Deep Feature Representation and Similarity Matrix based Noise Label Refinement Method for Efficient Face Annotation

A. Suruliandi[1], A. Kasthuri[1], S.P. Raja[2] *

[1] Department of Computer Science & Engineering, Manonmaniam Sundaranar University, Tirunelveli 627012 (India)
[2] Department of Computer Science & Engineering, Vellore Institute of Technology, Vellore, Tamilnadu (India)

Received 28 August 2020 | Accepted 19 March 2021 | Published 3 May 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Face annotation is a naming procedure that assigns the correct name to a person emerging from an image. Faces that are manually annotated by people in online applications include incorrect labels, giving rise to the issue of label ambiguity. This may lead to mislabelling in face annotation. Consequently, an efficient method is still essential to enhance the reliability of face annotation. Hence, in this work, a novel method named the Similarity Matrix-based Noise Label Refinement (SMNLR) is proposed, which effectively predicts the accurate label from the noisy labelled facial images. To enhance the performance of the proposed method, the deep learning technique named Convolutional Neural Networks (CNN) is used for feature representation. Several experiments are conducted to evaluate the effectiveness of the proposed face annotation method using the LFW, IMFDB and Yahoo datasets. The experimental results clearly illustrate the robustness of the proposed SMNLR method in dealing with noisy labelled faces.

Keywords

CNN, Deep Learning, Face Annotation, Noise Label Refinement, Similarity Matrix.

DOI: 10.9781/ijimai.2021.05.001

I. Introduction

Recent years have witnessed the rapid growth of digital cameras and mobile devices, powerful cloud computing facilities, Web 2.0 photo sharing portals and social networks. Social media repositories such as Facebook, Twitter, Flickr, YouTube and Picasa allow users to upload and share personal photos or videos. As a consequence, masses of images have been created, distributed and shared on the internet by millions of users today, resulting in a large quantum of image collections on online social networks. Consequently, image sharing sites have difficulty managing and retrieving huge aggregates of face images. The plethora of multimedia content accessible today demands that challenges in terms of its storage, organization and indexing for future search and access be addressed. Moreover, an important aspect of online social media services is that users can annotate face images with keywords called tags, labels or captions. This voluntary activity of users who annotate faces with labels is termed labelling. Such labels may, however, be incorrect, imprecise or incomplete. Studies [1]-[3] show that name labels provided by users are highly “noisy”, in the sense that only around 50% are actually appropriate to the corresponding person, because there are no restrictions or boundaries on assigning names to images on social media applications.

Due to the noisy nature of web facial images, early name labels of such web facial image databases were perhaps imperfect or damaged, in the absence of additional manual fine-tuning endeavours. A key technique that addresses this challenge is auto face annotation, which automatically assigns a name to the face of the corresponding person. Making an annotation reliable under noisy labeled facial images is a major challenge for real-life face annotation systems. To facilitate noise label refining and annotating huge facial image databases, several automatic face annotation methods have been proposed in the related work [4]-[9]. However, the labelling results reported fall short of the standards required of existing, reliable face annotation systems, especially in terms of real-time issues and noisy labels. Facial images normally have issues with variations in appearance, pose, illumination, occlusion, and noisy labels, all of which can result in mislabeling in face annotation. An efficient face annotation method must overcome these complications with innovative image mining abilities that capture discriminative and intrinsic information in faces. Moreover, sophisticated noise label refining capabilities are required to make the face annotation method robust. Hence, this paper proposes a new face annotation method, Similarity Matrix-based Noise Label Refinement (SMNLR), which concurrently deals with the problems of refining noise labels and assigning labels to facial images.

The face annotation method based on distance metric learning refines noisy labels powerfully and enhances the reliability of face

* Corresponding author.

E-mail addresses: suruliandi@yahoo.com (A. Suruliandi), kasthurianburajan@gmail.com (A. Kasthuri), avemariaraja@gmail.com (S.P.Raja).

annotation. The use of distance metric learning methods also implies that the appearance of facial features is not identical. Essentially, these methods are most appropriate for high-level noisy labels, and enhance the accuracy of face annotation. Thus, the proposed method refines human-provided unreliable labels by dropping inappropriate labels and adding missing ones. Additionally, the proposed method generates a suggested name list based on visual similarities for better face naming.

