IET Computer Vision

Review Article

IET Journals
The Institution of Engineering and Technology

A comparison of deep multilayer networks and Markov random field matching models for face recognition in the wild

ISSN 1751-9632
Received on 18th June 2015
Revised 25th December 2015
Accepted on 8th February 2016
E-First on 6th April 2016
doi: 10.1049/iet-cvi.2015.0222
www.ietdl.org

Shervin Rahimzadeh Arashloo[1] ✉

[1]*Department of Medical Informatics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran*
✉ *E-mail: Sh.Rahimzadeh@urmia.ac.ir*

Abstract: Robustness to a diverse range of image transformations and distortions has been an everlasting goal of visual pattern recognition. While there have been a huge number of efforts to advance the state-of-the art in this direction over the last decades, two prominent outstanding schemes, among others, are deep multilayer architectures and graphical models, providing some degree of robustness to undesired image perturbations. In this study, the authors aim at shedding some light on the underlying concepts, mechanisms, strengths and potentials of each methodology while discussing their relative merits from a practical point of view. In particular, they discuss the underlying motivations for the construction of deep multilayer architectures and undirected graphical models, also known as Markov random fields. The principles in the construction of each architecture, how invariance properties are achieved in each approach, the efficiency of each approach in terms of computations required during train and test as well as the degree of human labour required in each approach are discussed. Finally, an experimental comparison of the performances of the two frameworks is performed on a challenging problem of face recognition in unconstrained settings in the presence of a wide range of undesirable visual perturbations.

1 Introduction

In many practical applications, it is highly desirable to have the output of a classifier or other estimator unchanged when the input pattern undergoes various transformations. A notable example is that of object recognition from two dimensional images where the objects may be subject to rotation, scale, translation and so on. Ideally, during classification it is beneficial to first spatially transform the objects to be as similar to one another as possible and only then compute their similarity using a suitable measure [1]. However, a naive search over all possible sets of parameters for several transformations for each prototype during classification makes the computational burden of this procedure prohibitive. A plausible solution to the problem is offered by the image matching paradigm using graphical models, a.k.a Markov random fields (MRFs) [2, 3]. In this framework, images are first divided into their constituting parts and then a correspondence is sought between the two images at a lower level of representation. For classification, the established correspondences can be used in a variety of different ways. Some applications may consider geometrically normalising the images prior to classification. Other approach is to normalise the criterion function value of the match and use it directly for classification [4, 5]. Yet another alternative is to employ the correspondences in a standard feature extraction and matching scheme [6].

In a different methodology, geometric invariance can be achieved using deep multilayer architectures, best exemplified by multilayer neural networks [7]. There are different mechanisms used in deep multilayer networks to achieve robustness against geometric distortions. A common approach is to employ as training data images which are already undergone spatial transformations. Alternatively, one may use features which are less affected with unwanted image degradations as inputs to such networks. An alternative mechanism is regularising the objective function of optimisation in multilayer architectures so that certain spatial deformations would not change the target labels of the network [8]. A more common approach, however, is to build-in the invariance properties into the network structure. While higher order networks [9] in theory are capable to provide robustness against various geometric variations, the notion of convolutional architectures is much more common in practice [10–12]. The convolutional multilayer architectures are composed of multiple layers of filtering followed by pooling. The pooling operation in these networks serves to summarise the outputs of filter responses over small local subregions, which, when combined with replication of convolutional units' parameters over large fields makes them robust against small translations and distortions. Recognition in these networks can be performed jointly with the representation learning process or by using a standard classifier on the representations learned. This framework has been successfully applied to various visual recognition problems [11–15] and has recently exhibited impressive performances due to availability of large volume of training data and scalable computation resources such as thousands of CPU cores or graphical processing units (GPUs).

This paper aims at shedding some light on the underlying concepts and mechanisms of MRF-based image-to-image matching and the multilayer deep architectures, in particular deep convolutional networks. Strengths and potentials of each approach and their relative merits from a practical point of view in distortion invariant object recognition are analysed. The optimisation problems, the efficiency in terms of computations required during train and test and the number of training examples required for generalisation as well as the degree of human labour required in each approach are also discussed. Finally, a comparison on a challenging problem of unconstrained face recognition is made to quantitatively evaluate the two methodologies.

The remaining of this paper is organised as follows: In Section 2, the motivations in applying graphical models to image analysis problems and a description of the labelling problem in the context of deformable image matching method proposed in [4, 6, 16] is provided. The discussion is then followed by a description of spatial transformation invariant object recognition in this context. In Section 3, the multilayer deep architectures are introduced where the underlying principles and motivations are discussed followed by an explanation on how invariance can be achieved in these networks. A successful subcategory of deep nets known as convolutional networks for distortion invariant object recognition is explained next. Section 4 then compares more closely the two methodologies in practical settings for face recognition in wild conditions. Finally, the paper is drawn to conclusions in Section 5.

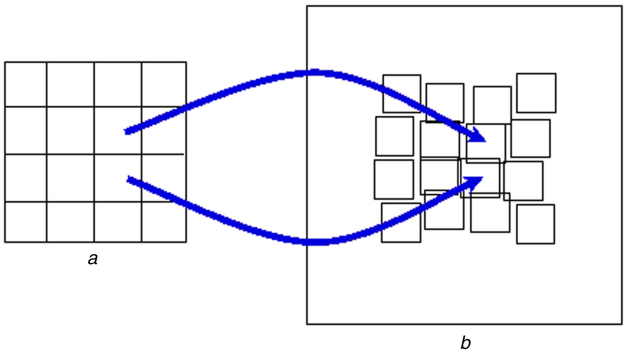

Fig. 1 *Mapping image A to image B on a block-by-block basis*

2 Graphical models in image analysis

Visual information in real images is exposed to noise resulting in various ambiguities. Noise and uncertainty may be introduced into images via various sources such as sensing device noise, illumination or pose changes, occlusion, inherent changes of a deformable object and so on. From an object recognition viewpoint, holistic approaches are more vulnerable to such changes in the appearance of objects. One approach to the problem in these situations is to model an object as a collection of smaller parts and consider image features at a lower level of representation. In many cases, the spatial configuration of image entities is a fundamental ingredient in the construction of an image. In these situations, it is necessary to impose some sort of prior knowledge on the problem so that the hypothesised model would favour some part of the solution space over another. This kind of prior knowledge can be forced on the solution by incorporating conditional dependencies between different entities of an image.

Context-dependent and correlated entities can be conveniently modelled by describing their interactions using graphical distributions and MRFs [2, 3]. In this framework, the state of each variable is dependent on its neighbours and independent of the states of the rest of variables. From a computational perspective, the local properties of MRFs and their sparsity in conditional dependencies are appealing as they support algorithms that can be implemented in efficient ways. In addition, such representations provide a suitable basis for multi-resolution analysis of images. These characteristics have motivated the application of MRFs to vision problems at all levels. In computational vision, this philosophical approach has been advocated in [17]. The MRF formalism facilitates the development of algorithms for a variety of problems in a systematic way rather than relying on *ad-hoc* heuristics. In this framework, solving each problem entails two tasks. First, one needs to decide upon the form of the posterior distribution. Second, the parameters of the distribution have to be determined. Another major ingredient of the approach is the optimisation method for inferring the mode of the posterior distribution (MAP inference).

