Labeled Faces in the Wild: A Survey

Erik Learned-Miller, Gary Huang, Aruni RoyChowdhury, Haoxiang Li, Gang Hua

Abstract In 2007, Labeled Faces in the Wild was released in an effort to spur research in face recognition, specifically for the problem of face verification with unconstrained images. Since that time, more than 50 papers have been published that improve upon this benchmark in some respect. A remarkably wide variety of innovative methods have been developed to overcome the challenges presented in this database. As performance on some aspects of the benchmark approaches 100% accuracy, it seems appropriate to review this progress, derive what general principles we can from these works, and identify key future challenges in face recognition. In this survey, we review the contributions to LFW for which the authors have provided results to the curators (results found on the LFW results web page). We also review the cross cutting topic of alignment and how it is used in various methods. We end with a brief discussion of recent databases designed to challenge the next generation of face recognition algorithms.

Erik Learned-Miller
University of Massachusetts, Amherst, Massachusetts, e-mail: elm@cs.umass.edu

Gary B. Huang
Howard Hughes Medical Institute, Janelia Research Campus, e-mail: gbhuang@cs.umass.edu

Aruni RoyChowdhury
University of Massachusetts, Amherst, Massachusetts, e-mail: aruni@cs.umass.edu

Haoxiang Li
Stevens Institute of Technology, Hoboken, New Jersey, e-mail: hli18@stevens.edu

Gang Hua
Stevens Institute of Technology, Hoboken, New Jersey, e-mail: ganghua@gmail.com

1 Introduction

Face recognition is a core problem and popular research topic in computer vision for several reasons. First, it is easy and natural to formulate well-posed problems, since individuals come with their own label, their name. Second, despite its well-posed nature, it is a striking example of *fine-grained classification*–the variation of two images within a class (images of a single person) can often exceed the variation between images of different classes (images of two different people). Yet human observers have a remarkably easy time ignoring nuisance variables such as pose and expression and focusing on the features that matter for identification. Finally, face recognition is of tremendous societal importance. In addition to the basic ability to identify, the ability of people to assess the emotional state, the focus of attention, and the intent of others are critical capabilities for successful social interactions. For all these reasons, face recognition has become an area of intense focus for the vision community.

This article reviews research progress on a specific face database, Labeled Faces in the Wild (LFW), that was introduced to stimulate research in face recognition for images taken in common, everyday settings. In the remainder of the introduction, we review some basic face recognition terminology, provide the historical setting in which this database was introduced, and enumerate some of the specific motivations for introducing the database. In Section 2, we discuss the papers for which the curators have been provided with results. We group these papers by the protocols for which they have reported results. In Section 3, we discuss alignment as a cross-cutting issue that affects almost all of the methods included in this survey. We conclude by discussing future directions of face recognition research, including new databases and new paradigms designed to push face recognition to the next level.

1.1 Verification and identification

In this article, we will refer to two widely used paradigms of face recognition: *identification* and *verification*. In identification, information about a specific set of individuals to be recognized (the *gallery*) is gathered. At test time, a new image or group of images is presented (the *probe*). The task of the system is to decide which of the gallery identities, if any, is represented by the probe. If the system is guaranteed that the probe is indeed one of the gallery identities, this is known as *closed set* identification. Otherwise, it is *open set* identification, and the system is expected to identify when an image does not belong to the gallery.

In contrast, the problem of *verification* is to analyze two face images and decide whether they represent the *same* person or two *different* people. It is usually assumed that neither of the photos shows a person from any previous training set.

Many of the early face recognition databases and protocols focused on the problem of identification. As discussed below, the difficulty of the identification problem was so great that researchers were motivated to simplify the problem by controlling the number of image parameters that were allowed to vary simultaneously. One of the salient aspects of LFW is that it focused on the problem of verification exclusively, although it was certainly not the first to do so.[1] While the use of the images in LFW originally grew out of a motivation to study learning from one example and fine-grained recognition, a side effect was to render the problem of face recognition in real-world settings signficantly easier–easier enough to attract the attention of a wide range of researchers.

1.2 Background

In the early days of face recognition by computer, the problem was so daunting that it was logical to consider a divide-and-conquer approach. What is the best way to handle recognition in the presence of lighting variation? Pose variation? Occlusions? Expression variation? Databases were built to consider each of these issues using carefully controlled images and experiments.[2] One of the most comprehensive efforts in this direction is the CMU Multi-PIE[3] database, which systematically varies multiple parameters over an enormous database of more than $750,000$ images [38].

Studying individual sources of variation in images has led to some intriguing insights. For example, in their efforts to characterize the structure of the space of images of an object under different lighting conditions, Belhumeur et al. [15] showed that the space of faces under different lighting conditions (with other factors such as expression and pose held constant) forms a convex cone. They propose doing lighting invariant recognition by examining the distance of an image to the convex cones defined for each individual.

Despite the development of methods that could successfully recognize faces in databases with well-controlled variation, there was still a gap in the early 2000's between the performance of face recognition on these controlled databases and results on real face recognition tasks, for at least two reasons:

- Even with two methods, call them A and B, that can successfully model two types of variation separately, it is not always clear how to combine these methods to produce a method that can address both sources of variation. For example, a method that can handle significant occlusions may rely on the precise registration of two face images for the parts that are not occluded. This might render the method ineffective for faces that exhibit both occlusions and pose changes. As another example, the method cited above to handle lighting variations [15] relies on all of the other parameters of variation being fixed.

[1] Other well-known benchmarks had previously used verification. See, for example, this benchmark [80].

[2] For a list of databases that were compiled before LFW, see the original LFW technical report [49].

[3] The abbreviation PIE stands for Pose, Illumination, and Expression.

- There is a significant difference between handling *controlled* variations of a parameter, and handling *random* or *arbitrary* values of a parameter. For example, a method that can address five specific poses may not generalize well to arbitrary poses. Many previously existing databases studied fixed variations of parameters such as pose, lighting, and decorations. While useful, this does not guarantee the handling of more general cases of these parameters. Furthermore, there are too many sources of variation to effectively cover the set of possible observations in a controlled database. Some databases, such as the ones used in the 2005 Face Recognition Grand Challenge [77], used certain "uncontrolled settings" such as an office, a hallway, or outdoor environments. However, the fact that these databases were built manually (rather than mining previously existing photos) naturally limited the number of settings that could be included. Hence, while the settings were uncontrolled in that they were not carefully specified, they were drawn from a small fixed set of empirical settings that were available to the database curators. Algorithms tuned for such evaluations are not required to deal with a large amount of previously unseen variability.

In 2006, while results on some databases were saturating, there was still poor performance on problems with real-world variation.

1.3 Variations on traditional supervised learning and the relationship to face recognition

In parallel to the work in the early 2000's on face identification, there was a growing interest in the machine learning community in variations of the standard supervised learning problem with large training sets. These variations included:

- learning from small training sets [68, 35],
- transfer learning–that is, sharing parameters from certain classes or distributions to other classes or distributions that may have less training data available [76], and
- semi-supervised learning, in which some training examples have no associated labels (e.g. [73]).

Several researchers chose face verification as a domain in which to study these new issues [21, 30, 36]. In particular, since face verification is about deciding whether two face images match (without any previous examples of those identities), it can be viewed as an instance of learning from a single training example. That is, letting the two images presented be I and J, I can be viewed as a single training example for the identity of a particular person. Then the problem can be framed as a binary classification problem in which the goal is to decide whether image J is in the same class as image I or not.

In addition, face verification is an ideal domain for the investigation of transfer learning, since learning the forms of variation for one person is important information that can be transferred to the understanding of how images of another person vary.

One interesting paper in this vein was the work of Chopra et al. from CVPR 2005 [30]. In this paper, a convolutional neural network (CNN) was used to learn a metric between face images. The authors specifically discuss the structure of the face recognition problem as a problem with a large number of classes and small numbers of training examples per class. In this work, the authors reported results on the relatively difficult AR database [66]. This paper was a harbinger of the recent highly successful application of CNNs to face verification.

1.3.1 Faces in the Wild and Labeled Faces in the Wild

Continuing the work on fine-grained recognition and recognition from a small number of examples, Ferencz et al. [57, 36] developed a method in 2005 for deciding whether two images represented the same object. They presented this work on data sets of cars and faces, and hence were also addressing the face verification problem. To make the problem challenging for faces, they used a set of news photos collected as part of the Berkeley "Faces in the Wild" project [19, 18] started by Tamara Berg and David Forsyth. These were news photos taken from typical news articles, representing people in a wide variety of settings, poses, expressions, and lighting. These photos proved to be very popular for research, but they were not suited to be a face recognition benchmark since a) the images were only noisily labeled (more than 10% were labeled incorrectly), and b) there were large numbers of duplicates. Eventually, there was enough demand that the data were relabeled by hand, duplicates were removed, and protocols for use were written. The data were released as "Labeled Faces in the Wild" in conjunction with the original LFW technical report [49].

There were several goals behind the introduction of LFW. These included

- stimulating research on face recognition in unconstrained images;
- providing an easy-to-use database, with low barriers to entry, easy browsing, and multiple parallel versions to lower pre-processing burdens;
- providing consistent and precise protocols for the use of the database to encourage fair and meaningful comparisons;
- curating results to allow easy comparison, and easy replication of results in new research papers.

In the following section, we take a detailed look at many of the papers that have been published using LFW. We do not review all of the papers. Rather we review papers for which the authors have provided results to the curators, and which are documented on the LFW results web page.[4] We now turn to describing results published on the LFW benchmark.

[4] http://vis-www.cs.umass.edu/lfw/results.html.

2 Algorithms and methods

In this section, we discuss the progression of results on LFW from the time of its release until the present. LFW comes with specific sets of image pairs that can be used in training. These pairs are labeled as "same" or "different" depending upon whether the images are of the same person. The specification of exactly how these training pairs are used is described by various protocols.

2.1 The LFW Protocols

Originally, there were two distinct protocols described for LFW, the *image-restricted* and the *unrestricted* protocols. The unrestricted protocol allows the creation of additional training pairs by combining other pairs in certain ways. (For details, see the original LFW technical report [49].)

As many researchers started using additional training data from outside LFW to improve performance, new protocols were developed to maintain fair comparisons among methods. These protocols were described in a second technical report [47].

The current 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.

In order to make comparisons more meaningful, we discuss the various protocols in three groups.

In particular, we start with the two protocols allowing no outside data. We then discuss protocols that allow outside data not related to identity, and then outside data with identity labels. We do not address the unsupervised protocol in this review.

2.1.1 Why study restricted data protocols?

Before starting on this task, it is worth asking the following question: Why might one wish to study methods that do not use outside data when their performance is clearly inferior to those that do use additional data? There are several possible answers to this question.

Utility of methods for other tasks. One reason to consider methods which use limited training data is that they can be used in other settings in which training data are limited. That is, it may be the case that in recognition problems other than face recognition, there may not be available the hundreds of thousands or millions of images that are available to train face recognizers. Thus, a method that uses less training data is more transportable to other domains.

Statistical efficiency versus asymptotic optimality. It has been known since the mid-seventies [87] that many methods, such as K–nearest neighbors (K-NN), continue to increase in accuracy with increasing training data until they reach optimal performance (also known as the *Bayes error rate*). In other words, if one only cares about accuracy with unlimited training data and unlimited computation time, there is no method better than K-NN.

Thus, we know not only that many methods will continue to improve as more training data is added, but that many methods, including some of the simplest methods, will achieve optimal performance. This makes the question of *statistical efficiency* a primary one. The question is not *whether* we can achieve optimal accuracy (the Bayes error rate), but rather, how fast (in terms of training set size) we get there. Of course, a closely related question is which method performs best with a fixed training set size.

At the same time, using equivalent data sets for training removes the question that plagues papers trained on huge, proprietary data sets: how much of their performance is due to algorithmic innovation, and how much is simply due to the specifics of the training data?

Despite our interest in fixed training set protocols, at the same time, the practical issues of how to collect large data sets, and find methods that can benefit from them the most, make it interesting to push performance as high as possible with no ceiling on the data set size. The protocols of LFW consider all of these questions.

Human learning and statistical efficiency. Closely related to the previous point is to note that humans solve many problems with very limited training data. While some argue that there is no particular need to mimic the way that humans solve problems, it is certainly interesting to try to discover the *principles* which allow them to learn from small numbers of examples. It seems likely that these principles will improve our ability to design efficient learning algorithms.

Allowed information → **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, Labeled Outside Data	yes	yes	yes	yes	yes	yes	yes

Table 1 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. The second LFW technical report gives additional details about these protocols [47].

2.1.2 Order of discussion

Within each protocol, we primarily discuss algorithms in the order with which we received the results. Note that this order does not always correspond to the official publication order.[5] We make every effort to remark on the first authors to use a particular technique, and also to refer to prior work in other areas or on other databases that may have used similar techniques previously. We apologize for any oversights in advance. Note that some methods, especially some of the commercial ones, do not give much detail about their implementations. Rather than devoting an entire section to methods for which we have little detail, we summarize them in Section 2.5. We now start with protocols incorporating labeled outside data.

2.2 Unrestricted with labeled outside data

This protocol allows the use of same and different training pairs from outside of LFW. The only restriction is that such data sets should not include pictures of people whose identities appear in the test sets. The use of such outside data sets has dramatically improved performance in several cases.

2.2.1 Attribute and simile classifiers for face verification, 2009 [53]

Kumar et al. [53] present two main ideas in this paper. The first is to explore the use of describable attributes for face verification. For attribute classifiers 65 describable visual traits such as gender, age, race, and hair color are used. At least 1000 positive and 1000 negative pairs of each attribute were used for training each attribute classifier. The paper gives the accuracy of each individual attribute classifier. Note that the attribute classifier does *not* use labeled outside data, and thus, when not used in conjunction with the simile classifier, qualifies for the *unlabeled outside data* protocols.

The second idea develops what they call *simile* classifiers, in which various classifiers are trained to rate face parts as “similar” or “not similar” to the face parts of certain reference individuals. To train these “simile” classifiers, multiple images of the same individuals (from outside of the LFW training data) are used, and thus this method uses outside labeled data.

The original paper [53] gives an accuracy of 85.29±1.23% for the hybrid system, and a follow-up journal paper [54] gives slightly higher results of 85.54±0.35%.

[5] Some authors have sent results to the curators before papers have been accepted at peer-reviewed venues. In these cases, as described on the LFW web pages, we highlight the result in our results table in red, indicating that it has not yet been published at a peer-reviewed venue. In most cases, the status of such results are updated once the work has been accepted at a peer-reviewed venue. However, we maintain the original order in which we received the results.