Generally, feature extraction techniques play a vital role in large collections of facial images by annotating them. Most of the existing face annotation methods [10]-[14] utilize the hand-crafted features for feature representation. Given that hand-crafted features are not adequate enough to handle the task, face annotation needs different levels of detailed descriptions to distinguish between faces in multi-granularity similarities. To tackle this problem, deep features are extracted from the deep network to describe face images. Deep networks, such as Convolutional Neural Networks (CNN) [15]-[18], offer superior multilevel facial representation. The CNN provides the highest number of descriptive features and is the least sensitive to real-time challenges. Recent researches [19]-[25] on facial image analysis state that deep features are more robust for such complex tasks. Hence, in this work, a CNN model is used for deep feature extraction. This CNN can effectively provide deep features from the face image and significantly improve annotation performance. The main contributions of this paper are, 1) A modified CNN architecture is introduced for deep feature extraction 2) A Similarity Matrix-based Noise Label Refinement (SMNLR) method is proposed to handle noisy labeled face images in a large-scale dataset. Inconsistent name labels can be effectively discovered by the probabilities of similarity measurements, and then fine-tuned or relabeled for training 3) The modified CNN with a proposed SMNLR method obtains state-of-the-art results on various face datasets, i.e., LFW, Yahoo, and IMFDB datasets.

A. Related Works

In recent years, Convolutional Neural Networks (CNNs) have shown an extraordinary ability for face feature representation in face annotation tasks. Several works [15], [26]-[28] on face applications indicate that deep feature extraction is more robust for such complex tasks. Ma et al. [29] combined the CNN model, AlexNet, with the proposed semantic extension model (SEM). CNN feature are provided as input for the proposed model. Problems with image tag refinement and assignment are overcome by using a self-defined Bayesian-based model which divides images with similar features into a semantic neighbor group. Venkatesh et al. [20] proposed the canonical correlation analysis (CCA) framework to facilitate a CNN feature and word-embedding vector. The CCA-KNN outperforms the Corel-5k, ESP-Game and IAPRTC-12 datasets. De Souza et al. [15] integrated the LBP feature descriptor with a modified Convolutional Neural Network (CNN) and proposed a new deep neural network called the LBPnet. An extended version of the LBPnet, called n-LBPnet, is also proposed. This method extracts deep features and outperforms other state-of-the-art techniques on the spoofing database. Kurban et al. [30] used the Eurecom Kinect Face dataset and Body Login Gesture Silhouettes dataset to create a virtual dataset of multimodal biometrics. Their study proves that Convolutional Neural Network (CNN)-based methods get better features and are also less sensitive to variations in pose, lighting and facial expressions in images.

Zeng et al. [31] have proposed a novel framework called Partial Permutation Matrix (PPM) for each image. In PPM, the samples of the same class from each image are related diagonally to the image set. SVM been introduced for labeling face images with names. Cour et al. [32] proposed a convex learning formulation based on minimizing a loss function suitable for partial label setting. The aim is to learn a classifier that can disambiguate partially labeled and ambiguously labeled images. Chen et al. [3] proposed a matrix completion for ambiguity resolution (MCAR) technique to calculate exact labels from unclearly labeled images. Noisy soft labeling vectors can, however, impact its performance. Consequently, iterative candidate elimination (ICE) procedure is applied to reduce the iterative ambiguity resolution by slowly eliminating parts of a vaguely labeled face. Liu et al. [33] proposed a self-error-correcting CNN (SECCNN) approach to work with noisy labels. The SECCNN develops a confidence policy that switches between the label of the sample and the max-activated output neuron of the CNN. Su et al. [34] have identified the difficulty of relating names with faces from large scale news images with captions. This problem was overcome by Person-based Subset Clustering which is mainly based on face clustering. This method provides the visual structural information all face images derived from the same name. Kumar et al. [35] proposed a two-step approach for both detection and recognition tasks. In the first step, a seed set is generated from the given image collection using detection and recognition algorithms. In the second step, the performance is improved by adapting the seed set. Maihani et al. [36] proposed a novel method for automatic image annotation wherein similar images are retrieved and a relative graph generated with tags. Finally, the tags of the dense community are chosen for the query image. Wang et al. [6] introduced an unsupervised label refinement (ULR) method to fine-tune weak labelled face images on online social networks. Their work uses a cluster-based approximation scheme for label refinement, while the majority voting approach is applied to tag names with facial images. The drawback of the ULR is that it cannot handle issues with duplicate names in real-life environments. Zhu et al. [8] proposed a knowledge transfer framework for face photo-sketch synthesis task. A new network architecture which allows to transfer knowledge from two teacher models to two student models are trained and knowledge has been transferred between two student models mutually. Two students network are trained using a small set of photo sketch pairs. Experimental results demonstrate that their proposed method performs better than other state-of-the-art methods. Zhu et al. [37] proposed a deep Convolutional Neural Network, to represent face photos. More precise person sketch patches and weight combination for sketch patch reconstruction could be obtained from the deep feature representations. Deep feature model based on the graphical representation is proposed to mutually discover weights for deep feature representations and reconstruction weights. Zhu et al. [38] proposed a deep collaborative framework with two opposite networks. These two networks perform the common communication between two opposite mappings. A collaborative loss is proposed in this work to limit the two contrary mappings and create them more balanced, as a result building the models more appropriate for photo–sketch synthesis task. Wang et al. [39] proposed a novel co-mining framework that utilizes two peer networks to identify the noisy faces, replaces the high-confidence clean faces and reassigns the clean faces in a mini-batch fashion.

