) 2DPCA-SVM [27]	2009	94.8	400/400	5-f CV
10) DIF [14]	2014	96.8	1722/1007	5-f CV (unclear)
11) ASR+ [33]	2014	95.0	602/448	LOO(880,400)$^{+}$
12) MA [29]	2008	87.1	212/199	74-26 HO^{+}
13) AAFD [13]	2010	88.9	1713/1009	80-20 HO^{+}
14) needle-map [50]	2010	84.3	100/100	70-30 HO^{+}
15) ERBF2/C4.5 [44]	2000	96.0	1906/1100	20 × HO, 30 male and 30 female for training
16) AdaBoost [52]	2006	92.0	3529?	HO^{+}, training: Chinese database
17) LDP [17]	2010	95.1	1100/900	not mentioned

* M: number of male images and F: number female images
$^{+}$ Unmixed evaluation: subject-disjoint train-and-test split.

26 different images of each of 50 women and 50 men. The 26 images represent different facial expressions, illumination conditions, and occlusions with sunglasses and scarf. The characteristics of this database (occlusion/no occlusion, facial expressions/neutral, etc.) allow a large number of different experiments. Thus, it is very difficult to make fair comparisons when the protocols are not exactly the same. For the 33 papers on AR database listed in Table 1, there appear to be 3 pairs of papers that used effectively the same protocol: lines 3 and 6, 4 and 7, and 8 and 9.

- **JAFFE – Expression recognition:** The JAFFE database [28] contains 7 expressions ('neutral' and six basic emotions: 'fear', 'happiness', 'sadness', 'surprise', 'anger' and 'disgust') captured from 10 Japanese women. For each subject, there are 3–4 face images for the non-neutral and one for the neutral expressions, *i.e.*, the database consists of 213 images. There are 27 different publications on JAFFE database represented in Table 2, and at least 13 different evaluation protocols. As a group, the papers that reported the HO and LOO protocols reported the highest estimated accuracies, 88.3% to 96.7%. The papers that used some variation of the CV protocol reported the next highest accuracies, 81% to 96.7%. And the papers that used some variation of the LOSO protocol reported the lowest accuracies, 55.2% to 93%. Probably the most important single piece of information in the Table is the 'unmixed' evaluation ('$^{+}$'), indicating whether the train and test division was person-disjoint.
- **FERET – Gender recognition:** The FERET database [39] contains more than 3,500 face images from women and men (with different races) involving different expressions and illumination conditions. There are many different experimental protocols reported in the literature, The protocols vary based on different number of images, number of females and males, and 'mixed' or 'unmixed' datasets. For the 17 papers on FERET database listed in Table 3, there is only 1 instance where a pair of papers used the same train-test protocol, including gender distribution and number of images: lines 6 and 7.

4. Experiments

In this Section, we show experimental results that demonstrate how widely the comparison of two algorithms' performance can vary based on the experimental protocol that is followed. We use two well-known recognition methods as examples: 1) LBP: local binary pattern features [1] with 6 × 6 partitions with a Naïve Bayes Nearest Neighbor (NBNN) classifier [4], and SRC: a sparse representation classifier [49] where the images were sub-sampled to 22 × 18 pixels. All face images of the galleries were resized to 110 × 90 pixels. We experiment on three different databases to show the accuracy in face recognition problems (AR) and face attributes recognition (JAFFE for expressions and FERET for gender).

Face recognition: In AR, training and testing images are selected randomly from the 26 available images. Results on galleries of 50 and 100 subjects, summarized for 50 randomized runs, are tabulated in Table 4 for 26 images per subject. A similar behavior is obtained when 20 and 10 images per subject are used (due to space considerations the tables are not shown), however, the greater the number of training images per subject, the greater the accuracy. One clear point from the results in Table 4 is that, for this dataset and these algorithm implementations, LBP consistently outperforms SRC. For each given number of images per subject, number of subjects, and train-and-test instance, the mean accuracy for LBP is always greater than that for SRC. However, for each of the HO instances, the maximum accuracy across the 50 SRC trials is regularly greater than the minimum across the 50 LBP trials. This shows that a single HO trial, even if paired for the same size of the train-test split, is not a sufficient basis for a comparison of algorithms. Comparison based on a single 2f or 3f CV trial can show an incorrect comparison, but comparisons based on a single 10f or 5f CV trial are consistent.

