Labeled Faces in the Wild: Updates and New Reporting Procedures

University of Massachusetts, Amherst Technical Report UM-CS-2014-003

Gary B. Huang and Erik Learned-Miller

***Abstract*— The Labeled Faces in the Wild (LFW) database has spurred significant research in the problem of unconstrained face verification and other related problems. While careful usage guidelines were established in the original technical report describing the database, certain unforeseen issues have arisen. One of the major issues is how to make fair comparisons among algorithms that use additional "outside data", i.e., data that is not part of LFW, for training. Another issue is the need for a clear definition of the "unsupervised paradigm" and the proper protocols for producing results under this paradigm. This technical report discusses these issues in detail and provides a new description of how we curate results and how we group algorithms together based on the details of the training data that they use. We encourage any authors who intend to publish their results on LFW to read both the original technical report and this one carefully.**

I. Introduction

The Labeled Faces in the Wild (LFW) database was published in 2007 and is described in a University of Massachusetts, Amherst, technical report from 2007 [6]. The database was designed to study the specific task in which a pair of two face images are presented, and a classifier is required to classify the pair as either "same" or "different" depending upon whether the images picture the same person or not.[1] While the original technical report referred to this problem as the *pair matching problem*, based on feedback from the vision community,[2] we have adopted the more widely used term of *face verification* for this problem.

The main purpose of this technical report is to update and clarify the specific paradigms of usage of the database, and the detailed protocols associated with those paradigms. In particular, several issues have arisen that make it unclear how various published methods should be compared to each other, and our primary goal is to resolve this issue. In addition, we describe how methods will be presented on the LFW results web page (`http://vis-www.cs.umass.edu/lfw/results.html`) and the criteria for being included therein.

The main issues covered by this report are as follows:

1) A definition of the *unsupervised paradigm*, which heretofore has not been defined by the curators of the database, but has emerged as a significant sub-area for publication of results.
2) An expansion of the number of categories of results, while maintaining backward compatibility of previously published results as much as possible. This includes a change in the definition of protocols that use "outside" training data, focused on whether that outside data uses same/different label information or not.
3) A discussion of the curators' method for determining whether results are eligible to be posted on the LFW website.

We start with a review of the original evaluation protocols, as described in the original LFW technical report [6].

[1]The terms "matched" or "mismatched" are also commonly used in place of "same" and "different".

[2]We thank Peter Belhumeur for bringing this issue to our attention.

II. The Original Evaluation Protocols

Training and testing data in LFW are presented as either matched or mismatched pairs. Other details of the database organization and provenance can be found in the original technical report [6], and we will not repeat them here.

In the original LFW technical report, just two different protocols were described: the image-restricted protocol and the unrestricted protocol. The essential difference between these was that in the unrestricted protocol, additional labeled training examples could be created by using the names associated with images. So if images A, B, C, and D were all George Bush, but the training data only contained the matched pairs (A, B) and (C, D), then under the *unrestricted* paradigm, one could add additional pairs such as (A, C) or (A, D) to increase the amount of data available for training an algorithm. This is not allowed under the image-restricted paradigm. For a more complete description of these protocols as they were originally defined, refer to the original LFW technical report [6]. Note, however, that from the time that this technical report is published, we will adopt the new protocol definitions given below. That is, this report supercedes the original report in any differing details.

A. Unforeseen uses of the database

There have been two major unforeseen uses of the database, which has created the need for new definitions and protocols. These are the use of the database in an "unsupervised" setting, and the use of outside training data to augment the training data of LFW.

III. The "unsupervised" setting

A variety of researchers have been interested in the question of how well the matched and mismatched pairs in the test sets

can be separated by a particular classifier that has not been tuned or trained on any same/different face pairs.

For example, one could associate a color histogram with each image in a pair, calculate the L_1 distance between the histograms, and set a threshold on this distance to categorize the images as "matched" (less than or equal to the threshold) or "mismatched" (above the threshold). In this example, and in many others of interest, the feature descriptor and distance function have been chosen without any label information about same/different pairs.

