EPML: Expanded Parts based Metric Learning for Occlusion Robust Face Verification

Gaurav Sharma, Frédéric Jurie, Patrick Pérez

EPML: Expanded Parts based Metric Learning for Occlusion Robust Face Verification

Gaurav Sharma[1], Frédéric Jurie[2] and Patrick Pérez[1]

[1]Technicolor
[2]GREYC CNRS UMR 6072, University of Caen Basse-Normandie

Abstract. We propose a novel Expanded Parts based Metric Learning (EPML) model for face verification. The model is capable of mining out the discriminative regions at the right locations and scales, for identity based matching of face images. It performs well in the presence of occlusions, by avoiding the occluded regions and selecting the next best visible regions. We show quantitatively, by experiments on the standard benchmark dataset Labeled Faces in the Wild (LFW), that the model works much better than the traditional method of face representation with metric learning, both (i) in the presence of heavy random occlusions and, (ii) also, in the case of focussed occlusions of discriminative face regions such as eyes or mouth. Further, we present qualitative results which demonstrate that the method is capable of ignoring the occluded regions while exploiting the visible ones.

1 Introduction

Face verification technology is critical for many modern systems. Handling occlusions is a major challenge in its real world application. In the present paper, we propose a novel Expanded Parts based Metric Learning (EPML) model which is capable of identifying many discriminative parts of the face, for the task of predicting if two faces are of the same person or not, especially in the presence of occlusions.

Metric Learning approaches [1–3] have recently shown promise for the task of face verification. However the face representation is usually fixed and is separate from the task of learning the model. The faces are usually represented as either features computed on face landmarks [3] or over a fixed grid of cells [4]. Once such fixed representations are obtained, a discriminative metric learning model is learned, with annotated *same* and *not-same* faces, for comparing faces based on identity. Since the representation is fixed, it is completely the model's responsibility to tackle the challenge of occlusions that might occur in in-the-wild applications. In the present, the proposed EPML model learns a collection of discriminative parts (of the faces) along with the discriminative model to compare faces with such a collection of parts. The collection of discriminative parts is automatically mined from a large set of randomly sampled candidate parts, in the learning phase. The distance function used considers the distances between all the n different parts in the model and computes the final distance between

the two faces with the closest (small number of) $k < n$ parts. Such `min` operation lends non-linearity to the model and allows it to selectively choose/reject parts at runtime, which is in contrast to the traditional representation where the model has no choice but to consider the whole face. This capability is specially useful in case of occlusion: while the traditional method is misguided by the occluded part(s), the proposed method can simply choose to ignore significantly occluded parts and use only the visible parts. We discuss this further later, along with qualitative results, in §3. In the following, we first set the context by briefly describing the traditional metric learning approaches (§1.1). We then motivate our method (§2) and present it in detail (§2.1) and finally we give experimental results (§3) and conclude the article (§4).

1.1 Background: Face verification using metric learning

Given a face image dataset $\mathcal{X}$ of positive (of the same person) pairs of images and negative pairs of images, represented with some feature vectors i.e.,

$$\mathcal{X} = \{(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) | i = 1, \ldots, l, j = 1, \ldots, m\}, \tag{1}$$

with $\mathbf{x}_i \in \mathbb{R}^D$ a face feature vector and $y_{ij} \in \{-1, +1\}$, the task is to learn a function to predict if two unseen face images, potentially of unseen person(s), are of the same person or not.

Metric learning approaches, along with standard image representations, have been recently shown to be well suited to this task [1–3]. Such approaches learn from $\mathcal{X}$ a Mahalanobis-like metric parametrized by matrix $M \in \mathbb{R}^{D \times D}$, i.e.,

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M (\mathbf{x}_i - \mathbf{x}_j). \tag{2}$$

To be a valid metric, M needs to be symmetric and positive semi-definite (PSD) and hence can also be factorized as

$$M = L^\top L, \tag{3}$$

with $L \in \mathbb{R}^{d \times D}$ and $d \leq D$. The metric learning can then be seen as an embedding learning problem: To compare two vectors, first project them on the d-dim row-space of L and then compare them with their ℓ^2 distance in the projected space, i.e.,

$$\begin{aligned} d^2(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^\top M (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top L^\top L (\mathbf{x}_i - \mathbf{x}_j) \\ &= (L(\mathbf{x}_i - \mathbf{x}_j))^\top (L(\mathbf{x}_i - \mathbf{x}_j)) \\ &= ||L\mathbf{x}_i - L\mathbf{x}_j||_2^2. \end{aligned} \tag{4}$$

Many regularized loss minimization algorithms have been proposed to learn such metrics with the loss functions arising from probabilistic (likelihood) or margin-maximizing motivations [1–3].