There are three main points to observe from the face recognition results on AR summarized in Table 4. The first point is simply that the accuracy estimate made by either the HO or the CV protocol increases as an increased proportion of the data is used for training. As the HO shifts from 50-50 to 80-20, and as the CV shifts from 2f to 5f, the estimated accuracy of each algorithm increases. This reflects the simple fact that as more data is used for training, the average accuracy increases. If this trend did not occur, it would suggest that the algorithm and/or the dataset is atypical in some important respect. The second point to note is that, if the mean of the 50 randomized trials is used to compare the two algorithms, the comparison is entirely stable. For each of the six train-and-test methods (columns in the table), and for either gallery size (50 or 100), the mean accuracy over 50 randomized trials is higher for LBP than it is for SRC. For these algorithm implementations and this dataset, LBP is clearly better than SRC. The third point to observe is that it is very easy, when not comparing algorithms based on the mean accuracy over 50 trials, or when comparing across different n-fold CV or different HO, to get an incorrect comparison. For the 100-subject gallery, comparing the 5f CV mean accuracy of SRC, which is 96.8%, to the 2f CV mean accuracy of LBP, which is 97.2%, might lead to the (wrong) conclusion that there is relatively little difference between the algorithms. The same could happen on comparing the 80-20 HO accuracy of SRC to the 50-50 HO accuracy of LBP. For the 50-subject gallery, the 80-20 HO accuracy of SRC is actually even slightly better than the 50-50 HO accuracy of LBP, 98.1% to 98%. This emphasizes that even when considering the mean accuracy across 50 trials, it is also essential that the size of the train-test split be the same in order to get a fair comparison.

Table 4: Experiments on face recognition (AR)

(Subjects)		80-20	67-33	50-50	5f	3f	2f
Method		HO	HO	HO	CV	CV	CV
(100)	max	100.0	99.8	98.9	99.9	99.5	98.0
LBP	mean	99.6	99.1	97.3	99.7	99.1	97.2
	min	98.8	98.1	95.5	99.4	98.5	96.0
	std	0.28	0.36	0.69	0.10	0.23	0.46
(100)	max	98.4	96.6	92.0	97.4	95.5	92.1
SRC	mean	97.0	94.5	90.8	96.8	94.7	90.8
	min	95.6	92.7	89.1	96.0	93.3	89.2
	std	0.69	0.81	0.80	0.30	0.44	0.61
(50)	max	100.0	100.0	99.7	99.9	100.0	99.5
LBP	mean	99.8	99.3	98.0	99.7	99.3	98.0
	min	99.2	97.8	94.6	99.5	98.7	95.9
	std	0.27	0.57	0.93	0.13	0.31	0.64
(50)	max	100.0	98.2	95.5	98.7	97.7	95.2
SRC	mean	98.1	96.4	93.5	97.9	96.5	93.4
	min	95.2	93.6	90.9	96.0	94.8	91.9
	std	0.91	1.09	1.04	0.52	0.67	0.76

Table 5: Experiments on expression recognition (JAFFE)

		90-10	80-20	50-50	10f	5f	2f	
Method		HO	HO	HO	CV	CV	CV	LOSO
LBP	max	100.0	100.0	90.5	93.0	90.6	82.8	72.7
	mean	90.7	86.7	69.6	90.7	86.7	69.6	45.4
	min	66.7	54.8	54.6	87.0	80.8	59.2	25.0
	std	7.02	6.33	6.27	1.47	2.41	5.16	17.78
SRC	max	100.0	100.0	85.2	93.4	93.4	82.6	72.7
	mean	90.6	88.3	72.4	90.6	88.3	72.4	44.9
	min	66.7	71.4	59.0	84.2	83.6	61.5	19.1
	std	6.05	5.33	5.22	1.67	2.05	4.50	16.27

Table 6: Experiments on gender recognition (FERET)