2.1 Deformable image matching

The problem of deformable matching involves estimating a deformation map between a pair of images subject to optimising a criterion function value. In an MRF formulation, the criterion function is defined as the probability of a configuration of nodes on a graphical model or inversely as the energy of the match corresponding to the negative log probability of the configuration.

In the method in [6, 18], the probability function involves unary and pairwise factors. The unary factors measure the fidelity of the estimated deformation to the data while the prior includes factors of arity two and imposes a smoothness constraint on the deformation. While there exists a variety of matching methods in the literature [19–32], the method proposed in [6, 18] is chosen for this study. The approach presents an MRF formulation to the image-to-image matching problem on a block-by-block basis. The model image is first divided into a number of non-overlapping blocks for each of which a corresponding block in the test image is sought, as shown in Fig. 1.

This choice is driven by its appealing characteristics outlined below. First, it can accommodate a considerable amount of image variations while providing a compact representation of the optimisation task in an efficient fashion. The efficiency draws on the idea of decomposing two-dimensional (2D) displacement vectors into two 1D disparities [16]. Second, the method uses a very successful scheme for optimisation, i.e. dual decomposition [33] which can be readily parallelised. In addition, the method is able to provide dense pixelwise correspondences between a pair of images in a relatively fast multiresolution fashion. Finally, the method has been extensively tested on a variety of different face recognition scenarios and has achieved very promising results. This method is overviewed briefly in the following sections.

2.2 Inference in graphical models

Once, the problem is formulated in an MRF context, the resultant energy function should be optimised to find the least energy labelling of graph nodes. The minimum energy corresponds to the maximum a posteriori (MAP) probability configuration of the graph. The MRF energy is comprised of different terms, representing different aspects of structural and texture variations as

$$E(x;\theta) = \sum_{v \in \mathcal{V}} \theta_v(x_v) + \sum_{c \in \mathcal{C}} \theta_c(x_c) \quad (1)$$

where $\mathcal{V}$ corresponds to the set of nodes of the model graph and $\mathcal{C}$ to the set of cliques (those comprising of more than one node). θ_v and θ_c stand for the unary and clique potentials of the graph capturing the texture and shape information while x denotes a configuration, i.e. a labelling of the graph.

There are exact algorithms for MAP inference in acyclic graphs (trees), such as max-product belief propagation [34]. Exact inference on cyclic (loopy) graphs with low tree-width is also feasible by converting the loopy graph into an equivalent tree and using junction tree algorithms [35]. However, the situation in large loopy graphs is much more complicated. In practice, many of the problems in vision naturally lead to a loopy graph with the number of nodes exceeding tens of thousands. The difficulty in this situation is manifested by slow convergence rates and poor local minima obtained by the early inference methods [36, 37]. However, in the recent years two important developments have changed the view to MRFs. The first one is that of emergence of very promising optimisation algorithms which could even reach global optimum of the target function under certain conditions [33, 38]. The proposed methods were not only applicable to pairwise MRFs, but in some cases were extendible to those graphs with higher order cliques [39, 40]. The second advancement is a technological one and corresponds to the advent of powerful parallel processing hardwares, especially GPUs. The computational advantages offered by GPUs when combined with effective inference methods adaptable to parallel processing have resulted in breakthroughs in using MRFs for different vision problems.

While there are a variety of inference algorithms for MRFs [33, 38, 41, 42] a main group among others is based on decomposing the original problem into smaller sub-problems on each of which exact inference is tractable. This approach, which is based on a Lagrangian dual of the original problem, leads to defining a lower bound on the original objective function which is then maximised. For inference, dual decomposition [33] is chosen in [6] for its perfect adaptability to parallel processing in addition to providing very good solutions in practice. Two layers model displacements in the horizontal and vertical directions in the work in [6]. The method then uses an edge-wise decomposition of the graph, resulting a large of sub-problems, for minimising the energy function. The MAP state of each of the resulting sub-problems is then found in a parallel on a GPU, as shown in Fig. 2.

2.3 Multi-resolution analysis

From a recognition point of view, the denser the correspondences between objects, the better the performance would be. In order to reduce the computational burden while increasing accuracy and

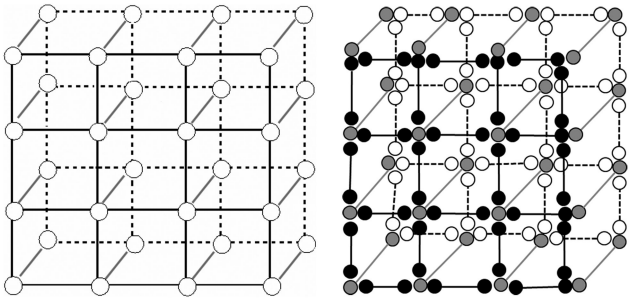

Fig. 2 *Left: Two-layer graph used for representing deformations in the horizontal and vertical directions in [6, 16], right: decomposition of the graph into edge-disjoint sub-problems used in [6]*

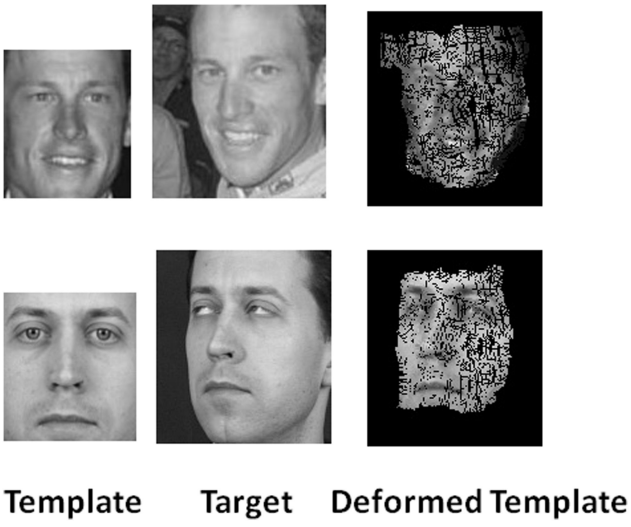

Fig. 3 *Results of image-to-image dense matching. In each row, the template image is deformed based on the displacement obtained after matching*

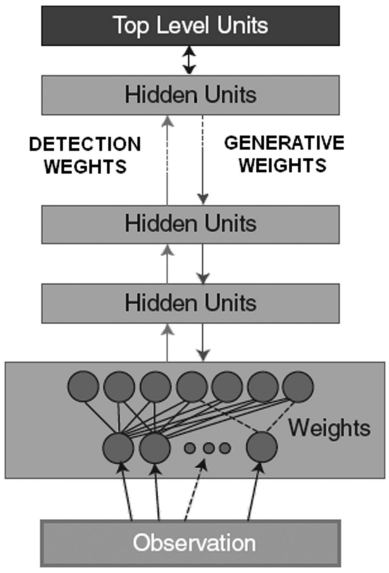

Fig. 4 *Sample multilayer network*

improving robustness to noise, a plausible technique is a multi-resolution approach. In a multi-resolution analysis, larger groups of nodes are used in coarser scales of the hierarchy while groups with a small number of nodes are processed in finer scales. There are two main considerations in applying multi-resolution techniques. The first one is how to coarsen the data and how to transform the posterior distribution in a way that the solution to the problem remains consistent across all scales. The second consideration is how to propagate the labelling obtained at a coarse resolution to the next finer scale so that the optimum at the finer level is reached more efficiently and it is consistent with the coarser scale solution.