1.2 Related work

The recognition of face under occlusions has a long history in computer vision literature. One of the pioneering work was that of Leonardis et al. [5] who proposed to make the *Eigenface* method more robust to occlusion by computing the coefficients of the eigenimages with a hypothesize-and-test paradigm using subsets of image points. Since then, more efficient face matching algorithms have been proposed, raising the question of how to make them more robust to occlusions. The best performing state-of-the-art methods (e.g. [6, 7]) are holistic in the sense that they represent the whole face by a single vector and are, hence, expected to be sensitive to occlusions.

The impact of occlusions on face recognition has been studied by Rama et al. [8], who evaluated three different approaches based on Principal Component Analysis (PCA) (i.e., the eigenface approach, a component-based method built on the eigen-features and an extension of the Lophoscopic Principal Component Analysis). They analysed the three different strategies and compared them when used for identifying partially occluded faces. The paper also explored how prior knowledge about occlusions, which might be present, can be used to improve the recognition performance.

Generally speaking, two different methodologies have been proposed in the literature. One consists of detecting the occlusions and reconstructing occluded parts prior to doing face recognition, while the other one relies on integrated approaches (of description and recognition together) i.e., those that are robust to occlusions by construction.

Within the family of *detect and reconstruct* approaches, one can mention the works of Colombo et al. [9, 10], who detect occlusions by doing a comparison of the input image with a generic model of face and reconstruct missing part with the Gappy Principal Component Analysis (GPCA) [11]. Lin et al. [12], propose another approach for automatically detecting and recovering the occluded facial regions. They consider the formation of an occluded image as a generative process which is then used to guide the procedure of recovery. More recently, Oh et al. [13] proposed to detect occluded parts by dividing the image into a finite number of disjoint local patches coded by PCA and then using 1-NN threshold classifier. Once detected, only the occlusion-free image patches are used for the later face recognition stage, where the matching is done with a nearest neighbor classifier using the Euclidean distance. Wright et al. [14] and Zhou et al. [15] explored the use of sparse coding, proposing a way to efficiently and reliably identify corrupted regions and exclude them from the sparse representation. Sparse coding has also been used efficiently by Ou et al. [16], while Morelli et al. [17] have proposed using compressed sensing. Min et al. [18] proposed to detect the occlusion using Gabor wavelets, PCA and support vector machines (SVM), and then do recognition with the non-occluded facial parts, using block-based Local Binary Patterns of Ojala et al. [19]. Tajima et al. [20] suggested detecting occluded regions using Fast-Weighted Principal Component Analysis (FW-PCA) and using the occluded regions for weighting the blocks for face representation. Alyuz et al. [21] proposed to deal with occlusions by using fully

automatic 3-D face recognition system in which face alignment is done through an adaptively selected model based registration scheme (where only the valid non-occluded patches are utilized), while during the classification stage, they proposed a masking strategy to enable the use of subspace analysis techniques with incomplete data. Min et al. [22] proposed to compute an occlusion mask indicating which pixel in a face image is occluded and to use a variant of local Gabor binary pattern histogram sequences (LGBPHS) to represent occluded faces by excluding features extracted from the occluded pixels. Finally, different from previous approaches, Colombo et al. [23] addressed the question of detection and reconstruction of faces using 3D data.

The second paradigm for addressing the recognition of occluded faces, which is to develop method that are intrinsically robust to occlusion, has received less attention in the past. Liao et al. [24] developed an alignment-free face representation method based on Multi-Keypoint Descriptors matching, where the descriptor size of a face is determined by the actual content of the image. Any probe face image (holistic or partial) can hence be sparsely represented by a large dictionary of gallery descriptors, allowing partial matching of face components. Weng et al. [25] recently proposed a new partial face recognition approach by aligning partial face patches to holistic gallery faces automatically, hence being robust to occlusions and illumination changes. Zhao [26] used a robust holistic feature relying on stable intensity relationships of multiple point pairs, being intrinsically invariant to changes in facial features, and exhibiting robustness to illumination variations or even occlusions.