This basic idea has given rise to what authors have been referring to as "the unsupervised paradigm" [1]–[3], [7]–[10], although such a paradigm was not discussed in the original LFW technical report, and has not been defined by the curators of the database. Our goal here is to provide a standard definition of such a paradigm that fits with commonly accepted definitions of unsupervised learning. Next, we present a standard unsupervised protocol for LFW. Note that this protocol is in conflict with the way some authors have defined it in previously published works.

A. Unsupervised learning

In *unsupervised learning*, an algorithm cannot have *any access whatsoever* to class labels of the data, statistics of those labels, or means of generating those labels. In the context of LFW, this means that an unsupervised algorithm cannot have access to whether any image pairs are "same" or "different", since these are the relevant class labels in this problem. Additionally, the algorithm cannot have access to the names or unique identifiers of any individuals, since this would allow the algorithm to create pairs of images that were "same" and "different" by pairing images of people whose names were the same or different. Finally, and this is a more subtle point, an algorithm cannot be told the *distribution* of labels in a training set, even when those labels are not provided for any particular image. For example, one might try to leverage the fact that approximately half of the pairs in a training set are matched and half are mismatched, by finding a threshold that splits the training data into two halves based on the distance between the images in each pair. While such a method does not use explicit pair labels, it does use the statistics of the pair labels, and hence is not allowed under the unsupervised learning paradigm.

Of course, if one relaxes the strict requirement of unsupervised learning, there are many ways of using weakly labeled data, noisily labeled data, partially labeled data, using the statistics of labels, and so on, but we do not create categories for these paradigms in our results. The reason is that there are so many flavors of semi-supervised learning that it is impractical to create categories for all of them. Instead we focus on clarifying the definition of an unsupervised learning paradigm in the context of LFW.

B. The LFW Unsupervised Paradigm: Definition

In the **unsupervised paradigm**, the practitioner should prepare a scalar-valued function $f(I, J)$ of two images I and J which returns a scalar d, such that increasing d implies a greater distance or dissimilarity between the images I and J. Any threshold θ then produces a binary classifier such that the class label is "same" when $f(I, J) \leq \theta$ and the class label is "different" when $f(I, J) > \theta$. This threshold θ can be varied to produce an ROC curve. **However, it is important to note that unsupervised learning gives no method for producing a specific threshold θ^* that can be used to define the scalar-valued accuracy of the classifier** (other than choosing it randomly). Instead, as we describe below, we adopt the area under the ROC curve (AUC) as a suitable scalar-valued measure of accuracy for the unsupervised paradigm.

In order to follow the protocol properly, the following restrictions apply:

- The function $f(\cdot, \cdot)$ should have no parameters that are set using any information about the LFW class labels of "same" and "different". Some such uses of label information are obvious. For example, we cannot apply a metric learning method to determine the function $f(\cdot, \cdot)$ by minimizing the empirical error on a training set. However, other uses of "labeled" data are more subtle, but are also not allowed. For example, one could adjust the function $f(\cdot, \cdot)$ so that half of the training data pairs were greater than some value θ and half were less than θ, without using the specific labels of those pairs. However, this is not unsupervised learning, since it takes advantage of the fact that half of the training data are "matched" and half are "unmatched". In other words, it is using information about the labels that is not allowed in an unsupervised setting.
- As mentioned above, $f(\cdot, \cdot)$ cannot be trained using images labeled with the individuals' names or any unique identifiers, since this is tantamount to training the classifier on "same" and "different" pairs.
- The threshold θ also *cannot be set by looking at training results or test results*. Some authors have argued that it is reasonable to name a method "unsupervised" as long as it is only the classification threshold θ that is set in a supervised manner. While we agree that this procedure may be of scientific interest, we do not agree that it represents unsupervised learning, and hence we disallow this in our unsupervised learning protocol.

As a rule of thumb, any procedure claiming to be unsupervised should make no use whatsoever of any "pairs" of images in which the pairs have been drawn from a distribution with a known label distribution. However, unsupervised methods **may appropriately use**:

- Statistics of the individual face images in a training portion of LFW, as long as the identities of those individuals are not used in any way.
- Statistics of other images outside of LFW, but again, without any information about whether pairs of images have the same identity or not. This would rule out collections of images of the same person, for example.