B. Motivation and Justification

Most of the existing methods [4], [10], [40], [41] are applied directly on labeled facial images for face annotation without fine-tuning the labels, culminating in noisy or incorrect labels in face-name association. Certain early studies [1], [6], [42] overcame this drawback using unsupervised clustering algorithms to refine noise labels. In these clustering algorithms, a face collection is divided into several groups based on the identity name. Noisy labels are refined by estimating the maximal cluster among the groups of faces. However, the algorithms cannot prove that a face image indisputably belongs to a particular identity name; rather, they simply state that there is a high probability of the face image corresponding to the identity in question. This kind of simple correlation between faces and labels is not effective enough to refine label ambiguity. Consequently, several researchers [3], [43] have attempted to resolve the incompatibility between faces and name labels with supervised distance metric learning approaches. Distance metric learning-based label refinement techniques have shown better results than other existing label refinement techniques. In complex cases, however, information transmission follows no standard form and varies in feature gaps, a drawback that limits face annotation. Therefore, a much more accurate and robust noise label refinement technique is essential for effective face annotation by refining noise from labeled facial images. Thus motivated, an effort is made in this work to address the issue, and a new distance metric learning-based noise label refinement method is proposed, called the Similarity Matrix-based Noise Label Refinement (SMNLR), it combines the Cosine and Mahalanobis distance measures.

At the same time, the number of variations in faces also gives rise to the issue of label ambiguity because facial images are generally captured under various issues such as illumination, occlusion, expressions, and variations in poses. Most of the existing face annotation methods [1], [2] consider only hand-crafted feature extraction techniques for feature representation. They effectively capture the most information from facial images, and try to resolve issues by using a single or double layer to extract facial features. But, in several difficult domains, such as twin persons, these hand-crafted features generate the similar features for different persons due to its limitations. Hence, the faces might attain association with irrelevant labels in the context of label refinement owing to the low quality facial features. Also, when it deals with misaligned faces, it generates the unwanted texture information of faces. Hence, a robust feature is to be extracted from face images by overcoming these issues to improve the reliability of proposed face annotation method. Instead of utilizing the hand-crafted features, in recent years, Convolutional Neural Networks (CNN) [15], [44] extracts facial features using multiple levels of layers, wherein every single layer extracts deep features from faces. The CNN's remarkable learning features have helped resolve a variety of computer vision problems. These include image annotation, face recognition, image classification, object detection and identification, indicating that using deep features in face annotation for feature representation would be most efficient. Therefore, in this work, a most effective deep feature is used for feature representation in proposed SMNLR method.

The proposed SMNLR method effectively explores noisy labels by utilizing a fusion of the two discriminative similarity matrices. From the point of view of the literature, it is observed that the Cosine [45], [46] and Mahalanobis [47], [48] distance metric learning methods represent the most powerful similarity information between faces, compared to other existing distance metric learning approaches. The Cosine distance metric provides the direction information between samples, based on a broad collection of orientations. The Cosine of the orientation has essential uniform information for the matching components of faces. However, it does not consider magnitude differences between samples. Consequently, in critical circumstances involving illumination and expression, the cosine distance metric is too complex to handle all of the matching similarity information in the samples. To overcome this shortcoming, the Mahalanobis distance metric activates the similarity matrix by incorporating the magnitude difference of the relationship between the samples. Generally, the Mahalanobis distance metric encodes more meaningful similarity measurements using the uniform distribution of the sample with respect to face reconstruction. Therefore, this work combines the direction-based cosine similarity matrix and the distribution-based Mahalanobis similarity matrix. Therefore, this work combines the direction-based Cosine similarity matrix and the distribution-based Mahalanobis similarity matrix. Since the fusion of the two discriminative matrices uses a normalization parameter, α, with a value of 0.5, it significantly eliminates noisy labels and reassigns correct labels, based on the distance of the least similarity value of the fused similarity matrix. Justified by this, a new distance metric learning-based face annotation method called the SMNLR is proposed to refine noise labels based on a fusion of the Cosine and Mahalanobis similarity matrices. In addition, when the corresponding test face is not found in the training dataset, the given test face image is annotated with a name, using the suggested labels list. The suggested name list contains a list of labels that are applied when the test face does not match with database images of the training set. Given the need to name unknown faces in the test image, the list of suggested names is considered. The procedure for creating a suggested list further enhances the reliability of the proposed SMNLR method.