Face verification in real world scenarios has recently attracted much attention, specially fueled by the availability of the excellent benchmark: Labeled Faces in the Wild (LFW) [27]. Many recent papers address the problem with novel approaches, e.g. discriminative part-based approach by Berg and Belhumeur [28], probabilistic elastic model by Li et al. [29], Fisher vectors with metric learning by Simonyan et al. [2], novel regularization for similarity metric learning by Cao et al. [30], fusion of many descriptors using multiple metric learning by Cui et al. [31], deep learning by Sun et al. [32], method using fast high dimensional vector multiplication by Barkan et al. [33]. Many of the most competitive approaches on LFW combine different features, e.g. [34–36] and/or use external data, e.g. [37, 38].

The works of Liao et al. [24] and Weng et al. [25] are the most closely related and competing works to the proposed EPML. They are based on feature set matching (via dictionaries obtained with sparse representation). Like in image retrieval, there are two ways of doing occlusion robust face matching: (i) match local features detected around keypoints from the two faces or (ii) aggregate the local features to make a global (per cell) feature vector and then match two image vectors. These works fall in the first category while the proposed method falls in the second. The first type of methods are robust to occlusions due to the matching process, while for the second type, the model and aggregation together have to account for the occlusion. Also, the first type of methods claim that they don't need alignment of faces. If a face detector is used then by the statistical

properties of the detector the faces will be already approximately aligned (LFW is made this way and strong models already give good results without further alignment). So the first type of methods are arguably more useful when an operator outlines a difficult 'unaligned' face manually and gives it as an input. In that case, we could also make her approximately align the faces as well. And in the case when the face detector is trained to detect faces in large variations in pose, then probably the pose will come out as a latent information from the detector itself, which can then be used to align the face approximately. In summary, we argue that both the approaches have merit and the second type, which is the subject of this paper, has the potential to be highly competitive when used with recently developed strong features with a model-and-aggregation designed to be robust to occlusion like the proposed EPML.

Our work could also be contrasted with feature selection algorithms, e.g. Viola and Jones [39], and many other works in similar spirit, where a subset of features (in a cascaded fashion) are selected from a very large set of features. The proposed method is similar to feature selection methods as we are selecting a subset of parts from among a large set of potential candidate parts. However, it is distinctly different as it performs a dynamic test time selection of most reliable parts, from among the parts selected at training, which are available for the current test pair.

Finally, our work is also reminiscent of the mid-level features stream of work, e.g. see Doersch et al. [40] and the references within, which aim at extracting visually similar recurring and discriminative parts in the images. In a similar spirit, we are interested in finding parts of faces which are discriminative for verification, after the learnt projection.

2 Motivation and Approach

A critical component in computer vision applications is the image representation. The state-of-the-art image representation methods first compute local image statistics (or features as they are usually called) and then *aggregate* them to form a fixed length representation of the images. This aggregation/pooling step reduces a relatively large number of local features to a smaller fixed length vector and there is a trade-off involved here, specially at a spatial level; it is now commonly accepted, e.g. for image classification, that, instead of a global image level aggregation, including finer spatial partitions of the image leads to better results [41]. Learning such partitions does better still [42, 43].

In the present paper, we would like to draw attention to some issues related to the fixed grid based spatial aggregation aspect in the context of facial verification tasks with metric learning algorithms. While the metric learning algorithms are expected to find and exploit (absence of) correlations between the various facial regions (or *parts*), they can only do so effectively if local features from different regions are not aggregated into the same (or a very small number of) dimension(s) of the final representation. Towards this issue of aggregation, there are two closely related points:

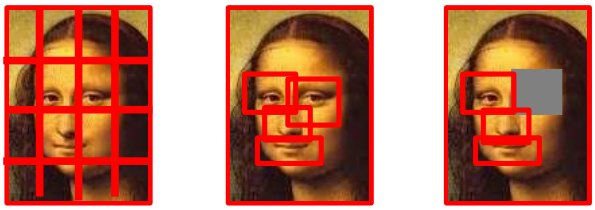

Fig. 1. While the uniform grid might aggregate features from two discriminative regions, e.g. nose and mouth (left) and thus make the final representation less effective, the proposed Expanded Parts based Metric Learning (EPML) model can optimally mine out the spatial bins required for the task (middle) leading to the most discriminative full representation. Further, in case a part of the face is occluded (marked gray in the figure), EPML can ignore the occluded part and take the decision based on the other visible parts (right) and hence be robust to occlusions.

