Self-Regulated Facial Image Annotation by Discrimination of the Facial Matrices from Weakly Annotated Images

[1] Kavitha G L; [2] D V Pranathi Suhasini; [3] Seema B Nikam

[1] Information Science, Atria Institute of Technology
Bangalore - 560024, Karnataka, India

[2] Information Science, Atria Institute of Technology
Bangalore - 560024, Karnataka, India

[3] Information Science, Atria Institute of Technology
Bangalore - 560024, Karnataka, India

Abstract - **An image may contain several faces captioned with their corresponding names. It may so happen that a facial image may be wrongly annotated. The self regulated image face naming technique that we propose aims at labeling a face in the image accurately. This is a challenging task because of the very large appearance variation in the images, as well as the potential mismatch between images and their captions. We propose this efficient face naming technique which is self regulated and aims at correctly labeling a face in an image. We first propose a new method called Unsupervised Regularized Low-Rank Depiction (URLRD) which productively employs the wrongly named image information to determine a low-rank matrix which is obtained by recreation along with examining many subspace structures of the data. Certain circumstances befall where a face is recreated by using its own facial image or from other subject's facial images. From the recreation method used we deduce a discriminatory matrix. Besides this we also deploy the Large Margin Nearest Neighbor (LMNN) method for face labeling an image which further leads to yet another kernel matrix and is based on the Mahalanobis distances of the data. We can note that the two corresponding facial matrices can be combined in such a way as to enhance the quality of each other. The fused matrix is used as a new reiterative plan to deduce the names of each facial image. Extensive analysis demonstrates the effectiveness of our accession**.

Keywords **- Facial matrix, Unsupervised Regularized Low Rank Depiction (URLRD), Large Margin Nearest Neighbor (LMNN), Unsupervised Label Refinement (ULR).**

1. Introduction

The Internet is a big hand of today's success of the people. The extensive growth of Internet based photo sharing has led to a large collection of photo images to be wrongly annotated. A few methods were proposed in the literature for this image annotation problem.

We aim at self regulated image naming which stand on the uncertain affiliated captions.

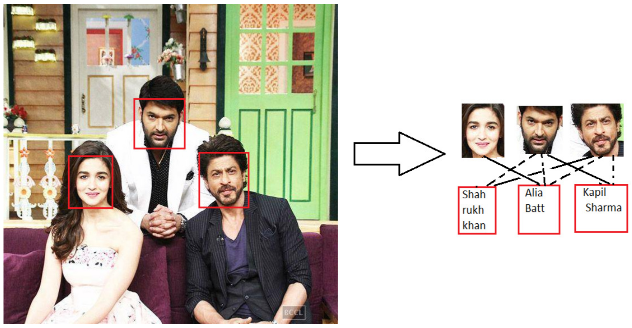


Fig. 1 Gives an inset to the face labeling problem based on images and corresponding labels. The solid lines represent the correctly named faces in the image and the dashed lines represent the weakly annotated faces.

A few initial steps used are the automated face detectors [1] and the labels are obtained using the label entity detector. The series of labels are expressed as the candidate label set. Despite these initial steps self-regulated face labeling is a challenging approach because of the very large appearance variation in the images, as well as the potential mismatch between images and their captions. Besides this, the candidate label set may sometimes be disturbed and incomplete and so a labeled image may not have the right labeled caption. Every face recognized would use only one label from the candidate label set or it may be set to null, indicating that the unidentified entity does not appear in the caption.

We introduce a new system of self-regulated face naming with label-based control. We obtain two corresponding facial matrices by determining the wrongly named images. These two matrices which are discriminated and merged into a single merged matrix based on which a reiterative plan is advanced for the self-regulated face naming.

We propose a new method called Unsupervised Regularized Low Rank Depiction to obtain the first facial matrix by consolidating wrongly labeled image information from the Unsupervised Label Refinement (ULR) method, so that the recreated matrix can be eventually obtained. To productively interpret the likeliness between the faces based on the visual appearance of the faces and the labels in the candidate label set, we accomplish the subspace structures [2] among faces based on the following inference that the faces of the same subject are present in the same subspace and the subspaces are linearly absolute.

Universal Label Refinement (ULR) [3] is devised to amplify the naming quality by using graph based and low-rank learning scheme. It is a scheme to refine the labels of the facial images by exploring machine learning techniques.

Introducing our proposed method, the URLRD is a new regularized approach which consolidates with the caption based weak supervision into the unbiased ULR in which we castigate the recreation of the faces using different subjects; based on the interpreted recreated matrix we can cipher the resemblance between each pair of faces.

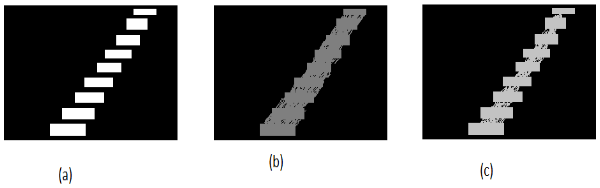


Fig. 2
(a) Original image W* according to the ground truth
(b) W* from ULR Algorithm
(c) W* from URLRD Algorithm (proposed system).

Furthermore, the kernel matrix is based on the Mahalanobis distances among the faces as another corresponding facial matrix.

Large Margin Nearest Neighbor (LMNN) [4] scheme uses the Mahalanobis distances consistently improving the kNN (k nearest neighbor) classification using Euclidean distances. LMNN classification works better with PCA Principle Component Analysis than Linear Discriminant Analysis when some form of dimensionality reduction is required for preprocessing.