		90-10	80-20	50-50	10f	5f	2f
Method		HO	HO	HO	CV	CV	CV
LBP	max	90.0	87.5	83.2	82.1	82.4	82.7
	mean	81.2	81.2	81.1	81.2	81.2	81.1
	min	71.0	74.5	78.0	80.5	80.1	79.9
	std	3.36	2.53	1.18	0.44	0.61	0.67
SRC	max	98.0	93.0	90.4	90.1	90.7	88.9
	mean	88.6	88.7	87.3	88.7	88.7	87.3
	min	80.0	82.0	83.8	87.2	86.8	85.3
	std	3.21	2.16	1.34	0.69	0.86	0.74

The problem is even greater if the comparison between algorithms is based on the accuracy of a single HO trial. For the 50-subject gallery, the maximum single-trial SRC accuracy of a given HO split is always greater than the minimum single-trial LBP accuracy, making it possible to generate an incorrect comparison result.

The conclusion from this example dataset and problem is that the most meaningful result to report for a possible comparison of algorithms is a 10f CV result averaged over some number (e.g., 50) of randomized trials.

It is well known that in face recognition, the estimated accuracy is strongly dependent on the number of subjects of the gallery and the number of face images per subject that are used in the training. This conclusion is evident in our experiments on AR: the greater the number of subjects to be recognized the lower the estimated accuracy; and the greater the number of training images per subject, the greater the accuracy. If a paper reports 'we obtain more than 95% accuracy on X database', it does not contain enough information about the conditions of the experiment to be able to compare to other published results. There are certain databases that allow so many experimental alternatives (*e.g.* the AR database) that is mandatory to report the details of the experiment. In addition, it is highly recommended that the experiments on known databases should use a known experimental protocol, in order to make fair comparisons possible from the literature. Reviewers should in general not accept a paper that presents results based on a single HO trial, without an explicit justification for why this is necessary. The highest standard of reliability would come from 10f CV averaged over some number of randomized train-test splits.

Expression recognition: Table 5 summarizes results on JAFFE of the two algorithmic approaches across 3 instances of H0 train-and-test, 3 instances of CV train and test, and the leave-one-subject-out (LOSO) variant of HO. Again, 50 randomized splits are generated for each train-and-test instance for each of the two algorithms.

As might be expected given the relatively large variation in the mean accuracy, the comparison of the two algorithms is not straightforward. Setting aside the LOSO result for the moment, LBP has higher mean accuracy than SRC for one of the HO instances and one of the CV instances, and SRC has higher mean accuracy than LBP for the other 2 HO instances and the other 2 CV instances. But this pattern is also correlated with the fraction of data used for training. For the HO and CV splits where less data is used for training, SRC has higher mean accuracy, and for the splits where more data is used for training, LBP has higher mean accuracy. This illustrates the danger in comparing algorithms based on performances obtained with different train-and-test instances. It also illustrates the danger in a comparison based on train-and-test instance that is chosen to economize on computational requirements rather than one chosen to maximize competency of the algorithms.

The most important point to take away from this experiment is the fundamental and huge difference between LOSO results and any instance of traditional HO or CV. With LOSO, the mean accuracy of both algorithmic approaches is basically 45%, whereas the minimum for either algorithm across all the HO and CV approaches is approximately 70%. The HO and CV instances of the algorithms trained with the largest fraction of training data, as shown in Table 5, obtain about 91% accuracy.

In a problem definition such as expression recognition, where an algorithm may be trained on images of some set of subjects and then applied to unseen subjects, a subject-disjoint train-and-test methodology inherently gives a more realistic estimate of accuracy. And LOSO is the instance of subject-disjoint that allows the maximum size of training data. In this example experiment, using an inappropriate train-and-test method could lead one to expect that the problem is solved with 91% accuracy, when in fact the performance for new persons will be about 45% accuracy.

We can conclude that when we are interested in recognizing a facial attribute like expression, it is very important that subjects that appear in the training set do not appear in the testing set. This is dramatically illustrated in the case of the expression recognition on JAFFE database, where LOSO (leave-one-subject-out) protocol separates the subject of the testing from the subjects of the testing. It is then possible, that the reported accuracy could be 100% (see Table 5, row 'SRC-max', column '80-23 HO'), and the reader could think that this algorithm is perfect for expression recognition, whereas a more realistic value is only approximately 45% (see Table 5, row 'SRC-mean', column 'LOSO'). In this example, there is a difference of 55%!