The renormalisation group transform (RGT) [43, 44] provides solutions to both problems in a principled way. In [6], a modified form of the RGT approach, a potential-based coarsening technique, referred to as super-coupling transform [44], is used. Using the supercoupling approach, it is shown [45] that the parameters of the energy should be chosen so that the data term of a block in a coarse scale equals the sum of the data terms of its corresponding blocks in a next finer scale. Regarding the pairwise term, the method prescribes a stronger interaction between sites in the coarser scales of the hierarchy. The employed multi-resolution approach, which decreasing the computational cost, facilitates achieving a pixelwise correspondence between a pair of images. Moreover, as a result of multi-resolution analysis, the robustness of the technique against image noise is reduced. Fig. 3 illustrates the results of pixelwise matching of two pairs of images in a multi-resolution fashion using the method of [6].

2.4 Classification

For classification purposes, as denoted earlier, the criterion function value, i.e. the match energy between a pair of images may be used for hypothesis selection. This methodology has been studied in [4, 5] where an energy normalisation scheme is proposed to make the energy of the match a better criterion for classification. In [4], the main shortcomings of the energy function value in an MRF model image matching approach for classification are identified as the effects of a spatial global rigid transformation, correlation of local distortions and last but not least the variations in textural contents of images. These shortcomings are compensated for and finally the classification is performed using the normalised match energy comprised of both shape and textural contents of images.

A different approach is just to take into account the dense correspondences in a standard feature extraction and classification scheme. An example of such an approach is the method of [46]. In [46], first a dense pixelwise alignment between a pair of images to be compared is obtained using an MRF image matching model described in the preceding sections. The matching has been performed in a symmetrised fashion by changing the roles of the model and the test images. Moreover, in order to further minimise the effects of pose variations, in addition to the original images, mirrored versions of both the model and the test images are employed. In the next step, textural contents of a pair of images are modelled using the multiscale local binary patterns (MLBP) [47], multiscale local phase quantisation [48] and the multiscale binarised statistical image features [49]. The features obtained are then fused together in a kernel space constructed by a new class-specific kernel discriminant analysis (CS-KDA) approach for decision making. The CS-KDA approach allowed unsupervised subspace learning for a binary classification problem based on kernel discriminant analysis (KDA) using a single instance of the positive class and a number of random instances for constructing the negative class. Further details about this approach may be found in [46].

3 Deep multi-layer architectures

Deep multi-layer architectures are composed of multiple modules of adaptive parameterised non-linearities stacked atop of each other where the parameters of each layer are subject to learning. In these networks, the data passes through many layers of non-linear functions to be used finally for classification or other predictions, as shown in Fig. 4. Prominent examples of such deep multi-layer architectures are neural networks with multiple hidden layers. Deep networks find applications in a variety of problems including speech and natural language processing [50, 51], object recognition [52, 53], face recognition [13], facial landmark detection [54], facial attributes analysis [55], object segmentation [56], pedestrian detection [57], optical flow estimation [58] and so on.

Deep networks automatically are able to learn from data which is particularly advantageous as the volume of the data and range of applications to machine learning methods is growing continually

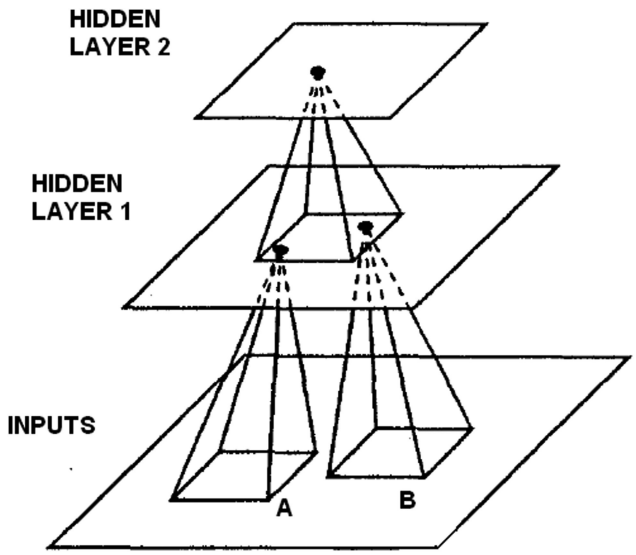

Fig. 5 *Sample architecture of a convolutional network for translation-invariant object recognition from 2D images [9]*

making it very challenging to tailor a bespoke feature and learning scheme to each individual task. In this respect, deep architectures can take advantage from unsupervised learning and data from similar tasks to improve the performance on those problems which ordinarily suffer from an insufficiency of labelled data.

There are different methods to design a deep network of non-linear modules. Such architectures may be constructed as generative models such as sigmoid belief networks [59, 60] or deep belief networks [61] or as discriminative approaches such as convolutional neural networks [10] including Fukushima's Neocognitron [11, 12]. Whether the deep network is generative or discriminative, there are various approaches to achieve robustness against the undesired geometric image perturbations in these networks. The first and probably the simplest approach is to train the network by example. For instance, in an object recognition scenario instances of an object which have undergone the transformations that the system must be robust against are provided as training data to the system. The transformed examples might be real images of an object or virtually generated instances. However, one drawback of such an approach is that generally a very large number of training images are required making the training phase computationally demanding. In addition, in such an approach the network would ideally provide robustness against the observed transformations and would not be generally able to extrapolate to unseen image transformations during test. It can be shown that training a network by distorted samples in effect changes the objective function of the network by adding a Tikhonov regularisation term [62].

A second approach to achieve invariance is to perform a preprocessing stage before any estimation. Such an approach, for example in an optical character recognition problem may be implemented by centring the input image, normalising its size and slant. However, as denoted earlier, this method would usually be computationally very demanding as inferring the correct transformation parameters for an object is itself an open problem. A different solution in this respect is to extract features which are known to be less affected by unwanted image transformations while preserving the discriminative capability of feature content. However, engineered feature construction may be highly application dependent and time consuming at times.

An alternative method to achieve spatial transformation invariance is to modify the error function of the network using the tangent propagation technique [8]. In this method, the error function is regularised using an additional term which penalises any changes in the error function with respect to the parameters of the transformations.

A final solution is to build-in the invariance properties into the structure of the network. One way to achieve this is through the use of higher-order networks where the output of a unit is a function of a polynomial of degree more than two of its inputs [9].

A different methodology which has been successfully applied to various spatial transformation invariant object recognition problems is the notion of convolutional architectures. The property which is exploited in these networks is that of local correlation in images, i.e. nearby pixels are more strongly correlated than pixels farther away. Many computer vision systems use this property and extract local features that depend only on local regions of an image. In order to obtain higher order and more abstract representations of an image, information can be combined in the later stages of the network. Local features which are useful in one part of an image are likely to be useful in other regions as well. These ideas are incorporated into convolutional architectures through three mechanisms known as local receptive fields, weight sharing and subsampling, discussed later.

3.1 Convolutional networks

For moderate size data, a deep fully connected multi-layer network would normally include several tens of thousands of parameters in each layer. By increasing the capacity of the system using such a large number of weights, a large training set is also required. In addition, certain hardware implementations may be ruled out as a result of the memory requirements to store a large number of parameters. Above all, the topology of the input data is completely ignored in a fully connected network. As a result, unstructured architectures do not seem to provide much built-in invariance against local image distortions. In contrast, a particularly successful subclass of multi-layer networks known as convolutional architectures can naturally accommodate moderate distortions. In these architectures, the layers are constructed as 2D arrays to be consistent with the structure of the image. While the first layer is usually fed with raw pixel intensities, each hidden unit in a next layer receives inputs from a predefined small region of the previous layer (local receptive field). The use of local receptive fields is partly motivated by the local correlation properties of an image.