In consideration of URLRD and LMNN we analyze the weak supervision in distinct and productive way. The two corresponding facial matrices are combined to obtain a merged facial matrix that is utilized for face labeling.
Summary:

1) We propose a new scheme URLRD by introducing a regularizer to ULR by which we articulate the first facial matrix using the resultant recreated matrix.
2) Introducing the LMNN which effectively refines the indistinct labels of the facial image. The kernel matrix is based on the Mahalanobis distances between all faces and is used as the second facial matrix.
3) Merging the two facial matrices obtained from URLRD and LMNN, we introduce an efficient scheme to name the facial images.
4) Extensive experimentation on fabricated information set and real world datasets exhibit the effectiveness of our scheme.

2. Related Work

Automatic face labeling is one of the major area of interests these days. Most of the research are focused on developing techniques for automatic image naming. Berg et al. [5] presented face clumping method to annotate the faces in news pictures. M Guillaumin [6] introduced the multiple-instance metric learning from automatically labeled bags of faces (MildML). Ozkan and Duygulu [7] developed a graph-based method by constructing the similarity graph of face. Zeng et al. [8] developed the low-rank SVM (LR-SVM) method which makes use of an assumption that the feature matrix of faces from the same subject is low rank. Luo and Orabona [9] developed learning from candidate labeling sets method for face naming.

Following is the comparison between our proposed method and existing systems:

Our proposed method URLRD is recounted to LR-SVM [8] and ULR [3]. In case of LR-SVM approach, LR-SVM considers distant supervision data in the permutation matrices, whereas URLRD utilizes regularizer that we have proposed, to deal with the recreation coefficients. In LR-SVM, data is not recreated by using itself as the base. In case of URLRD, it is related to the recreation-based approach of ULR. ULR is an unsupervised method that evaluates multiple subspace structures of data. Whereas, URLRD considers the image-level constraints to solve the face labeling problem in images.

Large-margin nearest neighbors (LMNN), is a traditional metric learning system. LMNN is constructed on appropriate supervision without any uncertainty. LMNN utilizes the hinge loss function. LMNN was proposed to learn distance metric M that supports the squared Mahalanobis distance between each training sample and its target neighbors to be smaller than those between this training sample and samples from other classes. The LMNN algorithm is built on the remark that the kNN will correctly classify an example if its k-nearest neighbors share the same label. The algorithm attempts to increase the number of training examples with this property by learning a linear transformation of the input space that precedes kNN classification using Euclidean distances
LMNN learns a distance metric that can be used to produce a facial matrix and can be fused with the facial matrix obtained from URLRD approach for the betterment of image labeling performance.

In the existing systems, such as MIL and MIML, data objects are represented as bags of instances. The distance between the data objects (bags) is a set-to-set distance. MIL makes use of class-to-bag distance, which assesses the relationships between the classes and the bags. The face labeling problem is solved by applying MIL and MIML method, in which each image is treated as a bag, faces in the image as the instances and names in the candidate name set as bag labels.
In some cases, the bag labels may be incorrect due to absence of names in the caption to which a face corresponds.

3. Discrimination of facial matrices for self-regulated image annotation

In this paper, we bring forward a new approach for self-regulated facial image annotation using a collection of images with captions. This is challenging because of the inherent mismatch between various facial images and their captions. We learn two facial matrices by making use of the equivocal labels, to perform image annotation based on the facial matrix obtained by fusing the two facial matrices. Further in the paper, we brief our new approach called Unsupervised Regularized Low-Rank Depiction (URLRD). The facial matrix obtained from this method is fused with the facial matrix obtained from the LMNN [4] method.

I_n is defined as the n×n similarity matrix, and 0_n , $1_n \in R_n$ as the n×1 column vectors of all zeros and ones, in the corresponding order. Also, we use I, 0 and 1 instead of I_n, 0_n, and 1_n in the case where the magnitudes are evident. tr(A) represents the trace of A and <A, B> means the dot product of two matrices. $A \circ B$ represents the element-wise multiplication of two matrices A and B ($a \circ b$ in case of vectors a and b). $\|A\|_\infty$ denotes the greatest absolute value of all the elements contained in matrix A. $\|A\|_F = (\sum_{i,j} A^2{}_{i,j})^{1/2}$ represents the Frobenious norm of the matrix A. $a \leq b$ implies that $a_i \leq b_i \; \forall \; i = 1,...,n$. $A \geq 0$ denotes that A is a positive semidefinite matrix (PSD matrix).

3.1 Problem Statement

Having a set of images, an image may contain several faces captioned with multiple names. It may so happen that a facial image might be wrongly annotated. This can happen due to the variation in the images and mismatch between the images and their captions. In this paper, we present methods for face naming using collection of images with captions. This is carried out in two steps: First, we retrieve all faces of a particular person from the data set. Second, establish the correct association between the names in the captions and faces in the image.

Let us assume that we have m images, each of which consists of r_i names and n_i faces, $\forall \; i = 1.....m$. Let $q \in \{1,...,p\}$ denote a name and $x \in R^d$ denote a face, where p is the total number of names in all the captions and *d* is the feature dimension. Thereafter, each image can be represented as (X^i , N^i), where $X^i = [x^i{}_1,...,x^i{}_{ni}] \in R^{d\times ni}$ is the data matrix for faces, that are in the *i*th image with each X^i_f being the *f*th face in the image (f = 1,...,N_i), and N^i={ q^i_1,..., $q^i_{r_i}$} is the corresponding set of candidate names with each $q^i_j \in$ { 1,...,p} being the *j*th name (j = 1,...,r_i). Further, let X =[X^1.....X^m] $\in \mathbb{R}^{d\times n}$ represent the data matrix of the faces from all m images, where n = $\sum_{i=0}^{m} n_i$.
After defining a binary label matrix Y =[Y^1,...,Y^m] $\in \{0,1\}^{(p+1)}$ with each $Y^i \in \{0,1\}^{(p+1)\times n_i}$ being the label matrix for each image X^i, the next step is to infer the facial label matrix Y based on the candidate name sets $\{N^i|_{i=1}^m\}$. When the ground-truth name of a face does not appear in the associated candidate name set N^i, we make

use of the (p+1)th name to denote null class, so that the face can be assigned to the (p+1)th name. The label matrix Y^i for each image should satisfy the following image-level constraints [8].