C. Outline of the Proposed Work

The outline of the face annotation process using the proposed SMNLR method is described in Fig. 1. The method comprises two phases, training and testing. The appropriate face region is chosen from the images to remove irrelevant information in the pre-processing step. In the training phase, deep features are extracted from the training images using the CNN. Two discriminative similarity matrices, the Cosine and Mahalanobis, are obtained using the training features and combined to create a fused similarity matrix. Noisy labels are refined and unambiguous labels reassigned, based on the similarity measurement of the fused similarity matrix. A suggested name list is also generated for face naming. In the testing phase, just as in the training phase, a feature extraction procedure is considered. The multi-class SVM classifier annotates the face images with their names.

D. Organization of the Paper

Section II explains the proposed SMNLR method in detail. Section III describes the databases and experimental results. Section IV discusses the performance analysis of the proposed method. Section V concludes the paper.

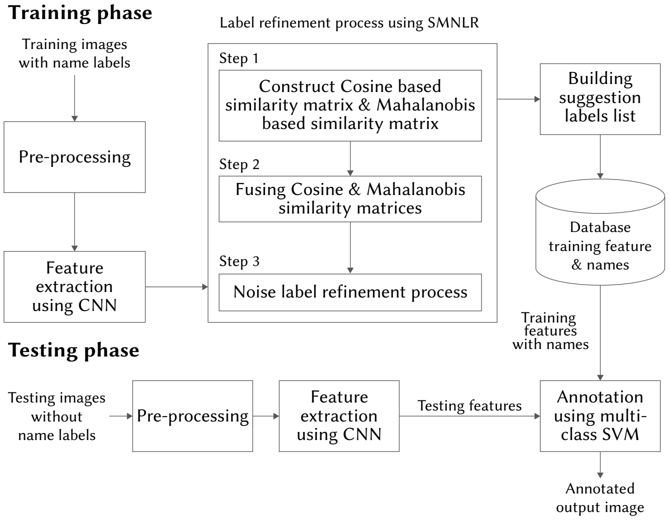

Training phase
Training images with name labels
Pre-processing
Feature extraction using CNN
Label refinement process using SMNLR
Step 1
Construct Cosine based similarity matrix & Mahalanobis based similarity matrix
Step 2
Fusing Cosine & Mahalanobis similarity matrices
Step 3
Noise label refinement process
Building suggestion labels list
Database training feature & names
Training features with names
Testing phase
Testing images without name labels
Pre-processing
Feature extraction using CNN
Testing features
Annotation using multi-class SVM
Annotated output image

Fig. 1. Process flow of the proposed SMNLR method.

II. The Proposed Method

A. Convolutional Neural Networks (CNN) Feature Extraction

Deep networks, such as Convolutional Neural Networks (CNN) [49], offer superior multilevel facial representation. The CNN model uses the output of a layer in the centre of the model as another description of the data, it is represented as a deep feature. Generally, CNN architecture consists of one or more convolution layers, often with a pooling layer, which are followed by one or more fully connected layers as in a neural network. The CNN uses this architecture to efficiently extract essential features from face image.

In this work, the CNN architecture consists of input layer, three convolutional layers, namely, convolution 1, convolution 2, and convolution 3; and three pooling layers, namely, pooling 1, pooling 2, and pooling 3. The input layer assigns an input image to the first convolutional layer. Convolutional layers play a major role in the CNN for feature extraction. Several convolutions can be performed on an input image, each utilizing a different filter and producing a unique feature map. Hence, the output of each layer describes a particular feature representation obtained from the input image. The convolution layer parameters contain several spatial and spectral learnable kernels or filters. In the first, second, and third convolutional layers, several feature maps are generated. Each convolutional layer is connected with a rectified linear unit (ReLU) and a pooling layer (down-sampling). A ReLU is an extensively used nonlinear activation function and presents a threshold operation to every component of the feature map. It assigns a value of 0 to the negative elements of the feature map. Pooling layers reduce the large number of features generated by the convolution layer. The convolved feature map is rendered more powerful and robust through the pooling layer. Max and average pooling are the most widely used techniques for the pooling process. The max pooling process is carried out by selecting the highest value of all the pixels in the receptive field to describe the output of the pooling feature map. The pooled (down-sampled) features generated in each pooling layer are provided as input to the next convolutional layer. Dropout layer is used to avoid the overfitting problem of features. Finally, fully connected layer generates deep feature values by combining all of the features learned from previous layers. A comprehensive demonstration of the CNN is shown in Fig. 2.