C. Protocol for the Unsupervised Paradigm

The following sequence of steps should be followed to follow the proper protocol under the unsupervised LFW paradigm:

- Define scalar-valued function $f(\cdot, \cdot)$ of two images.
- For each pair of images in a test set, compute the value of $f(\cdot, \cdot)$.
- Sort the image pairs according to their computed f value. Let these values of f, in sorted order be denoted $f_{(1)}, f_{(2)}, ..., f_{(N)}$.
- Now consider a set of thresholds equal to $f_{(1)}, f_{(2)},$, etc. Each threshold will generate a performance number and hence an ROC curve. (Note that there should be one additional threshold that is less than all of the f values, and classifies all pairs as mismatched.)
- To facilitate this process, we have provided code both for generating ROC curves and for computing the area under the ROC curve. These can be found on the LFW results page under the respective subheading.

We believe following these procedures will provide fair comparisons of results under this paradigm, while remaining consistent with the generally accepted meaning of "unsupervised". Note that we do not distinguish between unsupervised methods that use only unlabeled data within LFW, and those that use external, non-LFW unlabeled data. While this distinction may be of some scientific interest, there are currently not enough entries in this category to warrant splitting it further.

We now turn our attention to issues associated with supervised learning paradigms.

IV. Complications of using outside training data

While the original LFW protocols were described under the assumption that researchers would only use training data that was part of LFW itself, many researchers were interested in exploring additional sources of training data to improve performance. Examples of this "outside" training data have included or might include:

- Using faces (either non-LFW faces or training data from LFW) labeled with keypoints or parts (such as corners of the eyes) in order to produce pre-alignment algorithms that could improve performance of a face verification algorithm.
- Using an off-the-shelf face alignment algorithm that had been trained with non-LFW labeled data, and hence implicitly using this non-LFW data.
- Using (possibly large) sets of unlabeled face images from outside of LFW to study the statistics of face images, and potentially improve descriptors for LFW. For example, using a million non-LFW face images, one might define a face-based visual dictionary using Gaussian mixture modeling in order to build a novel feature representation for use in face verification. Of course, implementing this same idea by using the training sets provided with LFW separately for each training-test split would *not* constitute the use of outside data.
- Using additional matched or unmatched pairs from outside of LFW, taking care not to use any of the same people that are associated with a given LFW test set. (Using images of any person that appears in an LFW test set, during training, is disallowed under all LFW protocols.)

A. A brief note on humans in the loop

Note that there are no sanctioned LFW protocols that allow the hand-labeling of parts in a test image, the manual alignment of test images, or any sort of "human in the loop" processing. While such systems are certainly of scientific interest, they do not fit within any of the pre-defined protocols described for LFW.

B. Categories of outside data

Because there are so many ways that outside data could be used to enhance performance on face verification, it is challenging to find a way to compare algorithms fairly. When people started to use outside data, we initially divided the image-restricted results into two categories: those that used outside data and those that did not. Shortly thereafter, we decided that using outside data solely for the purposes of alignment was significantly different than using outside data to train a classifier. In addition, while many of these distinctions apply to both the image-restricted and unrestricted paradigms, most people who were using outside data early on were only reporting results on the image-restricted protocol, so we decided to split our image-restricted results, but not our unrestricted results, into multiple categories. Recently, in the face of a large number of new reports that use outside data in a variety of different ways, we decided it was time to revisit our protocols, and try to improve them.

C. Enumeration of protocols

From here forward, we plan to report results for six different protocols. The details of each protocol are discussed below, and Table I summarizes the allowable training data for each protocol.

Including the unsupervised protocol that was discussed above, the six protocols are:

1) Unsupervised.
2) Image-restricted with no outside data.
3) Unrestricted with no outside data.
4) Image-restricted with label-free outside data.
5) Unrestricted with label-free outside data.
6) Unrestricted with labeled outside data.

Please refer to the protocols by these exact names, as there will no doubt be confusion if different names are used. The reader may notice that there is no category "image-restricted with labeled outside data". The reason for this is that with arbitrary labeled outside data, there is no longer a useful distinction between the image-restricted and unrestricted paradigms, and so we collapse these into a single category. Since the unsupervised protocol is discussed above, we will continue our discussion with the second protocol.