1) Distinctiveness: In the same image, two faces cannot be annotated with the same name except the (p+1)th name, i.e., $\sum_{f=1}^{n_i} y_{j,f}^i \leq 1, \forall j \ = 1, \ldots, p$.

2) Expediency: the faces in the *i*th image should be tagged using the names from the set:
$\tilde{N}^i = N^i \cup \{(p+1)\}$, i.e., $Y_{j,f}^i = 0, \forall f = 1, \ldots, and$ j $\notin \tilde{N}^i$.

3) Non-Pleonastic: In the ith image, each face should be tagged exactly one name from the set $\tilde{N}^i$, i.e., $\sum_j Y_{j,f}^i = 1, \forall f \ = 1, \ldots, n_i$.

3.2 Face Naming Using Facial Matrix

The feasible set of **Y**$^{\mathbf{i}}$ for the *i*th image, based on image-level constraints can be defined as follows:

$$y^i = \left\{ Y^i \in \{0,1\}^{(p+1) \times n_i} \;\middle|\; \begin{array}{c} 1'_{(p+1)}(Y^i \circ T^i)1_{n_i} = 0, \\ 1'_{(p+1)} Y^i = 1'_{n_i} \\ Y^i 1_{n_i} \leq \left[1'_p, n_i\right]' \end{array} \right\} \quad (1)$$

The matrix $T^i \in \{0,1\}^{(p+1) \times n_i}$ has rows related to the indices of the names in $\tilde{N}^i$ are all zeros and rest of rows are all ones.
The feasible set for the label matrix can be represented as y $= \{Y = [Y^1, \ldots, Y^m] \mid Y^i \in y^i \ \forall_i = 1, \ldots, m \ \}$.
Let A $\in \mathbb{R}^{n \times n}$ be a facial matrix, which meets the condition A=A' and $A_{i,j} \geq 0, \forall i, j$. Each $A_{i,j}$ expresses the pair-wise similarity between the *i*th face and the *j*th face. Our goal is to learn a proper A such that $A_{i,j}$ is large if and only if the *i*th face and the *j*th face share the same ground-truth name. Then, the face naming problem can be solved based on the facial matrix A obtained. We solve the following, to annotate the faces in an image:

$$\max_{Y \in y} \sum_{c=1}^{p} \frac{y_c' A y_c}{1' y_c} \quad s.t \ \ Y = \left[y_1, y_2, \ldots, y_{(p+1)}\right]' \qquad (2)$$

$y_c \in \{0,1\}^n$ correlates to the *c*th row in Y. The faces with the same label are clustered as one group, and the sum of the average similarities for each group is maximized.
We propose URLRD method to learn the Unsupervised Regularized Low-Rank recreation matrix. We obtain our first facial matrix from our URLRD method. Also, we make use of LMNN method to obtain another facial matrix. Finally, we fuse these two facial matrices into one single facial matrix in order to perform image tagging.

3.3 Learning Discrimination of Facial Matrix with Unsupervised Regularized-Low-Rank Dipiction (URLRD)

We will first analyse ULR (unsupervised label refinement) and then present our proposed method which is URLRD.

3.3.1 Description of ULR

ULR was proposed to enhance the face labelling quality via a graph-based and low-rank learning (LRR) approach. ULR makes use of content-based image search face annotation, face annotation performance on database. LRR is designed to solve the subspace clustering problem. The goal of LRR is to evaluate the structure of subspace in the given data X =[$x_1, \ldots, x_n$]$\in \mathrm{R}^{d \times n}$. LRR attempts to obtain a recreation matrix W, which is based on an assumption that the subspaces have linearly independent vectors . This recreation matrix W is given by W = [$w_1, \ldots, w_n$]$\in \mathrm{R}^{n \times n}$, where each w_i denotes the representation of x_i using X as the base. Because X is used as the base to recreate itself, the ideal solution W∗ of LRR encodes the pair-wise resemblance between the data matrices. The efficiency problem of LRR is given as:

$$\min_{W,E} ||W||_* \ + \ \lambda \ \|E\|_{2,1} \qquad \text{s.t} \quad X = XW+E$$
(3)

where E $\in \mathrm{R}^{d \times n}$ is the recreation error, $\lambda > 0$ is a tradeoff parameter, $\|W\|_*$ which is the nuclear form, is used to replace rank(W) as commonly used in the rank minimization problems, and $\|E\|_{2,1} = \sum_{j=1}^{n} \ (\sum_{i=1}^{d} \ (E_{i,j})_2)^{1/2}$ is a regularizer that supports the recreation error E to be column-wise sparse. LRR performs better than sparse subspace clustering method, and hence produces better results in most of the real world applications that includes Faceprints.

Graph based method is proposed to determine the most relevant subset among the set of possible faces related to the query name, where the most relevant subset is likely to match with the faces of the queried person. Graph based method is implemented to rectify the correct faces of a queried person using both text and visual appearances. This approach eliminates the wrong tags, by applying geometrical constraint. The geometrical distance corresponding to the *i*th assignment refers to
$\sqrt{X^2 + Y^2}$ where,

$$X = \frac{\mathrm{locX(i)}}{\mathrm{sizeY(image1)}} - \frac{\mathrm{locX(match(i))}}{\mathrm{sizeX(image2)}}$$

$$Y = \frac{\mathrm{locY(i)}}{\mathrm{sizeY(image1)}} - \frac{\mathrm{locY(match(i))}}{\mathrm{sizeY(image2)}} \qquad (4)$$

And locX is the X coordinate and locY is Y coordinate of the feature points in the images, sizeX and sizeY hold X and Y sizes of the images and match(i) corresponds to the matched keypoint in the second image of the ith feature point in the first image.