An important attribute of convolutional architectures is that parameters are replicated across space. This property of convolutional networks is inspired by the observation that basic feature detectors which have a beneficial use in one part of an image are probably useful in other regions. The property of local invariance to distortions in these architectures is achieved through a mechanism known as pooling. In a pooling process, using a max, sum or other operations, the outputs of one layer (feature map) over a small region is summarised and sub-sampled next in order to achieve robustness against distortions, as shown in Fig. 5. Combining multiple layers of pooling can provide more and more invariance to distortions. The layer which performs pooling typically has a lower resolution compared to the previous convolutional layer. A convolutional network is very often realised by alternately stacking convolutional and subsampling layers.

In Fukushima's Neocognitron method [11, 12], an early version of the realisation of the convolution and sub-sampling layers which was first introduced in [63] was proposed.

The distortion invariance properties offered by the convolutional networks together with their efficient training procedures due to the severely constrained parameterisation makes them a suitable candidate for object detection and recognition. As a result of the small number of weights used in convolutional networks, the common optimisation problems of deep multi-layer networks do not occur in convolutional networks. However, there are also certain drawbacks in using convolutional architectures. For instance, different parts of an image can have different local properties and as a result the spatial stationary assumption of feature detectors may not always be the best option. The problem might, however, be circumvented by *locally* sharing weights [13, 64]. A further disadvantage of multiple layers of convolution and pooling is that such an approach may cause the information about the precise position of detailed structure and micro-texture be lost. As a result, some works limited the number of pooling layers to just a few [65].

4 Experimental evaluation: face verification in the wild

In this section, a quantitative evaluation of various deep networks and the method of [46] as a representative of graph-based methods on the challenging problem of unconstrained face recognition on the labelled faces in the wild (LFW) database [66] is provided. The LFW dataset [66] is a large database including real world variations of facial images such as pose, illumination, expression, occlusion, low resolution, blur and so on. It contains 13,233 images of 5749 subjects. Evaluation of a method on this dataset is performed by determining whether a pair of images belongs to the same person or not. 'View 2' of the dataset which consists of 3000 matched and 3000 mismatched pairs and is divided into 10 sets is commonly used for performance reporting. The evaluation is performed in a leave-one-out cross validation scheme on the entire test sets. Different evaluation protocols on this database are the image 'restricted', 'unrestricted' and 'unsupervised' settings. The restricted setting provides training data for the image pairs as 'same' or 'not same'. The image unrestricted setting in addition provides the identities of the subjects in each pair. In the 'unsupervised' setting, no training data in the form of same/not same pairs is provided. Each of the restricted and unrestricted settings is further divided into other sub-settings discussed in the following. The overall performance of the method over ten test folds is reported as the mean accuracy and the standard deviation of the mean for the restricted and unrestricted settings while the area under curve (AUC) measure is used for performance reporting in the unsupervised setting.

It should be mentioned that despite the fact that different methods use slightly different versions of the LFW data in terms of image alignment (e.g. the aligned version, funnelled version etc.) the effects of such discrepancies are well dominated by other factors including the source of the data used for training and the degree of human intervention in the training procedure. As a result, different evaluation protocols defined for the LFW database are based on the source of the training data and the degree of supervision incorporated rather than other factors. For example, in terms of alignment, this can be justified by the maturity of image alignment methods achieved in recent years. In this respect, the evaluation protocol of the database includes unsupervised setting, image-restricted setting: no outside training data, image-restricted setting: label-free outside data and unrestricted setting: labelled outside data discussed next. Any testing conditions beyond those covered by the evaluation protocol of the LFW database have negligible effects on system performance as can be evidenced by a simple comparison between the results reported on the LFW results web page (http://www.vis-www.cs.umass.edu/lfw/results.html) which are also reported in the following sections.

Table 1 Comparison of the performances of various methods on the LFW database in the unsupervised setting

Method	AUC
SD-MATCHES, 125 × 125, aligned [67]	0.5407
H-SX-40, 81 × 150, aligned [67]	0.7547
GJD-BC-100, 122 × 225, aligned [67]	0.7392
LARK unsupervised, aligned [68]	0.7830
LHS, aligned [69]	0.8107
pose adaptive filter (PAF) [70]	0.9405
MRF-MLBP [18]	0.8994
MRF-fusion-CS-KDA [46]	0.9894

Table 2 Comparison of the performances of various methods on the LFW database in the image restricted setting, no outside training data

Method	$\mu \pm S_E$
eigenfaces, original [71]	0.6002 ± 0.0079
Nowak, original [72]	0.7245 ± 0.0040
Nowak, funnelled [72]	0.7393 ± 0.0049
hybrid descriptor-based, funnelled [73]	0.7847 ± 0.0051
3 × 3 multi-region Histograms(1024) [74]	0.7295 ± 0.0055
pixels/MKL, funnelled [75]	0.6822 ± 0.0041
V1-like/MKL, funnelled [75]	0.7935 ± 0.0055
APEM (fusion), funnelled [76]	0.8408 ± 0.0120
MRF-MLBP, funnelled [18]	0.7908 ± 0.0014
Fisher vector faces [77]	0.8747 ± 0.0149
MRF-Fusion-CSKDA [46]	0.9589 ± 0.0194

4.1 Unsupervised setting

In the unsupervised setting, no information in the form of class labels or pair labels (same/not same) is available. Generally, deep networks are either trained in an unsupervised fashion followed by a supervised fine tuning stage or directly trained in a supervised mode. As a result, no deep learning method has been evaluated under the unsupervised protocol. On the other hand, the class-specific kernel discriminant approach proposed in [46] can be operated in an unsupervised fashion. The evaluation results in this setting for the method of [46] along with other approaches are provided in Table 1. In this scenario, the image-to-image matching scheme followed by a kernel fusion approach of multiple descriptors achieves the best average performance by a large margin.

4.2 Image restricted setting: no outside training data

In this setting, LFW training data is provided as same/not same without the identity of subjects. Moreover, no outside data is allowed to be used. Despite the fact that the MRF-fusion-CS-KDA approach of [46] uses no data in the form of same/not same, the results obtained by this method are comparable to other approaches using such training data.

As deep networks commonly require a very large training data, the data provided by the LFW protocol in this setting does not seem to be adequate to train such architectures. More specifically, the large capacity of deep nets necessitates very large training data to be available. Consequently, to the best of our knowledge, no deep network has been evaluated in this setting. In contrast, the MRF image-to-image matching method requires no training data to establish correspondences. In the construction of the class-specific KDA model, roughly 1000 random images are utilised to build the mode. Moreover, reducing the size of the training set in this method down to 300 has been found to have negligible effect on the performance. The performance of the MRF-fusion-CS-KDA approach [46] is compared to other approaches under this protocol in Table 2. The image matching approach followed by the class-specific kernel fusion method ranks first in this scenario again by a large margin.

4.3 Image-restricted setting: label-free outside data

In this scenario, LFW training data is only provided in the form of same/different. In addition, outside training data, provided that no information in the form of same/different is given, can be used. Although the MRF-fusion-CS-KDA method does not utilise same/different LFW or other training data, its performance is comparable to other approaches under this setting.