These numbers should be adjusted downward slightly to 84.52% and 84.78% since there was an error in how their accuracies were computed.[6]

This paper was also notable in that it gave results for human recognition on LFW (99.2%). While humans had an unfair advantage on LFW since many of the LFW images were celebrities, and hence humans have seen prior images of many test subjects, which is not allowed under any of the protocols, these results have nevertheless been widely cited as a target for research. The authors also noted that humans could do remarkably well using only close crops of the face (97.53%), and even using only "inverse crops", including none of the face, but portions of the hair, body, and background of the image (94.27%).

2.2.2 Face Recognition with Learning-Based Descriptor, 2010 [26]

Cao et al. [26] develop a visual dictionary based on unsupervised clustering. They explore K-means, principal components analysis (PCA) trees [37] and random projection trees [37] to build the dictionary. While this was a relatively early use of learned descriptors, they were not learned discriminatively, i.e. to optimize performance.

One of the other main innovative aspects of this paper was building verification classifiers for various combinations of poses such as frontal-frontal, or rightfacing-leftfacing, to optimize feature weights conditioned on the specific combination of poses. This was done by finding the nearest pose to training and test examples using the Multi-PIE data set [38]. Because the Multi-PIE data set uses multiple images of the same subject, this paper is put in the category with outside labeled data. However, it seems plausible that this method could be used on a subset of multi-PIE that did not have images of the same person, as long there was a full range of labeled poses. Such a method, if pursued would qualify these techniques for the category *image-restricted with label-free outside data*.

The highest accuracy reported for their method was 84.45±0.46%.

2.2.3 An Associate-Predict Model for Face Recognition, 2011 [110]

This paper was one of the first systems to use a large additional amount of outside labeled data, and was, perhaps not coincidentally, the first system to achieve over 90% on the LFW benchmark.

The main idea in this paper (similar to some older work [13]) was to *associate* a face with one person in a standard reference set, and use this reference person to *predict* the appearance of the original face in new poses and lighting conditions.

[6] The authors reported that their classifier failed to complete, due to a failed preprocessing step, in 53 out of 6000 cases. According to the footnote in their journal paper, they scored about 85% of these cases as correct. However, according to the protocol, if an answer is not given, the test sample must be considered incorrect.

Building on the previous work by one of the co-authors [26], this paper also uses different strategies depending upon the relative poses of the presented face pair. If the poses of the two faces are deemed sufficiently similar, then the faces are compared directly. Otherwise, the associate-predict method is used to try to map between the poses. The best accuracy of this system on the *unrestricted with labeled outside data* was 90.57±0.56%.

2.2.4 Leveraging Billions of Faces to Overcome Performance Barriers in Unconstrained Face Recognition, 2011 [95]

This proprietary method from `Face.com` uses 3D face frontalization and illumination handling along with a strong recognition pipeline and achieves 91.30±0.30% accuracy on LFW. They report having amassed a huge database of almost 31 billion faces from over a billion persons.

They further discuss the contribution of effective 3D face alignment (or *frontalization* to the task of face verification, as this is able to effectively take care of out-of-plane rotation, which 2D based alignment methods are not able to do. The 3D model is then used to render all images into a frontal view. Some details are given about the recognition engine – it uses non-parametric discriminative models by leveraging their large labeled data set as exemplars.

2.2.5 Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification, 2012 [16]

This work presented two significant innovations. The first was to do a new type of non-affine warping of faces to improve correspondences while preserving as much information as possible about identity. While previous work had addressed the problem of non-linear pose-normalization (see, for example, the work by Asthana et al. [10, 11]), it had not been successfully used in the context of LFW.

In particular, as the authors note, simply warping two faces to maximize similarity may reduce the ability to perform verification by eliminating discriminative information between the two individuals. Instead, a warping should be done to maximize similarity while maintaining identity information. The authors achieve this identity-preserving warping by adjusting the warping algorithm so that parts with informative deviations in geometry (such as a wide nose) are preserved better (see the paper for additional details). This technique makes about a 2% (91.20% to 93.10%) improvement in performance relative to more standard alignment techniques.

This paper was also one of the first evaluated on LFW to use the approximate symmetry of the face to its advantage. Since using the above warping procedure tends to distort the side of the face further from the camera, the authors reflect the face, if necessary, such that the side closer to the camera is always on the right side of the photo. This results in the right side of the picture typically being more faithful to the appearance of the person. As a result, the learning algorithm which is subsequently applied to the flipped faces can learn to rely more on the more faithful side of the face. It should be noted, however, that the algorithm pays a price when the person's face is asymmetric to begin with, since it may need to match the left side of a person's face to their own right side. Still this use of facial symmetry improves the final results.

The second major innovation was the introduction of so-called *Tom-vs-Pete* classifiers as a new type of learned feature. These features were developed by using external labeled training sets (also labeled with part locations) to develop binary classifiers for pairs of identities, such as two individuals named Tom and Pete. For each of the $\binom{n}{2}$ pairs of identities in the external training set, k separate classifiers are built, each using SIFT features from a different region of the face. Thus, the total number of Tom-vs-Pete classifiers is $k \times \binom{n}{2}$. A subset of these were chosen by maximizing discriminability.

The highest accuracy of their system was $93.10 \pm 1.35\%$. However, they increased accuracy (and reduced the standard error) a bit further by adding attribute features based upon their previous work, to $93.30 \pm 1.28\%$.

2.2.6 Bayesian Face Revisited: A Joint Formulation, 2012 [28]

One of the most important aspects of face recognition in general, viewed as a classification problem, is that all of the classes (represented by individual identities) are highly similar. At the same time, within each class is a significant amount of variability due to pose, expression, and so on. To understand whether two images represent the same person, it can be argued that one should model both the distribution of identities, and also the distribution of variations within each identity.

This basic idea was originally proposed by Moghaddam et al. in their well-known paper "Bayesian face recognition" [69]. In that paper, the authors defined a difference between two images, estimated the distribution of these differences conditioned on whether the images were drawn from the same identity or not, and then evaluated the posterior probability that this difference was due to the two images coming from different identities.

Chen et al. [28] point out a potential shortcoming of the probabilistic method applied to image differences. They note that by forming the image difference, information available to distinguish between two classes (in this case the "same" versus "different" classes of the verification paradigm) may be thrown out. In particular, if $\mathbf{x}$ and $\mathbf{y}$ are two image vectors of length N, then the pair of images, considered as a concatenation of the two vectors, contains $2N$ components. Forming the difference image is a linear operator corresponding to a projection of the image pair back to N dimensions, hence removing some of the information that may be useful in deciding whether the pair is "same" or "different". This is illustrated in Figure 1. To address this problem, Chen et al. focus on modeling the joint distribution of image pairs (of dimension $2N$) rather than the difference distribution (of dimension N). This is an elegant formulation that has had a significant impact on many of the follow-up papers on LFW.

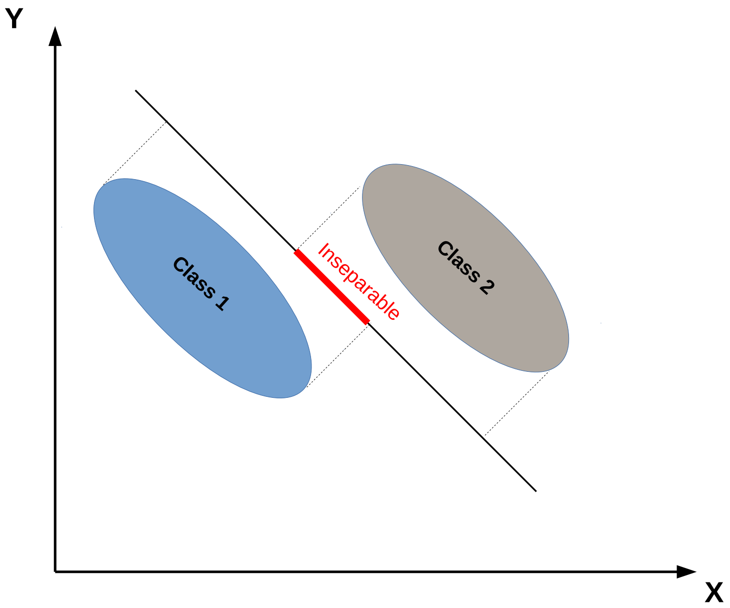

Fig. 1 When the information from two images is projected to a lower dimension by forming the difference, discriminative information may be lost. The joint Bayesian approach [28] strives to avoid this projection, thus preserving some of the discriminative information.

Another appealing aspect of this paper is the analysis that shows the relationship between the joint Bayesian method and the reference-based methods, such as the simile classifier [53], the multiple one-shots method [96], and the associate-predict method [110]. The authors show that their method can be viewed as equivalent to a reference method in the case that there are an infinite number of references, and that the distributions of identities and within class variance are Gaussian.

The accuracy of this method while using outside data for training (*unrestricted with labeled outside data*) was 92.42±1.08%.

2.2.7 Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification, 2013 [29]

This paper argues that high-dimensional descriptors are essential for high performance, and also describes a method for compression termed as *rotated sparse regression*. They construct the high-dimensional feature using local binary patterns (LBP), histograms of oriented gradients (HOG) and others, extracted at 27 facial landmarks and at five scales on 2D aligned images. They use principal components analysis (PCA) to first reduce this to 400 dimensions and use a supervised method such as linear discriminant analysis (LDA) or a joint Bayesian model [28] to find a discriminative projection. In a second step, they use L1-regularized regression to learn a sparse projection that directly maps the original high-dimensional feature into the lower-dimensional representation learned in the previous stage.

They report accuracies of 93.18±1.07% under the *unrestricted with label-free outside data* protocol and 95.17±1.13% using their WDRef (99,773 images of 2,995 subjects) data set for training following the *unrestricted with labeled outside data* protocol.

2.2.8 A Practical Transfer Learning Algorithm for Face Verification, 2013 [25]

This paper applies transfer learning to extend the high performing joint Bayesian method [28] for face verification. In addition to the data likelihood of the target domain, they add the KL-divergence between the source and target domains as a regularizer to the objective function. The optimization is done via closed-form updates in an expectation-maximization framework. The source domain is the non-public WDRef data set used in their previous versions [28, 29] and the target is set to be LFW. They use the high-dimensional LBP features from [29], reducing its size from over 10,000 dimensions to 2,000 by PCA.

They report 96.33±1.08% accuracy on LFW in the *unrestricted with labeled outside data* protocol, which improves over the results from using joint Bayesian without the transfer learning on high dimensional LBP features [29].

2.2.9 Hybrid Deep Learning for Face Verification, 2013 [91]

This method [91] uses an elaborate hybrid network of convolutional neural networks (CNNs) and a Classification-RBM (restricted Boltzmann machine), trained directly for verification. A pair of 2D aligned face images are input to the network. At the lower part, there are 12 groups of CNNs, which take in images each covering a particular part of the face, some in colour and some in grayscale. Each group contains five CNNs that are trained using different bootstrap samples of the training data. A single CNN consists of four convolutional layers and a max-pooling layer. Similar to [48], they use local convolutions in the mid- and high-level layers of the CNNs. There can be eight possible "input modes" or combinations of horizontally flipping the input pair of images and each of these pairs are fed separately to the networks. The output from all these networks is in layer L0, having $8*5*12$ neurons. The next two layers average the outputs, first among the eight input modes and then the five networks in a group. The final layer is a classification RBM (models the joint distribution of class labels, binary input vectors and binary hidden units) with two outputs that indicate same or different class for the pairs, which is discriminatively trained by minimizing the negative log probability of the target class given the input, using gradient descent. The CNNs and the RBM are trained separately; then the whole model is jointly fine-tuned using back-propagation. Model averaging is done by training the RBM with five different random sets of training data and averaging the predictions. They create a new training data set, "CelebFaces", consisting of 87,628 images of 5,436 celebrities collected from the web. They report 91.75±0.48% accuracy on the LFW in the *unrestricted with label-free outside data* protocol and 92.52±0.38% following the *unrestricted with labeled outside data* protocol.

2.2.10 POOF: Part-Based One-vs-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation, 2013 [17]

When annotations of parts are provided, this method learns highly discriminative features between two classes based on the appearance at a particular landmark or part that has been provided. They formulate face verification as a fine-grained classification task, for which this descriptor is designed to be well suited.

For training a single "POOF" or Part-Based One-vs-One Feature, it is provided a pair of classes to distinguish and two part locations - one for alignment and the other for feature extraction. All the images of the two classes are aligned with respect to the two locations using similarity transforms with 64 pixels horizontal distance between them. A crop of 64×128 at the mid-point of the two locations is taken and grids of 8×8 and 16×16 are placed on it. Gradient direction histograms and color histograms are used as base features for each cell and concatenated. A linear support vector machine (SVM) is trained on these to separate the two classes. These SVM weights are used to find the most discriminative cell locations and a mask is obtained by thresholding these values. Starting from a given part location as a seed, its connected component is found in the thresholded mask. Base features from cells in this connected component are concatenated and another linear SVM is used to separate the two classes. The score from this SVM is the score of that part-based feature.

They learn a random subset of 10,000 POOFs using the database in [16], getting two 10,000-dimensional vectors for each LFW pair. They use both absolute difference ($|f(A) - f(B)|$) and product ($f(A).f(B)$) of these vectors to train a same-versus-different classifier on the LFW training set. They report 93.13±0.40% accuracy on LFW following the *unrestricted with labeled outside data* protocol.

2.2.11 Learning Discriminant Face Descriptor for Face Recognition, 2014 [58]

This approach learned a "Discriminative Face Descriptor" (DFD) based upon improving the LBP feature (which are essentially differences in value of a particular pixel to its neighbours). They use the Fisher criterion for maximizing between class and minimizing within class scatter matrices to learn discriminative filters to extract features at the pixel level as well as find optimal weights for the contribution of neighbouring pixels in computing the descriptor. K-means clustering is used to find the most dominant clusters among these discriminant descriptors (typically of length 20). They reported best performance using K=1024 or 2048.

They used the LFW-a images and cropped the images to 150 × 130. They further used a spatial grid to encode separate parts of the face separately into their DFD representation and also apply PCA whitening. The descriptors themselves were learned using the FERET data set (*unrestricted with labeled outside data*), however the authors note that the distribution of images in FERET is quite different from that of LFW – performance on LFW is an indicator of the generalizable power of their descriptor. They report an LFW accuracy of 84.02±0.44%.