Unsupervised Label Refinement (ULR) task is to learn a refined label matrix $F^{*} \in R^{n\times m}$ to improve the initial raw label matrix Y . ULR makes use of an assumption called "label smoothness". i.e., the more similar the visual contents of two facial images, the more likely they share the same labels. The label smoothness principle is formulated as an idealization problem of reducing the following loss function $E_s(F,W)$:

$$E_s(F,W) = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j} \| F_{i*} - F_{j*}\|^2_F = tr(F^{T}LF) \qquad (5)$$

Where W is a weight matrix of a sparse graph, $\| \cdot \|_F$ denotes the Frobenius norm, L = D−W denotes the Laplacian matrix where D is a diagonal matrix with diagonal elements as $D_{ii} = \sum_{j=1}^{n} W_{i,j}$ and tr denotes a trace function.

In the method we introduce, although the names from captions are equivocal and corrupted, they still provide us with the weak supervision that is useful for improving the performance of face naming in terms of computation time and accuracy.

Motivated by this, we implement a new term $\|W \circ H\|^2_F$, which is called regularization term that includes the weak supervised information. Definition of $H \in \{0,1\}^{n\times n}$ depends on the candidate name sets $\{N^{i}|^{m}_{i=1}\}$. $H_{i,j} = 0$ if the following two conditions satisfy:

1) the ith face and the jth face has at least one name in common, in the corelated candidate name sets and

2) i = j. If not, $H_{i,j} = 1$.

And so forth, non-zero entries in W, where the corelated pair of faces have no names in common in their candidate name sets, and the entries that corelate to the situations where a face is recreated by itself, are penalized. Therefore, the resultant facial matrix W is expected to be more distinguishable, with information related to weak supervision encoded in H.

By implementing the new regularizer $\|W\circ H\|^2_F$ into ULR, the new optimization problem is achieved as follows:

$$\min_{W,E,} \|W\|_{*}+\lambda \|E\|_{2,1} + \frac{\gamma}{2} \|W\circ H\|^2_F \quad \text{s.t. } X = XW+E$$
(5)

where γ ≥ 0 is a used to balance the new regularizer with the other term. This problem is referred to as URLRD. By setting the parameter γ to zero, the URLRD problem in Eq(5) can be reduced to the ULR problem .

Once we obtain the ideal solution W* after solving Eq(6), the facial matrix A_W can be computed as $A_W = \frac{1}{2}(W^{*}+W^{*'})$.

3) Optimization: To obtain equivalent optimization problem , an intermediate variable J is introduced in Eq(6):

$$\min_{W,E,J} \|J\|_{*} + \lambda \|E\|_{2,1} + \frac{\gamma}{2} \|W\circ H\|^2_F \quad \text{s.t. } X = XW+E, W = J.$$
(6)

We Consider the following augmented Lagrangian function from Augmented Lagrangian Method (AML):

$$L= \| J \|_{*} + \lambda \|E\|_{2,1} + \frac{\gamma}{2} \|W\circ H\|^2_F + <U, X-XW-E > + <V,W-J> + \frac{\rho}{2} (\| X - XW - E \|^2_F + \|W-J\|^2_F \qquad (7)$$

where ρ is a positive penalty parameter and $U \in R^{d\times n}$ and $V \in R^{n\times n}$ are the Lagrange multipliers. Notably, we set the following parameters as follows:

$E_0 = X - XW_0$, $W_0 = (1/n)(1_n 1'_n - H)$, $J_0 = W_0$ and U_0, V_0 as zero matrices. The following steps are performed recursively at the *t*th iteration, until convergence is achieved.

1) Fix the others and update J_{t+1} by

$$\min_{J_{t+1}} \|J_{t+1}\|_{*} + \frac{\rho_t}{2} \left\| J_{t+1} - \left(W_t + \frac{V_t}{\rho_t} \right) \right\|^2_F$$

which can be solved in closed form using the singular value thresholding method.

2) Fix the others and update W_{t+1} by

$$\min_{W_{t+1}} \frac{\gamma}{2} \|W_{t+1} \circ H\|^2_F + (U_t, X - XW_{t+1} - E_t) + (V_t, W_{t+1} - J_{t+1}) + \frac{\rho_t}{2} \|X - XW_{t+1} - E_t\|^2_F$$

$$+\frac{\rho_t}{2}\|W_{t+1}-J_{t+1}\|_F^2$$
(8)

Due to the new regularizer $||W\circ H||_F^2$ this problem cannot be solved as in [2] by using pre-computed SVD. We use the gradient descent method to efficiently solve (7), where the gradient

with respect to W_{t+1} is

$$(H\circ H)\circ W_{t+1}+\rho_t(X'X+I)W_{t+1}+V_t-\rho_t J_{t+1}-X'(\rho_t(X-E_t)+U_t)$$

3) Fix the others and update E_{t+1} by

$$\min_{E_{t+1}}\frac{\lambda}{\rho_t}\|E_{t+1}\|_{2,1}+\frac{1}{2}\left\|E_{t+1}-\left(X-XW_{t+1}+\frac{U_t}{\rho_t}\right)\right\|_F^2$$

4) Update U_{t+1} and V_{t+1} by respectively using

$$U_{T+1}=U_t+\rho_t(X-XW_{t+1}-E_{t+1})$$
$$V_{t+1}=V_t+\rho_t(W_{t+1}-J_{t+1}).$$

5) Update ρ_{t+1}using
$\rho_{t+1}=min(\rho_t(1+\Delta\rho),\rho_{max})$ where Δ_ρ and ρ_{max} are the constant parameters.
6) The iterative algorithm stops if the two convergence conditions are both satisfied
$\|X-XW_{t+1}-E_{t+1}\|_\infty\leq\epsilon$
$\|W_{t+1}-J_{t+1}\|_\infty\leq\epsilon$
where ϵ is a constant parameter.