(i) At *what resolutions and locations* such parts should appear?
(ii) *How many* of such parts are optimal for the task?

The current face verification methods usually split the face using uniform grids, possibly at multiple scales, and leave the rest to the metric learning stage. We, instead, propose a novel Expanded Parts based Metric Learning (EPML) model, inspired by the recently proposed Expanded Parts Model for image classification [44], for the task of face verification. The proposed method is capable of mining out the parts from a set of large number of randomly sampled candidate parts. The distance function used in EPML is a non-linear function which uses a subset of best matching parts from among all the parts present in the model. Hence, in the case of occlusions, while the traditional metric learning methods have no choice but to use the fixed full image representation, the proposed EPML can choose to ignore the occluded parts and select the next best matching visible parts. Fig. 1 illustrates the points.

2.1 Proposed method

The goal of the algorithm is to learn a collection of n parts and match a pair of face images using only $k < n$ best matching parts.

In contrast to the metric learning approaches, discussed above, we define the distance between a pair of face images as

$$d_e^2(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{k} \min_{\boldsymbol{\alpha}_{ij} \in \{0,1\}^n} \sum_{p=1}^{n} \alpha_{ij}(p) \|L_p \mathbf{x}_{i|p} - L_p \mathbf{x}_{j|p}\|^2 \tag{5}$$

$$\text{s.t. } \|\boldsymbol{\alpha}_{ij}\|_1 = k \;\; \text{and } O_v(\boldsymbol{\alpha}_{ij}) < \theta,$$

where, $\boldsymbol{\alpha}_{ij} = (\alpha_{ij}(1), \ldots, \alpha_{ij}(n)) \in \{0,1\}^n$ is the binary indicator vector specifying which of the parts are being used and which are not, L_p is the projection

Algorithm 1 Stochastic gradient descent for learning EPML

Given: Number of candidate parts (N), rate (r), parameters k, β_p, d'
`parts` $\leftarrow$ Randomly sample N candidate parts
for $p = 1, \ldots, N$ **do**
  $L_p \leftarrow$ Whitened-PCA($\{\mathbf{x}_{i|p}, \forall \mathbf{x}_i$ in training set$\}, d'$)
end for
for iter $= 1, 2, 3, \ldots, 10^6$ **do**
  Randomly sample a pos or neg training pair $(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$, with equal probability
  Compute distance (Eq. 5) to get $d_e^2(\mathbf{x}_i, \mathbf{x}_j)$ and $\boldsymbol{\alpha}_{ij}$
  if $y_{ij}(b - d_e^2(\mathbf{x}_i, \mathbf{x}_j)) < 1$ **then**
    for all p such that $\boldsymbol{\alpha}_{ij}(p) = 1$ **do**
      $L_p \leftarrow L_p - rL_p(\mathbf{x}_{i|p} - \mathbf{x}_{j|p})(\mathbf{x}_{i|p} - \mathbf{x}_{j|p})^\top$
    end for
    $b \leftarrow b + ry_{ij}$
  end if
  `parts_image_map` $\leftarrow$ *note_used_parts* ($\boldsymbol{\alpha}_{ij}$)
  if mod(iter, 10^5) $= 0$ **then**
    (`parts`, $\{L_p\}$) $\leftarrow$ *prune_parts* (`parts_image_map`, β_p, `parts`, $\{L_p\}$)
  end if
  Output: (`parts`, $\{L_p\}, b$) // n parts left after pruning
end for

matrix for the p^{th} part and $\mathbf{x}_{i|p}$ is the feature vector of the region corresponding to the part p for face image i. We also ensure that the parts do not overlap more than a threshold θ, captured by the second constraint based on the overlap function O_v, to encourage coverage of the faces. To mine out the parts and learn the parameters $\{L_p | p = 1, \ldots, n\}$, we propose to solve the following margin maximization problem:

$$\min_{\{L_p\}, b} F(\mathcal{X}; \{L_p\}, b) = \sum_{\mathcal{X}} \max\left(0, 1 - y_{ij}\{b - d_e^2(\mathbf{x}_i, \mathbf{x}_j)\}\right), \tag{6}$$

$$\text{where,} \quad \mathcal{X} = \{(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) | i = 1, \ldots, l, j = 1, \ldots, m\} \tag{7}$$

is the given (annotated) training set, and the b parameter is an offset/threshold to which the distances between the two examples $\mathbf{x}_i$ and $\mathbf{x}_j$ is to be compared (to decide same or not-same) and is learnt (Alg. 1, more below).

