CP-mtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval

Binod Bhattarai, Gaurav Sharma, Frédéric Jurie

▶ **To cite this version:**

Binod Bhattarai, Gaurav Sharma, Frédéric Jurie. CP-mtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval. 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Jun 2016, Las Vegas, NV, United States. hal-01301381

HAL Id: hal-01301381
https://hal.science/hal-01301381

Submitted on 12 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CP-mtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval

Binod Bhattarai[1,*] Gaurav Sharma[2,3,†] Frederic Jurie[1,*]

[1]University of Caen, France [2]MPI for Informatics, Germany [3]IIT Kanpur, India

Abstract

We propose a novel Coupled Projection multi-task Metric Learning (CP-mtML) method for large scale face retrieval. In contrast to previous works which were limited to low dimensional features and small datasets, the proposed method scales to large datasets with high dimensional face descriptors. It utilises pairwise (dis-)similarity constraints as supervision and hence does not require exhaustive class annotation for every training image. While, traditionally, multi-task learning methods have been validated on same dataset but different tasks, we work on the more challenging setting with heterogeneous datasets and different tasks. We show empirical validation on multiple face image datasets of different facial traits, e.g. identity, age and expression. We use classic Local Binary Pattern (LBP) descriptors along with the recent Deep Convolutional Neural Network (CNN) features. The experiments clearly demonstrate the scalability and improved performance of the proposed method on the tasks of identity and age based face image retrieval compared to competitive existing methods, on the standard datasets and with the presence of a million distractor face images.

1. Introduction

Many computer vision algorithms heavily rely on a distance function over image signatures and their performance strongly depends on the quality of the metric. Metric learning (ML) i.e. learning an optimal distance function for a given task, using annotated training data, is in such cases, a key to good performance. Hence, ML has been a very active topic of interest in the machine learning community and has been widely used in many computer vision algorithms for image annotation [11], person re-identification [2] or face matching [12], to mention a few of them.

*GREYC CNRS UMR 6072. Supported by projects ANR–12–CORD–014–SECULAR ANR-12-SECU-005–PHYSIONOMIE

†Currently with CSE, Indian Institute of Technology Kanpur. Majority of the work was done at Max Planck Institute for Informatics.

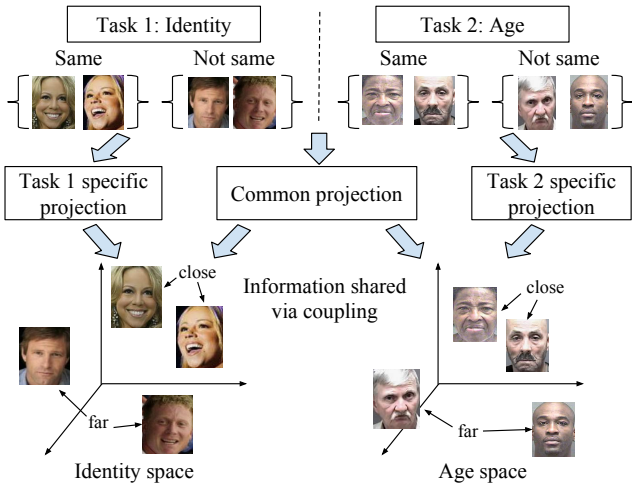

Figure 1. Illustration of the proposed method. We propose a multi-task metric learning method which learns a distance function as a projection into a low dimensional Euclidean space, from pairwise (dis-)similarity constraints. It learns two types of projections jointly: (i) a common projection shared by all the tasks and (ii) task related specific projections. The final projection for each task is given by a combination of the common projection and the task specific projection. By coupling the projections and learning them jointly, the information shared between the related tasks can lead to improved performance.

This paper focuses on the task of face matching i.e. comparing images of two faces with respect to different criteria such as identity, expression or age. More precisely, the task is to retrieve faces similar to a query, according to the given criteria (e.g. identity) and rank them using their distances to the query.

One key contribution of this paper is the introduction of a cross-dataset multi-task ML approach. The main advantage of multi-task ML is leveraging the performance of single task ML by combining data coming from different but related tasks. While many recent works on classification have shown that learning metrics for related tasks together using multi-task learning approaches can lead to improvements in performance [1, 6, 19, 21, 28, 43], most of earlier works on face matching are based on a single task. In addition, there

are only a few works on multi-task ML [25, 37, 41], with most of the multi-task approaches being focussed on multi-task classification. In addition, the previous multi-task ML methods have been shown to work on the same dataset but not on cross dataset problems. Finally, none of the mentioned approaches have been showed to be scalable to millions of images with features of thousands of dimensions.

In the present paper, our goal is hence to develop a scalable multi-task ML method, using linear embeddings for dimensionality reduction, able to leverage related tasks from heterogeneous datasets/sources of faces. Such challenging multi-task heterogeneous dataset setting, while being a very practical setting, has received almost negligible attention in the literature. Towards that goal, this paper presents a novel Coupled Projection multi-task Metric Learning method (CP-mtML) for learning better distance metrics for a main task by utilizing additional data from related auxiliary tasks. The method works with pairwise supervision of similar and dissimilar faces – in terms of different aspects e.g. identity, age and expression – and does not require exhaustive annotation with presence or absence of classes for all images. We pose the metric learning task as the one of learning coupled low dimensional projections, one for each task, where the final distance is given by the Euclidean distance in the respective projection spaces.

The projections are coupled with each other by enforcing them to be a combination of a common projection and a task specific one. The common projection is expected to capture the commonalities in the different tasks, while the task specific components are expected to specialize to the specificities of the corresponding tasks. The projections are jointly learned using, at the same time, training data from different datasets containing different tasks.

The proposed approach is experimentally validated with challenging publicly available datasets for facial analysis based on identity, age and expression. The task of semantic face retrieval is evaluated in a large scale setting, i.e. in the presence of order of millions of distractors, and compared with challenging baselines based on state-of-the-art unsupervised and supervised projection learning methods. The proposed model consistently improves over the baselines. The experimental section also provides qualitative results visually demonstrating the improvement of the method over the most challenging baselines.

2. Related Work

As said in the introduction, because of its key role in many problems, ML has received lot of attention in the literature. The reader can refer to [3, 18] for comprehensive surveys on ML approaches in general. Among the possible classes of distances, the Mahalanobis-like one is certainly the most widely studied [22, 29, 38, 39] and has been very successful in variety of face matching tasks [4, 11, 12, 31].