3.4 Large Margin Nearest Neighbor Classification (LMNN)

Weinberger and Saul [4] proposed the LMNN method to learn a distance metric M that promotes the squared Mahalonobis distances between each training sample and its target neighbours to be smallers than the distance between this training sample and samples from other classes. In LMNN, the metric is trained with the goal that the k-nearest neighbors always belong to the same class and the examples from various classes are separated by a large margin. The algorithm is based on an observation that an example will be classified correctly by KNN decision rule, if its K-nearest neighbors share the same label. Large Margin Nearest Neighbor (LMNN) metric learning algorithm has been used widely in many applications and has produced promising results.

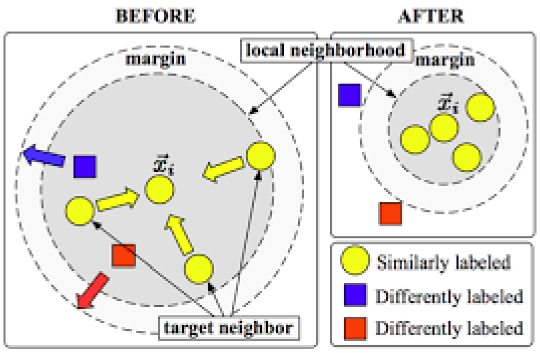


Fig. 3 Before(left) - Simplified representation of an input's neighborhood before training. After (Right)- Simplified representation of an input's neighbourhood after training.

LMNN optimizes matrix M with the help of semidefinite programming. The objective is twofold: For every data point $\vec{x}_i$, the target neighbours should be close and imposters (differntly labelled) should be far away. The learned metric causes the input vector $\vec{x}_i$ to be surrounded by training instances of the same class. This optimization is illustrated in figure 3.

Let $\{(x_i, y_i)|^n{}_{i=1}\}$ be the n labeled samples: $x_i \in R^d$ denotes the i^{th} sample, with d being the feature dimension, and $y_i \in \{1,...,z\}$ denotes the label of this sample, with z being the total number of classes. $\eta_{i,j} \in \{0,1\}$ indicates whether x_j is a target neighbor of x_i. i.e, $\eta_{i,j} = 1$ if x_j is a target neighbour of x_i, and $\eta_{i,j} = 0$ if x_j is a target neighbor of x_i ,$\forall i,j \in\{1..n\}$. $\nu_{i,\ell} \in\{0,1\}$ indicates whether x_ℓ and x_i are from different classes. i.e, $\nu_{i,\ell} = 1$ if $y_\ell \neq \mathbf{y_i}$, and $\nu_{i,\ell} = 0$ if $y_\ell = \mathbf{y_i}$, $\forall$ i,l $\in\{1,...,n\}$. The squared Mahalanobis distance between t-wo samples x_i and x_j can be defined as:

$$d^2{}_M(x_i,x_j)=(x_i-x_j)'M(x_i-x_j).$$

LMNN minimizes the following idealization problem:

$$\min_{M\geq 0}\sum\nolimits_{(i,j):\eta ij=1} d^2{}_M(x_i,x_j)+\mu\sum\nolimits_{(i,j,l)\in S}\xi_{i,j,l}$$
$$\text{s.t } d^2{}_M(x_i,x_l)-d^2{}_M(x_i,x_j)\geq 1-\xi_{i,j,l}\ ,\ \forall(i,j,l)\in S,\ \xi_{i,j,l}\geq 0\ ,\ \forall(i,j,l)\in S \quad (9)$$

where $\xi_{i,j,l}$ is a slack variable, μ is a tradeoff parameter and $S=\{(i, j,l)|\eta_{i,j}=1,\ \nu_{i,l}=1,\ \forall i, j,l \in\{1,...,n\}\}$. Therefore, $d^2{}_M(x_i,x_j)$ is the squared Mahalanobis distance between x_i and its target neighbor x_j, and $d^2{}_M(x_i,x_l)$ is the squared Mahalanobis distance between x_i and x_j that belong to different classes. The slack variable can condone the cases when $d^2{}_M(x_i,x_l)-d^2{}_M(x_i,x_j)$ is smaller than one. The LMNN problem in Eq. (9) can be equivalently reformulated as the idealization problem as follows:

$\min_{M\geq 0} \sum_{(i,j):\eta i,j=1} d^2_M(x_i,x_j) + \mu \sum_{(i,j,l)\in S} |1 - d^2_M(x_i,x_\ell) + d^2_M(x_i,x_j)|_+$

Where $|\cdot|_+$ is the truncation function.

Algorithm 1 summarizes the entire learning process.

Algorithm 1: LMNN

```
Input: Data samples ${x_i,y_i}^N_{i=1}$,
number of target neighbors K,
output dimension m,
maximum number of optimization iterations T.
Result: matrix $L \in R^{dxm}$
Initialize L with the first m leading eigen vectors
of the covariance matrix of the data samples ${x_i}^N_{i=1}$;
For t=1 to T do
Randomly generate subsamples S;
Calculate the descending direction d;
Use line search algorithm to find the step length $\lambda$;
Update $L \leftarrow L + \lambda d$;
if the termination condition satisfies then
break;
```