Two deep learning networks have reported results under this setting. The first is the method of [78], where a metric learning approach based on a three-layer neural network is proposed. Different image representations such as LBP and SIFT features are used as inputs to the network and then the network is trained as a large margin classifier to produce small distances between images of the same subject while producing large distances for images of different individuals. The training is performed using back propagation without any unsupervised pre-training. Other work in [64], uses local convolutional restricted Boltzman machines for face verification. While being scalable, the proposed method is robust to small misalignments. The proposed deep architecture was applied both to raw pixel values as well as LBP codes. By taking a convolutional approach, the authors were able to train the model directly on full size images relying less on manual alignment. A

comparison of the performances of different methods under this protocol is provided in Table 3. As it is evident from the table the MRF-fusion-CS-KDA ranks first in this setting while the best performing deep network achieves a lower performance by a margin of more than 5%.

4.4 Unrestricted setting: labelled outside data

This is known to be the most permissive protocol, which allows training on many types of external training data in addition to the LFW training set augmented with the names of individuals. As deep multilayer architectures typically require very large labelled training data, the majority of these methods are evaluated under this protocol. In the following, we briefly overview some of the methods which have been evaluated in this setting. In [100], a hybrid network composed of multiple layers of convolutional and restricted Boltzman machine layers is proposed to extract relational features from a pair of faces. Different mid or high level features are extracted from different face regions using the idea of locally shared weights. The improvement offered by average pooling is experimentally validated on LFW. Other work in [99] used a very deep nine-layer network composed of different types of layers. This work initially employed a 2D fiducial point detection followed by a 3D model to correct for any pose deviations from frontal. Three convolutional layers were then used along with a subsampling layer to make the system more robust to registration errors. As relatively accurate alignment has already been achieved via a combined 2D+3D approach, only a single subsampling layer was employed to retain information about the feature locations. In addition, weight sharing was adopted in only the first layers of the network to provide basic representations for the upper layers which were designed to provide different representations in different regions of face images. In [103], a deep convolutional network is proposed to learn effective representations for face images. Multiple levels of convolution and pooling are employed. In order to learn a diverse set of features, weight sharing is not employed in the last stages of the network. The novelty of the approach is to learn deep representations of face images under the supervision of a large number of classes. The entire network is then trained using back propagation. In a later study [65], the authors proposed to use two supervisory signals of identification and verification concurrently. The identification signal served to enlarge the inter-class variations while the verification signals minimised the intra-class variations. The use of two different supervisory signals improved the accuracy of the system compared to their earlier work. In [102], a convolutional net is proposed to generate virtual frontal views of subjects in the wild from their non-frontal images. The network comprises three convolutional layers, the first two followed by a max-pooling layer. For recognition, a different deep architecture is run on the transformed images (Table 4).

In addition to the deep architectures reviewed earlier, there are other works employing a multilayer deep architecture for face recognition [105–107]. However, as these methods have not been evaluated on the LFW data set, a fair comparison regarding their performance would not be feasible. In this setting, the deep learning method of [65] achieves the best performance, via a commercial system. The best published method in this setting is that of [103] achieving an accuracy of 0.9745 ± 0.0026 compared to the MRF-fusion-CS-KDA with an accuracy of 0.9589 ± 0.0194.

Table 3 Comparison of the performances of various methods on the LFW database in the image restricted setting, label-free outside training data

Method	$\mu \pm S_E$
MERL, [79]	0.6002 ± 0.0079
MERL+Nowak, funnelled [79]	0.7618 ± 0.0058
LDML, funnelled [80]	0.7927 ± 0.0060
hybrid, aligned [81]	0.8398 ± 0.0035
combined b/g s-based, aligned [73]	0.8683 ± 0.0034
NReLU [82]	0.8073 ± 0.0134
single LE + holistic [83]	0.8122 ± 0.0053
LBP + CSML, aligned [84]	0.8557 ± 0.0052
CSML + SVM, aligned [84]	0.8800 ± 0.0037
High-Thr. Brain-Inspired F., aligned [85]	0.8813 ± 0.0058
LARK supervised, aligned [68]	0.8510 ± 0.0059
DML-eig SIFT, funnelled [86]	0.8127 ± 0.0230
DML-eig combined, funnelled and aligned [86]	0.8565 ± 0.0056
convolutional DBN [64]	0.8777 ± 0.0062
SFRD+PMML [87]	0.8935 ± 0.0050
pose adaptive filter (PAF) [70]	0.8777 ± 0.0051
sub-SML [88]	0.8973 ± 0.0038
VMRS [89]	0.9110 ± 0.0059
DDML [78]	0.9068 ± 0.0141
MRF-fusion-CS-KDA [46]	0.9589 ± 0.0194

Table 4 Comparison of the performances of various methods on the LFW database in the unrestricted setting, labelled outside data

Method	$\mu \pm S_E$
attribute classifiers [90]	0.8525 ± 0.0060
simile classifiers [90]	0.8414 ± 0.0041
attribute and simile classifiers [90]	0.8554 ± 0.0035
multiple LE + comp [83]	0.8445 ± 0.0046
associate-predict [91]	0.9057 ± 0.0056
Tom-versus-Pete [92]	0.9310 ± 0.0135
Tom-versus-Pete + Attribute [92]	0.9330 ± 0.0128
combined joint Bayesian [93]	0.9242 ± 0.0108
high-dim LBP [94]	0.9517 ± 0.0113
DFD [95]	0.8402 ± 0.0044
TL joint Bayesian [96]	0.9633 ± 0.0108
face.com r2011b [97]	0.9130 ± 0.0030
Face++ [98]	0.9727 ± 0.0065
DeepFace-ensemble [99]	0.9735 ± 0.0025
ConvNet-RBM [100]	0.9252 ± 0.0038
POOF-gradhist [101]	0.9313 ± 0.0040
POOF-HOG [101]	0.9280 ± 0.0047
FR+FCN [102]	0.9645 ± 0.0025
DeepID [103]	0.9745 ± 0.0026
Gaussian face [104]	0.9852 ± 0.0066
DeepID2 [65]	0.9915 ± 0.0013
MRF-fusion-CS-KDA [46]	0.9589 ± 0.0194

4.5 Discussion

A number of observations from the previous experiments are in order. First, both architectures (deep nets and graphical models represented by the MRF-fusion-CS-KDA method of [46]) can achieve very high performances in the presence of various geometric distortions and misalignments in images. Second, as deep networks generally require a supervised training stage, their applicability is limited to special settings where large labelled data is available. This can be considered as a major drawback in their application to various problems. Third, even in the presence of moderate size labelled data deep networks do not seem to be trained effectively. This stems from the fact that such networks have a large capacity and as a result require very large training data. The lack of any deep networks in the image restricted setting with no outside training data is illustrative of this fact. Fourth, when very large unlabelled training data in addition to a moderate size labelled data is available, deep networks may be applicable. However, even in this case (the image restricted setting, label-free outside training data) their performance is inferior to many other methods including MRF-fusion-CS-KDA approach. The potential of deep multilayer networks is best exploited when very large labelled data are available. This is best represented by the near perfect results obtained in the unrestricted setting when labelled outside data is available in addition to the LFW database. On the

other hand, graph-based methods are very efficient in terms of the number of training data. Such methods do not require large labelled data sets and good performance can be achieved even with a single face sample.