The various Mahalanobis-like methods differ in their objective functions which are themselves related to the type of constraints provided by the training data. The constraints can be given at class level (i.e. same-class vectors have to be close from one another after projection) [29], under the form of triplet constraints i.e. $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ with $\mathbf{x}_i$ relatively closer to $\mathbf{x}_j$ compared to $\mathbf{x}_k$ [38], or finally by pairwise constraints $(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$ such that $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar (dissimilar) if $y_{ij} = +1$ ($y_{ij} = -1$) [22, 31].

While the above mentioned works considered only a single task, multi-task ML has recently been shown to be advantageous, allowing to learn the metrics for several related tasks jointly [25, 40, 41]. Multi-task Large Margin Nearest Neighbor (mt-LMNN) [25], which is an extension of the (single task) LMNN method [38], was one of the earliest multi-task ML methods. Given T related tasks, mt-LMNN learns $T + 1$ Mahlanobis-like metrics parametrized by matrices $M_0, \{M_t\}_{t=1}^{T}$.

M_0 encodes the general information common to all tasks while M_t's encode the task specific information. Since a full rank matrix is learned, the method scales poorly with feature dimensions. Pre-processing with unsupervised compression techniques such as PCA is usually required, which potentially leads to loss of information beforehand. Similarly, Wang et al. [37] proposed a multi-feature multi-task learning approach inspired by mt-LMNN. In general, mt-LMNN suffers from overfitting. To overcome overfitting, Yang et al. [40] proposed a regularizer based on Bregman matrix divergence [8]. In contrast with these works, Yang et al. [41] proposed a different but related approach aiming at learning projection matrices $L_t \in \mathbb{R}^{d \times D}$ with $d \ll D$. They factorized these matrices as $L_t = R_t^\top L_0$, where L_0 is common transformation matrix for all the tasks and R_t are task specific matrices. Their method is an extension of the Large Margin Component Analysis (LMCA) [34]. It is important to note that LMCA requires k-nearest neighbors for every classes in their objective function, and hence does not allow to handle tasks in which only pairwise (dis-)similarity constraints are available. Furthermore, computing the k-nearest neighbors is computationally expensive.

In contrast to the works exploiting related tasks, Romera-Paredes et al. [28] proposed a multitask learning method which utilises a set of unrelated tasks, enforcing via constraints that these tasks must not share any common structure. Similarly, Du et al. [9] used age verification as an auxiliary task to select discriminative features for face verification. They use the auxiliary task to remove age sensitive features, with feature interaction encouraged via an orthogonal regularization. Other works such as [15, 20, 26] discourage the sharing of features between the unrelated set of tasks.

The application considered in this paper, i.e. face retrieval, requires encoding face images by visual descriptors.

This is another problem, widely addressed by the literature. Many different and successful face features have been proposed such as [14, 24, 30, 33]. In the present work, we use signatures based on (i) Local Binary Patterns (LBP) [24] which are very fast to compute and have had a lot of success in face and texture recognition, and (ii) Convolutional Neural Networks (CNN) [17] which have been shown to be very effective for face matching [32]. The computation of face signatures is usually done after cropping and normalizing the regions of the images corresponding to the faces. We do it by first locating face landmarks using the approach of Cao et al. [5].

3. Approach

As stated in the introduction, the proposed method aims at jointly learning Mahalanobis-like distances for T different but related tasks, using positive and negative pairs from the different tasks. The motivation is to exploit the relations between the tasks and potentially improve performance. In such a case, the distance metric between vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ can be written as

$$d^2_{M_t}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M_t (\mathbf{x}_i - \mathbf{x}_j) \tag{1}$$

where $M_t \in \mathbb{R}^{D\times D}$ is a task specific parameter matrix (in the following, subscript t denotes task t). To be a valid metric, M must be positive semi-definite and hence can be factorized as $M = L^\top L$. Following [22, 38] we decompose M as the square of a *low rank* matrix $L \in \mathbb{R}^{d\times D}$, with $\mathrm{rank}(L) \leq d \ll D$. This has the advantage that the distance metric can now be seen as a projection to a Euclidean space of dimension $d \ll D$ i.e.

$$d^2_{L_t}(\mathbf{x}_i, \mathbf{x}_j) = \|L_t\mathbf{x}_i - L_t\mathbf{x}_j\|^2, \tag{2}$$

thus resulting in a discriminative task-adaptive compression of the data. However, it has the drawback that the optimization problem becomes non-convex in L $\forall d < D$, even if it was convex in M [38]. Nonetheless, it has been observed that even if convergence to global maximum is not guaranteed anymore, the optimization of this cost function is usually not an issue and, in practice, very good results can be obtained [12, 22].

We consider an unconstrained setting with diverse but related tasks, coming from possibly different heterogenous datasets. Training data consists of sets of annotated positive and negative pairs from the different task related training sets, denoted as $\mathcal{T}_t = \{(\mathbf{x}_i, \mathbf{x}_j, y_{ij})\} \subset \mathbb{R}^D \times \mathbb{R}^D \times \{-1, +1\}$. In the case of face matching, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the face signatures while $y_{ij} = +1$ (-1) indicates that the faces are similar (dissimilar) for the considered task e.g. they are of the same person (for identity retrieval) or they are of the same age (for age retrieval) or they both are smiling (for expression retrieval).

Algorithm 1 SGD for proposed CP-mtML

Given: $\{\mathcal{T}_t | t = 1, \ldots, T\}, \eta_0, \eta$
Initialize: $b_t = 1$, $L_i \leftarrow \mathrm{wpca}(\mathcal{T}_i)$, $L_0 \leftarrow L_1$
for all $i = 0, \ldots,$`niters`-1 **do**
  for all $t = 0, \ldots, T-1$ **do**
    if $\mathrm{mod}(i, T) == t$ **then**
      Randomly sample $(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) \in \mathcal{T}_t$
      Compute $d^2_t(\mathbf{x}_i, \mathbf{x}_j)$ using Eq. 3
      if $y_{ij}(b_t - d^2_t(\mathbf{x}_i, \mathbf{x}_j)) < 1$ **then**
        $L_0 \leftarrow L_0 - \eta_0 y_{ij} L_0 (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$
        $L_t \leftarrow L_t - \eta y_{ij} L_t (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$
        $b_t \leftarrow b_t + 0.1 \times \eta y_{ij}$
      end if
    end if
  end for
end for

The main challenge here is to exploit the common information between the tasks e.g. learning for age matching might rely on some structure which is also beneficial for identity matching. Such structures may or may not exist, as not only the tasks but also the datasets themselves are different.