B. Convolution Layers

In convolution 1, 4 convolution filters with size of 4 × 4 are applied for the convolution process to generate feature maps. The convolution filter is applied with the stride of 1 to the input image. The convolution process is performed using Equation (1).

$$FM_p(x) = \sum_{\forall y \epsilon N(x)} G(y) * K_p(m) \tag{1}$$

where FM_p (x) is the output feature map of the convolution process, where p = 1, 2, ..., 4 represents the p number of feature maps. G(y), is the input image, and y = (i, j), represents the position of the pixel value corresponding to the neighbourhood of value x = (i, j), i.e., yϵN(x) in the input image; Here, K_p (m), also with p = 1, 2, ..., 4, belongs to the value in the pth convolution filter in the corresponding position of y, and m = (1, 1), (1, 2)..., (4, 4) means the position of elements in the convolution filter.

C. ReLu Layers

The 4 feature maps generated from convolution 1 are provided as input to the next ReLU layer. This layer activates the non-linear function to each element of the feature maps using Equation (2).

$$\mathrm{ReLU}(x) = f\begin{Bmatrix} x, x \geq 0 \\ 0, x < 0 \end{Bmatrix} \tag{2}$$

D. Pooling Layers

In pooling 1, the rectified feature maps are down-sampled to find the local maxima in the neighborhood, using the max-pooling process. The feature maps are down-sampled using Equation (3).

$$D_p(z) = \max\,\{FM_p(x)\}_{\forall x \epsilon N(z)} \tag{3}$$

Here, $D_p(z)$ means the outputs of the pooling processes corresponding to the feature maps, FM_p. The feature map element at x = (i, j) is belonging to the neighborhood of the value of z = (i, j) , i.e., xϵN(z) in the down-sampled feature map. The down-sampled features generated from pooling 1 are provided as input to the next convolution layer 2. The three processes mentioned above such as convolution, ReLU and pooling are repeated in the second and the third layers of CNN. In convolution 2, 6 filters are applied for the convolution process so as to extract feature information from the faces. 24 feature maps are generated from convolution 2 and the features down-sampled in pooling 2.

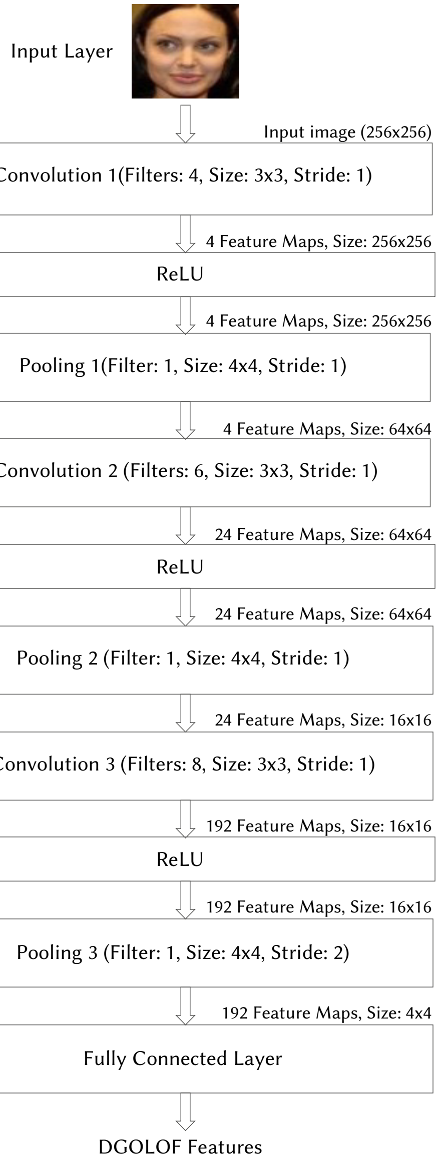

Input Layer
Input image (256x256)
Convolution 1(Filters: 4, Size: 3x3, Stride: 1)
4 Feature Maps, Size: 256x256
ReLU
4 Feature Maps, Size: 256x256
Pooling 1(Filter: 1, Size: 4x4, Stride: 1)
4 Feature Maps, Size: 64x64
Convolution 2 (Filters: 6, Size: 3x3, Stride: 1)
24 Feature Maps, Size: 64x64
ReLU
24 Feature Maps, Size: 64x64
Pooling 2 (Filter: 1, Size: 4x4, Stride: 1)
24 Feature Maps, Size: 16x16
Convolution 3 (Filters: 8, Size: 3x3, Stride: 1)
192 Feature Maps, Size: 16x16
ReLU
192 Feature Maps, Size: 16x16
Pooling 3 (Filter: 1, Size: 4x4, Stride: 2)
192 Feature Maps, Size: 4x4
Fully Connected Layer
DGOLOF Features

Fig. 2. Deep feature extraction using CNN.