Protocol	Same/Different Labels for LFW training pairs allowed?	Identity info for LFW training images allowed?	Annotations for LFW training data allowed?	Non-LFW images allowed?	Non-LFW annotations allowed?	Same/Different labels for non-LFW pairs allowed?	Identity info for non-LFW images allowed?
Unsupervised	**no**	**no**	yes	yes	yes	**no**	**no**
Image-Restricted, No Outside Data	yes	**no**	**no**	**no**	**no**	**no**	**no**
Unrestricted, No Outside Data	yes	yes	**no**	**no**	**no**	**no**	**no**
Image-Restricted, Label-Free Outside Data	yes	**no**	yes	yes	yes	**no**	**no**
Unrestricted, Label-Free Outside Data	yes	yes	yes	yes	yes	**no**	**no**
Unrestricted With Labeled Outside Data	yes	yes	yes	yes	yes	yes	yes

TABLE I: This table summarizes the new LFW protocols. There are six protocols altogether, shown in the left column. The allowability for each category of data is shown to the right. Refer to the main text for additional details.

D. Image-restricted with no outside data

This is the original "image-restricted" protocol described in the original LFW technical report. It assumes that no data from outside LFW will be used, including additional images, or tools such as eye-detectors, alignment methods, or feature extractors that have been trained on outside data.[3] Of course the same/different training labels are expected to be used, but as described above in Section II, one is disallowed from using the names of people to generate additional training examples through transitivity of identity.

Note that some tools like face alignment algorithms or facial feature detectors may be usable under this paradigm provided that they adhere to the following rules:

- They do not use any data outside of LFW, even for their original training.
- They do not rely on any additional annotations, such as the manual localization of facial landmarks.
- They are developed using the training sets only, and none of the test data.

In other words, such tools must be completely unsupervised in nature, and the unsupervised data on which they operate must be entirely within the training sets of LFW. For example, the congealing algorithms for face alignment obey these rules [4], [5]. The first of these [4] was used to produce the "funneled" version of LFW. The second [5] was used to produce the "deep funneled" version of LFW. Both of these alternative versions of LFW may be used legitimately under this paradigm. **Note, however, that the LFW-a data set, which relies on training data outside of LFW, may not be used under this protocol.**

[3]One may reasonably argue that even highly generic descriptors such as SIFT have been tuned using data that is outside of LFW, and hence, by this definition, would not be allowed under the *image-restricted with no outside data* paradigm. While this is technically correct, we will allow descriptors that have been trained for a completely different purpose to be used under this paradigm. It is quite difficult to define a clear line between what types of descriptors and preprocessing methods could be used, since many descriptors may have been tuned on "natural images". As such, we will reserve the right to categorize a method as using outside data or not based upon our judgment of whether the descriptors or methods have been specifically adapted to faces or the face verification problem. In general, if a descriptor or method is "pre-existing", and was developed without regard to its use in any face processing task, then we will allow it under the *image-restricted with no outside data* paradigm. If a descriptor or method has been built using face data, face parts, face models, etc. as input, we will not allow it under this paradigm.

E. Unrestricted with no outside data

Like the previous protocol, no outside data is allowed, either in the form of images or pre-trained functions such as detectors or alignment algorithms. The only difference between this protocol and the previous protocol is that additional training data, in the form of new pairs of "same" and "different" images may be used by leveraging the names of people associated with the LFW training data. For details on how to do this, refer to the original technical report.

F. Image-restricted with label-free outside data

This protocol is image-restricted, in that using names associated with the LFW training images is not allowed. However, using additional data from outside LFW, or certain types of LFW annotation *are allowed*, provided that:

- The outside data cannot contain any information about whether two images are "same" or "different".
- The outside data cannot contain the identity of any individual, since this can be used to create "same" and "different" pairs.

The outside data may legitimately include:

- Images, patches, or other data sources from outside LFW, including face images with no name or identity labels. A subtle point here is whether unlabeled movies of faces should be allowed. In this case, it is trivial to produce pairs of "same" faces by tracking a face in a video, and for this reason, we disallow movies of faces under this protocol, even if they are not specifically labeled.
- Annotations of data sources from outside of LFW, as long as those annotations do not include person names or other information that would allow the creation of "same" or "different" face pairs.
- Annotations of LFW *training* images, such as the location of features, or segmentations.