Towards this goal, we propose to couple the projections as follows: we define a generic global projection L_0 which is common for all the tasks, and, in addition, we introduce T additional task-specific projections $\{L_t | t = 1, \ldots, T\}$. The distance metric for task t is then given as

$$\begin{aligned} d^2_t(\mathbf{x}_i, \mathbf{x}_j) &= d^2_{L_0}(\mathbf{x}_i, \mathbf{x}_j) + d^2_{L_t}(\mathbf{x}_i, \mathbf{x}_j) \\ &= \|L_0\mathbf{x}_i - L_0\mathbf{x}_j\|^2 + \|L_t\mathbf{x}_i - L_t\mathbf{x}_j\|^2. \end{aligned} \tag{3}$$

With this definition of d_t we learn the projections $\{L_0, L_1, \ldots, L_t\}$ *jointly* for all the tasks.

Learning the parameters of our CP-mtML model, i.e. the projection matrices $\{L_0, L_1, \ldots, L_t\}$, is done by minimizing the total pairwise hinge loss given by:

$$\underset{L_0, \{L_t, b_t\}_{t=1}^T}{\mathrm{argmin}} \sum_{t=1}^{T} \sum_{\mathcal{T}_t} [1 - y_{ij}(b_t - d^2_t(\mathbf{x}_i, \mathbf{x}_j))]_+, \tag{4}$$

with $[a]_+ = \max(0, a)$, $b \in \mathbb{R}$ being the bias, for all training pairs from all tasks. We optimize this function jointly w.r.t. all the projections, ensuring information sharing between the different tasks.

In practice, stochastic gradient descent (SGD) is used for doing this optimization. In each iteration, we randomly pick a pair of images from a task, project them in (i) the common and (ii) the corresponding task specific spaces and then compute the square of the Euclidean distance between image descriptors in the respective sub-spaces. If the sum of distances violates the true (dis-)similarity constraint, we

update both matrices. To update the matrices, we use the closed-form expression of the partial derivatives of the distance function d_t w.r.t. L_0, L_t, given by

$$\frac{\partial d_t^2(\mathbf{x}_i, \mathbf{x}_j)}{\partial L_k} = L_k(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \forall k = 0, \ldots, T \quad (5)$$

Alg. 1 summarizes this learning procedure.

The learning rates of the different projections are set as explained in the following. Regarding the update of the common projection matrix, we can note that the update is done for every violating training example of every task, while other projection matrices are updated much less frequently. Based on this observation, the learning rate for task specific projection matrices is set to a common value denoted as η while the learning rate for the common projection matrix, denoted as η_0, is set as a fractional multiple of η i.e. $\eta_0 = \gamma\eta$, where, $\gamma \in [0, 1]$ is a hyper-parameter of the model. The biases b_t are task specific and are the thresholds on the distances separating positive and negative pairs.

Advantage over mt-LMNN [25]

The proposed distance function (Eq. 3) can be rearranged and written as $d_t^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top (L_0^\top L_0 + L_t^\top L_t)(\mathbf{x}_i - \mathbf{x}_j)$ and thus bears resemblance to the distances learned with mt-LMNN [25], where $d_t^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top (M_0 + M_t)(\mathbf{x}_i - \mathbf{x}_j)$. However, the proposed model as well as the learning procedure are significantly different from [25]. First, the objective function of mt-LMNN is based on triplets (while our is based on pairs) i.e. after projection a vector should be closer to another vector of the same class than to a vector of a different class. The learning procedure of mt-LMNN requires triplets which is in general more difficult to collect and annotate than pairs. Second, despite the fact that mt-LMNN leads to a semidefinite program which is convex, the proposed model has many practical advantages. Since a low rank projection is learnt, there is no need for an explicit regularization as limiting the rank acts as a regularizer. Another advantage is that the low dimensional projections lead to a discriminative task-adaptive compression, which helps us do very efficient retrieval. Third, the proposed SGD based learning algorithm is highly scalable and can work with tens of thousands of examples in thousands of dimensional spaces, without any compression/pre-processing of the features. Finally, another big advantage of our approach is that it can work in an online setting where the data streams with time.

4. Experimental Results

We now report the experiments we conducted to validate the proposed method for the task of face retrieval based on traits which can be inferred from faces, including identity, age and expressions. Such a task constitutes an important application domain of face based visual analysis methods. They find application in security and surveillance systems as well as searching large human centered image collections. In our experiments we focus on the two main tasks of identity and age based face retrieval. For the former, we use age and expressions prediction tasks as auxiliary tasks while for the later, we use identity prediction as the auxiliary task. We also evaluate identity based retrieval at a very large scale, by adding a million of distractor faces collected independently from the web.

We now give details of the datasets we used for the evaluation, followed by the features and implementation details and then discuss the results we obtain.

CASIA Web [42] dataset consists of 494,414 images with weak annotations for 10,575 identities. We use this dataset to train Convolutional Neural Network (CNN) for faces.

Labeled Faces in the Wild (LFW) [13] is a standard benchmark for faces, with more than 13,000 images and around 5,000 identities.

MORPH(II) [27] is a benchmark dataset for age estimation. It has around 55,000 images annotated with both age and identity. There are around 13,000 identities, with an average of 4 images per person, each at different ages. We use a subset of around 13,000 images for our experiment. We use this dataset for age matching across identities and hence randomly subsample it and select one image per identity.

FACES [10] is a dataset of facial expressions with 2052 images of 171 identities. Each identity has 6 different expressions (neutral, happy, angry, in fear, disgusted, and sad) with 2 images of each. Here again, we sample one image from each of the expression of every person, and carefully avoid identity based pairings.

SECULAR [4] is a dataset having one million face images extracted from Flickr. These are randomly crawled images and these images are not biased to any of the tasks or datasets. We use these images as distractors during retrieval.

4.1. Implementation details

All our experiments are done with grayscale images. The CNN model (details below) is trained with normalized images of CASIA dataset. We use Viola and Jones [36] face detector for other datasets. For detecting facial key points and aligning the faces, we use the publicly available implementation[1] of the facial keypoints detector of [5]. Faces are encoded using the following two features.