4. Performing Face Annotation

The first facial matrix A_w can be calculated as, $A_W = \frac{1}{2}(W^* + W^{*\prime})$, using coefficient matrix W^* learned from URLRD, and regularize A_W to the range [0,1]. The second facial matrix can be calculated from learnt distance metric M of LMNN as $A_K = K$, where K is a kernel matrix depending upon the Mahalanobis distance. These two facial matrices use weak supervision information in different ways. Therefore, the two facial matrices contain interdependent information which is beneficial for face annotation. We combine the two facial matrices obtained from our URLRD and LMNN to attain better accuracy, and we call this fused facial matrix as URLRD, which is our proposed method. This fused facial matrix A is the linear combination of the two facial matrices derived from URLRD and LMNN, where A is given by, $A=(1-\alpha)A_W + \alpha A_K$, where α is a parameter in the range [0, 1]. Lastly, the image face naming tagging is carried out based on A. We work on image face annotation by solving the following idealization problem:

$\max_{Y\in y} \sum_{c=1}^{p} \frac{Y'c\,AYc}{1'Yc}$ S.T, Y = [y_1……….$y_{(p+1)}$]’. (10)

But, the above problem is computationally expensive to solve. To solve this problem, we propose an iterative method. At each iteration, an objective function is approximated using $\hat{y}'_c$ Ay_c /1’$\hat{y}_c$ that can substitute $y'_c Ay_c/1'y_c$, where $\hat{y}_c$ is the solution for y_c inferred from the previous iteration. Therefore, we solve the linear programming problem at each iteration, as follows:

$\max_{y\in Y} \sum_{c=1}^{p}$ $b'_c y_c$, s.t. Y = [$y_1,...,y_{(p+1)}$]′ (11)

where $b_c = A\hat{y}_c / 1'\hat{y}_c, \forall c = 1,...,p$. In some cases, the candidate name set may be incomplete, due to which some of the faces may not annotated with their correct name. Hence, in addition a vector $b_{p+1} = \theta_1$ is defined, which allows some of the faces to be assigned to a null class, where the predefined parameter is 0.

The problem in Eq. (11) can be reformed by defining $B \in R^{(p+1)\times n}$ as B=[$b_1,...,b_{p+1}$]. The reformulated form is as follows:

$\max_{y\in Y}$ <B,Y>

(12)

The viable set for Y is defined as $Y = \{Y = [Y^1,...,Y^m] | Y^i \in Y^i, \forall i = 1,...,m\}$. Matrix B can be expressed as B=[$B^1,...,B^m$] , where each $B^i \in R^{(p+1)\times ni}$ correlates to Y^i. Then, the objective function in Eq. (12) can be conveyed as <B,Y> = $\sum_{i=1}^{m}$ <B^i,Y^i >. Therefore, Eq. (12) can be optimized by solving m sub-problems, with each sub-problem related to one image in the following form:

$\max_{yi\in Yi}$ <B^i,Y^i> $\forall i = 1,...,m$

(13)

The ith problem in Eq. (13) can be reformulated as a minimization problem as follows:

$$\min_{Y^i_{q,f}\in\{0,1\}} \sum_{q\in N^i} \sum_{f=1}^{n_i} -B^i_{q,f} Y^i_{q,f}$$

S.T $\sum_{q\in N^i} Y^i_{q,f} = 1 \quad \forall f = 1,...,n_i$

$$\sum_{f=1}^{n_i} Y^i_{q,f} \leq 1 \quad \forall q \in N^i$$

$\sum_{f=1}^{n_i} Y^i_{(p+1),f} \leq n_i$

(14)

in which we have left out the elements $\{Y^i_{q,f}|_{q\notin N^i}\}$, because these elements are zeros according to the feasibility constraint in Eq. (1). In this paper, we adopt the Hungarian algorithm to efficiently solve the problem in (14). Certainly, for an i^{th} image, the cost c(f, p +1) for assigning a face X^i_f to the corresponding null name is set to $-B^i_{(p+1),f}$ and the cost c(f , q) for assigning a face X^i_f to a real name q is set to $-B^i_{q,f}$.

Synopsis: We iteratively solve the linear programming problem in Eq. (12), to deduce the facial label matrix for all faces in an image. The face tagging can be performed effectively with the Hungarian algorithm. Consider Y(t) as the facial label matrix at *t*th iteration. Y(1) is the initial facial label matrix. The iterative process is carried out until the convergence condition is realized. The iterative face naming algorithm is as follows:

Algorithm 2: Face Naming Algorithm

```
Input: The feasible label sets $\{y^i|_{i=1}^m\}$, the affinity matrix A, the initial label matrix Y(1) and the parameters $\bar{N}_{iter}, \theta$.
for t= $1{:}\,\bar{N}_{iter}$ do
    Update B by using B =$[b_1,\ldots,\ b_{p+1}]'$,
    where $b_c = \frac{A\tilde{y}_c}{1'\tilde{y}_c}, \forall c = 1,\ldots,p$
    with $\tilde{y}_c$ being the c-th column of Y$(t)'$, and $b_{p+1} = \theta 1$.
    Update Y(t +1) by solving m sub problems in Eq (14).
    break if Y(t +1)=Y(t).
end for
Output: the label matrix Y(t +1)
```

Table 1: Details of the datasets we have used

Dataset	#images	#faces	#names	#face /image	#names /caption	Ground truth ratio
Synthetic	850	701	31	1.10	1.06	0.90
Movie face database	4512	430	151	2.02	1.02	0.63
SCface	4160	130	95	2.50	1.56	0.52

5. Experiments

We analyze our proposed schemes URLRD, LMNN algorithms for face labeling using one synthetic dataset and two real-world datasets.

5.1 Datasets

5.1.1 Synthetic dataset

Faces and poses are collected as the synthetic dataset. Top 15 popular names first found and then for each name we casually sample 70 images where this name appears as the image tag. The synthetic dataset contains 701 faces in 850 images, with a total of 31 names appearing in the corresponding tags, which includes these top 15 popular names and other names associated with these 850 images.