Regarding the degree of hand coding required, in the image-to-image matching approach there is only a single parameter controlling the trade-off between data term and the continuity prior which has to be set manually. However, LBP and LPQ representations employed require a certain degree of human prior knowledge. In contrast, in deep networks the representation learning is performed almost automatically. However, as denoted earlier, these networks require the designer to decide upon many important hyper-parameters that very often in practice there exists not much insight on how to tune them other than trial and error.

As mentioned earlier, training deep architectures is another drawback of these networks, which requires loads of labelled samples in addition to high performance computing resources. The MAP inference in MRF although challenging, has witnessed major improvements in the last few years. However, computationally, it is still demanding which necessitates certain hardware implementations for practical use. On the other hand, deep networks once trained, can produce the output relatively fast.

The idea of pooling in deep network, although instrumental in providing a degree of invariance against geometric distortions, causes the network to lose information about the precise locations of features. In contrast, image-to-image matching methods provide pixelwise precision in feature locations.

Finally, a closer look into the method of [46] reveals that in fact it can be considered as a multilayer system. In the lowest layer, different representations (LBP, LPQ and BSIF) are extracted. Next, the obtained representations are subject to non-linear geometric transformations using the pixelwise correspondences of the MRF model. The geometrically transformed representations are then locally summarised using histograms. Finally, the obtained histograms are projected into the kernel space for final decision making. Thus, the whole system resembles a four-layer network. However, in contrast to deep convolutional architectures where the parameters of all layers are optimised together, in the MRF-fusion-CS-KDA approach, each layer is designed separately. In this case, if the whole system could be fine-tuned in a similar fashion as the deep networks, better performance is expected.

5 Conclusion

MRF image-to-image matching and deep multilayer architectures (in particular convolutional networks) are compared from an object recognition perspective under geometric misalignments. Both architectures possess effective mechanisms to achieve a reasonable degree of geometric invariance in practical settings. However, their potentials are best revealed in different test scenarios. While graph-based matching methods can operate with a low number of unlabelled training data, deep multilayer architectures require very large labelled training sets for training which limits their applicability to certain problems. Although the training phase in deep networks can take several hours or days, the testing phase can be performed relatively fast. In contrast, inference in MRF matching models seems to be relatively slower in the test phase. From the point of view of prior hand-coding required, the MRF-based approach relied on a high degree of human hand-coding while deep architectures in comparison require the designer to decide upon many hyper-parameters which does not seem to be trivially feasible.

6 Acknowledgments

The author gratefully acknowledge the labelled faces in the wild database web page (http://www.vis-www.cs.umass.edu/lfw/results.html) for providing the evaluation results of different methods on the LFW database.

7 References

[1] Duda, R.O., Hart, P.E., Stork, D.G.: '*Pattern classification*' (New York, 2001, 2nd edn.)

[2] Li, S.: '*Markov random field modeling in image analysis*' (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001)

[3] Blake, A., Kohli, P., Rother, C. (Eds.): '*Markov random fields for vision and image processing*' (MIT Press, 2011)

[4] Arashloo, S.R., Kittler, J.: 'Energy normalization for pose-invariant face recognition based on mrf model image matching', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (6), pp. 1274–1280

[5] Arashloo, S.R., Kittler, J.: 'Pose-invariant face matching using mrf energy minimization framework'. Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), 2009 (LNCS, 5681), pp. 56–69

[6] Arashloo, S.R., Kittler, J.: 'Fast pose invariant face recognition using super coupled multiresolution Markov random fields on a gpu', *Pattern Recognit. Lett.*, 2014, **10**, (11), pp. 2396–2407

[7] Bengio, Y.: 'Learning deep architectures for ai', *Found. Trends Mach. Learn.*, 2009, **2**, (1), pp. 1–127

[8] Simard, P., Victorri, B., LeCun, Y.*, et al.*: 'Tangent prop – a formalism for specifying selected invariances in an adaptive network'. Neural Information Processing Systems (NIPS), 1991, pp. 895–903

[9] Bishop, C.M.: '*Neural networks for pattern recognition*' (Oxford University Press, 1995)

[10] Lecun, Y., Bottou, L., Bengio, Y.*, et al.*: 'Gradient-based learning applied to document recognition', *Proc. IEEE*, 1998, **86**, pp. 2278–2324

[11] Fukushima, K., Miyake, S., Ito, T.: 'Neocognitron: a neural network model for a mechanism of visual pattern recognition', *IEEE Trans. Syst. Man Cybern.*, 1983, **SMC-13**, pp. 826–834

[12] Fukushima, K.: 'Neocognitron: a hierarchical neural network capable of visual pattern recognition', *Neural Netw.*, 1988, **1**, pp. 119–130

[13] Sun, Y., Wang, X., Tang, X.: 'Hybrid deep learning for face verification'. IEEE Int. Conf. on Computer Vision (ICCV), 2013, December 2013, pp. 1489–1496

[14] Yaniv, T., Ming, Y., MarcAurelio, R.*, et al.*: 'Deepface: closing the gap to human-level performance in face verification'. Computer Vision and Pattern Recognition (CVPR), 2014

[15] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet Classification with Deep Convolutional Neural Networks', in Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C.*, et al.* (eds.), '*NIPS*'. 2012, pp. 1106–1114

[16] Shekhovtsov, A., Kovtun, I., Hlavac, V.: 'Efficient mrf deformation model for non-rigid image matching', *Comput. Vis. Image Underst.*, 2008, **112**, pp. 91–99

[17] Geman, S., Geman, D.: 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1984, **PAMI-6**, (6), pp. 721–741

[18] Arashloo, S.R., Kittler, J.: 'Efficient processing of mrfs for unconstrained-pose face recognition'. IEEE Sixth Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), 2013, September 2013, pp. 1–8

[19] Konrad, J., Dubois, E.: 'Bayesian estimation of motion vector fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992, **14**, (9), pp. 910–927

[20] Heitz, F., Bouthemy, P.: 'Multimodal estimation of discontinuous optical flow using Markov random fields', *IEEE Trans. Pattern Anal. Mach.Intell.*, 1993, **15**, (12), pp. 1217–1232

[21] Boykov, Y., Veksler, O., Zabih, R.: 'Fast approximate energy minimization via graph cuts'. The Proc. of the Seventh IEEE Int. Conf. on Computer Vision, 1999, 1999, vol. **1**, pp. 377–384

[22] Kumar, M., Torr, P., Zisserman, A.: 'Learning layered motion segmentations of video'. Tenth IEEE Int. Conf. on Computer Vision, 2005 (ICCV 2005), 17–21 October 2005, vol. **1**, pp. 33–40

[23] Jiang, H., Drew, M., Li, Z.: 'Matching by linear programming and successive convexification', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (6), pp. 959–975

[24] Tobias, H., Sascha, M., Hans-Peter, M., Ivo, W.A.: 'Shape-Guided Deformable Model with Evolutionary Algorithm Initialization for 3D Soft Tissue Segmentation', in Karssemeijer, N., Lelieveldt, B. (Eds.): Information Processing in Medical Imaging: 20th International Conference, IPMI 2007, Kerkrade, The Netherlands, July 2–6, 2007. (Springer Berlin Heidelberg, 2007), pp. 1–12

[25] Glocker, B., Komodakis, N., Tziritas, G.*, et al.*: 'Dense image registration through mrfs and efficient linear programming', *Med. Image Anal.*, 2008, **12**, (6), pp. 731–741, (special issue on Information Processing in Medical Imaging 2007)