Local Binary Patterns (LBP). We use the publicly available `vlfeat` [35] to compute descriptors. We resized the aligned face images to 250×250 and centre cropped to

[1] https://github.com/soundsilence/FaceAlignment

170×100. We set cell size equal to 10 for a descriptor of dimension 9860.

Convolutional Neural Networks (CNN). We use model trained on CASIA dataset with the architecture of Krizhevsky et al. [17] to compute the feature of faces. Before computing the features, the images are normalized similar to CASIA. We use the publicly available Caffe [16] deep learning framework to train the model. The weights of the `fc7` layer are taken as the features (4096 dimensions) and are ℓ_2 normalized. As a reference, our features give a verification rate of 88.4 ± 1.4 on the LFW dataset with unsupervised training setting ($+10\%$ compared to Fisher Vectors (FV) [31]) and 92.9 ± 1.1 with supervised metric learning with heavy compression (4096 dimensions to 32 dimensions) cf. 91.4% for $16\times$ longer FVs.

4.2. Compared methods.

We compared with the following three challenging methods for discriminative compression, using the same features, same compressions and same experimental protocol for all methods for a fair comparison.

WPCA has been shown to be very competitive method for facial analysis – even comparable to many supervised methods [14]. We compute the Whitened PCA from randomly sampled subset of training examples from the main task.

Single Task Metric Learning (stML) learns a discriminative low dimensional projection for each of the task independently. In Alg. 1, we only have a global projection, with no tasks, i.e. $T = 0$, reducing it to single task metric learning which we use as a baseline. This is one of the state-of-art stML methods [31] for face verification.

Metric Learning with Union of Tasks (utML). We also learn a metric with the union of all tasks to verify that we need different metrics for different tasks instead of a global metric. We take all pairwise training data from all tasks and learn a single metric as in stML above.

mtLMNN. We did experiments with publicly available code of [25] but obtained results only slightly better than WPCA and hence do not report them.

4.3. Experimental Protocol

We report results on two semantic face retrieval tasks, (i) identity based face retrieval and (ii) age based face retrieval. We now give the details of the experimental protocol i.e. details of metric used, main experiments and how we create the training data for the tasks.

Performance measure. We report the 1-call@Kmetric averaged over all the queries. n-call@$K \in [0, 1]$ is an information retrieval metric [7] which is 1 when at least n of the top K results retrieved are relevant. With $n = 1$, this metric is relevant for evaluating real systems, e.g. in security and surveillance applications, where at least one of the top scoring K retrievals should be the person of interest, which can be further validated and used by an actual operator.

Identity based retrieval. We use the LFW as the main dataset for identity based retrieval experiments and MORPH (for age matching) and FACES (for expressions matching) as the auxiliary datasets. We use $10,000$ (positive and negative) training pairs from LFW, disjoint from the query images. For auxillary tasks, of expression and age matching, we randomly sample $40,000$ positive and negative pairs, each. This setting is used to demonstrate performance improvements, when the data available for auxiliary task is more than that for the main task. To compare our identity retrieval performance with existing state-of-art rank boosting metric learning [23], we randomly sampled $25,000$ positive and negative pairs (cf. $\sim 32,000$ by [23]) and take the same sets of constraints as before from auxiliary tasks.

Following Bhattarai et al. [4], we choose one random image from the identities which have more than five images, as query images and the rest as training images. This gives us 423 query images in total. We use these images to do Euclidean distance, in the projection space, based nearest neighbor retrieval from the rest of the images, one by one. The non-query images are used to make identity based positive and negative pairs for the main task. We use two auxiliary tasks, (i) age matching using MORPH and (ii) expressions matching using FACES.

Age based retrieval. We use the MORPH dataset as the main dataset and the LFW dataset as the auxiliary dataset. We randomly split the dataset into two disjoint parts as train+validation and test sets. In the test set, one image from each age class is taken as the probe query while the rest make the gallery set for retrieval. We take $10,000$ age pairs and $30,000$ of identity pairs.

Large scale retrieval with 1M distractors. We use the SECULAR dataset for distractors. We make the assumption that, as these faces are crawled from Flickr accounts of randomly selected common users, they do not have any identity present in LFW, which is a dataset of famous people. With this assumption, we can use these as distractors for the large scale identity based retrieval task and report performances with the annotations on the main dataset, since all of the distractors will be negatives. However, we can not make the same assumption about age and hence we do not use distractors for age retrieval experiments.

Parameter settings. We choose the values for the parameters (η, η_0, `niters`) by splitting the train set into two parts and training on one and validating on the other i.e. these sets were disjoint from all of the test sets used in the experiments.

		No distractors				1M distractors			
Method	Aux	$K=2$	5	10	20	$K=2$	5	10	20
WPCA	n/a	30.0	37.4	43.3	51.3	24.6	28.8	33.8	39.0
stML	n/a	38.1	51.1	60.5	69.3	26.0	37.4	43.3	48.7
utML	expr	31.0	38.1	48.5	57.9	20.3	25.8	31.9	38.5
CP-mtML	expr	**43.5**	**55.6**	**63.6**	**69.5**	**33.1**	**43.3**	**51.1**	**55.3**
utML	age	21.7	31.4	41.1	53.0	12.8	18.9	24.6	31.7
CP-mtML	age	**46.1**	**56.0**	**63.4**	68.3	**35.7**	**43.5**	**47.8**	**52.2**

Table 2. Identity based face retrieval performance (1-call@K for different K) with and without distractors with LBP features. Auxiliary task is either Age or Expression matching. Projection dimension, $d = 64$

		No distractors				1M distractors			
Method	Aux	$K=2$	5	10	20	$K=2$	5	10	20
WPCA	n/a	72.1	80.4	83.7	89.1	65.2	72.1	75.9	78.7
stML	n/a	**76.8**	85.1	89.6	92.0	70.7	78.0	82.0	84.2
utML	expr	73.5	82.3	87.2	90.3	67.1	76.8	79.0	82.0
CP-mtML	expr	**76.8**	**86.5**	**90.3**	**93.4**	**71.2**	**79.7**	**83.2**	**85.3**
utML	age	73.0	82.0	88.2	91.0	68.1	76.1	81.1	82.7
CP-mtML	age	**76.8**	**85.8**	**90.3**	**93.6**	**71.2**	**79.0**	**83.0**	**85.1**

Table 3. Identity based face retrieval performance (1-call@K for different K) with and without distractors with CNN features. Auxiliary task is either Age or Expression matching. Projection dimension, $d = 64$

Projection	$K=2$	5	10	20
L_0	30.3	38.1	43.3	51.8
L_1	35.0	46.6	55.8	64.8
L_2	4.5	7.6	10.4	13.0
$L_0 + L_1$	**43.5**	**55.6**	**63.6**	**69.5**

Table 1. Performance (1-call@K) of different projections matrices learned with proposed CP-mtML (LBP features, $d = 64$) for identity retrieval with auxiliary task of expression matching.