2.2.12 Face++, 2014

We discuss two papers from the Face++/Megvii Inc. group here, both involving supervised deep learning on large labeled data sets. These, along with Facebook's DeepFace [97] and DeepID[92], exploited massive amounts of labeled outside data to train deep convolutional neural networks (CNNs) and reach very high performance on LFW.

In the first paper from the Face++ group, a new structure, which they term the pyramid CNN [34] is used. It conducts supervised training of a deep neural network one layer at a time, thus greatly reducing computation. A four-level Siamese network trained for verification was used. The network was applied on four face landmarks and the outputs were concatenated. They report an accuracy of 97.3% on the LFW *unrestricted with labeled outside data* protocol.

The Megvii Face Recognition System [113] was trained on a data set of 5 million labeled faces of around 20,000 identities. A ten-layer network was trained for identification on this data set. The second-to-last layer, followed by PCA, was used as the face representation. Face verification was done using the L2 norm score, achieving 99.50±0.36% accuracy. With the massive training data set size, they argue that the advantages of using more sophisticated architectures and methods become less significant. They investigate the long tail effect of web-collected data (lots of persons with very few image samples) and find that after the first 10,000 most frequent individuals, including more persons with very few images into the training set does not help. They also show in a secondary experiment that high performance on LFW does not translate to equally high performance in a real-world security certification setting.

2.2.13 DeepFace: Closing the Gap to Human-Level Performance in Face Verification, 2014 [97]

This paper from Facebook [97] has two main novelties - a method for 3D face frontalization[7] and a deep neural net trained for classification. The neural network

[7] See Section 3 for a discussion of previous work on 3D frontalization.

featured 120 million parameters, and was trained on 4,000 identities having 4 million images (the non-public *SFC* data set). This paper was one of the first papers to achieve very high accuracies on LFW using CNNs. However, as mentioned above, other papers that used deep networks for face recognition predated this by several years [70, 48]. Figure 2 shows the basic architecture of the DeepFace CNN, which is typical of deep architectures used on other non-face benchmarks such as ImageNet.

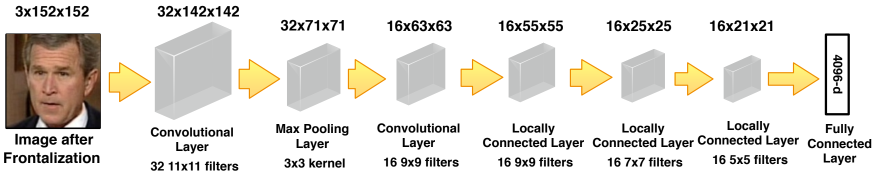

Fig. 2 The architecture of the DeepFace convolutional neural network [97]. This type of architecture, which has been widely used in other object recognition problems, has become a dominant presence in the face recognition literature.

3-D frontalized RGB faces of size 152 × 152 are taken as input, followed by 32 11 × 11 convolution filters (C1), a max-pooling layer (2 × 2 size with stride of two pixels, M2) and another convolutional layer with 16 9 × 9 filters (C3). The next three layers (L4-6) are locally connected layers [48], followed by two fully connected layers (F7-8). The 4096-dimensional F7 layer output is used as the face descriptor. ReLU activation units are used as the non-linearity in the network and *dropout regularization* is applied to F7 layer. L_2-normalization is applied to the descriptor. Training the network for 15 epochs took three days. The weighted χ^2 distance is used as the verification metric. Three different input image types (3D aligned RGB, grayscale with gradient magnitude and orientation and 2-D aligned RGB) are used, and their scores are combined using a kernel support vector machine (SVM). Using the *restricted* protocol, this reaches 97.15% accuracy. Under the *unrestricted* protocol, they train a Siamese network (initially using their own SFC data set, followed by two epochs on LFW pairs), reaching 97.25% after combining the Siamese network with the above ensemble. Finally, adding four randomly-seeded DeepFace networks to the ensemble a final accuracy of 97.35±0.25% is reached on LFW following the *unrestricted with labeled outside data* protocol.

2.2.14 Recover Canonical-View Faces in the Wild with Deep Neural Networks, 2014 [115]

In this paper, the authors train a convolutional neural network to recover the canonical view of a face by training it on 2D images without any use of 3D information. They develop a formulation using symmetry and matrix-rank terms to automatically select the frontal face image for each person at training time. Then the deep network is used to learn the regression from face images in arbitrary view to the canonical (frontal) view.

After this canonical pose recovery is performed, they detect five landmarks from the aligned face and train a separate network for each patch at each landmark along with one network for the entire face. These small networks (two convolutional and two pooling layers) are connected at the fully connected layer and trained on the CelebFaces data set [91] with the cross-entropy loss to predict identity labels. Following this, a PCA reduction is done, and an SVM is used for the verification task, resulting in an accuracy of 96.45±0.25% under the *unrestricted with labeled outside data* protocol.

2.2.15 Deep Learning Face Representation from Predicting 10,000 Classes, 2014 [92]

In this approach, called "DeepID" [92], the authors trained a network to recognize 10,000 face identities from the "CelebFaces" data set [91] (87,628 face images of 5436 celebrities, non-overlapping with LFW identities). The CNNs had four convolutional layers (with 20, 40, 60 and 80 feature maps), followed by max-pooling, a 160-dimensional fully-connected layer (DeepID-layer) and a softmax layer for the identities. The higher convolutional layers had locally shared weights. The fully-connected layer was connected to both the third and fourth convolutional layers in order to see multi-scale features, referred to as a "skipping" layer. Faces were globally aligned based on five landmarks. The input to a network was one out of 60 patches, which were square or rectangular and could be both colour or grayscale. Sixty CNNs were trained on flipped patches, yielding a $160*2*60$ dimensional descriptor of a single face. PCA reduction to 150 dimensions was done before learning the joint Bayesian model, reaching an accuracy of 96.05%. Expanding the data set (CelebFaces+ [88]) and using the joint Bayesian model for verification gives them a final accuracy of 97.45±0.26% under the *unrestricted with labeled outside data* protocol.

2.2.16 Surpassing Human-Level Face Verification Performance on LFW with GaussianFace, 2014 [65]

This method uses multi-task learning and the discriminative Gaussian process latent variable model (DGP-LVM) [55, 100] to be one of the top performers on LFW. The DGP-LVM [100] maps a high-dimensional data representation to a lower-dimensional latent space using a discriminative prior on the latent variables while maximizing the likelihood of the latent variables in the Gaussian process (GP) framework for classification. GPs themselves have been observed to be able to make accurate predictions given small amounts of data [55] and are also robust to situations when the training and test data distributions are not identical. The authors were motivated to use DGP-LVM over the more usual GPs for classifications as the former, by virtue of its discriminative prior, is a more powerful predictor.

The DGP-LVM is reformulated using a kernelized linear discriminant analysis to learn the discriminative prior on latent variables and multiple source domains are used to train for the target domain task of verification on LFW. They detail two uses of their *Gaussian Face* model - as a binary classifier and as a feature extractor. For the feature extraction, they use clustering based on GPs [51] on the joint vectors of two faces. They compute first and second order statistics for input joint feature vectors and their latent representations and concatenate them to form the final feature. These GP-extracted features are used in the GP-classifier in their final model.

Using 200,000 training pairs, the "GaussianFace" model reached $98.52 \pm 0.66\%$ accuracy on LFW under the *unrestricted with labeled outside data* protocol, surpassing the recorded human performance on close-cropped faces (97.53%).

2.2.17 Deep Learning Face Representation by Joint Identification-Verification, 2014 [88]

Building on the previous model, DeepID [92], "DeepID2" [88] used both an *identification signal* (cross-entropy loss) and a *verification signal* (L2 norm verification loss between DeepID2 pairs) in the objective function for training the network, and expanded the CelebFaces data set to "CelebFaces+", which has 202,599 face images of 10,177 celebrities from the web. 400 aligned face crops were taken to train a network for each patch and a greedy selection algorithm was used to select the best 25 of these. A final 4000 (25*160) dimensional face representation was obtained, followed by PCA reduction to 180-dimensions and joint Bayesian verification, achieving 98.97% accuracy.

The network had four convolutional layers and max-pooling layers were used after the first three convolutional layers. The third convolutional layer was locally connected, sharing weights in 2×2 local regions. As mentioned before, the loss function was a combined loss from identification and verification signals. The rationale behind this was to encourage features that can discriminate identity, and also reduce intra-personal variations by using the verification signal. They show that using either of the losses alone to train the network is sub-optimal and the appropriate loss function is a weighted combination of the two.

A total of seven networks are trained using different sets of selected patches for training. The joint Bayesian scores are combined using an SVM, achieving 99.15±0.13% accuracy under the *unrestricted with labeled outside data* protocol.

2.2.18 Deeply Learned Face Representations are Sparse, Selective and Robust, 2014 [93]

Following on with the DeepID "family" of models, "DeepID2+" [93] increased the number of feature maps to 128 in the four convolutional layers, the DeepID size to 512 dimensions and expanded their training set to around 290,000 face images from 12,000 identities by merging the CelebFaces+ [88] and WDRef [29] data sets. Another interesting novelty of this method was the use of a loss function at multiple layers of the network, instead of the standard supervisory signal (loss function) in the top layer. They branched out 512-dimensional fully-connected layers at each of the 4 convolutional layers (after the max-pooling step) and added the loss function (a joint identification-verification loss) after the fully-connected layer for additional supervision at the early layers. They show that removal of the added supervision lowers their performance, as well as some interesting analysis on the sparsity of the neural activations. They report that only about half the neurons get activated for an image, and each neuron activates for about half the images. Moreover they found a difference of less than 1% when using a binary representation by thresholding, which led them to state that the fact that a neuron is activated or not is more important than the actual value of that activation.

This report an accuracy of 99.47±0.12% (*unrestricted with labeled outside data*) using the joint Bayesian model trained on 2000 people in their training set and combining the features from 25 networks trained on the same patches as DeepID2 [88].

2.2.19 DeepID3: Face Recognition with Very Deep Neural Networks, 2015 [89]

"DeepID3" uses a deeper network (10 to 15 feature extraction layers) with Inception layers [94] and stacked convolution layers (successive convolutional layers without any pooling layer in between) on a similar overall pipeline to DeepID2+ [93]. Similar to DeepID2+, they include unshared weights in later convolutional layers, max-pooling in early layers and the addition of joint identification-verification loss functions to branched-out fully connected layers from each pooling layer in the network.

They train two networks, one using the stacked convolution and the other using the recently-proposed Inception layer used in the GoogLeNet architecture, which was a top-performer in the ImageNet challenge in 2015 [94]. The two networks reduce the error rate of DeepID2+ by 0.81% and 0.26%, respectively.

The features from both the networks on 25 patches is combined into a vector of about 30,000 dimensions. It is PCA reduced to 300 dimensions, followed by learning a joint Bayesian model. It achieved 99.53±0.10% verification accuracy on LFW (*unrestricted with labeled outside data*).

2.2.20 FaceNet: A unified embedding for face recognition and clustering, 2015 [82]

This model from Google, called the FaceNet [82], uses 128-dimensional representations from very deep networks, trained on a 260-million image data set using a *triplet loss* at the final layer - the loss separates a positive pair from a negative pair by a margin. An online hard negative exemplar mining strategy within each mini-batch is used in training the network. This loss directly optimizes for the verification task and so a simple L2 distance between the face descriptors is sufficient.

They use two variants of networks. In NN1, they add $1 \times 1 \times d$ convolutional layers between the standard Zeiler&Fergus CNN [112] resulting in 22 layers. In NN2, they use the recently proposed Inception modules from GoogLeNet [94] which is more efficient and has 20 times lesser parameters. The L2-distance threshold for verification is estimated from the LFW training data. They report results, following the *unrestricted with labeled outside data* protocol, on central crops of LFW (98.87±0.15%) and when using a proprietary face detector (99.6±0.09%) using the NN1 model, which is the highest score on LFW in the *unrestricted with labeled outside data* protocol. The scores from using the NN2 model were reported to be statistically in the same range.

2.2.21 Tencent-BestImage, 2015 [8]

This commercial system followed the *unrestricted with labeled outside data* protocol and built their system combining an alignment system, a deep convolutional neural network with 12 convolution layers, and the joint Bayesian method for verification. The whole system was trained on their data set - "BestImage Celebrities Face" (BCF), which contains about 20,000 individuals and 1 million face images and is identity-disjoint with respect to LFW. They divided the BCF data into two subsets for training and validation. The network was trained on the BCF training set with 20 face patches. The features from each patch were concatentated, followed by PCA and the joint Bayesian model learned on BCF validation set. They report an accuracy of 99.65±0.25% on LFW under the *unrestricted with labeled outside data* protocol.

2.3 Label-free outside data protocols

In this section, we discuss two of the LFW protocols together–*image-restricted with label-free outside data* and *unrestricted with label-free outside data*. While these results are curated separately for fairness on the LFW page, conceptually they are highly similar, and are not worth discussing separately.

These protocols allow the use of outside data such as additional faces, landmark annotations, part labels, and pose labels, as long as this additional information does

Method	Net. Loss	Outside data	# models	Aligned	Verif. metric	Layers	Accu.
DeepFace [97]	ident.	4M	4	3D	wt. chi-sq.	8	97.35±0.25
Canon. view CNN [115]	ident.	203K	60	2D	Jt. Bayes	7	96.45±0.25
DeepID [92]	ident.	203K	60	2D	Jt. Bayes	7	97.45±0.26
DeepID2 [88]	ident. + verif.	203K	25	2D	Jt. Bayes	7	99.15±0.13
DeepID2+ [93]	ident. + verif.	290K	25	2D	Jt. Bayes	7	99.47±0.12
DeepID3 [89]	ident. + verif.	290K	25	2D	Jt. Bayes	10-15	99.53±0.10
Face++ [113]	ident.	5M	1	2D	L2	10	99.50±0.36
FaceNet [82]	verif. (triplet)	260M	1	no	L2	22	99.60±0.09
Tencent [8]	-	1M	20	yes	Jt. Bayes	12	99.65±0.25

Table 2 CNN top results: As some of the highest results on LFW have been from using supervised convolutional neural networks (CNNs), we compare the details of the top-performing CNN methods in a separate table. N.B. – unknown parameters that were not mentioned in the corresponding papers are denoted with a "-".

not contain any information that would allow making pairs of images labeled "same" or "different". For example, a set of images of a single person (even if the person were not labeled) or a video of a person would not be allowed under these protocols, since any pair of images from the set or from the video would allow the formation of a "same" pair.

Still, large amounts of information can be used by these methods to understand the general structure of the space of faces, to build supervised alignment methods, to build attribute classifiers, and so on. Thus, these methods would be expected to have a significant advantage over the "no outside data" protocols.