Note that using an alignment algorithm that has been trained using outside data also constitutes the (legitimate) use of outside data. In summary, arbitrary additional data can be used as long as this data does not allow the creation of same/different pairs not found in LFW. Note that additional same/different pairs *are allowable* under the final protocol (unrestricted with labeled outside data).

G. Unrestricted with label-free outside data

This protocol is the same as the previous one, except the names of the LFW training images can be used to create additional same/different pairs, as described in Section II. Note that names of other, non-LFW face images are *not allowed* under this protocol.

H. Unrestricted with labeled outside data

This is the most permissive protocol, and allows training on many types of external training data, including:

- Additional pairs of faces labeled as "same" or "different", as long as they do not contain individuals in an LFW test set.
- The names of any individual face image, whether outside LFW or in the LFW training sets, as long as they do not contain individuals in the LFW test sets.
- Arbitrary annotations such as feature localizations, segmentations, or attributes, for either external data or LFW training data.

V. Summary of protocol types and inclusion of results on the LFW Results Page

Table I provides a summary of the protocol types for LFW. We will group performance numbers by these six categories, so researchers who publish results should carefully specify which categories they are reporting for, using the terms in the leftmost column of the table.

To ensure that their results are put in the proper category, authors should declare that their training data does not contain any of the disallowed categories of data as described in this report. For example, if a set of experiments is done under the *unrestricted with no outside data* paradigm, the authors might write:

> We trained our classifier in the unrestricted setting, i.e., we created all possible same/different pairs from the LFW training sets. However, no supplementary annotations or other labels were used, and no images, annotations, or other data sources from outside LFW were used. As a result we are publishing our results under the *unrestricted with no outside data* protocol.

If, upon reading a publication, we are unable to determine which protocol was used, we will either

- contact the authors for clarification,
- publish the results under the most permissive protocol (unrestricted with labeled outside data),
- refrain from publishing the results.

To the extent that time allows, we will work with authors to clarify issues around which protocols were followed. Of course, such issues will be minimized if authors are as clear and explicit as possible about exactly how their face verification systems were developed and trained.

A. Additional protocols not defined by this document

Of course, as the curators of LFW results, we welcome the publication of additional experiments that use other protocols to train and test classifiers on LFW, such as "human-in-the-loop" protocols. However, if these protocols do not conform to the procedures laid out in this document, we will not be able to put them on our results page. There are simply too many types of results for us to track all of them. We encourage authors to give detailed explanations of their alternative protocols, and eventually, if there are enough examples of such a protocol, we may add it to our list of curated results. This is what gave rise to the unsupervised protocol, for example.

References

[1] Shervin Rahimzadeh Arashloo and Josef Kittler. Efficient processing of MRFs for unconstrained-pose face recognition. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8. IEEE, 2013.

[2] Oren Barkan, Jonathan Weill, Lior Wolf, and Hagai Aronowitz. Fast high dimensional vector multiplication face recognition. In *International Conference on Computer Vision*, pages 1–8, 2013.

[3] Javier Ruiz del Solar, Rodrigo Verschae, and Mauricio Correa. Recognition of faces in unconstrained environments: A comparative study. *EURASIP Journal on Advances in Signal Processing (Recent Advances in Biometric Systems: A Signal Processing Perspective*, 2009(184617):19, 2009.

[4] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *International Conference on Computer Vision*, 2007.

[5] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems 25*, pages 773–781, 2012.

[6] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] Sibt ul Hussain, Thibault Napoléon, and Frédéric Jurie. Face recognition using local quantized patterns. In *British Machine Vision Conference*, pages 1–11, 2012.

[8] Hae Jong and Peyman Milanfar. Face verification using the LARK representation. *IEEE Transactions on Information Forensics and Security*, 6(4):1275–1286, 2011.

[9] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie. Local higher-order statistics (LHS) for texture categorization and facial analysis. In *European Conference on Computer Vision*, pages 1–12, 2012.

[10] Dong Yi, Zhen Lei, and Stan Z Li. Towards pose robust face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2013.