4.4. Quantitative Results

We now present the quantitative results of our experiments. We first evaluate the contributions of the different projections learnt, i.e. the common projection L_0 and the task specific projection L_t, in terms of performance on the main task. We then show the performance of the proposed CP-mtML w.r.t. the compared methods on the two experiments on (i) identity based and (ii) age based face retrieval. We mention the auxiliary task in brackets e.g. CP-mtML (expr) means that the auxiliary task was expression matching, with the main task being clear from context.

Contributions of projections. Tab. 1 gives the performance of the different projections for the task of identity based retrieval task with expression matching as the auxiliary task. We observe an expected trend; the combination of the common projection L_0 with the task specific one L_1 performs the best at 69.5 at $K = 20$. The projection for the auxiliary task L_2 expectedly does comparatively badly at 13.0, as it specializes on the auxiliary task and not on the main task. The projection L_1 specializing on the main task is better than the common projection L_0 (64.8 vs. 51.8) while their combination is the best (69.5). The trend was similar for the auxiliary task. This demonstrates that the projection learning follows the expected trend, the global projection captures commonalities and in combination with the task specific projections performs better for the respective tasks.

Identity based retrieval. We evaluate identity based face retrieval with two different features i.e. LBP and CNN, both with and without one million distractors. Tab. 2 and 3 give the performances of the different methods for different values of K (the number of top scoring images considered). First of all we notice the general trend that the performances are increasing with K, which is expected. We see that, both in the presence and absence of distractors, the proposed method performs consistently the best compared to all other methods. In the case of LBP features, the performance gains are slightly more when the auxiliary task is age prediction e.g. 46.1 for CP-mtML (age) vs. 43.5 for CP-mtML (expr) at $K = 2$, both these values are much better than WPCA and stML (30.0 and 38.1) respectively. Interestingly, when we take all the tasks together and learn only a single projection, i.e. utML, it degrades for both age and expression as auxiliary tasks, but more so for age (21.7 vs. 31). This happens because the utML projection brings similar age people closer and hence confuses identity more, as age is more likely to be shared compared to expressions which are characteristic of different people. The proposed CP-mtML is not only able to recover this loss but also leverages the extra information from the auxiliary task to improve performance of the main task.

When distractors are added the performances generally

Method	Aux	No distractor			1M distractors		
		$d = 32$	64	128	$d = 32$	64	128
WPCA	-	34.3	43.3	52.5	23.4	33.8	40.4
stML	-	50.1	60.5	63.6	33.3	43.3	51.3
utML	expr	44.2	48.5	57.4	25.3	31.9	31.9
CP-mtML	expr	**55.6**	**63.6**	**70.2**	**37.6**	**51.1**	**54.6**
utML	age	37.6	41.1	51.5	17.5	24.6	34.0
CP-mtML	age	**52.5**	**63.4**	**69.0**	**34.3**	**47.8**	**53.9**

No distractor			1M distractors		
$d = 32$	64	128	$d = 32$	64	128
83.9	83.7	85.6	74.5	75.9	75.2
88.4	89.6	88.7	80.6	82.0	**81.6**
85.1	87.2	86.3	73.0	79.0	78.3
88.7	**90.3**	**89.4**	**81.3**	**83.2**	81.1
85.3	88.2	86.5	76.6	81.1	79.2
88.2	**90.3**	**89.6**	**80.9**	**83.0**	**81.6**

Table 4. Identity based face retrieval, 1-call@10 at different projection dimension, d, (left) using LBP and (right) CNN features.

go down e.g. 68.3 to 52.2 for LBP and 93.6 to 85.1 for CNN with CP-mtML (age). However, even in the presence of distractors the performance of the proposed CP-mtML is better than all other methods, particularly stML e.g. 43.3 for CP-mtML (expr) vs. 37.4 for stML at $K = 5$ with LBP and 79.7 for CP-mtML (expr) vs. 78.0 for stML with CNN.

The performances of the two different features are quite different. The lightweight unsupervised LBP features perform lower than the more discriminative CNN features, which are trained on large amounts of extra data e.g. 86.5 vs. 55.6 at $K = 5$ for CP-mtML (expr). The performance gains for the proposed method are larger for LBP compared to CNN features e.g. $+4.5$ vs. $+1.4$ at $K = 5$ for CP-mtML (expr) cf. stML. While such improvements are modest for CNN features, they are consistent for all the cases. Parallely, the improvements for LBP features are substantial, especially in the presence of distractors e.g. $+7.8$ for CP-mtML (expr) vs. stML at $K = 10$. While it may seem that using stronger feature should then be preferred over using a stronger model, we note that this may not be always preferable. In a surveillance scenario, for instance, where a camera just records hours of videos and we need to find a specific face after some incident, using time efficient features as a first step for filtering and then using the stronger feature on a sufficiently small set of filtered examples is advantageous. This is highlighted by the time complexities of the features; in practice LBP are much faster than CNN to compute. While CNN features roughly take 450 milliseconds, the LBP features take only a few milliseconds on a 2.5 GHz processor.

Further, Tab. 4 presents the 1-call@10 while varying the projection dimension, which is directly proportional to the amount of compression. We observe that all methods gain performance when increasing the projection dimension, however, with diminishing returns. In the presence of one million distractors, CP-mtML (expr) improves by $+13.5$ when going from $d = 32$ to $d = 64$ and $+3.5$ when going from $d = 64$ to $d = 128$ for LBP. The results for larger d were saturated for LBPs with a slight increase. The performance changes with varying d in the presence of distractors for CNN features are more modest. CNN with distractors gets $+1.9$ for $d = 32$ to $d = 64$ and -2.1 for $d = 64$ to $d = 128$ i.e. the algorithm starts over-fitting at higher dimensions for the stronger CNN features. As an idea of space complexity, at compression to $d = 32$ dimensional single precision vector per face, storing ten million faces would require one gigabytes of space, after projection. Interestingly, the proposed method is better than stML in all but one case (CNN features with $d = 128$) which is a saturated case anyway.