2.3.1 Face recognition using boosted local features, 2003 [50]

One of the earliest methods applied to LFW was developed at Mitsubishi Electric Research Labs (MERL) by Michael Jones and Paul Viola [50]. This work built on the authors' earlier work in boosting for face detection [101], adapting it to learn a similarity measure between face images using a modified AdaBoost algorithm. They use filters that act on a pair of images as features, which are a set of linear functions that are a superset of the "rectangle" filters used in their face detection system. A threshold on the absolute difference of the scalar values returned by a filter applied on a pair of faces can be used to determine valid or invalid variation of a particular property or aspect of a face (the validity being with respect to whether the faces belong to the same identity).

The technical report was released before LFW, and so does not describe application to the database, but the group submitted results on LFW after publication, achieving $70.52 \pm 0.60\%$.

2.3.2 LFW Results Using a Combined Nowak Plus MERL Recognizer, 2008 [46]

This early system combined the method of Nowak et al. [74] with an unpublished method [46] from Mitsubishi Electric Research Laboratory (MERL), and thus technically counts as a method whose full details are not published. However, some details are given in a workshop paper [46].

The MERL system initially detects a face using a Viola-Jones frontal face detector, followed by alignment based on nine facial landmarks (also detected using a Viola-Jones detector). After alignment, some simple lightning normalization is done. The score of the MERL face recognition system [50] is then averaged with the score from the best-performing system of that time (2007), by Nowak and Jurie [74].

The accuracy of this system was 76.18±0.58%.

2.3.3 Is that you? Metric Learning Approaches for Face Identification, 2009 [39]

This paper presents two methods to learn robust distance measures for face verification, the logistic discriminant-based metric learning (LDML) and marginalized K–nearest neighbors (MkNN) classifier. The LDML learns a Mahalanobis distance between two images to make the distances between positive pairs smaller than the distances between negative pairs and obtain a probability that a pair is positive in a standard linear logistic discriminant model. The MkNN classifies an image pair belongs to the same class with the marginal probability that both of them are assigned to the same class using a k–nearest neighbor classifier. In the experiments, they represent the images as stacked multi-scale local descriptors extracted at nine facial landmarks. The facial landmarks detector is trained with outside data. Without using the identify labels for the LFW training data, the LDML achieves $79.27 \pm 0.6\%$ under the *image-restricted with label-free outside data* protocol. They obtain this accuracy by fusing the scores from eight local features including LBP, TPLBP, FPLBP, SIFT and their element-wise square root variants with a linear combination. This multiple feature fusion method is shown to be effective in a number of literatures. Under the *unrestricted with label-free outside data* protocol, they show that the performance of LDML is significantly improved with more training pairs formed using the identity labels. And they obtain their best performance $87.50 \pm 0.4\%$ accuracy by linearly combining the 24 scores with the three methods LDML, large margin nearest neighbor (LMNN) [102] and MkNN over the eight local features.

2.3.4 Multiple One-Shots for Utilizing Class Label Information, 2009 [96]

The authors extend the one-shot similarity (OSS) introduced in [104] which we will describe under the *image-restricted, no outside data* protocol. In brief, the OSS for an image pair is obtained by training a binary classifier with one image in the pair as the positive sample and a set of pre-defined negative samples to classify the other image in the pair. This paper extends the OSS to be multiple one-shots similarity vector by producing OSS scores with different negative sample sets. Each set reflecting either a different subject or a different pose. In their face verification system, the faces are firstly aligned with a commercial face alignment system. The aligned faces are published as the “aligned” LFW data set or LFW-a data set. The face descriptors are then constructed by stacking local descriptors extracted densely over the face images. The information theoretic metric learning (ITML) method is adopted to obtain a Mahalanobis matrix to transform the face descriptors and a linear SVM classifies a pair of faces to be matched or not based on the multiple OSS scores. They achieve their best result $89.50 \pm 0.51\%$ accuracy by combining 16 multiple OSS scores including eight descriptors (SIFT, LBP, TPLBP, FPLBP and their square root variants) under two settings of the multiple OSS scores (the subject-based negative sample sets and pose-based negative sample sets).

2.3.5 Attribute and Simile Classifiers for Face Verification. 2009 [53]

We discussed the attribute and simile classifiers for face verification [53] under the *unrestricted with labeled outside data* protocol. The authors’ result with the attribute classifier qualifies for the *unrestricted with label-free outside data* protocol. They reported their result with the attribute classifier on LFW as $85.25 \pm 0.60\%$ accuracy in their follow-up journal paper [54].

2.3.6 Similarity Scores based on Background Samples, 2010 [105]

In this paper, Wolf et al. [105] extend the one-shot similarity (OSS) introduced in [104] to the two-shot similarity (TSS). The TSS score is obtained by training a classifier to classify the two face images in a pair against a background face set. Although the TSS score by itself is not discriminative for face verification, they show that the performance is improved by combining the TSS scores with OSS scores and other similarity scores. They extend the OSS and TSS framework to use linear discriminant analysis instead of an SVM as the online trained classifier. In addition to OSS and TSS, they propose to represent each image in the pair with its rank vector obtained by retrieving similar images from the background face set. The correlation between the two rank vectors provides another dimensionality of the similarity measure of the face pair. In their experiments, they use the LFW-a data set to handle alignment. Combining the similarities introduced above with eight variants of local descriptors, they obtain an accuracy of $86.83 \pm 0.34\%$ under the *image-restricted with label-free outside data* protocol.

2.3.7 Rectified Linear Units Improve Restricted Boltzmann Machines, 2010 [70]

Restricted Boltzmann machines (RBMs) are often formulated as having binary-valued units for the hidden layer and Gaussian units for the real-valued input layer. Nair and Hinton [70] modify the hidden units to be "noisy rectified linear units" (NReLUs), where the value of a hidden unit is given by the rectified output of the activation and some added noise, i.e. $max(0,x+N(0,V))$, where x is the activation of the hidden unit given an input, and $N(0,V)$ is the Gaussian noise. RBMs with 4000 NReLU units in the hidden layer are first pre-trained generatively, then discriminatively trained as a feed-forward fully-connected network using back-propagation (in the latter case the Gaussian noise term is dropped in the rectification).

In order to model face pairs, they use a "Siamese" network architecture, where the same network is applied to both faces and the cosine distance is the symmetric function that combines the two outputs of the network. They show that that NReLUs are *translation equivariant* and *scale equivariant*(the network outputs change in the same way as the input), and combined with the *scale invariance* of cosine distance the model is analytically invariant to the rescaling of its inputs. It is not translation invariant. LFW images are center-cropped to 144 × 144, aligned based on the eye location and sub-sampled to 32 × 32 3-channel images. Image intensities are normalized to be zero-mean and unit-variance. They report an accuracy of 80.73±1.34% (*image-restricted with label-free outside data*). It should be noted that because the authors use manual correction of alignment errors, this paper does not conform to the LFW protocols, and thus need not be used as a comparison against fully automatic methods.

2.3.8 Face Recognition with Learning-based Descriptor, 2010 [26]

This paper was discussed under the *unrestricted with labeled outside data* protocol. With the holistic face as the only component, the method qualifies for the *image-restricted with label-free outside data* protocol, under which the authors obtain an accuracy of 81.22±0.35%.

2.3.9 Cosine Similarity Metric Learning for Face Verification, 2011 [72]

This paper proposes cosine similarity metric learning (CSML) to learn a transformation matrix to project faces into a subspace in which cosine similarity performs well for verification. They define the objective function to maximize the margin between the cosine similarity scores of positive pairs and cosine similarity scores of negative pairs while regularizing the learned matrix by a predefined transformation matrix. They empirically demonstrate that this straightforward idea works well on LFW and that by combining scores from six different feature descriptors their method achieves an accuracy of $88.00 \pm 0.38\%$ under the *image-restricted with label-free* *outside data* protocol. Subsequent communication with the authors revealed an error in the use of the protocol. Had the protocol been followed properly, our experiments suggest that the results would be about three percent lower, i.e., about 85%. Still, this method has played an important role in subsequent research as a popular choice for the comparison of feature vectors.

2.3.10 Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition, 2011 [31]

This method [31] uses the biologically-inspired V1-like features that are designed to approximate the initial stage of the visual cortex of primates. It is essentially a cascade of linear and non-linear functions. These are stacked into two and three layer architectures, HT-L2 and HT-L3 respectively. These models take in 100×100 and 200×200 grayscale images as inputs. A linear SVM is trained on a variety of vector comparison functions between two face descriptors. Model selection is done on 5,915 HT-L2 and 6,917 HT-L3 models before the best five were selected. Multiple kernels were used to combine data augmentations (rescaled crops of 250×250, 150×150 and 125×75), blend the top five models within each "HT class", and also blend models across HT classes. The HT-L3 gives an accuracy of 87.8% while combining all of the models gives a final accuracy of 88.13±0.58% following the *image-restricted with label-free outside data* protocol.

2.3.11 Face Verification Using the LARK Representation, 2011 [84]

This work extends previous work [83] in which two images are represented as two local feature sets and the matrix cosine similarity (MCS) is used to separate faces from backgrounds. All kinds of visual variations are addressed implicitly in the MCS which is the weighted sum of the cosine similarities of the local features. In this work, the authors present the locally adaptive regression kernel (LARK) local descriptor for face verification. LARK is defined as the self-similarity between a center pixel and its surroundings. In particular, the distance between two pixels is the geodesic distance. They consider an image as a 3D space which includes the 2D coordinates and the gray-scale value at each pixel. The geodesic distance is then the shortest path on the image surface. PCA is then adopted to reduce the dimensionality of the local features. They further apply an element-wise logistic function to generate a binary-like representation to remove the dominance of large relative weights to increase the discriminative power of the local features. They conduct experiments on LFW under both the unsupervised protocol and the image restricted protocol.

In the unsupervised setting, they compute LARKs of size 7×7 from each face image. They evaluate various combinations of different local descriptors and similarity measures and report that the LBP with Chi-square distance achieves the best 69.54% accuracy among the baseline methods. Their method achieves 72.23% accuracy. Under the *image-restricted with label-free outside data* protocol, they adopt the OSS with LDA for face verification and achieve an accuracy of 85.10±0.59% by fusing scores from 14 combinations of local descriptors (SIFT, LBP, TPLBP and pcaLARK) and similarity measures (OSS, OSS with logistic function, MCS and MCS with logistic function).

2.3.12 Probabilistic Models for Inference About Identity, 2012 [62]

This paper presents a probabilistic face recognition method. Instead of representing each face as a feature vector and measuring the distances between faces in the feature space, they propose to construct a model in which identity is a hidden variable in a generative description of the image data. Other variations in pose, illumination and etc., is described as noise. The face recognition is then framed as a model comparison task.

More concretely, they present a probabilistic latent discriminant analysis (PLDA) model to describe the data generation. In PLDA, the data generation depends on the latent identity variable and an intra-class variation variable. This design helps factorize the identity subspace and within-individual subspace. The model is learned by expectation-maximization (EM) and the face verification is conducted by looking at the likelihood ratio of an image pair generated by a similar pair model over a dissimilar pair model. The PLDA model is further extended to be a mixture of PLDA models to describe the potential non-linearity of the face manifold. Extensive experiments are conducted to evaluate the PLDA and its variants in face analysis. Their face verification result on LFW is $90.07 \pm 0.51\%$ under the *unrestricted with label-free outside data* protocol.

2.3.13 Large Scale Strongly Supervised Ensemble Metric Learning, with Applications to Face Verification and Retrieval, 2012 [43]

High-dimensional overcomplete representations of data are usually informative but can be computationally expensive. This paper proposes a two-step metric learning method to enforce sparsity and to avoid features with little discriminability and improve computational efficiency. The two-step design is motivated by the fact that straightforwardly applying the group lasso with row-wise and column-wise L_1 regularization is very expensive in high-dimensional feature spaces. In the first step, they iteratively select μ groups of features. In each iteration, the feature group which gives the largest partial derivative of the loss function is chosen and the Mahalanobis matrix of a weak metric for the selected feature group is learned and assembled into a sparse block diagonal matrix $A_\dagger$. With an eigenvalue decomposition, they obtain a transformation matrix to reduce the feature dimensionality. After that, in the second step, another Mahalanobis matrix is learned to exploit the correlations between the selected feature groups in the lower-dimensional subspace. They adopt the projected gradient descent method to iteratively learn the Mahalanobis matrix.

In their experiments, they use the LFW-a data set and center crop the face images to 110×150. By concatenating two types of features (covariance matrix descriptors and soft local binary pattern histograms) after the first step, they achieve $92.58 \pm 1.36\%$ accuracy under the *image-restricted with label-free outside data* protocol.

2.3.14 Distance Metric Learning with Eigenvalue Optimization, 2012 [111]

In this paper, the authors present an eigenvalue optimization framework for learning a Mahalanobis metric. They learn the metric by maximizing the minimal squared distances between dissimilar pairs while maintaining an upper bound for the sum of squared distances between similar pairs. They further show that this is equivalent to an eigenvalue optimization problem. Similarly, the previous metric learning method LMNN can also be formulated as a general eigenvalue decomposition problem.

They further develop an efficient algorithm to solve this optimization problem, which will only involve the computation of the largest eigenvector of a matrix. In the experiments, they show that the proposed method is more efficient than other metric learning methods such as LMNN and ITML. On LFW, they evaluate this method with both the LFW *funneled* data set and the LFW-a data set. They use SIFT features computed at the fiducial points for faces on the *funneled* LFW data set and achieve $81.27 \pm 2.30\%$ accuracy. On the "aligned" LFW data set, they evaluate three types of features including concatenated raw intensity values, LBP and TPLBP. Combining the scores from the four different features with a linear SVM, they achieve $85.65 \pm 0.56\%$ accuracy under the *image-restricted with label-free outside data* protocol.

2.3.15 Learning Hierarchical Representations for Face Verification with Convolutional Deep Belief Networks, 2012 [48]

In this work [48], a local convolutional deep belief network is used to generatively model the distribution of faces. Then, a discriminatively learned metric (ITML) is used for the verification task. The shared weights of convolutional filters (10×10 in size) in the CRBM (convolutional RBM) makes it possible to use high-resolution images as input. Probabilistic max-pooling is used in the CRBM to have local translation invariance and still allow top-down and bottom-up inference in the model.