A ReLU is comprised between the convolution 2 and the pooling 2 operation. The features of pooling 2 are given as input to the convolution 3. In convolution 3, 8 filters are applied to the convolution process. There are 192 feature maps are generated in convolution 3 and applied to ReLU process. The linearly rectified features are down-sampled in pooling 3. The filter size of 4 × 4 and stride of 1 are applied to all convolution and poling layers. Finally, the deep features are obtained from the fully connected layer of the CNN.

The CNN architecture is trained using a widely used gradient descent method, called stochastic gradient descent (SGD). Table I shows the parameters of the CNN architecture designed in this work for deep feature extraction. By this way, the deep feature of faces is given to the proposed face annotation method.

TABLE I. Parameters for CNN Model

Layer name	No. of filters	Filter size	Stride/ padding	No.of feature maps	Output size
Input layer	n/a	n/a	n/a	1	256x256
Convolution 1	4	3x3	1/0	4	256x256
ReLU	n/a	n/a	n/a	4	256x256
Pooling 1	1	4x4	1	4	64x64
Convolution 2	6	3x3	1/0	24	64x64
ReLU	n/a	n/a	n/a	24	64x64
Pooling 2	1	4x4	1	24	16x16
Convolution 3	8	3x3	1/0	192	16x16
ReLU	n/a	n/a	n/a	192	16x16
Dropout	n/a	n/a	n/a	192	16x16
Pooling 3	1	4x4	2	192	4x4
Fully connected	n/a	n/a	n/a	n/a	3072

The CNN feature enriches spatial localization and effectively exploits minute texture information to resolve real-time issues affecting face images. The convolution and pooling layers of CNN are able to obtain enough information such as edges, orientations, and corner features from the facial images. Edge filters help identify difficult structures caused by facial images. When a face is rotated, key texture features like the eyes, nose and mouth (i.e., non-frontal face) are likely to be lost, but orientation filters help identify enough information from the rest of the face. When elderly faces are considered, corner features help identify the (key point localization) shape of the mouth, nose, eyes and cheeks better than other textures, and effectively differentiate between such faces and other faces. Fig. 3 shows the sample of feature maps generated from the convolutional layers of CNN.

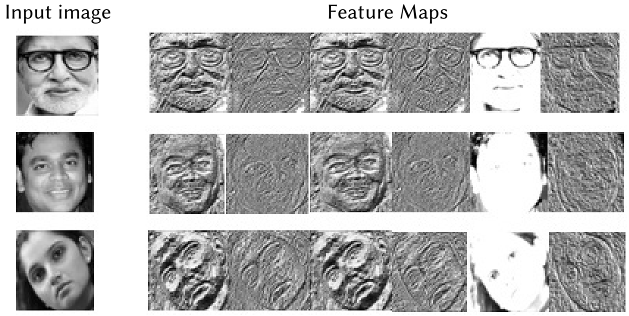


Fig. 3. A sample of CNN feature maps.

III. The Proposed Similarity Matrix-based Noise Label Refinement (SMNLR)

In this section, a new method called the Similarity Matrix-based Noise Label Refinement (SMNLR) is proposed for face annotation. Particularly, two different learning schemes are introduced to obtain two discriminative similarity matrices by learning from noisy labeled faces. The two similarity matrices are further combined to produce a fused similarity matrix, and the noisy labels refined, based on the fused affinity matrix. Section III(A) below introduces a new procedure to generate the Cosine-based similarity matrix. Section III(B) below introduces a new procedure to generate the Mahalanobis-based similarity matrix. Section III(C) describes the fusion of the Cosine and Mahalanobis similarity matrices. Section III(D) introduces a noise label refinement process to refine the noisiness of the labeled faces. The fused matrix effectively discovers the noise labels in labeled facial images. Section III(E) and Section III(F) describe the suggested list generation procedure and face naming procedure respectively.

A. Learning the Cosine-based Similarity

This section explains a new procedure to generate the Cosine-based similarity matrix. The collection of facial features is divided into several subsets, based on their names. The mean feature is calculated, from among the features for each subset, to make a set of effective mean features.The first similarity matrix is calculated between each training facial feature and each subset means feature, based on the Cosine distance.