Method	Aux	No distractors		1M distractors	
		K=10	20	10	20
MLBoost	n/a	54.1	63.4	34.3	39.5
CP-mtML	expr	**58.9**	**69.5**	**38.1**	**45.6**
CP-mtML	age	**61.5**	**70.7**	**39.7**	**47.8**

Table 5. Performance comparison with existing MLBoost [23] (for LBP features and $d = 32$).

Tab. 5 gives the comparisons (with LBP features and $d = 32$) with MLBoost [23]. At $K = 10$ CP-mtML obtains 61.5, 58.9 with age and expressions as auxiliary taks, respectively, while the MLBoost method stays at 54.1. Hence the proposed method is better than the results reported in the literature. As said before, we also used the publicly available code of mtLMNN [25]. We obtained results only slightly better than WPCA and hence do not report them.

With the above results we conclude the following. The proposed method effectively leverages the additional complementary information in the related tasks of age and expression matchings, for the task of identity based face retrieval. It consistently improves over the unsupervised WPCA, supervised stML which does not use additional tasks and also utML which combines all the data. It is also better than these methods at a range of projection dimensions (i.e. compressions), deteriorating only at the saturated case of high dimensions with strong CNN features.

Age based retrieval. Fig. 3 presents some results for face retrieval based on age for the different methods, with the auxiliary task being that of identity matching. In this task CP-mtML outperforms all other methods by a significant margin with LBP features. These results are different and interesting from the identity based retrieval experiments above, as they show the limitation of CNN features, learnt on identities, to generalize to other tasks — the performances with LBP features are higher than those with CNN features.

Figure 2. The 5 top scoring images (LBP & no distractors) for three queries for the different methods (auxiliary task in brackets). True (resp. False) Positive are marked with a green (resp. red) border (best viewed in color).

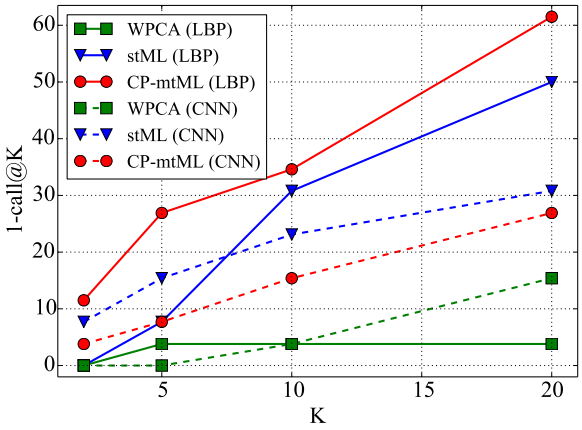

Figure 3. Age retrieval performance (1-call@K) for different K with auxiliary task of identity matching. The dimension of projection is $d = 32$

While the trend is similar for LBP features i.e. CP-mtML is better than stML, it is reversed for CNN features. With CNN features, stML learns to distinguish between ages when trained with such data, however, CP-mtML ends up being biased, due to its construction, towards identity matching and degrades age retrieval performance when auxiliary task is identity matching. However, the performance of CPmtML with LBP features is much higher than of any of the methods with CNN features.

4.5. Qualitative results

We now present some qualitative comparisons between the proposed CP-mtML, with age and expression matching as auxiliary tasks, with the competitive stML method. Fig. 2 shows the top five retrieved faces for three different queries for stML and the proposed CP-mtML with age and expression matching as auxiliary tasks. The results qualitatively demonstrate the better performance obtained by the proposed method. In the first query (left) all the methods were able to find correct matches in the top five. While stML found two correct matches at ranks 1 and 4, CP-mtML (age) also found two but with improved ranks i.e. 1 and 2 and CP-mtML (expression) found three correct matches with ranks $1, 2$ and 5. While the first query was a relatively simple query, i.e. frontal face, the other two queries are more challenging due to non-frontal pose and deformations due to expression. We see that stML completely fails in these cases (for $K = 5$) while the proposed CP-mtML is able to retrieve one correct image with ranks $1, 3$ (middle) and $5, 2$ (right) when used with age and expression matching as auxiliary tasks, respectively. It is interesting to note that with challenging pose and expression the appearances of the faces returned by the methods are quite different (right query) which demonstrates that CP-mtML projection differs from that learned by stML.

5. Conclusions

We presented a novel Coupled Projection multi-task Metric Learning (CP-mtML) method for leveraging information from related tasks in a metric learning framework. The method factorizes the information into different projections, one global projection shared by all tasks and T task specific projections, one for each task. We proposed a max-margin hinge loss minimization objective based on pairwise constraints between training data. To optimize the objective we use an efficient stochastic gradient based algorithm. We jointly learn all the projections in a holistic framework which leads to sharing of information between the tasks. We validated the proposed method on challenging tasks of identity and age based image retrieval with different auxiliary tasks, expression and age matching for the former and identity matching in the later. We showed that the method improves performance when compared to competitive existing approaches. We analysed the qualitative results, which also supported the improvements obtained by the method.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. 1

[2] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *IVC*, 32(4):270–286, 2014. 1

[3] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv:1306.6709*, 2013. 2

[4] B. Bhattarai, G. Sharma, F. Jurie, and P. Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *ECCV Workshops*, 2014. 2, 4, 5

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *Intl. Journal of Computer Vision*, 107(2):177–190, 2014. 3, 4

[6] R. Caruana. Multitask learning. *JMLR*, 1997. 1

[7] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *ACM SIGIR*, 2006. 5

[8] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007. 2

[9] L. Du and H. Ling. Cross-age face verification by coordinating with cross-face age verification. In *CVPR*, 2015. 2

[10] N. C. Ebner, M. Riediger, and U. Lindenberger. Faces–a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 2010. 4

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *CVPR*, 2009. 1, 2

[12] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *ICCV*, 2009. 1, 2, 3

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4

[14] S. U. Hussain, T. Napoléon, and F. Jurie. Face recognition using local quantized patterns. In *BMVC*, 2012. 3, 5