The authors argue that in images like faces, that exhibit clear spatial structure, the weights of a hidden unit being shared across the locations in the whole image is not desirable. On the other hand, using a layer with fully-connected weights may not be computationally tractable without either subsampling the input image or first applying several pooling layers. In order to exploit this structure, the image is divided into overlapping regions and the weight-sharing in the CRBM is restricted to be local. Contrastive divergence is used to train the local CRBM. Two layers of these CRBMs are stacked to form a deep belief network (DBN). The *local CRBM* is used in the second layer of their network. In addition to using raw pixels, the uniform LBP descriptor is also used as input to the DBN. The two features are combined at the score level by using a linear SVM. The LFW-a face images are used as input, with three croppings at sizes $150 \times 150, 125 \times 75, 100 \times 100$, resized to the same size before input to the DBN. The deep learned features give competitive performance (86.88±0.62%) to hand-crafted features (87.18±0.49%), while combining the two gives the highest of 87.77±0.62% (*image-restricted with label-free outside data*).

2.3.16 Bayesian Face Revisited: A Joint Formulation, 2012 [28]

We discussed this paper under the *unrestricted with labeled outside data* protocol. The authors also present their result under the *unrestricted with label-free outside data* protocol. Combining scores of four descriptors (SIFT, LBP, TPLBP and FPLBP), they achieve an accuracy of $90.90 \pm 1.48\%$ on LFW.

2.3.17 Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification, 2013 [29]

We discussed this paper under the *unrestricted with labeled outside data* protocol. Without using the WDRef data set for training, they report an accuracy of $93.18 \pm 1.07\%$ under the *unrestricted with label-free outside data* protocol.

2.3.18 Fisher Vector Faces in the Wild, 2013 [85]

In this paper, Simonyan et al. [85] adopt the Fisher vector (FV) for face verification. The FV encoding had been shown to be effective for general object recognition. This paper demonstrates that this encoding is also effective for face recognition. To address the potential high computational expense due to the high dimensionality of the Fisher vectors, the authors propose a discriminative dimensionality reduction to project the vectors into a low dimensional subspace with a linear projection.

To encode a face image with FV, it is first processed into a set of densely extracted local features. In this paper, the dense local feature of an image patch is the PCA-SIFT descriptor augmented by the normalized image patch location in the image. They train a Gaussian mixture model (GMM) with diagonal covariance over all the training features. As shown in Figure 3, to encode a face image with FV, the face image is first aligned with respect to the fiducial points. The Fisher vector is then the stacked, average first and second order differences of the image features over each GMM component center. To construct a compact and discriminative face representation, the authors propose to adopt a large-margin dimensionality reduction step after the Fisher vector encoding.

In their experiments, they report their best result as $93.03 \pm 1.05\%$ accuracy on LFW under the *unrestricted with label-free outside data* protocol.

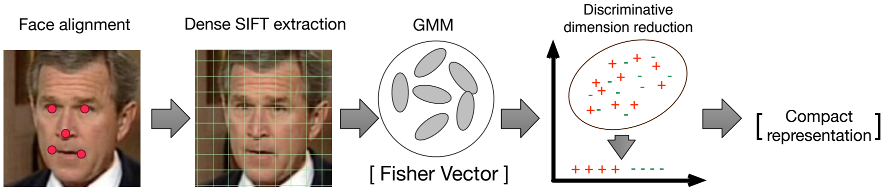

Fig. 3 The Fisher vector face encoding work-flow [85].

2.3.19 Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild, 2013 [32]

In this paper, the authors present a region-based face representation. They divide each face image into spatial blocks and sample image patches from a fixed grid of positions. The patches are then represented by nonnegative sparse codes and sum pooled to construct the representation for the block. PCA whitening is then applied to reduce its dimensionality. After processing each image into a sequence of block representations, the distance between two images are the fusion of pairwise block-to-block distances. They further propose a metric learning method to jointly learn the sequence of Mahalanobis matrices for discriminative block-wise distances. Their best result on LFW is 89.35 ± 0.50% (*image-restricted with label-free outside data*) fusing 8 distances from two different scales of face images and four different spatial partitions of blocks.

2.3.20 Towards Pose Robust Face Recognition, 2013 [108]

This paper presents a pose adaptive framework to handle pose variations in face recognition. Given an image with landmarks, they present a fitting algorithm to fit a 3D shape of the given face. The 3D shape is used to project the pre-defined 3D feature points to the 2D image to reliably locate facial feature points. They then extract descriptors around the feature points with Gabor filtering and concatenate local descriptors to represent the face. In their method, an additional technique to address self-occlusion is to use descriptors from the less-occluded half face for matching. In their experiments, they show that this pose adaptive framework can handle pose variations well in unconstrained face recognition. They obtain 87.77 ± 0.51% accuracy on LFW (*image-restricted with label-free outside data*).

2.3.21 Similarity Metric Learning for Face Recognition, 2013 [23]

This paper presents a framework to learn a similarity metric for unconstrained face recognition. The learned metric is expected to be robust to the large intra-personal variation and discriminative in order to differentiate similar image pairs from dissimilar image pairs. The robustness is introduced by projecting the face representations into the intra-personal subspace, which is spanned by the top eigenvectors of the intra-personal covariance matrix after the whitening process. After mapping the images to the intra-personal subspace, the discrimination is incorporated in learning the similarity metric. The similarity metric is defined as the difference of the image pair similarity against the distance measure, parameterized by two matrices respectively. The matrices are learned by minimizing the hinge loss and regularizing the two matrices to identity matrices. In their experiments, they use LBP and TBLBP descriptors on the LFW-a data set and SIFT descriptors on the LFW *funneled* data set computed at nine facial key points. Under the *image-restricted with label-free outside data* protocol, combining six scores from the three descriptors and their square roots variants they achieve $89.73 \pm 0.38\%$ accuracy. Under the *unrestricted with label-free outside data* protocol, they generate more training pairs with the identity labels and improve the accuracy to $90.75 \pm 0.64\%$.

2.3.22 Fast High Dimensional Vector Multiplication Face Recognition, 2013 [12]

In this method, the authors propose the over-complete LBP (OCLBP) descriptor, which is the concatenation of LBP descriptors extracted with different block and radius sizes. The OCLBP based face descriptor is then processed by Whiten-PCA and LDA. They further introduce a non-linear dimensionality reduction technique Diffusion Maps (DM) with the proposed framework. Extensive experiments are conducted with different local features and dimensionality reduction methods combinations. They report $91.10 \pm 0.59\%$ accuracy under the *image-restricted with label-free outside data* protocol and $92.05 \pm 0.45\%$ under the *unrestricted with label-free outside data* protocol.

2.3.23 Discriminative deep metric learning for face verification in the wild, 2014 [41]

In this method, referred to as DDML, a verification loss between pairs of faces is directly incorporated into a deep neural network, resulting in a non-linear distance metric that can be trained end-to-end using the back-propagation algorithm. The rationale for the verification loss is that the squared Euclidean distance between positive pairs is smaller than that between negative pairs, formulated as a large margin metric learning problem. The network is initialized randomly with three layers and $tanh$ as the nonlinear activation function. They use 80×150 crops of the *LFW-a* (aligned) data set and extract Dense SIFT (45 SIFT descriptors from 16×16 non-overlapping patches, resulting in a 5760-dimensional vector), LBP features (10×10 non-overlapping blocks to get a 7080-dimensional vector) and Sparse SIFT (SIFT computed on 9 fixed landmarks at 3 scales on the *funneled* LFW images, resulting in a 3456-dimensional vector). These features are projected down to 500 dimensions using PCA Whitening. Multiple features are fused at the score level by averaging. Their final accuracy is 90.68±1.41% (*image-restricted with label-free outside data*).

2.3.24 Large Margin Multi-Metric Learning for Face and Kinship Verification in the Wild, 2014 [42]

In this paper, a large margin multi-metric learning (LM^{3}L) method is proposed to exploit discriminative information from multiple features of the same face image. Extracting multiple features from the face images, the distance between two face images is the weighted sum of Mahalanobis distances in each image feature. LM^{3}L jointly learn the distance metrics in different features and the weights of the features by optimizing each distance metric to be discriminative while minimizing the difference of distances in different features of the same image pair. They evaluate the method on the LFW-a data set with SIFT, LBP and Sparse SIFT features. With all the features, they achieve an accuracy of 89.57 ± 1.53% under the *image-restricted with label-free outside data* protocol.

2.3.25 Effective Face Frontalization in Unconstrained Images, 2014 [40]

To show the importance of 3D frontalization to the task of face verification, in this paper Hassner et al. [40] evaluate their *alignment technique* using an earlier face recognition method [104], so that the impact of 3D alignment is not subsumed by the representation power of a more powerful model like the deep network.

Prior work in 3D frontalization of faces [97] would try to reconstruct the 3D surface of a face and then use this 3D model to general views, usually of a canonical pose. This paper explores the alternative of using a single 3D reference surface, without trying to modify the 3D head model to fit every query face's appearance. Although the exact head shape of a query face would be containing discriminative information regarding identity, the final 3D shape fitted to the query face would be an approximation largely dependent upon the accuracy of facial landmark localization. Solving the simpler problem by using an unmodified 3D shape model is shown to give qualitatively equivalent frontalization results, and performance improvement over 2D keypoint alignment methods is demonstrated on face verification and gender estimation tasks.

The frontalized faces of LFW, termed "LFW3D", provided a 3% boost over the LFW-a aligned images. By combining multiple feature descriptors and models by stacking linear SVM scores, they reach an accuracy of 91.65±1.04% on the *image-restricted with label-free outside data* protocol of LFW.

2.3.26 Multi-scale Multi-descriptor Local Binary Features and Exponential Discriminant Analysis for Robust Face Authentication, 2014 [75]

In this paper, the authors represent an face image as the concatenation of region based descriptors which are stacked histograms of local descriptors over multiple scales. They further utilize the exponential discriminant analysis (EDA) to address the small-sample-size problem in LDA to learn a discriminative subspace for the face image feature. And they adopt the within class covariance normalization to project the feature after EDA into a subspace, in which the directions contribute to large intra-class distances have lower weights. They obtain their best result $93.03 \pm 0.82\%$ on LFW (*image-restricted with label-free outside data*) by fusing scores from three different local features.

2.4 No outside data protocols

The most restrictive LFW protocols are the "no outside data" protocols, including *image restricted with no outside data* and *unrestricted with no outside data*. We present these results together as well.

2.4.1 Face Recognition Using Eigenfaces, 1991 [99]

Turk et al. [99] introduce the eigenpicture method by Sirovich [86] to face recognition. The eigenfaces approach they developed is a very important face recognition method in early years. The eigenfaces are the eigenvectors spanning the PCA subspace of a set of training faces. To recognize the unseen face, it is projected to a low-dimensional subspace with the eigenfaces and compared to the average face of each person.

As an early work on face recognition, it is mainly for recognizing frontal faces. Because it assumes faces are well aligned, the PCA subspace keeps mostly variations related to the identity which spans a good "face space". And after projecting faces into the "face space", the representations are all low-dimensional weight vectors. As a result, they can build a near-real-time face recognition system with this eigenface approach for both face detection and recognition. This is an impressive progress considering the limited computational power in early years.

The eigenface approach is designed for well-aligned frontal faces. For the real-world faces in LFW, it achieves $60.02 \pm 0.79\%$ verification accuracy in the *image-restricted with no outside data* protocol.

2.4.2 Learning visual similarity measures for comparing never seen objects, 2007 [74]

Nowak et al. [74] present a method to recognize general objects. Without having the class labels in training stage, they present a method to learn to differentiate if two images are for the same object from training image pairs with only “same” and “different” labels. This is a typical setting for the *image-restricted with no outside data* protocol on LFW.

In the proposed method, they first extract corresponded image patches from the image pair. Then the differences between the corresponded image patches are quantized with an ensemble of randomized binary trees to obtain a vectorized representation for the image pair. A binary linear SVM is applied to the vectorized representation to predict whether the image pair is the “same” or “different”.

This method achieves $72.45 \pm 0.40\%$ accuracy on the original LFW data set and $73.93 \pm 0.49\%$ with the *funneled* LFW data set.

2.4.3 Unsupervised joint alignment of complex images, 2007 [45]

This is the method that generated the *funneled* LFW data set. In this paper, Huang et al. [45] present a method to align images unsupervisedly in the *image-restricted with no outside data* setting of LFW. It is observed that the face recognition accuracy is improved when the recognition method is applied after an alignment stage. The method extends the congealing-style [56] method to handle real-world images. Compared with other domain specific alignment algorithms, congealing does not require manual labeling of specific parts of the object in the training stage. In congealing, a distribution field is defined as the sequence of feature values at a pixel location across a sequence of images. The congealing process is to iteratively minimize the entropy of the distribution field by applying affine transformations to the images. In this work, they use soft quantized SIFT features in congealing.

It shows that with the images aligned by this proposed method, the verification accuracy is improved. For example, the method by Nowak et al. [74] achieves $72.45 \pm 0.40\%$ accuracy on the original LFW data set but is improved to $73.93 \pm 0.49\%$ after aligning images with the proposed method.

2.4.4 Descriptor Based Methods in the Wild, 2008 [104]

In this paper, Wolf et al. [104] evaluate the descriptor-based methods on LFW with LBP descriptor, Gabor filter and two variants of LBP descriptor named Three-Patch LBP (TPLBP) and Four-Patch LBP (FPLBP). The TPLBP and FPLBP are produced by comparing the values of three or four patches to produce a bit value in the code assigned to each pixel. For each descriptor, they use both the euclidean distance and hellinger distance to evaluate the similarity of a face pair. Then they train a linear

SVM to fuse the 8 kinds of prediction scores and achieve an improved performance after fusion.

Besides the evaluation of these descriptor-based methods, they also adopt the one-shot learning for face verification. In this method, a binary classifier is learned online using one face in the given face pair as positive example with a set of negative examples. The binary classifier then evaluates the other face image to obtain the one-shot similarity (OSS). The same process is applied for each face in the pair to obtain an average similarity score of the face pair. They evaluate this method on LFW with the four descriptors and their element-wise square root variants. Combining the 8 scores also improve the accuracy.

Their best result on LFW is $78.47 \pm 0.51\%$ by fusing all 16 scores with a linear SVM, under the *image-restricted with no outside data* protocol.

2.4.5 Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference, 2009 [81]

Sanderson et al. [81] present a region-based face representation. They divide each face into several fixed regions. 2D DCT (Discrete Cosine Transform) features are extract densely from each region. Then a soft quantized histogram is constructed for each region with a Gaussian mixture model as the visual dictionary. The distance between two faces are defined as the average L_1 distances of the corresponded region histograms.