Each face is characterized as a d-dimensional feature vector using the CNN. For an image x_1 being represented whose CNN feature is defined as f_1, the feature group F is shown as expression (4). The CNN features of each face image, x_p, where p = 1, 2, ..., N in the training dataset, X, can be represented as

$$F = \{f_1, f_2, f_3, \dots \dots f_N\}. \tag{4}$$

Here, N is the total number of features in training set. The CNN features of each face image containing 3072 feature values. Hence the limit for N is specified as 3072 .These training features are grouped into subsets, based on the M names, using Equation (5).

$$S_{ji} = \left[\cup_{M=j}^{N} f_{ji}\right] \tag{5}$$

where $S_{ji} = \{f_{11}, f_{12}, \dots \dots f_{MN}\}$ is a subset and j = 1, 2, ... M is the number of subsets based on person names in the training set and i = 1, 2, ... N represents the number of features in each subset.

The mean feature, MF_j, is calculated for each subset using Equation (6).

$$MF_j = \frac{1}{N}\sum_{i=1}^{N} S_{ji} \tag{6}$$

where number of mean feature, $MF_j = [MF_1, MF_2, \dots MF_M]$, is calculated for all subsets.

The first similarity matrix can be calculated using Equation (7). The Cosine similarity is calculated between each face feature, f_i, in the training set and the mean features of each subset, MF_j.

$$SM1_{ij} = f_i^T . MF_j / (\| f_i \| . \| MF_j \|) \tag{7}$$

where $SM1_{ij}$ represents an element in the ith row and jth column of the cosine similarity matrix, SM1. T is the transpose of the distance value.

B. Learning the Mahalanobis-based Similarity Matrix

This section introduces a new procedure to generate the Mahalanobis-based similarity matrix. Like the first similarity matrix, the mean feature of each subset is calculated, but in contrast, here the mean feature is calculated differently. The collection of facial features is evenly partitioned into several subsets, based on their names. For each subset, the most similar nearest neighbours of each feature among the subset are found using the KNN. The set of minimum distances are calculated in each subset. Finally, the new subset is produced and the mean is calculated. The second similarity matrix is calculated, based on the Mahalanobis distance between each training facial feature and each subset mean feature.

The distance, $d(f_x, f_y)$ between feature f_x and its target neighbours, f_y is calculated using Equation (8).

$$d(f_x, f_y) = \sqrt{\sum_{x=1}^{N} (f_x - f_y)^2} \tag{8}$$

where x = 1, 2, ... N is the number of features in the subset and y = 1, 2, ... T is the number of target nearest neighbours. The set of minimum distances, NF, is formed by using the most similar images with the minimum distance value $d(f_x, f_y)$ using Equation (9).

$$NF=\left\{\min\left(d(f_1,f_y)\right)\cup\min\left(d(f_2,f_y)\right)\ldots\ldots\cup\ \min\left(d(f_x,f_T)\right)\right\} \tag{9}$$

where T represents the number of nearest neighbours. The new subset NF = $\{f_1, f_2, f_3, \ldots \ldots \ldots f_N\}$ is generated, and the process repeated with all other features in other subsets. The mean of each new subset, NMF_j, is calculated using Equation (10).

$$NMF_j = \frac{1}{N}\sum_{i=1}^{N} f_{ji} \tag{10}$$

where N is the number of features in the new subset, NF. The Mahalanobis distance is calculated between each training set facial feature, f_i and the mean features of each new subset, NMF_j using the following Equation (11).

$$SM2_{ij} = \left((f_i - NMF_j)\right)^T C^{-1}(f_i - NMF_j) \tag{11}$$

where $SM2_{ij}$ represents an element of the ith row and jth column of the second similarity matrix, where N is the number of features in the training dataset, M the total number of subsets, f_i the feature of the ith image in the training dataset, NMF_j the mean feature of the jth subset, C^{-1} the inverse covariance matrix, and T the transpose of the distance value.

C. Learning the Fusion of the Cosine and Mahalanobis-based Similarity Matrices

The first similarity matrix, SM1, is learned from Equation 9 and the second, SM2 , from Equation (12). The two are merged to ake a fused similarity matrix. The fused similarity matrix effectively discovers noise labels, since both matrices contain complementary details of the faces and the discriminative relationship between the faces.

$$FSM_{ij} = \alpha SM1_{ij} + (1-\alpha)SM2_{ij} \tag{12}$$

where FSM_{ij} is the fused similarity matrix, and α the normalization parameter in the range [0, 1] For an enhanced of the label refinement process performance, the normalization parameter value of α is fixed at a range between 0 and 1, respectively, throughout the experiments.