[15] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014. 2

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 5

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 5

[18] B. Kulis. Metric learning: A survey. *FTML*, 5(4):287–364, 2012. 2

[19] M. Lapin, B. Schiele, and M. Hein. Scalable multitask representation learning for scene classification. In *CVPR*, 2014. 1

[20] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *ECCV*, 2014. 2

[21] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013. 1

[22] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2, 3

[23] R. Negrel, A. Lechervy, and F. Jurie. Boosted metric learning for efficient identity-based face retrieval. In *BMVC*, 2015. 5, 7

[24] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002. 3

[25] S. Parameswaran and K. Q. Weinberger. Large margin multitask metric learning. In *NIPS*, 2010. 2, 4, 5, 7

[26] J. Pu, Y.-G. Jiang, J. Wang, and X. Xue. Which looks like which: Exploring inter-class relationships in fine-grained visual categorization. In *ECCV*, 2014. 2

[27] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FGR*, 2006. 4

[28] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *AISTATS*, 2012. 1, 2

[29] R. Salakhutdinov and G. E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *AISTATS*, 2007. 2

[30] G. Sharma, S. U. Hussain, and F. Jurie. Local higher-order statistics (LHS) for texture categorization and facial analysis. In *ECCV*, 2012. 3

[31] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013. 2, 5

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 3

[33] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *TIP*, 19(6):1635–1650, 2010. 3

[34] L. Torresani and K.-c. Lee. Large margin component analysis. In *NIPS*, 2007. 2

[35] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 4

[36] P. Viola and M. J. Jones. Robust real-time face detection. *Intl. Journal of Computer Vision*, 57(2):137–154, 2004. 4

[37] S. Wang, S. Jiang, Q. Huang, and Q. Tian. Multi-feature metric learning with knowledge transfer among semantics and social tagging. In *CVPR*, 2012. 2

[38] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009. 2, 3

[39] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002. 2

[40] P. Yang, K. Huang, and C.-L. Liu. Geometry preserving multi-task metric learning. *Machine learning*, 92(1):133–175, 2013. 2

[41] P. Yang, K. Huang, and C.-L. Liu. A multi-task framework for metric learning with common subspace. *Neural Computing and Applications*, 22(7-8):1337–1347, 2013. 2, 10

[42] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 4

[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 1

6. Additional Results

In this section we present additional both quantiative and qualitative results.

6.1. Quantitative Results

In this section, we compare performance of existing state-of-art multitask metric learning method, mtLMCA of Yang et al. [41] with the performance of the proposed method and other baselines. In addition to it, we present the in-depth analysis of the proposed algorithm such as it's time complexity and scalability. We then present the optimization curves of loss functions of our method and mtLMCA.

Comparisons with mtLMCA. We implemented the existing mtLMCA and compare the performance with the proposed method. For mtLMCA, we initialized the the common projection, L_0 and task specific, R_t matrices with identity matrices as explained in the paper. Whereas, for rest of the cases, as stated in the Alg. 1 with the WPCA.

Tab. 6 shows the performance comparison. In comparison with mtLMCA, we observe that the proposed CP-mtML outperforms mtLMCA by a significant margin. We explain it as follows. Without loss of generality consider task 1 (e.g. identity matching), the projection by proposed method is given by a common L_0 and a task specific L_1 while that by mtLMCA is given by common L_0 and task specific R_1 . While L_0 , L_1 are both $d \times D$ matrices R_1 is $d \times d$. Hence in CP-mtML there are dD common (across tasks) parameters and dD task specific parameters, while mtLMCA has same dD common parameters but only d^2 task specific parameters. We suspect that with equal number of task specific and common parameters CP-mtML is able to exploit the shared as well as task specific information well while for mtLMCA the small number of task specific parameters are not able to do so effectively e.g. for the specific case of 9860D LBP features projected to 64D, while 50% of the parameters are task specific for CP-mtML, only $64^2/(9860 \times 64) = 0.7\%$ are task specific in mtLMCA. In addition to it, we could see this method as utML with a very small fraction of task specific parameters. As mentioned before, utML learns a common projection matrix taking training examples from both the domains. From the performance also, it supports our argument. We can see that the performance of mtLMCA is slightly better than utML. This is due to the small separate task specific parameters in mtLMCA. Our proposed method, CP-mtML is capable of learning large task specific parameters maintaining the same projection dimension as that of other methods, which ultimately gives the improved performance.

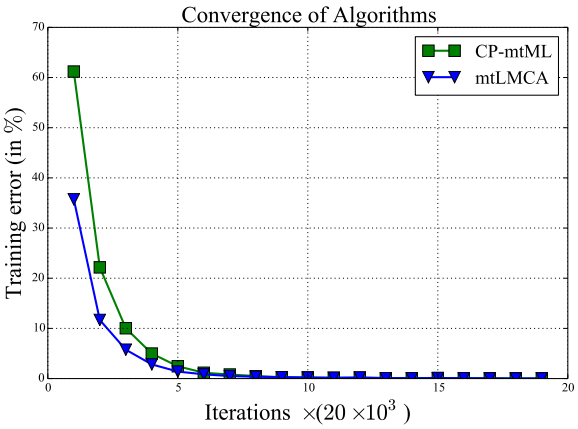

Figure 4. Optimization of loss functions

Time Complexity and Scalability. CP-mtML is about 2.5× slower to train than stML – specifically it takes 40 minutes to train CP-mtML with 50, 000 training pairs while compressing 9860D LBP features to 64D on a single core of 2.5 Ghz system running Linux. The training time is linear in the number of training examples. As the 64D features are real vectors it takes 256 bytes (with 4 bytes per real) to index one face or about a manageable 1.8 TB to index the current human population of about 7 billion people; hence we claim scalability.

Convergences of Algorithms. Fig. 4 shows the convergences of CP-mtML and mtLMCA. From the figure, we see that both the algorithms are converged well.

6.2. Qualitative Results

We present some more qualitative results to compare the proposed Coupled Projection multi-task Metric Learning (CP-mtML) with the most competitive baseline i.e. Single Task Metric Learning (stML). The main task here is that of identity based face retrieval while the auxiliary tasks are expression (expr) and age (age) based matching.

We can make the following observations

1. Fig. 5 shows some queries for which CP-mtML (age) does better than CP-mtML (expr) and stML. The results suggest that adding information based on age matching makes identity matching more robust to high variations due to challenging pose (left) and occlusions (hair and hand in the middle and right examples).