Computing the distance. We solve the distance computation (5) by resorting to an approximate greedy forward selection. We first compute the distances between each of the parts in the two images and then recursively (i) select the best matching parts and (ii) discard the remaining parts which overlap more than a threshold with the combined area of the already selected parts.

Solving the optimization with SGD. The optimization (6) is non-convex due to the presence of the `min` function in the distance (5). We use stochastic gradient descent to solve it. The analytical subgradients of the objective function,

Fig. 2. Example positive image pairs from the LFW [27, 35] dataset.

w.r.t. a single training pair $(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$, are given by

$$\nabla_{L_p} F = y_{ij} L_p (\mathbf{x}_{i|p} - \mathbf{x}_{j|p})(\mathbf{x}_{i|p} - \mathbf{x}_{j|p})^\top. \tag{8}$$

The algorithm we use for learning is given in Alg. 1. We use no regularization and a small step size (rate r) with a fixed number of one million iterations.

Parts mining. We divide the whole set of one million learning iterations into 10 sets and after each set (i.e., 100 thousand iterations), we prune the parts by removing those parts which were used for less than β_p fraction of iterations in that set. So, if there are T iterations after which pruning happens, a part has to be utilized in at least $\lceil \beta_p T \rceil$ iterations. Such pruning helps in removing redundant and/or non-discriminative parts. We start with N candidate parts, which are randomly sampled, and prune them to n parts in the final model. N, β_p are free parameters of the algorithm.

Initialization with WPCA. We use Whitened Principal Component Analysis (WPCA) to initialize the projection matrices for each part. For each part, we crop all the faces, in the training set, corresponding to that part and perform WPCA on them, which is used to initialize the projection matrix L_p for the respective part.

3 Experimental results

Database used. We use the aligned version [35] of the Labeled Faces in the Wild (LFW) database by Huang et al. [27]. The dataset has more than 13000 images of over 4000 persons. The evaluation is done on the task of face verification in the unrestricted setting. The test set of LFW contains 10 folds of 300 positive and 300 negative pairs. The evaluation is done by using 1 fold for testing and remaining 9 folds for training and reporting the average of the accuracies

obtained for 10 folds. The identities of the persons in the training set are used to make positive pairs (of same person) and negative pairs (of different persons) of face images, for training. The training is done with unoccluded images and the testing is done with one of the test pair images occluded by one of the methods discussed below. The evaluation simulates the case when the database has unoccluded images of the known persons and the test images come with unexpected occlusions.

Image description. To describe the face (part) images, we resort to the Local Binary Pattern (LBP) descriptors of Ojala et al. [19]. We use grayscale images and centre crop them to size 150×100 pixels and do not do any other preprocessing. We utilize the publicly available `vlfeat` [45] library for LBP, with cell size of 10 pixels, resulting in $D = 9860$.

Baseline method (ML). The baseline metric learning method, denoted ML in the following, is a max margin optimization similar to Eq. 6 with the distance function as explained in §1.1. The training is done with stochastic gradient based algorithm, same as that of the proposed EPML.

Parameters. We fixed the projection dimension for the baseline method to be 64 after doing preliminary experiments and observing saturation of performance for large values of the projection dimension. The number of randomly generated candidate parts was fixed to $N = 500$. The parts were sampled randomly, to have between 20% to 80% of the widths and heights of the face image with random locations. Other parameters were fixed to as $k = 20, d = 20$, because of diminishing returns for higher values of these parameters which makes learning and testing slower. Each candidate part was itself represented by LBP similar to the baseline method.

Horizontal train/test-time flipping. Since the face images are taken 'in the wild' and have highly varied expressions, poses and appearances, we average out some of the variations by replacing every face pair by four pairs through horizontal flip of each image. Doing this at training should ideally make the system invariant to such variations and make such flipping redundant at test time. However, as the training set is limited, we investigate test time flipping as well.