They also propose a two-step method in constructing the soft histogram for acceleration. The Gaussian components are clustered into K clusters. The K Gaussian components nearest to the cluster center are evaluated first in the histogram construction to obtain K likelihoods Then the Gaussian components are evaluated cluster by cluster in the descending order with respect to the likelihoods until the total number of evaluated Gaussian components exceeds the threshold.

The above distance between two faces is normalized by dividing the average pairwise distance of the two faces and a set of cohort faces. They observe this distance normalization method is effective that it brings additional 2.57% average accuracy. This work achieves $72.95 \pm 0.55\%$ accuracy on LFW (*image-restricted with no outside data*).

2.4.6 How far can you get with a modern face recognition test set using only simple features?, 2009 [78]

Pinto et al. [78] present that it is possible to achieve a good recognition performance on LFW by combining several low-level simple features. They extract 48 variants of V1-like features by varying parameters such as the size of Gabor filters and spatial frequencies. To combine the effectiveness of different features, they adopt the multiple kernel learning (MKL) to jointly learn a weighted linear combination of the 48 kernels and the parameters of the kernel SVM for classification. Their best result on LFW is $79.35 \pm 0.55\%$ following the *image-restricted with no outside data* protocol.

2.4.7 Probabilistic Elastic Matching for Pose Variant Face Verification, 2013 [60]

Li et al. [60] present an elastic matching method to handle the pose variations in face verification, reporting results under the *image-restricted with no outside data* protocol of LFW. Without relying on a sophisticated face alignment system, they resort to identify the corresponded regions to compare with in matching two faces. As long as the selected corresponded regions are from a semantically consistent face part, the matching could be invariant to pose variations. In their method, a set of face part models as a Gaussian mixture model (GMM) is learned over all training features. The feature is densely extracted local descriptor augmented by the spatial locations of the image patch in the image. Incorporating the spatial information at the feature-level make each Gaussian component of the GMM capture the joint spatial-appearance distribution of certain face structure. With this GMM, a face can be represented as a sequence of features each of which induces the highest probability on a Gaussian component of the GMM.

In the experiments, they center crop the face images to 150 × 150 and densely extract local descriptors. Given an image, they concatenate the selected sequence of features to be its face representation. An image pair is then represented as the element-wise difference of the two face representations. A SVM is trained from matched and mismatched face pairs for face verification. In their following-up work, they name the GMM the Probabilistic Elastic Part (PEP) model and the face representation is named PEP-representation. The work-flow is illustrated in Figure 4. The best result reported in the paper is $84.08 \pm 1.20\%$ on the *funneled* LFW fusing the prediction scores obtained with the SIFT and LBP features with a linear SVM.

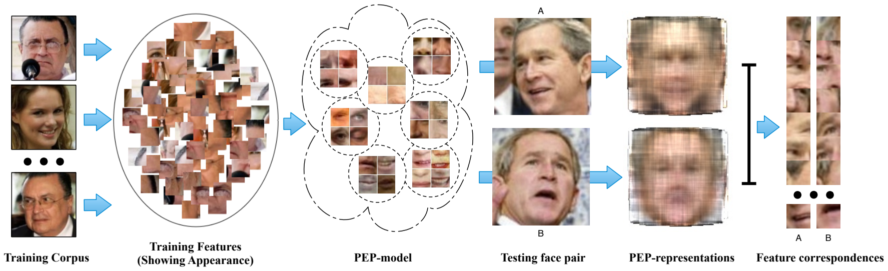

Fig. 4 The training and testing work-flow of the probabilistic elastic matching [60].

2.4.8 Efficient Processing of MRFs for Unconstrained-Pose Face Recognition, 2013 [9]

Arashloo et al. [9] present a method to handle the pose variations via dense pixel matching across face images with MRFs. They propose to reduce the processing time of inference in MRF-based image matching by parallelizing the computation on GPU cores. The major contribution of this paper is how it parallelizes the computation on GPU cores. After adopting the dual decomposition for the MRF optimization, the original problem is decomposed into a set of subproblems. To efficiently solve the subproblems, they further propose several techniques such as incremental subgradient updates and multi-resolution analysis. After obtaining the image matching, multi-scale LBP descriptors are extracted from matched image regions. They stack the descriptors, apply PCA and use the cosine similarity score for face verification. Their best result on LFW is $79.08 \pm 0.14\%$ under the *image-restricted with no outside data* protocol.

2.4.9 Fisher Vector Faces in the Wild, 2013 [85]

We discussed this paper under the *unrestricted with label-free outside data* protocol. They also report their result under the restricted protocol, in which they obtain $87.47 \pm 1.49\%$ accuracy on LFW.

2.4.10 Eigen-PEP for Video Face Recognition, 2014 [61]

Li et al. [61] develop the Eigen-PEP method upon their early work [60]. With the probabilistic elastic matching, a face image or a face track can be represented as a set of face parts. Since the faces are implicitly aligned in a part-based representation, the similar idea from the eigenfaces [99] is adopted here to build a low-dimensional face representation. They use the joint Bayesian classifier [28] for verification. They construct a two-frame face track for each image by adding the mirrored face and achieve $88.97 \pm 1.32\%$ accuracy on the *funneled* LFW in this paper (*image-restricted with no outside data*).

2.4.11 Class-Specific Kernel Fusion of Multiple Descriptors for Face Verification Using Multiscale Binarised Statistical Image Features, 2014 [79]

In this paper, Arashloo et al. [79] address the pose variations via dense pixel matching with their prior work [9]. They then extract three kinds of descriptors, the multi-scale binarized statistical image feature, the multi-scale LBP and the multi-scale local phase quantization feature from the matched image regions. The image representations are embedded into a discriminative subspace with a class-specific kernel discriminant analysis approach. Their best result on the *funneled* LFW data set is $95.89 \pm 1.94\%$ achieved by combining the results of the three image representations (*image-restricted with no outside data*).

2.4.12 Hierarchical-PEP Model for Real-world Face Recognition, 2015 [59]

In this paper, Li et al. [59] present a Hierarchical-PEP model to hierarchically apply the probabilistic elastic part (PEP) model combined with a PCANet [27] to achieve an improved face verification accuracy. They point out that the parts selected after the elastic matching could still present significant visual appearance variations due to the pose variations of the faces. Applying the PEP model to the parts could further introduce pose-invariance in the part representations. After that, the dimensionality of the part representation is discriminatively reduced by a net of PCA and Linear Discriminant Embedding (LDE). They achieve $91.10 \pm 1.47\%$ accuracy in this paper on the *funneled* LFW data set under the *image-restricted with no outside data* protocol.

2.5 Other methods

Here we include those methods for which details are too brief to merit a separate section. These are usually proprietary methods from commercial systems where in-depth detail is not available.

2.5.1 Colour & Imaging Technology (TCIT), 2014 [4]

TCIT calculates the average position of the facial area and judges the identical person or other person by face recognition using the facial area. Face Feature Positioning is applied to get the face data template which is used to verify different faces. They report an accuracy of 93.33±1.24% *(unrestricted with labeled outside data).*

2.5.2 betaface.com, 2014 [1]

They have used original LFW images, converted to grayscale, auto-aligned with their alignment system and followed unrestricted protocol with labeled outside data. LFW data was not used for training or fine-tuning. Their reported accuracy is 98.08±0.16% *(unrestricted with labeled outside data).*

2.5.3 insky.so, 2015 [2]

They used original LFW images to run the test procedure, without doing any training on the LFW images. They report 95.51±0.13% accuracy *(unrestricted with labeled outside data)*.

2.5.4 Uni-Ubi, 2015 [5]

They used original LFW images, converted to grayscale, auto-aligned with their face detector and alignment system. LFW was not used for training or fine-tuning. They report 99.00±0.32% accuracy *(unrestricted with labeled outside data)*.

2.5.5 VisionLabs ver. 1.0, 2013 [6]

The method makes use of metric learning and dense local image descriptors. External data is only used implicitly for face alignment. They report 92.90±0.31% accuracy for the unrestricted training setup *(unrestricted with label-free outside data)*, using LFW-a aligned images.

2.5.6 Aurora Computer Services Ltd: Aurora-c-2014-1, 2014 [7]

The face recognition technology is comprised of Aurora's proprietary algorithms, machine learning and computer vision techniques. They report results using the *unrestricted with label-free outside data* training protocol , achieving 93.24±0.44%. The aligned and funneled sets and some external data were used solely for alignment purposes.

3 Pose and Alignment

One of the most significant issues in face verification is how to address variations in pose. For instance, consider the restricted case in which both faces are guaranteed to be from the same pose, but the pose may vary. The most informative features for comparison will likely change if presented with two profile faces versus two frontal faces. An ideal verification system would presumably account for these differences.

Even more vexing than the above case of how to select features conditioned on pose, however, is the more general problem of how to compare two images that exhibit significantly different poses. Many of the errors seen in the top systems show that these situations are among the most difficult to address (see Figure 5 and Figure 6). Because pose is a cross-cutting issue that virtually every verification system must address in some fashion, we treat it as a separate topic in this section.

LFW was designed to fit into what we call the Detection-Alignment-Recognition pipeline. In particular, by including in LFW only images from the OpenCV Viola-Jones face detector, the designers facilitated the building of end-to-end face recognition systems. Given a recognizer that works well on LFW, the practitioner can pair this with the Viola-Jones face detector to produce an end-to-end system with more predictable performance.

A consequence of the decision to use only faces detected by this specific detector, however, is that most LFW faces are within 20 degrees of frontal, and just a small percentage show some greater degree of yaw angle. This makes addressing pose in LFW a bit easier than it might be for databases with even greater pose variation, such as the recent IJB-A database [52]. Still, the techniques used on LFW to address pose encompass a wide range of strategies and can be expected to be incorporated into systems designed for new and more difficult benchmarks.

There are many approaches to addressing pose in verification problems. These include

1. aligning the input images, either by transforming both to a canonical pose, or by transforming one of them to the other;
2. building mappings that allow inference of what one view looks like given another view;
3. conditioning on pose, such as building separate classifiers for each category of pose pairs;
4. having no explicit mechanism for addressing pose, but rather providing a learning algorithm, and enough training data, so that a model can learn to compare images across pose.

In this section, we review some of the mechanisms that authors have used to address pose variation in LFW, and their relative successes and drawbacks. Tables 3 and 4 enumerate all of the alignment methods used in the papers reviewed in this survey. They are grouped by strategy of alignment (alignment type). The papers using a specific method are given in the rightmost column of the tables.

3.1 Alignment, transformation, and part localization

Probably the most common way of addressing pose changes is to attempt to transform images to a canonical pose or position as a pre-processing step. Because LFW images are the results of detections of the Viola-Jones face detector [101], they are already roughly centered and scaled. However, it seems intuitive that improving the consistency of the head position in preprocessing should improve verification performance. Huang et al. [45] were the first authors to show that alignment improves verification performance on LFW, for at least two different alignment methods.

Landmark-based methods. One common way to align face images is to find landmarks, such as the corners of the eyes and the mouth. Once the landmarks have been detected, one can either transform the image such that the landmarks are placed

Alignment type	**Common name**	**Brief Description**	**Method reference**	**Method usage in LFW**
None	-	No pre-processing. Use of raw LFW images.	-	Eigenfaces, 1991 [99], Nowak, 2007 [74], Multi-Region histograms, 2009 [81], Learning-based descriptor, 2010 [26] (no global alignment, but parts aligned), Associate-predict, 2011 [110], FaceNet, 2015 [82]
Manual alignment	-	Used Machine Perception Toolbox from MPLab, UCSD to detect eye location, manually corrected eye coordinates for worst 2000 detections, used coordinates to rotate and scale images.	Nair et al., 2010 [70]	NReLU, 2010 [70]
Congeal1	Funneling or SIFT-congealing	1) GMM on SIFT features. 2) Jointly align images to minimize entropy of SIFT cluster IDs.	Huang et al., 2007 [45]	Nowak, 2007 [74], Funneling, 2007 [45], Hybrid descriptor-based, 2008 [104], HT Brain-inspired, 2009 [78], LDML-MkNN, 2009 [39], PEP, 2013 [60], Eigen-PEP, 2014 [61], POP-PEP, 2015 [59].
Congeal2	Deep funneling or deep congealing	1) Boltzmann machine model of unaligned face images. 2) Adjust image to maximize its likelihood under model.	Huang et al., 2012 [44]	-
MRF based	MRF-based alignment	MRF for matching is done using Daisy features and multi-scale LBP histograms. It starts by using images from LFW-a or LFW-funneled (Congeal1).	Arashloo et al., 2013 [9].	MRF-MLBP, 2013 [9], MRF-Fusion-CSKDA, 2014 [79].
Landmark1	Buffy	1) Build classifiers for each of K different landmarks. 2) Similarity transform of detected landmarks.	Everingham et al., 2006 [33].	SFRD + Multiple-metric, 2013 [32], Fisher Vector Faces, 2013 [85].
Landmark2	MERL Alignment	1) Nine landmarks located using Viola-Jones detector. 2) Similarity transform puts landmarks in standard position.	Nowak-MERL, 2008 [46].	Nowak-MERL, 2008 [46].
Landmark3	LFW-a	This is a slightly modified version of "Buffy" used by face.com	Wolf et al., 2011 [106]	Wolf et al., 2011 [106] (Journal verson of [96] and [105]), Cosine, 2011 [72], Large scale feature search, 2011 [31], LARK, 2011 [84], DML-eigen, 2012 [111], Conv-DBN, 2012 [48], CMD (Ensemble metric), 2012 [43], LBP PLDA, 2012 [62], MRF-MLBP, 2013 [9], Pose-robust, 2013 [108], Similarity metric, 2013 [23], DFD, 2014 [58], VisionLabs [6], DDML, 2014 [41], Face and Kinship, 2014 [42], Multi-scale LBP, 2014 [75].
Landmark4	Component-based discriminative search	1) Detect possible modes or positions of face components. 2) "Direction classifiers" used to find best alignment direction between image patch and face component.	Liang et al., 2008 [63]	Learning-based, 2010 [26], Joint-Bayesian, 2012 [28].
Landmark5	Explicit shape regression (5 landmark rectification)	Coarse to fine regression. Note: the primary benefit of this method is not really to "align" the image but rather to find landmarks which are used as conditional feature locations.	Cao et al., 2014 [24]	"Blessing" of dimensionality, 2013 [29], TL-Joint Bayesian, 2013 [25].

Table 3 Alignment techniques: Part 1. This table and the one on the following page summarize the various alignment techniques used in conjunction with LFW.