Gender recognition: In our experiments, we used 1,000 unmixed subjects (600 males and 400 females) of FERET database. Table 6 summarizes results of the two algorithmic approaches across 50 randomized splits for 3 instances of H0 train-and-test, and 3 instances of CV train and test. These splits are all unmixed, meaning that the train and test portions of the data are subject-disjoint.

One point to note in these results is that the SRC algorithm has a significantly higher mean accuracy compared to the LBP algorithm. Across the 6 different train-and-test instances, the mean accuracy of SRC varies in a relatively narrow band of 87% to 89%. In contrast, the mean accuracy of LBP varies in a relatively narrow band of just over 81%. Despite the fact that the SRC algorithm clearly out-performs the LBP algorithm on this problem, using a single trial of a HO methodology to compare the algorithms could easily lead to the opposite conclusion. For example, with a 90-10 HO split, the maximum LBP accuracy is 90% whereas the minimum SRC accuracy is only 80%. This problem does not occur across any of the 3 instances of the CV methodology. This result reinforces that reporting accuracy results based on a single HO trial should generally be considered unacceptable.

A final remark can be given: if we compare two different algorithms, *e.g.* LBP and SRC, we could erroneously conclude –if we don't use the same protocol– that one method is much better than another (see in Table 6, 90% for LBP selecting the maximal accuracy of 90-10 HO and 87.2% for SRC selecting the minimal accuracy of 10-f CV). However, if we had used the same protocol we could observe the opposite because in Table 6 SRC is always better than LBP using the same protocol.

5. Conclusions

In our results, it is clear that the variation of the estimated accuracy of an algorithm can be enormous. Using the same classification algorithm, the estimated accuracy can be totally different depending on *i)* the selection of training and testing data, *ii)* the number of samples of the used dataset and *iii)* the number of single accuracies used to estimate the average of final accuracy (for instance in 10-fold cross-validation the average of 10 single accuracies were used instead of only one in case of hold-out protocol).

Based on the published literature, it is rare to find two papers published on the same problem that use the same experimental protocol in all important elements. This is seen clearly in the works summarized in Tables 1, 2 and 3.

For problems where a subject-disjoint train-and-test split is essential in order to obtain a useful accuracy estimate, papers are often published using a non-disjoint split. This is seen clearly in Tables 2 and 3. For problems of this type, a leave-one-subject-out protocol would seem to be the default recommendation for useful experimental results. When the dataset is so large as to truly present computational challenges, a subject-disjoint LOO protocol might be used.

A single simple HO accuracy estimate, for example the often-used 80-20 HO, does not result in an accuracy estimate that allows confident comparison of two different algorithms for solving the same problem on the same dataset. This is clear in the results in Tables 4-6. In our experiments, a 10f CV protocol generally results in an accuracy estimate that would allow comparison of algorithms that use the same protocol on the same dataset.

We believe that the research community is sometimes over-focused on accuracy. But showing improved accuracy is still perceived to be a major requirement for publication. In a way, our results should help to shift the focus away from purely accuracy, because many papers may not be able to show statistically significantly improved accuracy using a cross-validation protocol, and so would need to better justify the other advantages of their approach.

The quality of experiment results in the biometrics literature could be improved if authors, reviewers and editors follow our EPD methodology and paid closer attention to the details of the protocol used to obtain the reported accuracy estimates.

Acknowledgment: This work was funded by Seed Grant Program of The College of Engineering at the Pontificia Universidad Catolica de Chile and the College of Engineering at the University of Notre Dame (ND-PUC 201506).

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE-PAMI*, 28(12):2037–2041, 2006.

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

[3] S. Baluja and H. A. Rowley. Boosting sex identification performance. *IJCV*, 2007.

[4] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, pages 1–8, 2008.

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

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

[7] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[8] W. Deng, J. Hu, and J. Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE-PAMI*, 34(9):1864–1870, 2012.

[9] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *CVPR*, pages 399–406, 2013.