Random occlusions. To test the robustness of the system to occlusions, we overlay uniform rectangular patches at randomly sampled locations *on one of the faces* (randomly selected) of a test pairs. This simulates the case when an unoccluded image is present in the dataset and an occluded image, captured by a system on the field, has to be matched to it. Such cases can happen when there is natural occlusion due to damage or dust on the camera lens or front glass, especially in the case of surveillance cameras. We sample such patches with areas ranging from 20% to 80% of the face area. Fig. 3 (top row) shows examples of such randomly occluded faces used in the experiments.

Focussed occlusions. To stress the system further, we manually mark the parts which are frequently found discriminative, e.g. eyes, central part of the

Fig. 3. Examples images showing the kind of occlusions used in the experiments. The top row shows random occlusions while the bottom row shows focussed, and fixed, occlusion of eyes (one, both) and mouth+nose.

face around the nose and the mouth. We then occlude these regions, one at a time and in combinations, by overlaying uniform patches. As with the random occlusions above, we do it for one of the, randomly selected, faces in the pair. We test the system for robustness to occlusion of (i) left/right eye, (ii) both eyes, (iii) central face part around the nose, (iv) mouth and (v) nose and mouth. Fig. 3 (bottom row) shows examples of such occlusions.

3.1 Quantitative results

Fig. 4 shows the results for the proposed Expanded Parts based Metric Learning (EPML) method vs. the baseline Metric Learning (ML), in the case of random occlusions. The four graphs correspond to the cases when the horizontal flipping is used, or not, for train and test image pairs (marked 'train flip' and 'test flip' in the graph titles). We observe that the proposed EPML model clearly outperforms the metric learning (ML) method with fixed grid image representation. The improvements are significant, from 2% to 6% absolute across the range of occlusions. The standard error on the mean are always relatively small (less than 0.5%) and hence the improvements are statistically significant.

Tab. 1 gives the results of the proposed EPML model vs. ML, in the case of focussed occlusions. We find again that the proposed method is robust to occlusions especially for the discriminative parts, e.g. the performance drops much more significantly (to 61.7, for ML) when the eyes are occluded, compared to when the nose and mouth are occluded (to 73.5, for ML) and the gain in such cases is larger for the proposed method, e.g. +7.5 absolute for both eyes occluded vs. only +2 for nose and mouth both occluded. The method, thus, seems to recover gracefully from the occlusion of highly discriminative face regions compared to the traditional ML methods.