[26] Komodakis, N., Tziritas, G., Paragios, N.: 'Fast, approximately optimal solutions for single and dynamic mrfs'. Computer Vision and Pattern Recognition (CVPR), 2007

[27] Felzenszwalb, P., Huttenlocher, D.: 'Efficient belief propagation for early vision'. Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2004 (CVPR 2004), 2004, vol. **1**, pp. I-261–I-268

[28] Duchenne, O., Joulin, A., Ponce, J.: 'A graph-matching kernel for object categorization'. Int. Conf. on Computer Vision (ICCV), 2011, pp. 1792–1799

[29] Keysers, D., Deselaers, T., Gollan, C.*, et al.*: 'Deformation models for image recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (8), pp. 1422–1435

[30] Liu, K., Zhang, J., Huang, K.*, et al.*: 'Deformable object matching via deformation decomposition based 2d label mrf', IEEE Conference Computer Vision and Pattern Recognition (CVPR), June 2014

[31] Zhou, F., la Torre, F.D.: 'Deformable graph matching'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2922–2929

[32] Wiskott, L., Fellous, J.-M., Krger, N.*, et al.*: 'Face recognition by elastic bunch graph matching', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 775–779

[33] Komodakis, N., Paragios, N., Tziritas, G.: 'Mrf optimization via dual decomposition: message-passing revisited'. Int. Conf. on Computer Vision (ICCV), 2007, pp. 1–8

[34] Pearl, J.: '*Probabilistic reasoning in intelligent systems: networks of plausible inference*' (Morgan Kaufmann, 1988)

[35] Wainwright, M., Jordan, M.: '*Graphical models, exponential families, and variational inference*' (Now Publishers Inc., Hanover, MA, USA, 2008), vol. **1**, (1–2)

[36] Besag, J.: 'On the statistical analysis of dirty pictures', *J. R. Stat. Soc.*, 1986, **48**, (3), pp. 259–302

[37] Kirkpatrick, S., , C.D.G.Jr., Vecchi, M.P.: 'Optimization by simulated annealing', *Science*, 1983, **220**, pp. 671–680

[38] Kolmogorov, V.: 'Convergent tree-reweighted message passing for energy minimization', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (10), pp. 1568–1583

[39] Werner, T.: 'High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF)'. Proc. of the 2008 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2008), Madison, USA, June 2008, pp. 109–116

[40] Ramalingam, S., Kohli, P., Alahari, K., *et al.*: 'Exact inference in multi-label crfs with higher order cliques'. Computer Vision and Pattern Recognition (CVPR), 2008

[41] Boykov, Y., Veksler, O., Zabih, R.: 'Fast approximate energy minimization via graph cuts', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (11), pp. 1222–1239

[42] Werner, T.: 'A linear programming approach to max-sum problem: a review', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (7), pp. 1165–1179

[43] Gidas, B.: 'A renormalization group approach to image processing problems', *PAMI*, 1989, **11**, (2), pp. 164–180

[44] Bober, M., Petrou, M., Kittler, J.: 'Nonlinear motion estimation using the supercoupling approach', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (5), pp. 550–555

[45] Arashloo, S.R., Kittler, J., Christmas, W.J.: 'Pose-invariant face recognition by matching on multi-resolution mrfs linked by supercoupling transform', *Comput. Vis. Image Underst.*, 2011, **115**, (7), pp. 1073–1083, (special issue on Graph-based Representations in Computer Vision)

[46] Arashloo, S., Kittler, J.: 'Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features', *IEEE Trans. Inf. Forensics Sec.*, 2014, **9**, (12), pp. 2100–2109

[47] Chan, C.-H., Kittler, J., Messer, K.: 'Multi-scale local binary pattern histograms for face recognition'. Proc. of Int. Conf. on Biometrics, August 2007, pp. 809–818

[48] Chan, C.H., Tahir, M., Kittler, J., *et al.*: 'Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (5), pp. 1164–1177

[49] Kannala, J., Rahtu, E.: 'Bsif: binarized statistical image features'. Proc. 21st Int. Conf. on Pattern Recognition (ICPR 2012), Tsukuba, Japan, 2012, pp. 1363–1366

[50] Socher, R., Pennington, J., Huang, E.H., *et al.*: 'Semi-supervised recursive autoencoders for predicting sentiment distributions'. Empirical Methods on Natural Language Processing (EMNLP), 2011, pp. 151–161

[51] Bordes, A., Glorot, X., Weston, J., *et al.*: 'Joint learning of words and meaning representations for open-text semantic parsing'. Artificial Intelligence and Statistics (AISTATS), 2012, vol. **22**, pp. 127–135

[52] Ciresan, D.C., Meier, U., Schmidhuber, J.: 'Multi-column deep neural networks for image classification'. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3642–3649

[53] Ranzato, M.A., Huang, F.J., Boureau, Y.L., *et al.*: 'Unsupervised learning of invariant feature hierarchies with applications to object recognition'. IEEE Conf. on Computer Vision and Pattern Recognition, 2007 (CVPR '07), 2007, pp. 1–8

[54] Sun, Y., Wang, X., Tang, X.: 'Deep convolutional network cascade for facial point detection'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3476–3483

[55] Luo, P., Wang, X., Tang, X.: 'A deep sum-product architecture for robust facial attributes analysis'. Int. Conf. on Computer Vision (ICCV), 2013, pp. 2864–2871

[56] Chen, F., Yu, H., Hu, R., *et al.*: 'Deep learning shape priors for object segmentation'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1870–1877

[57] Ouyang, W., Wang, X.: 'Joint deep learning for pedestrian detection'. Int. Conf. on Computer Vision (ICCV), 2013

[58] Weinzaepfel, P., Revaud, J., Harchaoui, Z., *et al.*: 'Deepflow: large displacement optical flow with deep matching'. Int. Conf. on Computer Vision (ICCV), 2013, pp. 1385–1392

[59] Titov, I., Henderson, J.: 'Constituent parsing with incremental sigmoid belief networks'. The Association for Computational Linguistics (ACL), 2007

[60] Saul, L.K., Jaakkola, T., Jordan, M.I.: 'Mean field theory for sigmoid belief networks', *CoRR*, 1996, **cs.AI/9603102**

[61] Hinton, G.E., Osindero, S., Teh, Y.W.: 'A fast learning algorithm for deep belief nets', *Neural Comput.*, 2006, **18**, (7), pp. 1527–1554

[62] Bishop, C.: '*Pattern recognition and machine learning*' (Springer, New York, 2006), vol. **4**

[63] Hubel, D., Wiesel, T.N.: 'Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex', *J. Physiol.*, 1962, **160**, pp. 106–154

[64] Huang, G.B., Lee, H., Learned-Miller, E.G.: 'Learning hierarchical representations for face verification with convolutional deep belief networks'. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2518–2525

[65] Sun, Y., Wang, X., Tang, X.: 'Deep learning face representation by joint identification-verification', *CoRR*, 2014, **abs/1406.4773**

[66] Huang, G., Mattar, M., Berg, T., *et al.*: 'Labeled faces in the wild, a database for studying face recognition in unconstrained environments', 2008, (faces in real life images workshop in ECCV)

[67] del Solar, J.R., Verschae, R., Correa, M.: 'Recognition of faces in unconstrained environments: a comparative study', *EURASIP J. Adv. Signal Process.*, 2009, **2009**, pp. 1–19