Alignment type	**Common name**	**Brief Description**	**Method reference**	**Method usage in LFW**
Landmark6	Associate-Predict face alignment	Four landmarks are detected using a standard facial point detector and used to determine twelve facial components.	Yin et al., 2011 [110].	Associate-Predict Model, 2011 [110].
Landmark7	Consensus of exemplars	Performs the alignment based not on the part locations in the image itself, but on “generic” parts - where the parts would be for an average person (using 120 reference faces) with the same pose and expression as the test image. Avoids over-alignment which could distort identity information.	Belhumeur et al., 2013 [14].	Tom-vs-Pete, 2012 [16], POOF, 2013 [17].
Landmark8	CNN for facial landmark	3-level cascaded CNN, with the input as the face region from a face detector, and each level regressing to the 5 output key-points.	Sun et al., 2013 [90].	Hybrid CNN-RBM, 2013 [91], DeepID, 2014 [92].
Landmark9	SDM (Intraface)	At training, SDM learns the sequence of descent directions that minimizes the mean of sampled NLS (non-linear least squares) functions. At test time, these learned directions are used instead of the Jacobean or Hessian, which makes it much faster computationally.	Xiong et al., 2013 [107].	DeepID2, 2014 [88], DeepID2+, 2014 [93], DeepID3, 2015 [89]
Landmark10	TCIT	Commercial system for face alignment	TCIT, 2014 [4]	TCIT, 2014 [4]
Landmark11	betaface	”	betaface.com, 2014 [1]	betaface.com, 2014 [1]
Landmark12	Uni-Ubi	”	Uni-Ubi, 2015 [5]	Uni-Ubi, 2015 [5]
Landmark13	Tencent-BestImage	”	Tencent-BestImage, 2015 [8]	Tencent-BestImage, 2015 [8]
Landmark14	OKAO Vision	”	OMRON, 2009 [3]	OMRON, 2009 [3]
3D-1	3D pose normalization.	Uses a 3D head model for normalizing pose. Very similar to the DeepFace approach.	3D pose normalization, 2011 [11].	-
3D-2	DeepFace	1) Landmark-based 2d alignment (6 landmarks) 2) Dense landmark identification (67 landmarks) 3) Iterative projection from 3-D mask to estimate pose. 4) Reproject landmarks and image from frontal pose.	Taigman et al., 2014 [97]	Billion Faces, 2011 [95], DeepFace, 2014 [97], Effective face frontalization, 2014 [40].
3D-3	Fast 3D Model Fitting	1) Detects landmarks using a three-view Active Shape Model. 2) Solves for pose and shape by matching 34 landmarks to 3D vertex index on a deformable face model	Yi et al., 2013 [108].	Towards Pose Robust FR, 2013 [108].
3D-4	3D Morphable Model	1) Detects landmarks using SDM (Landmark11) on image. 2) Does pose-adaptive filtering of the 3DMM to handle non-correspondences between 2D and 3D landmarks	Zhu et al., 2015 [114].	High-Fidelity Pose Normalization, 2015 [114].
view	Recover Canonical-view face	Recovers the canonical view of a face using a deep neural network (commercial system).	Zhu et al, 2014. [115].	CNN view-recovery, 2014 [115]

Table 4 Alignment techniques: Part 2. This table and the one on the previous page summarize the various alignment techniques used in conjunction with LFW.

into a standard position, or simply sample patches or features at the landmark locations. This approach has been taken by many authors. These methods are shown in Tables 3 and 4 under the alignment type of *Landmark* [33, 46, 106, 63, 26, 110, 14, 90, 107, 4, 1, 5, 8, 3].

In particular, the LFW-a alignment [106] was widely used by many verification systems. These images were produced by aligning seven fiducial points to fixed locations with a similarity transform. Subsequent methods explored improving the accuracy of the landmark detectors. For instance, Sun et al. [90] performed detection using a deep convolutional network cascade, which allowed for using larger context and implicit geometric constraints, leading to better performance in difficult conditions due to factors such as occlusion and extreme pose angles. Other methods have explored fitting a larger number of landmarks (generally more than 50) to face images, using techniques such as boosted regression in Cao et al. [24], or through approximate second order optimization in Xiong et al. [107].

As one moves from similarity transforms to more complex classes of transformations for producing alignments, a natural question is whether discriminative verification information may be lost in the alignment process. For instance, if an individual's face has a narrow nose, and landmarks are placed at the extremes of the width of the nose, then positioning these landmarks into a canonical position will remove this information.

Berg and Belhumeur [16] addressed this issue in the context of their piecewise affine alignment using 95 landmarks. In order to preserve identity information, they warped the image not based on the detected landmarks themselves, but rather by the inferred landmarks of a generic face in the same pose and expression as the test image to be aligned. This is accomplished by using a reference data set containing 120 individuals. For each individual, the image whose landmark positions most closely match the test image is found, and these landmark positions are then averaged across all the subjects to yield the generic face landmarks. By switching from a global affine alignment to a piecewise alignment, they increase the accuracy of their system from 90.47% to 91.20%, and by additionally using their identity-preserving generic warp, they achieve a further increase in accuracy to 93.10%.

Note that since all of these landmark-based methods rely on the training of landmark detectors, they require additional labeling beyond that provided by LFW, and hence require any verification methods which use them to abandon the category of *no outside data*. The unsupervised methods, discussed next, do not have this property.

Two-dimensional unsupervised joint alignment methods (congealing). In contrast to methods that rely on trained part localizers, other methods are unsupervised and attempt to align methods using image similarity. One group of such methods is known as *congealing* [68, 45, 44]. In congealing, a set of images are *jointly aligned* by transforming each image to maximize a measure of similarity to the other images. This can be viewed as maximizing the likelihood of each image with respect to all of the others, or alternatively, as minimizing the entropy of the full image set. Once a set of images has been aligned, it can be used to produce a "machine" that aligns new image samples efficiently. This new machine is called a *funnel*. Thus, images aligned with congealing are referred to as *funneled* images. Since congealing can be done using only the training set images for a particular test set, it relies on no additional annotations, and is compatible with the *no outside data* protocols.

The LFW web site provides a two additional versions of the original LFW images that have been aligned using congealing. The first is referred to as *funneled*. In this version of the database, each image was processed with the congealing method of Huang et al. [45]. This method was shown to improve classification rates over some of the early landmark-based alignment methods, but was not as effective as some of the later landmark methods, such as the one used in LFW-a.

An improved version of congealing was developed [44], and was used to produce another version of LFW, known as the *deep-funneled* version. This method used a feature representation learned from a multi-layer Boltzmann machine to align images under the congealing framework. This unsupervised method appears to be comparable to most of the landmark-based methods with respect to the final classification accuracy, and has the advantage of being unsupervised.

One other notable unsupervised method was presented by Arashloo et al. [79, 9] in two separate papers. They start from the *funneled* LFW images and use a Markov random field to further warp the images so that they are more similar.

Frontalization and other methods using 3D information. Another idea to handle differences in views is to attempt to transform views to a canonical frontal pose, sometimes known as *frontalization*. This is clearly beyond the abilities of methods which only perform affine or landmark-based alignment, since the process of transforming a profile face view to a frontal view requires an implicit understanding of the geometry of the head, occluded areas, and the way other features, such as hair, appear from different perspectives.

Early work along these lines was done at Mitsubishi [11, 10], although this did not result in state-of-the-art results on LFW. More recently, Taigman et al. [97] developed a frontalization method that contributed a modest improvement to accuracy on LFW, although most of their gains are attributable to their CNN architecture and the large training sets. Finally, two other methods are essentially landmark-based, but used 3D models to fit 3D landmark coordinates to 2D images [108, 114].

3.2 Conditioning on pose explicitly

Rather than transforming images so that they are all approximately frontal, another approach to dealing with pose variability is to apply strategies separately to different types of image pairs. For example, if one classifies each input image as left-facing (A), frontal (B), or right-facing (C), then we can define nine types of input pairs: AA, AB, AC, BA, BB, BC, CA, CB, CC. One approach is to train separate classifiers for each group of these images, focusing on the peculiarities of each group. By reflecting right-facing images (C), to be left-facing (B), we can reduce the total number of pair categories to just four: AA, AB, BA, BB, although doing this may eliminate information about asymmetric faces. As mentioned in Section 2.2.2, this approach was used by Cao et al. [26].

The associate-predict model proposed by Yin et al. [110] uses the above strategy to separate pairs of test images into those pairs that have similar pose (AA, BB, and CC), which the authors refer to as *comparable* images, and those that do not have similar pose. For the comparable images, the authors run a straightforward computation of an image distance. For the non-comparable images, the authors "associate" features of a face with the features of a set of reference faces, and "predict" the appearance of the feature from a new viewpoint by using the feature appearance from the closest matching reference person, in the desired view.

3.3 Learning our way out of the pose problem

As discussed in Section 2, almost all of the current dominant methods use some CNN architecture and massive training sets. One of the original motivations for using convolutional neural networks was to introduce a certain amount of invariance to the position of the inputs. In addition, max-pooling operators, which take the maximum feature response over a neighborhood of filter responses, also introduce some invariance to position.

However, the invariance introduced by CNNs and max-pooling can also eliminate important positional information in many cases, and it may be difficult to analyze whether the subtle geometrical information required to discriminate among faces is preserved through these types of operations. While many deep learning approaches have shown excellent robustness to small misalignments, all that we are aware of continue to show modest improvements by starting with aligned images. Even the highest performing system (FaceNet [82]) improves from 98.87% without explicit alignment to 99.63% by using a trained alignment system. This seems to suggest that a system dedicated to alignment may relieve a significant burder on the discriminative system. Of course, given enough training data, such an advantage may dissolve, but at this point it still seems worthwhile to produce alignments as a separate step in the process.

4 The Future of Face Recognition

As this article is being written, the highest reported accuracy on LFW described by a peer-reviewed publication stands at $99.63 \pm 0.09\%$, by Schroff et al. [82]. This method reported only 22 errors on the entire test set of 6000 image pairs. These errors are shown in Figure 5 and Figure 6. Furthermore, five of these 22 errors correspond to labeling errors in LFW, meaning that only 17 pairs represent real errors. Accounting for the five ground-truth errors in LFW, the highest accuracy should not go above $\frac{5995}{6000} \approx 99.9\%$, so the results for the protocol *unrestricted with labeled* *outside data* are very close to the maximum achievable by a perfect classifier.[8] With accuracy rates this high, it is time to ask the question "What next?" High accuracy on verification protocols does not necessarily imply high accuracy on other common face recognition protocols such as identification. In addition, some real-world applications of face recognition involve imaging that is significantly more challenging than LFW. Next, we explore some aspects of face recognition that still need to be addressed.

4.1 Verification versus identification

As discussed in Section 1, the LFW protocols are defined for the face verification problem. Even for such realistic images, the problem of verification, for some image pairs, can often be quite easy. It is not uncommon that two random individuals have large differences in appearance. In addition, given two images of the same person taken randomly from some distribution of "same" pairs, it is quite common that such images are highly similar. Thus, verification is, by its nature a problem in which many examples are easy.

For identification, on the other hand, the difficulty of identifying a person is directly related to the number of people in the gallery. With a small gallery, identification can be relatively easy. On the other hand, with a gallery of thousands or millions of people, identifying a probe image can be extremely difficult. The reason for this is simple and intuitive–the more people in a gallery, the greater the chance that there are two individuals that are highly similar in appearance.

It is for this reason that many standard biometric benchmarks use evaluation criteria that are independent of the gallery size, using a combination of the *true accept rate* (TAR) and *false accept rate* (FAR) for open set recognition. The true accept rate is defined to be the percentage of probes which, when compared to the matching gallery identity, are identified as matches. The false accept rate is the percentage of incorrect identities to which a probe is matched. Because it is defined as a percentage, it is independent of the gallery size. It is common to fix the FAR and report the TAR at this fixed FAR, as in "a TAR of 85% at a FAR of 0.1%".

To understand the relationship between accuracies on verification and identification, it is instructive to consider how a high-accuracy verification system might perform in a realistic identification scenario. In particular, consider a verification system that operates at 99.0% accuracy. On average, for 100 matched pairs, and 100 mismatched pairs, we would expect it to make only two errors. Now consider such

[8] For a classifier to get more than $5,995$ of the $6,000$ test examples correct according to the benchmark, it must actually report the wrong answer on at least one of the five incorrectly labeled examples in LFW. Of course it is always possible that a classifier could get extremely lucky and "miss" just the right five examples that correspond to labeling errors in the database while getting all of the other examples, corresponding to correctly labeled test data, correct. However, a method that has a very low error rate overall, and at the same time "accidentally" reports the correct answers for the labeled errors, is likely to be fitting to the test data in some manner.

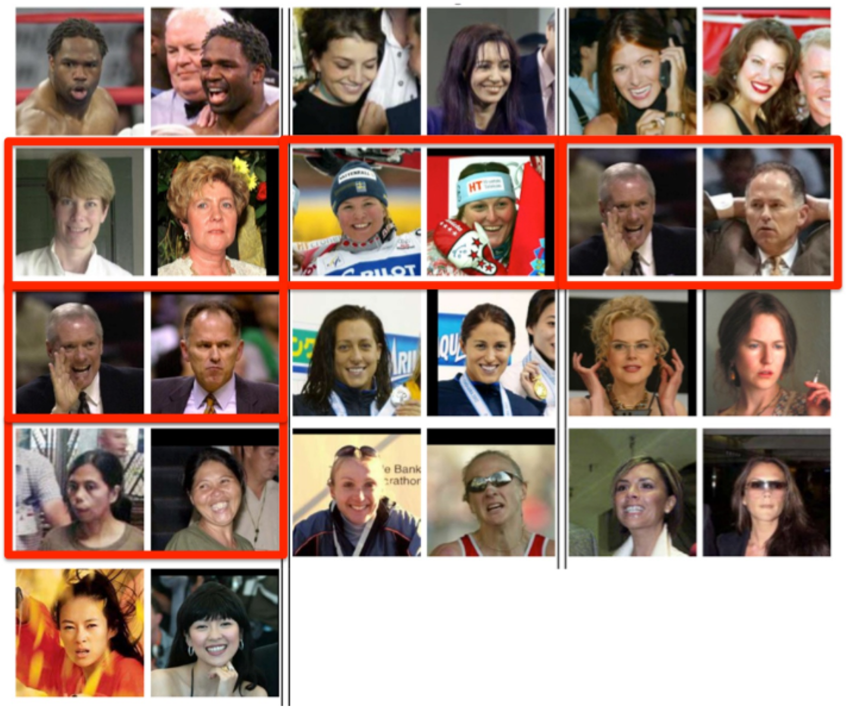

Fig. 5 All of the errors produced by the FaceNet verification algorithm [82] on matched pairs. The pairs surrounded in red are labeling errors in LFW. Thus, while these were flagged as errors, the FaceNet system actually gave the correct answer (correctly identifying these pairs as mismatches). The remainder of the pairs (without red boxes) were identified incorrectly as mismatches. They are in fact matches. It is interesting to note that the rightmost pair in the third row shows the actress Nicole Kidman, but in the rightmost image of this pair, she is wearing an artificial nose, in order to appear more like Virginia Woolf in the film *The Hours* [103]. Thus, this case represents an extreme variation of an individual that would not normally be encountered in everyday life, and it is not clear that one should train a system until this example is evaluated correctly.

a system used in a closed set identification scenario with 901 gallery subjects. For example, this might represent a security system in a large office building.