5.1.2 Real-world Datasets

Movie Face Database (MFD):

Database primarily as a benchmark for face recognition algorithms in unconstrained settings. MFD is built from frames extracted from movies of different languages. MFD database consists of 4512 facial images corresponding to 430 actors collected from approximately 103 movies. MFD consists of 67 male and 33 female actors with at least 200 images for each actor. MFD comes with detailed annotation in terms of age, bounding box, movie release, expression, gender, pose, makeup, and possible kind of occlusion.

SCface – Surveillance Cameras face database

SCface is a database of static images of human faces. Images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. Database contains 4160 static images (in visible and infrared spectrum) of 130 subjects. Images from different quality cameras mimic the real-world conditions and enable robust face recognition algorithms testing, emphasizing different law enforcement and surveillance use case scenarios.

5.2 Metric for Facial image identification

1) ULR

Metric for facial image Universal Label Refinement(ULR) is devised to amplify the naming quality by using graph based and low-rank learning scheme. It is a scheme to refine the labels of the facial images by exploring machine learning techniques.

2) LR-SVM [8]

SVM organizers are informed for each name to deal with the out-of-sample cases.LR-SVM concurrently learns the partial permutation matrices for grouping the faces and the rank of the data matrices are minimized from each group.

3) MildML [6]
Analyzes the Mahalanobis distance metric such that the images with common names are pulled closer, while the images that do not share any common label are pushed apart.

4) MMS [9]
Determining algorithm: decipher the face naming issue by learning SVM classifiers for each name.

5) Constrained Gaussian mixture model (cGMM) [10] [11]
Each name is related with a Gaussian density function in the feature space with the parameters estimated from the data and each face is estimated to be independently generated from the related Gaussian function. The overall assignments are chosen to achieve the maximum log.

5.3 Experimental Results

1) Results on the Synthetic Dataset:
The productiveness of our proposed method URLRD can be certified as we compare the facial matrices obtained from ULR and URLRD with the ideal facial matrix W* according to the ground truth as shown in the fig 2 (a)
The white points in the fig 2 (a) corresponds to the faces belonging to the same entity with the same names. The entries are set to zero to avoid self reconstruction.

The facial matrix of ULR obtained from W* in fig 2 (b) shows the following:

- The face is reconstructed by itself as the elements in the figure seem to be large, this should be evaded.
- Generally, Coefficients between the faces of the same entity are not significantly larger than the ones between faces from different subjects.

From the fig 2(c) we acquire the facial matrix by applying the URLRD method which has smaller values for the elements and has a similar facial matrix obtained from fig 2(a) as more obvious diagonal structures are exhibited in the facial matrix.

2) Results on the Real-World Datasets:
We consider accuracy and precision as the two criteria for the evaluation of our performance. The percentage of correctly annotated faces out of all the other over all faces is the percentage of accuracy, while the percentage of precision is faces that are rightly annotated as real names. Deducing names based on the faces in the image with ambiguous captions, we make use of all the images of the dataset. The real name ratio is the percentage of faces that are labeled as real names using all the steps over the faces in the dataset.

Table 2: The accuracies and precisions of different methods are as shown in the following table

Method	Movie Face Database (MFD)		SCface – Surveillance Cameras face database	
	Accuracy	Precision	Accuracy	Precision
MildML	0.543	0.512	0.633	0.648
LR-SVM	0.488	0.409	0.612	0.644
MMS	0.567	0.522	0.578	0.590
cGMM	0.556	0.587	0.662	0.644
LMNN	0.590	0.593	0.689	0.694
ULR	0.566	0.567	0.623	0.674
URLRD	0.578	0.589	0.654	0.681

Using bipartite graph for deducing the names of faces and changing the hyperparameter θ to tune the real name ratio.
For cGMM, set $c(f, q) =- \ln N(xif;\mu q,q)$, and $c(f, p + 1) = -\ln N(xif;\mu(p+1),(p+1)) + \theta$, as in [10], where μq and q (resp. μ(p+1) and (p+1)) are the mean and covariance of the faces assigned to the qth class (resp., the null class) in cGMM.
For MildML, set $c(f,q) =- x\in Sq\ w(xif ,x)$ and $c(f, p + 1) = \theta$, as in [6], where w(xif ,x) is the similarity between xif and x and Sq contains all faces assigned to the name q while inferring the names of the faces.
For MMS and LR-SVM, set $c(f,q) =-decq(xif)$ and $c(f, p + 1) =-decnull(xif) + \theta$, where decq(xif) and decnull(xif) are the decision values of SVM organizers from the qth name and the null class, respectively [11].