		No distractors				1M distractors			
Method	Aux	$K = 2$	5	10	20	$K = 2$	5	10	20
WPCA	n/a	30.0	37.4	43.3	51.3	24.6	28.8	33.8	39.0
stML	n/a	38.1	51.1	60.5	69.3	26.0	37.4	43.3	48.7
utML	expr	31.0	38.1	48.5	57.9	20.3	25.8	31.9	38.5
mtLMCA	expr	29.3	40.7	48.0	61.0	19.9	28.4	34.8	40.0
CP-mtML	expr	**43.5**	**55.6**	**63.6**	**69.5**	**33.1**	**43.3**	**51.1**	**55.3**
utML	age	21.7	31.4	41.1	53.0	12.8	18.9	24.6	31.7
mtLMCA	age	27.4	39.7	50.4	61.0	18.7	24.6	29.8	35.5
CP-mtML	age	**46.1**	**56.0**	**63.4**	68.3	**35.7**	**43.5**	**47.8**	**52.2**

Table 6. Identity based face retrieval performance (1-call@K for different K) with and without distractors with LBP features. Auxiliary task is either Age or Expression matching. Projection dimension, $d = 64$

2. Fig. 6 shows some queries for which CP-mtML (expr) does better than CP-mtML (age) and stML. The results suggest that adding information based on expression matching makes identity matching more robust to challenging expressions.

3. Fig. 7 shows some queries for which CP-mtML (expr) and CP-mtML (age) do better than stML. These cases are really challenging and the results retrieved by stML, while being sensible, are incorrect. Adding more information based on age and/or expression matching improves results.

4. Fig. 8 shows some queries for which all three methods do well. These are queries with either neutral expression and frontal pose or with characteristic appearances e.g. moustache, baseball cap, glasses, hairstyle etc. which occur for the same person in the gallery set as well.

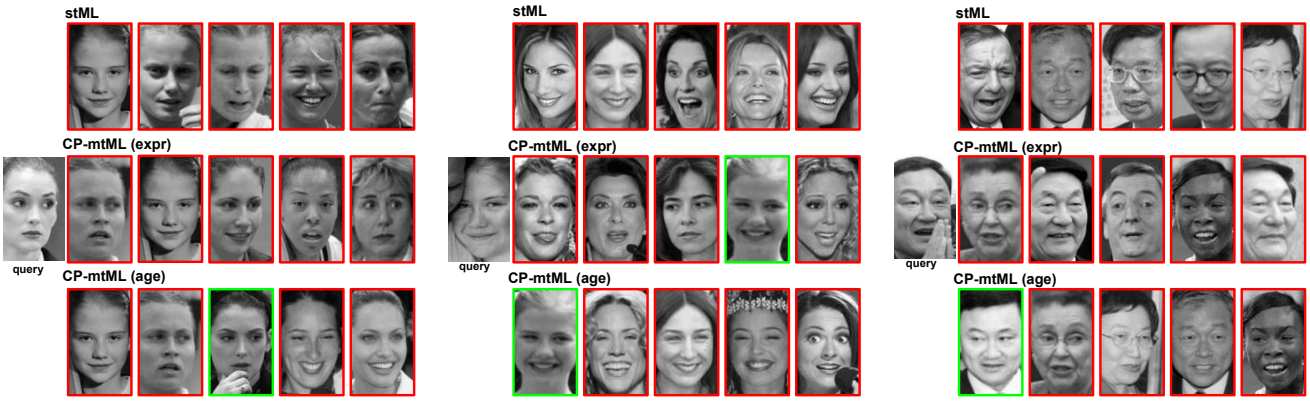

Figure 5. **Sample set of queries for which CP-mtML (age) performs better than CP-mtML (expr) and stML.** The 5 top scoring images (LBP & no distractors) for the queries for the different methods. True (resp. false) positives are marked with a green (resp. red) border. Best viewed in color.

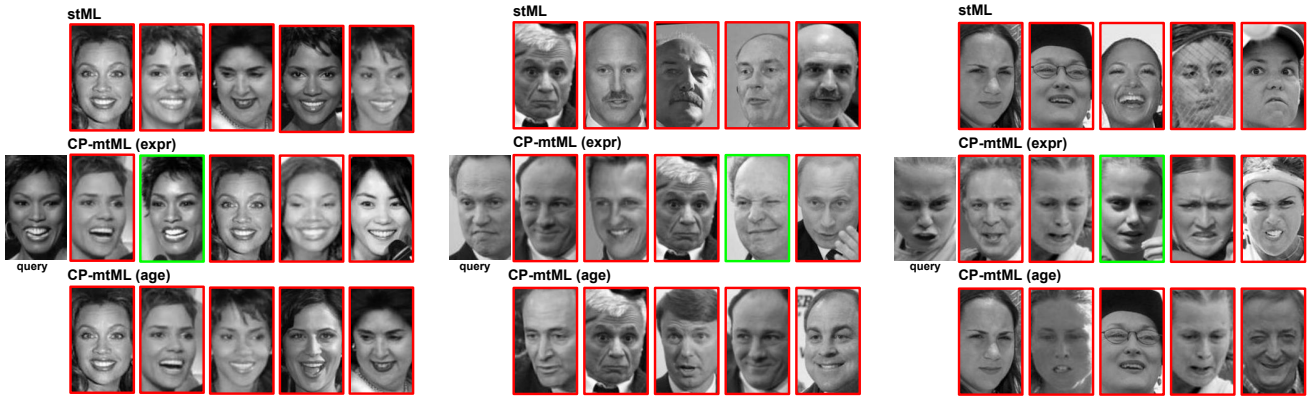

Figure 6. **Sample set of queries for which CP-mtML (expr) performs better than CP-mtML (age) and stML.** The 5 top scoring images (LBP & no distractors) for the queries for the different methods. True (resp. false) positives are marked with a green (resp. red) border. Best viewed in color.

Figure 7. **Sample set of queries for which CP-mtML (expr) and CP-mtML (age) both perform better than stML.** The 5 top scoring images (LBP & no distractors) for the queries for the different methods. True (resp. false) positives are marked with a green (resp. red) border. Best viewed in color.

Figure 8. **Sample set of queries for which all of CP-mtML (expr), CP-mtML (age) and stML perform well.** The 5 top scoring images (LBP & no distractors) for the queries for the different methods. True (resp. false) positives are marked with a green (resp. red) border. Best viewed in color.