In a typical case of identification under these parameters, in addition to matching the correct subject, we would expect 1% of the 900 mismatched gallery identities to be rated as a "match" with the probe image by the verification classifier. That is, we would expect to have one correct identity and nine incorrect identities to be above the match threshold of our verification system. The job of the identification system would then be to sort these in the correct order by selecting the one true match as the "best match" from among the 10 that were above threshold. This is quite difficult since by definition the 10 selected images look like the probe identity. If we are successful at selecting the correct match from this set of 10 similar identities 50%

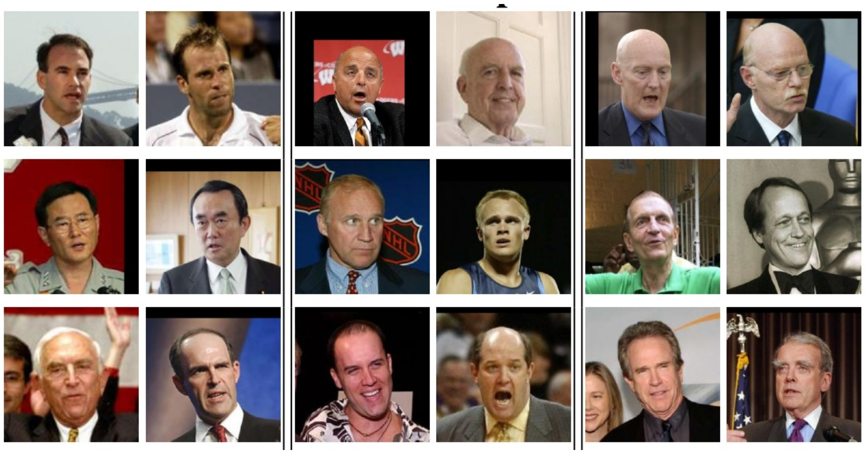

Fig. 6 All of the errors produced by the FaceNet verification algorithm [82] on mismatched pairs. These are the only pairs of mismatched images that were incorrectly reported as matched pairs by [82].

of the time, which is already quite impressive, then the total identification rate is merely 50%.

In larger galleries, the problem of course becomes even more difficult. In a pool of 9901 gallery subjects, achieving 50% identification with a 99% accurate identification system would require finding the correct identity from among 100 examples that looked similar to the probe. This informal analysis demonstrates why identification can be so much harder than verification. In addition, these examples describe closed set identification. Open set identification is even more difficult, as one must try to determine whether the probe is in the gallery at all.

4.2 New databases and benchmarks

In order to study the identification problem with a gallery and probe images, one needs a data set established for this purpose. Some authors have developed protocols from the images in LFW for this purpose, e.g. [20], sometimes by augmenting LFW images with other image sets [64]. Other authors have augmented the images in LFW to study image retrieval with large numbers of distractors [22]. However, the time is ripe for new databases and benchmarks designed specifically for new problems, especially identification problems. Several new databases aim to address these needs.

In this section, we discuss several new face recognition databases and benchmarks, and the new issues they allow researchers to address. We only include discussions of publicly available databases. These include the CASIA database of faces, the FaceScrub database, IJB-A database and benchmark from IARPA, and the MegaFace database from the University of Washington.

4.2.1 IJB-A database and benchmark

The recently announced IJB-A database [52] is designed to study the problems of open set identification, verification, and face detection. It includes both images and videos of each subject and establishes standard protocols.

The database includes images and videos from 500 subjects in unconstrained environments, and all media have creative commons licensing. In order to get a wider range of poses and other conditions than LFW, the images were identified and localized by hand, rather than using an automatic face detector (as with LFW) which is likely to be biased towards easier-to-detect faces.

One interesting element of the protocols provided with this database is that a distinction is made (for identification protocols) about whether a classifier was trained on gallery images or not. Another interesting aspect of this database is that probes are presented as media collections rather than single images. Thus, a probe may consist of a combination of individual images and video. Thus, this encourages exploration of how to best use multiple probe images at the same time to increase accuracy.

4.2.2 The FaceScrub and CASIA data sets

This section describes two distinct databases known as FaceScrub and CASIA. The FaceScrub data set [71] contains $107,818$ images of celebrities automatically collected from the web, and verified using a semi-automated process. It contains 530 different individuals, with an average of approximately 200 images per person. As such it is an important example of a *deep* data set, rather than a *broad* data set, meaning that it has a large number of images per individual. The data set is released under a creative commons license, and the URLs, rather than the images themselves are released.

The CASIA-WebFace data set, or simply CASIA, consists of $494,414$ images, and is similar in spirit to the FaceScrub data set. It is described here [109].

The automatic processing of the images in these databases has two important implications:

- First, because images that are outliers are automatically rejected, there is a limit to the degree of variability seen in the images. For example, heavily occluded images may be marked as outliers, even if they contain the appropriate subject.
- Second, it is difficult to know the percentage of correct labels in the database. While the authors could presumably estimate this fairly easily, they have not reported these numbers in either FaceScrub or CASIA.

Despite these drawbacks, these large and deep databases are two that are currently available to researchers to train large face recognition systems with large numbers of parameters, and because of that, they are valuable resources.

4.2.3 MegaFace

Another new database designed to study large scale face recognition is MegaFace [67], a database of one million face images derived from the Yahoo 100 Million Flickr creative commons data set [98]. This database, which contains one image each of one million *different* individuals, is designed to be used with other databases to allow the addition of large numbers of distractors.

In particular, the authors describe protocols that are used in conjunction with FaceScrub, described in the previous section. All of the images in MegaFace are first registered in a gallery, with one image each. Then, for each individual in FaceScrub, a single image of that person is also registered in the gallery, and the remaining images are used as test examples in an identification paradigm. That is, the goal is to identify the single matching individual from among the $1,000,001$ identities in the gallery.

The paradigms discussed in this work are important in addressing the ability to identify individuals in very large galleries, or in the open set recognition problem. The authors show that several methods that perform well on the standard LFW benchmark quickly deteriorate as distractors are added. A notable exception is the FaceNet system [82], which shows remarkable robustness to distractors.

5 Conclusions

In this article, we have reviewed the progress on the Labeled Faces in the Wild database from the time it was released until the current slew of contributions, which are now coming close to the maximum possible performance on the database. We analyzed the role of alignment and noted that current algorithms can perform almost as well without any alignment after the initial face detection, although most algorithms do get a small benefit from alignment preprocessing. Finally, we examined new emerging databases that promise to take face recognition, including face detection and multimedia paradigms, to the next level.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

Research reported in this publication was also partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work is also partly supported by US National Science Foundation Grant IIS 1350763 and GH's start-up funds from the Stevens Institute of Technology.

References

1. betaface.com. http://betaface.com.
2. inksy.so. http://www.insky.so/.
3. Omron from okao vision. http://www.omron.com/technology/index.html.
4. Taiwan colour & imaging technology (tcit). http://www.tcit-us.com/.
5. Uni-ubi. http://uni-ubi.com.
6. Visionlabs ver. 1.0. http://www.visionlabs.ru/face-recognition.
7. Aurora computer services ltd: Aurora-c-2014-1. http://www.facerec.com/, 2014.
8. Tencent-bestimage. http://bestimage.qq..com, 2014.
9. S. R. Arashloo and J. 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.
10. A. Asthana, M. J. Jones, T. K. Marks, K. H. Tieu, and R. Goecke. Pose normalization via learned 2D warping for fully automatic face recognition. In *BMVC*. Citeseer, 2011.
11. A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937–944. IEEE, 2011.
12. O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1960–1967. IEEE, 2013.
13. E. Bart and S. Ullman. Class-based feature matching across unrestricted transformations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1618–1631, 2008.
14. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2930–2940, 2013.
15. P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.
16. T. Berg and P. N. Belhumeur. Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In *BMVC*, volume 2, page 7. Citeseer, 2012.
17. T. Berg and P. N. Belhumeur. POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 955–962. IEEE, 2013.
18. T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Whos in the picture? In *NIPS*, 2005.
19. T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848. IEEE, 2004.

20. L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *Information Forensics and Security, IEEE Transactions on*, 9(12):2144–2157, 2014.
21. D. Beymer and T. Poggio. Face recognition from one example view. In *Computer Vision, 1995. Proceedings, Fifth International Conference on*, pages 500–507. IEEE, 1995.
22. B. Bhattarai, G. Sharma, F. Jurie, and P. Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *Computer Vision-ECCV 2014 Workshops*, pages 160–172. Springer, 2014.
23. Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2408–2415. IEEE, 2013.
24. X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.
25. X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun. A practical transfer learning algorithm for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3208–3215. IEEE, 2013.
26. Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2707–2714. IEEE, 2010.
27. T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. PCANet: A simple deep learning baseline for image classification? *ArXiv e-prints*, 2014.
28. D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.
29. D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE, 2013.
30. S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.
31. D. Cox and N. Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 8–15. IEEE, 2011.
32. Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3554–3561. IEEE, 2013.
33. M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy”–Automatic naming of characters in TV video. In *BMVC*, volume 2, page 6, 2006.
34. H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou. Learning deep face representation. *arXiv preprint arXiv:1403.2802*, 2014.
35. L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611, 2006.
36. A. Ferencz, E. G. Learned-Miller, and J. Malik. Building a classification cascade for visual identification from one example. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 286–293. IEEE, 2005.
37. Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems*, pages 473–480, 2007.
38. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.
39. M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 498–505. IEEE, 2009.
40. T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. *arXiv preprint arXiv:1411.7964*, 2014.
41. J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1875–1882. IEEE, 2014.

42. J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multimetric learning for face and kinship verification in the wild. In *Proc. ACCV*, 2014.
43. C. Huang, S. Zhu, and K. Yu. Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval. *arXiv preprint arXiv:1212.6094*, 2012.
44. G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, pages 764–772, 2012.
45. G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Computer Vision (ICCV), IEEE International Conference on*, pages 1–8. IEEE, 2007.
46. G. B. Huang, M. J. Jones, and E. Learned-Miller. LFW results using a combined Nowak plus MERL recognizer. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.
47. G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.
48. G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.
49. 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.
50. M. Jones and P. A. Viola. Face recognition using boosted local features. Technical Report TR2003-25, Mitsubishi Electric Research Laboratory, 2003.
51. H.-C. Kim and J. Lee. Clustering based on Gaussian processes. *Neural computation*, 19(11):3088–3107, 2007.
52. B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1931–1939. IEEE, 2015.
53. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.
54. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1962–1977, 2011.
55. N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16(3):329–336, 2004.
56. E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006.
57. E. G. Learned-Miller, A. Ferencz, and J. Malik. Learning hyper-features for visual identification. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, volume 17, page 425. MIT Press, 2005.
58. Z. Lei, M. Pietikainen, and S. Z. Li. Learning discriminant face descriptor. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(2):289–302, 2014.
59. H. Li and G. Hua. Hierarchical-PEP model for real-world face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4055–4064, 2015.
60. H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE, 2013.
61. H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-PEP for video face recognition. *Asian Conference on Computer Vision (ACCV)*, 2014.
62. P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince. Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):144–157, 2012.
63. L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *Computer Vision–ECCV 2008*, pages 72–85. Springer, 2008.

64. S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.
65. C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with Gaussianface. *arXiv preprint arXiv:1404.3840*, 2014.
66. A. Martinez and R. Benavente. The AR face database. Technical Report CVC Technical Report 24, Ohio State University, 1998.
67. D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. MegaFace: A million faces for recognition at scale. *arXiv preprint arXiv:1505.02108*, 2015.
68. E. G. Miller, N. E. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 464–471. IEEE, 2000.
69. B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.
70. V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.
71. H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 343–347. IEEE, 2014.
72. H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Computer Vision–ACCV 2010*, pages 709–720. Springer, 2011.
73. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2-3):103–134, 2000.
74. E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
75. A. Ouamane, B. Messaoud, A. Guessoum, A. Hadid, and M. Cheriet. Multi scale multi descriptor local binary features and exponential discriminant analysis for robust face authentication. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 313–317. IEEE, 2014.
76. S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.
77. P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.
78. N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2591–2598. IEEE, 2009.
79. S. Rahimzadeh Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *IEEE Transactions on Information, Forensics, and Security*, 9(12):2100–2109, 2014.
80. S. A. Rizvi, P. J. Phillips, and H. Moon. The FERET verification testing protocol for face recognition algorithms. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 48–53. IEEE, 1998.
81. C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in Biometrics*, pages 199–208. Springer, 2009.
82. F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 815–823. IEEE, 2015.
83. H. J. Seo and P. Milanfar. Training-free, generic object detection using locally adaptive regression kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.
84. H. J. Seo and P. Milanfar. Face verification using the LARK representation. *Information Forensics and Security, IEEE Transactions on*, 6(4):1275–1286, 2011.
85. K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, volume 60, 2013.

86. L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 1987.
87. C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.
88. Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.
89. Y. Sun, D. Liang, X. Wang, and X. Tang. DeepID3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.
90. Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.
91. Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496. IEEE, 2013.
92. Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1891–1898. IEEE, 2014.
93. Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014.
94. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.
95. Y. Taigman and L. Wolf. Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. *arXiv preprint arXiv:1108.1122*, 2011.
96. Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *BMVC*, pages 1–12, 2009.
97. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014.
98. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.
99. M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.
100. R. Urtasun and T. Darrell. Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, pages 927–934. ACM, 2007.
101. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.
102. K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *In NIPS*, 2006.
103. Wikipedia. The Hours (film) — Wikipedia, The free encyclopedia, 2015. [Online; accessed 30-June-2015].
104. L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.
105. L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *Computer Vision–ACCV 2009*, pages 88–97. Springer, 2010.
106. L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1978–1990, 2011.
107. X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.
108. D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3539–3545. IEEE, 2013.

109. D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
110. Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 497–504. IEEE, 2011.
111. Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13(1):1–26, 2012.
112. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.
113. E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.
114. X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.
115. Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.