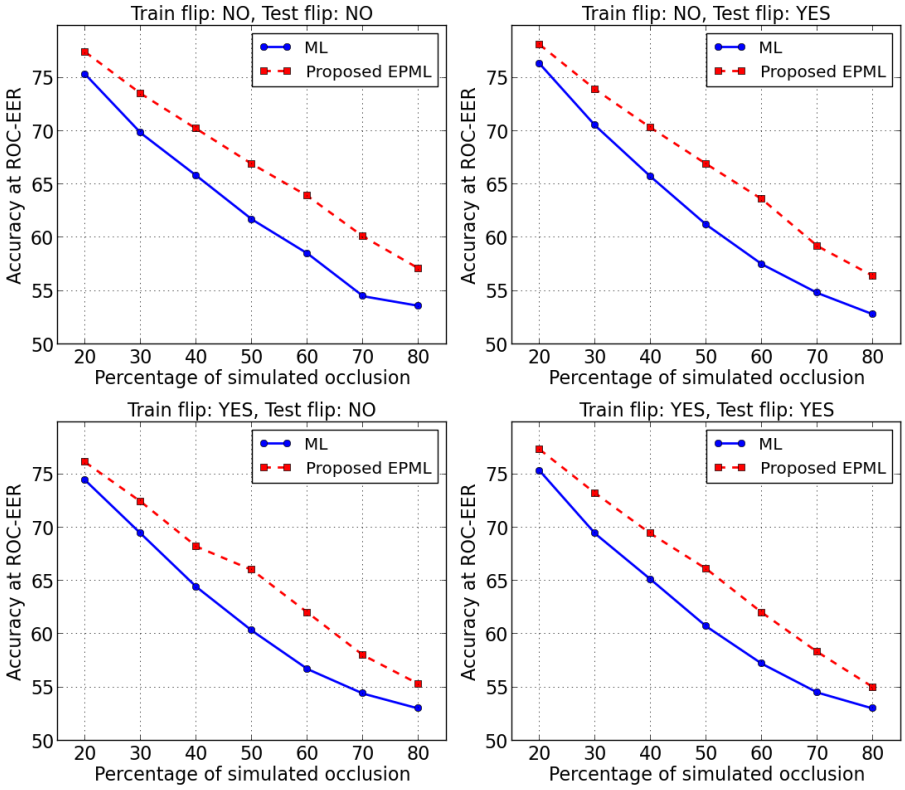

Fig. 4. The performance of the proposed EPML vs. the traditional metric learning methods in the presence of different level of random occlusions. The image pairs are flipped or not during training and/or testing. See §3 for discussion.

Hence we conclude that the proposed method is more robust to occlusion and learning a localized parts based model instead of a model based on a global face representation is beneficial for obtaining occlusion robustness.

3.2 Qualitative results

Fig. 5 shows the visualization of the scoring by the proposed EPML model in the presence of significant occlusions. We can see that the model is capable of ignoring the occluded parts and using the visible regions for scoring the face images. Based on the part occurrences, it may be inferred that the discriminative regions are mostly around the eyes and mouth of the faces. Note that there appears to be a preference by the model to the left region of the face, particularly the left eye, while it seems that it is ignoring the right part of the face. Since the scoring is done by averaging over the four possible pairs formed by horizontally flipping the two images, every part should be perceived along with its horizontally mirror

Fig. 5. Visualization of the parts used by the proposed EPML model, for matching pairs of faces in the presence of significant occlusions. The top 5 parts out of 20 (with randomly selected colors for better visibility) selected by the model for scoring the face image pairs are shown. We observe that the method quite successfully ignores the occluded parts. The parts used are also quite diverse and have good coverage as ensured by the model. See §3.2 for discussion.

Table 1. Results with focussed occlusion on the LFW dataset. See §3.1 for discussion.

	Left eye	Right eye	Both eyes	Nose	Mouth	Nose + mouth
ML	75.5	73.4	61.7	78.0	77.3	73.5
EPML	78.9	77.0	69.2	79.1	78.5	75.5

version (about the centre vertical axis) and hence, there is no such preference by the model. We conclude that the model seems to counter occlusions well.

3.3 Discussion w.r.t. state-of-the-art methods

The focus of this paper was to propose a novel method based on localized parts and to evaluate it specially in the context of occlusion robust face verification. Face verification is a very active topic and many features have been proposed obtaining from about around 70% to up to 93% performance on the LFW dataset without using external data and a near perfect 99% while using large amounts of external data[1]. Our implementation with Local Binary Pattern (LBP) features obtains 86% on the test set of LFW (ROC-EER without occlusion) which is competitive with other methods using similar features.

The EPML algorithm localizes parts and mines them out from a large set of candidate parts. It uses only a subset of parts to match a given pair of images, which allows it to be robust to occlusion. We demonstrated the benefits of the model for the highly popular and lightweight LBP features and we believe that the proposed EPML model will similarly benefit other strong, but global, image representations as well, especially in the case of occlusions.

4 Conclusion

We proposed a novel Expanded Parts based Metric Learning algorithm. The proposed method is capable of mining out the discriminative parts, from among a large set of randomly sampled candidate parts, at the appropriate locations and scales. While the traditional metric learning algorithms use a fixed grid based image representation and are strongly misguided by occlusions, the proposed method has the flexibility to be able to ignore the occluded parts and work with the next best matching visible parts and hence has better robustness to occlusions. The effectiveness of the proposed method w.r.t. the traditional metric learning methods was verified by experiments on the challenging Labeled Faces in the Wild (LFW) dataset with a single feature channel. In the future we would like to use the method with multiple channels of features and perhaps use similar principle to do feature selection as well.

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

Acknowledgement. This work was partially supported by the FP7 European integrated project AXES and by the ANR project PHYSIONOMIE.