[10] 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.

[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.

[12] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang. Single Sample Face Recognition via Learning Deep Supervised Auto-Encoders. *IEEE-TIFS*, 10(10):2108–2118, 2015.

[13] 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 Conference on System Science and Engineering*, 2010.

[14] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic Estimation from Face Images: Human vs. Machine Performance. *IEEE-PAMI*, 37(6):1148, 2015.

[15] 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.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

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

[18] A. Jain, B. Klare, and A. Ross. Guidelines for best practices in biometrics research. In *IEEE International Conf. Biometrics, Phuket, Thailand*, 2015.

[19] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *ECCV*, pages 331–344. Springer, 2012.

[20] Z. Jiang, Z. Lin, and L. S. Davis. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE-PAMI*, 35(11):2651–2664, 2013.

[21] A. Koutlas and D. I. Fotiadis. An automatic region based methodology for facial expression recognition. In *IEEE-SMC*, pages 662–666, 2008.

[22] H.-W. Kung, Y.-H. Tu, and C.-T. Hsu. Dual Subspace Nonnegative Graph Embedding for Identity-Independent Expression Recognition. *IEEE-TIFS*, 10(3):626–639, March 2015.

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

[24] 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, 2010.

[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, 2006.

[26] P. Liu, S. Han, Z. Men, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, 2014.

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

[28] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[29] E. Makinen and R. Raisamo. Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE-PAMI*, 30(3):514–547, March 2008.

[30] A. Martinez and R. Benavente. The AR face database, June 1998. CVC Tech. Rep, No. 24.

[31] J. Matey, G. Quinn, P. Grother, E. Tabassi, and C. Watson. Modest proposals for improving biometric recognition papers. In *IEEE 7th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2015)*, 2015.

[32] D. Mery and K. Bowyer. Face Recognition via Adaptive Sparse Representations of Random Patches. In *IEEE-WIFS*, 2014.

[33] D. Mery and K. Bowyer. Recognition of facial attributes using adaptive sparse representations of random patches. In *ECCV-Workshop on Soft Biometrics*, 2014.

[34] D. Mery and K. Bowyer. Automatic facial attribute analysis via adaptive sparse representation of random patches. *Pattern Recognition Letters*, 2015.

[35] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE-PAMI*, 24(5):707–711, 2002.

[36] I. Naseem, R. Togneri, and M. Bennamoun. Linear Regression for Face Recognition. *IEEE-PAMI*, 32(11):2106–2112, Nov 2010.

[37] H.-T. Nguyen and A. Caplier. Local Patterns of Gradients for Face Recognition. *IEEE-TIFS*, 10(8):1739–1751, 2015.

[38] 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.

[39] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[40] N. Poh, A. Martin, and S. Bengio. Performance generalization in biometric authentication using joint user-specific and sample bootstraps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):492–498, 2007.

[41] 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.

[42] Y. Shinohara and N. Otsu. Facial Expression Recognition Using Fisher Weight Maps. In *IEEE-FG 2004*, pages 499–504, 2004.

[43] 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-TIP*, 24(4):1386–1398, April 2015.

[44] 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 TNN*, 11(4):948–960, July 2000.

[45] 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. *IEEE-TIFS*, 8(3):488–499, 2013.

[46] 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.

[47] X. Wei, C. T. Li, and Y. Hu. Face recognition with occlusion using Dynamic Image-to-Class Warping (DICW). In *IEEE-FG 2013*, pages 1–6, 2013.

[48] R. Weng, J. Lu, J. Hu, G. Yang, and Y.-P. Tan. Robust Feature Set Matching for Partial Face Recognition. In *ICCV*, pages 601–608, Dec 2013.

[49] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE-PAMI*, 31(2):210–227, 2009.

[50] 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.

[51] M. Yang, D. Dai, L. Shen, and L. Van Gool. Latent dictionary learning for sparse representation based classification. In *CVPR*, pages 4138–4145, 2014.

[52] Z. Yang, M. Li, and H. Ai. An Experimental Study on Automatic Face Gender Classification. In *ICPR*, pages 1099–1102, 2006.

[53] 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.

[54] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010.

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

[56] 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.