[68] Seo, H.J., Milanfar, P.: 'Face verification using the lark representation', *IEEE Trans. Inf. Forensics Sec.*, 2011, **6**, (4), pp. 1275–1286

[69] Sharma, G., ul Hussain, S., Jurie, F.: 'Local higher-order statistics (lhs) for texture categorization and facial analysis'. Proc. of the 12th European Conf. on Computer Vision – Volume Part VII, ECCV'12, Berlin, Heidelberg, 2012, pp. 1–12

[70] Yi, D., Lei, Z., Li, S.: 'Towards pose robust face recognition'. IEEE Computer Vision and Pattern Recognition, 2013

[71] Turk, M.A., Pentland, A.P.: 'Face recognition using eigenfaces'. Proc. Conf. on Computer Vision and Pattern Recognition, 1991, pp. 586–591

[72] Nowak, E., Jurie, F.: 'Learning visual similarity measures for comparing never seen objects'. IEEE Conf. on Computer Vision and Pattern Recognition, 2007 (CVPR '07), June 2007, pp. 1–8

[73] Wolf, L., Hassner, T., Taigman, Y.: 'Descriptor based methods in the wild'. Post ECCV workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition, 2008

[74] Sanderson, C., Lovell, B.C.: 'Multi-region probabilistic histograms for robust and scalable identity inference'. ICB, 2009 (LNCS, 5558), pp. 199–208

[75] Pinto, N., DiCarlo, J.J., Cox, D.D.: 'How far can you get with a modern face recognition test set using only simple features?'. IEEE Computer Vision and Pattern Recognition, 2009

[76] Li, H., Hua, G., Lin, Z., *et al.*: 'Probabilistic elastic matching for pose variant face verification'. IEEE Computer Vision and Pattern Recognition, 2013

[77] Simonyan, K., Parkhi, O., Vedaldi, A., *et al.*: 'Fisher vector faces in the wild'. British Machine Vision Conf. (BMVC), 2013

[78] Hu, J., Lu, J., Tan, Y.-P.: 'Discriminative deep metric learning for face verification in the wild'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2014

[79] Huang, G.B., Jones, M.J., Learned-miller, E.: 'Lfw results using a combined nowak plus merl recognizer'. Faces in Real-Life Images Workshop in European Conf. on Computer Vision (ECCV), 2008

[80] Guillaumin, M., Verbeek, J., Schmid, C.: 'Is that you? metric learning approaches for face identification'. Int. Conf. on Computer Vision, September 2009, pp. 498–505. Available at http://www.lear.inrialpes.fr/pubs/2009/GVS09

[81] Taigman, Y., Wolf, L., Hassner, T.: 'Multiple one-shots for utilizing class label information'. British Machine Vision Association (BMVC), 2009

[82] Nair, V., Hinton, G.E.: 'Rectified linear units improve restricted Boltzmann machines'. Int. Conf. on Machine Learning (ICML), 2010, pp. 807–814

[83] Cao, Z., Yin, Q., Tang, X., *et al.*: 'Face recognition with learning-based descriptor'. Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2707–2714

[84] Nguyen, H.V., Bai, L.: 'Cosine similarity metric learning for face verification'. Asian Conf. on Computer Vision (ACCV), 2010 (LNCS, 6493), pp. 709–720

[85] Cox, D.D., Pinto, N.: 'Beyond simple features: a large-scale feature search approach to unconstrained face recognition'. Face and Gesture Recognition (FG), 2011, pp. 8–15

[86] Ying, Y., Li, P.: 'Distance metric learning with eigenvalue optimization', *J. Mach. Learn. Res.*, 2012, **13**, pp. 1–26

[87] Cui, Z., Li, W., Xu, D., *et al.*: 'Fusing robust face region descriptors via multiple metric learning for face recognition in the wild'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3554–3561

[88] Cao, Q., Ying, Y., Li, P.: 'Similarity metric learning for face recognition'. The IEEE Int. Conf. on Computer Vision (ICCV), December 2013

[89] Barkan, O., Weill, J., Wolf, L., *et al.*: 'Fast high dimensional vector multiplication face recognition'. The IEEE Int. Conf. on Computer Vision (ICCV), December 2013

[90] Kumar, N., Berg, A.C., Belhumeur, P.N., *et al.*: 'Attribute and simile classifiers for face verification'. Int. Conf. on Computer Vision (ICCV), 2009, pp. 365–372

[91] Yin, Q., Tang, X., 0001, J.S.: 'An associate-predict model for face recognition'. Computer Vision and Pattern Recognition (CVPR), 2011, pp. 497–504

[92] Berg, T., Belhumeur, P.N.: 'Tom-vs-Pete classifiers and identity-preserving alignment for face verification'. British Machine Vision Conf. (BMVC), 2012, pp. 1–11

[93] Chen, D., Cao, X., Wang, L., *et al.*: 'Bayesian face revisited: a joint formulation'. European Conf. on Computer Vision (ECCV), 2012 (LNCS, 7574), pp. 566–579

[94] Chen, D., Cao, X., Wen, F., *et al.*: 'Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3025–3032

[95] Lei, Z., Pietikainen, M., Li, S.: 'Learning discriminant face descriptor', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (2), pp. 289–302

[96] Cao, X., Wipf, D., Wen, F., *et al.*: 'A practical transfer learning algorithm for face verification'. Int. Conf. on Computer Vision (ICCV), 2013

[97] Taigman, Y., Wolf, L.: 'Leveraging billions of faces to overcome performance barriers in unconstrained face recognition', *CoRR*, 2011, **abs/1108.1122**

[98] Fan, H., Cao, Z., Jiang, Y., *et al.*: 'Learning deep face representation', *CoRR*, 2014, **abs/1403.2802**

[99] Taigman, Y., Yang, M., Ranzato, M., *et al.*: 'Deepface: closing the gap to human-level performance in face verification'. Conf. on Computer Vision and Pattern Recognition (CVPR), 2014

[100] Sun, Y., Wang, X., Tang, X.: 'Hybrid deep learning for face verification'. The IEEE Int. Conf. on Computer Vision (ICCV), December 2013

[101] Berg, T., Belhumeur, P.N.: 'Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation'. Computer Vision and Pattern Recognition (CVPR), 2013, pp. 955–962

[102] Zhu, Z., Luo, P., Wang, X., *et al.*: 'Recover canonical-view faces in the wild with deep neural networks', *CoRR*, 2014, **abs/1404.3543**
[103] Sun, Y., Wang, X., Tang, X.: 'Deep learning face representation from predicting 10,000 classes'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2014
[104] Lu, C., Tang, X.: 'Surpassing human-level face verification performance on lfw with Gaussian face', *CoRR*, 2014, **abs/1404.3840**
[105] Chopra, S., Hadsell, R., Lecun, Y.: 'Learning a similarity metric discriminatively, with application to face verification'. Proc. of Computer Vision and Pattern Recognition Conf., 2005, pp. 539–546
[106] Zhu, Z., Luo, P., Wang, X., *et al.*: 'Deep learning identity-preserving face space'. IEEE Int. Conf. on Computer Vision (ICCV), 2013, December 2013
[107] Kan, M., Shan, S., Chang, H., *et al.*: 'Stacked progressive auto-encoders (spae) for face recognition across poses'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2014, pp. 1883–1890