References

1. Mignon, A., Jurie, F.: PCCA: A new approach for distance learning from sparse pairwise constraints. In: CVPR. (2012)
2. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: BMVC. (2013)
3. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: ICCV. (2009)
4. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: CVPR. (2013)
5. Leonardis, A., Bischof, H.: Dealing with occlusions in the eigenspace approach. In: CVPR. (1996)
6. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep fisher networks for large-scale image classification. In: NIPS. (2013)
7. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. (2014)
8. Rama, A., Tarres, F., Goldmann, L., Sikora, T.: More robust face recognition by considering occlusion information. In: FG. (2008)
9. Colombo, A., Cusano, C., Schettini, R.: Detection and restoration of occlusions for 3d face recognition. In: ICME. (2006)
10. Colombo, A., Cusano, C., Schettini, R.: Recognizing faces in 3d images even in presence of occlusions. In: BTAS. (2008)
11. Everson, R., Sirovich, L.: Karhunen–loeve procedure for gappy data. JOSA A **12** (1995) 1657–1664
12. Lin, D., Tang, X.: Quality-driven face occlusion detection and recovery. In: CVPR. (2007)
13. Oh, H.J., Lee, K.M., Lee, S.U.: Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. IVC **26** (2008) 1515–1523
14. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. PAMI **31** (2009) 210–227
15. Zhou, Z., Wagner, A., Mobahi, H., Wright, J., Ma, Y.: Face recognition with contiguous occlusion using markov random fields. In: CVPR. (2009)
16. Ou, W., You, X., Tao, D., Zhang, P., Tang, Y., Zhu, Z.: Robust face recognition via occlusion dictionary learning. PR **47** (2014) 1559–1572
17. Morelli Andrés, A., Padovani, S., Tepper, M., Jacobo-Berlles, J.: Face recognition on partially occluded images using compressed sensing. PRL **36** (2014) 235–242
18. Min, R., Hadid, A., Dugelay, J.: Improving the recognition of faces occluded by facial accessories. In: FG. (2011)
19. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24** (2002) 971–987
20. Tajima, Y., Ito, K., Aoki, T., Hosoi, T., Nagashima, S., Kobayashi, K.: Performance improvement of face recognition algorithms using occluded-region detection. In: ICB. (2013)
21. Alyuz, N., Gokberk, B., Akarun, L.: 3-d face recognition under occlusion using masked projection. IEEE Transactions on Information Forensics and Security **8** (2013) 789–802

22. Min, R., Hadid, A., Dugelay, J.L.L.: Efficient detection of occlusion prior to robust face recognition. Scientific World Journal **2014** (2014) 519158
23. Colombo, A., Cusano, C., Schettini, R.: Three-dimensional occlusion detection and restoration of partially occluded faces. Journal of mathematical imaging and vision **40** (2011) 105–119
24. Liao, S., Jain, A.K., Li, S.Z.: Partial face recognition: Alignment-free approach. PAMI **35** (2013) 1193–1205
25. Weng, R., Lu, J., Hu, J., Yang, G., Tan, Y.P..P.: Robust feature set matching for partial face recognition. In: ICCV. (2013)
26. Zhao, X., He, Z., Zhang, S., Kaneko, S., Satoh, Y.: Robust face recognition using the GAP feature. PR **46** (2013) 2647–2657
27. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
28. Berg, T., Belhumeur, P.N.: POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR. (2013)
29. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: CVPR. (2013)
30. Cao, Q., Ying, Y., Li, P.: Similarity metric learning for face recognition. In: ICCV. (2013)
31. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X.: Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In: CVPR. (2013)
32. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: ICCV. (2013)
33. Barkan, O., Weill, J., Wolf, L., Aronowitz, H.: Fast high dimensional vector multiplication face recognition. In: ICCV. (2013)
34. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: ICCV. (2009)
35. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: ACCV. (2009)
36. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: ACCV. (2010)
37. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
38. Berg, T., Belhumeur, P.N.: Tom-vs-pete classifiers and identity-preserving alignment for face verification. In: BMVC. (2012)
39. Viola, P., Jones, M.J.: Robust real-time face detection. Intl. Journal of Computer Vision **57** (2004) 137–154
40. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: NIPS. (2013)
41. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
42. Sharma, G., Jurie, F.: Learning discriminative representation image classification. In: BMVC. (2011)
43. Jiang, L., Tong, W., Meng, D., Hauptmann, A.G.: Towards efficient learning of optimal spatial bag-of-words representations. In: ICMR. (2014)
44. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: CVPR. (2013)
45. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)