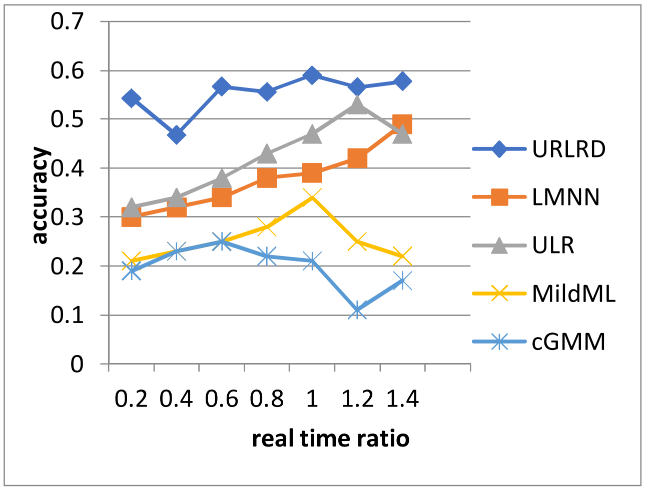


Fig. 4 (a) Accuracy versus real name ratio on the Movie Face Database

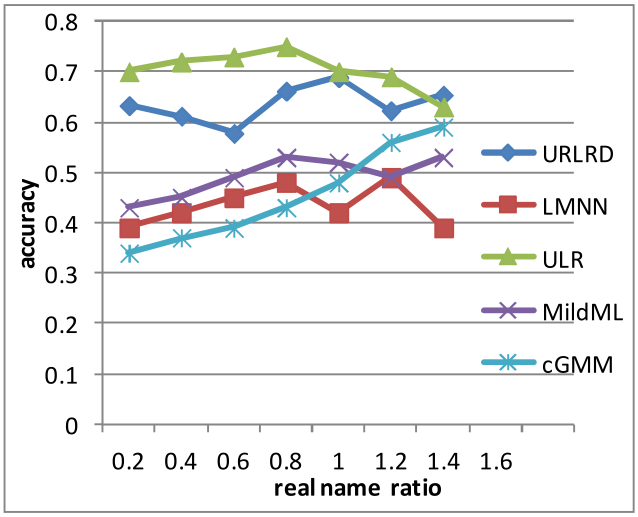
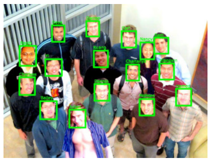


Fig. 4 (b) Accuracy versus real name ratio on the SCface database

Table 3: Two document examples with their naming results for LRR, ULR and URLRD, shows the maximum number of accurately named faces in an image.

#images	LRR	ULR	URLRD
	Shah Rukh khan, Kareena	Kareena, Shah Rukh khan, Alia	Karishma, Kareena, Alia,Shah Rukh Khan
	Dennis, Mohsen, Hamed, Yi,Sam	Deva,Carl, Kuang, Julian, Bailey, Ragib	Nancy, Dennis, Deva, Mohsen, Hamed, Yi,Sam, Carl, Kuang, Julian, Bailey, Ragib, Xian

5.4 Observations

1) Considering the 4 baseline algorithms which are MMS, cGMM, LR-SVM, and MildML we can say that there is no consistency in the values of the Movie Face Database and for the SCFace Database, MildML gives the best precision and cGMM gives the best accuracy.
2) Comparing LMNN and MildML we can say that LMNN outperforms MildML as seen from the table 2 both in accuracy and precision on both the datasets used. This indicates that LMNN utilizes ambiguous information to determine the discriminative distance metric.
3) Faces in a common subspace should belong to the same entity which is indicated by the ULR method. The method URLRD proposed by us achieves better consistency in performance than ULR on both the data sets used. This indicates that URLRD is beneficial while exploring the subspace structure among faces.
4) URLRD outperforms on both the datasets in terms of accuracy and precision. Compared to all other algorithms used URLRD achieves the best results. The fused facial matrix is more discerning for face labeling.
5) The Movie Face Database is challenging and interesting as it has worse results on comparison with the SCFace dataset this may be due to the reason that there are more faces in an image which is labeled with many captions in the MFD.

5. Conclusions

In this paper, we present an approach for face detection and naming which minimizes computation time while achieving high detection accuracy. To productively employ the face naming of the facial images we introduce

URLRD by using this scheme we increase the evaluation of auto face annotation performance. We also intensify the LMNN algorithm which delves on discriminating Mahalanobis distance metric. Two facial matrices are obtained by merging the matrices acquired by URLRD and LMNN. Our proposed methods focus on tackling the critical problem of enhancing the label quality and accurately naming the facial images. We analyze the two challenging and interesting real-world datasets from which we can certify that our URLRD and LMNN outperforms ULR and kNN respectively and several other baseline algorithms. Our future work will further speed up the current solution for very large applications and investigate other techniques to improve the label refinement task.

References

[1] P. Viola and M. J. Jones, “Robust real-time face detection,” Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, 2004

[2] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in Proc. 27th Int. Conf. Mach. Learn., Haifa, Israel, Jun. 2010, pp. 663–670.

[3] Dayong WANG, Chu Hong HOI, Ying HE, Jianke ZHU “Mining Weakly-Labeled Web Facial Images for Search-Based Face Annotation”,2014.

[4] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” J. Mach. Learn. Res., vol. 10, pp. 207–244, Feb. 2009.

[5] T. L. Berget al., “Names and faces in the news,” in Proc. 17th IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Washington, DC, USA, Jun./Jul. 2004, pp. II-848–II-854.

[6] M. Guillaumin, J. Verbeek, and C. Schmid, “Multiple instance metric learning from automatically labeled bags of faces,” in Proc. 11th Eur. Conf. Comput. Vis., Heraklion, Crete, Sep. 2010, pp. 634–647.

[7] D. Ozkan and P. Duygulu, “A graph based approach for naming faces in news photos,” in Proc. 19th IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., New York, NY, USA, Jun. 2006, pp. 1477–1482

[8] Z. Zeng et al., “Learning by associating ambiguously labeled images,” in Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 708–715

[9] J. Luo and F. Orabona, “Learning from candidate labeling sets,” in Proc. 23rd Annu. Conf. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2010, pp. 1504–1512

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, “Face recognition from caption-based supervision,” Int. J. Comput. Vis., vol. 96, no. 1, pp. 64–82, 2012

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, “Automatic face naming with caption-based supervision,” in Proc. 21st IEEE Conf. Comput. Vis. Pattern Recognit., Anchorage, AL, USA, Jun. 2008, pp. 1–8.

Kavitha G L is an MTech (CSE) graduate, pursuing PhD in cloud computing. She is currently working as Asst. Professor in Atria Institute of Technology, Bangalore. Her area of interests includes machine learning and image processing.

D V Pranathi Suhasini is pursuing B.E in Atria Institute of Technology, Bangalore. Her area of interests includes image processing.

Seema B Nikam is pursuing B.E in Atria Institute of Technology, Bangalore. Her area of interests includes image processing.