D. Noise Label Refinement Process

The initial noisy name label matrix is refined and reassigned the correct labels, based on the similarity measurement of the fused similarity matrix. The noise labels are replaced with their corresponding subsets, based on the minimum distance between each face and the faces in each subset. Hence, each noise-labeled subset is transformed into a fine-tuned labeled subset, and all faces with their corresponding labels can be relied on for face naming. The noise labels are refined using Equation (13).

$$NL_i=\min\left(FSM_{i,1}, FSM_{i,2}, FSM_{i,3,}, \ldots\ldots\ldots, FSM_{N,M}\right) \tag{13}$$

where $FSM_{i,1}$ is the similarity value between the ith face and 1st subset.

Fig. 4 shows an example of the label refinement process wherein, for instance, the training features are partitioned into three subsets, based on a person's name. Subset 1, Subset 2, and Subset 3 consist of the sample names P1, P2, and P3 respectively. In each subset, the three different labeled samples are represented by three different shapes, such as a circle, triangle, and square respectively. Subset 1 has three noisy labels. The three samples, which are actually of different persons, are ambiguously labeled P1. This means that the three samples are incorrectly grouped in Subset 1, while the images are grouped on the basis of the name. Similarly, Subset 2 and Subset 3 contain three and two ambiguously labeled samples respectively. The noise labels are rearranged in appropriate subsets using the proposed SMNLR method, which efficiently enhances ambiguously labeled faces with the fused Cosine and Mahalanobis matrices.

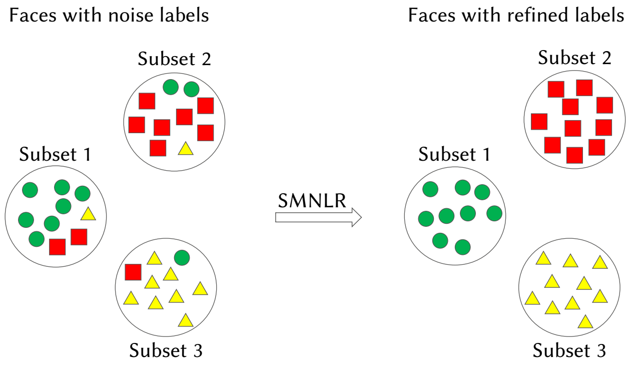


Fig. 4. SMNLR refines the noise labels.

E. Building a Suggested Labels List

In the testing stage, if a corresponding face that is similar to the test face does not occur in the training dataset, it could degrade the face naming capability of the proposed method. To resolve this problem, it is critical to name the unknown face in the test image and, therefore, a suggested label list is created for each instance of the training set. The similarity of each face image and face image collection of all relevant faces are computed. These similarity measurements are sorted in ascending order. The names are retrieved where appears in the labels associated with relevant face images. In training dataset, the suggested labels list, SNL_i is generated for each feature, f_i. The fused similarity matrix, FSM_{ij} is sorted in ascending order using Equation (14).

$$SNL_i = \text{sort}\left(FSM_{i,1}, FSM_{i,2}, FSM_{i,3,} \ldots\ldots\ldots\ldots, FSM_{i,M}\right) \tag{14}$$

where $FSM_{i,1}$ is the similarity value of the i^{th} training feature corresponding to subset 1, and M is the total number of subsets.

F. Face Naming Using the Multi-class SVM

The face image is annotated with its correct name, using the Multi-class SVM. In the training phase, all the faces with their noise labels are refined, using the proposed method. In the testing phase, the test features are compared with the features of the training set, using the Multi-class SVM classifier. The SMNLR applies the following conditions for face naming:

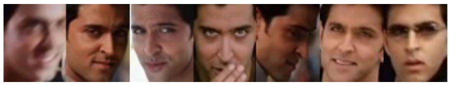
(1) When the multi-class SVM classification result is predicted as positive, a name is assigned to the input face with its corresponding predicted class name.

(2) When the multi-class SVM classification result is predicted as negative, the SMNLR suggests a name list for the input face image.

IV. Database Description

This section describes about the publicly available datasets for face annotation. In this research, the experiments are conducted using the three different datasets, namely, Labeled Faces in the Wild (LFW), Indian Movie Face Database (IMFDB), and Yahoo! News. The LFW dataset is publicly available and it can be collected from http://vis-www.cs.umass.edu/lfw/#explore. IMFDB dataset is publicly available from http://cvit.iiit.ac.in/projects/IMFDB/. Yahoo dataset is available from http://goo.gl/2XlES. It contains the news images with captions. The samples of faces are shown in Table II.

TABLE II. Database Description

Database	Training Set (faces & names)	Testing Set (faces)	Sample face images with various issues
LFW	12500	10450	
Yahoo	8900	7050	
IMFDB	10300	11500	

V. Experimental Results and Analysis

A. The Proposed SMNLR Face Annotation Results for Various Datasets

The performance of the proposed SMNLR method was evaluated with experiments conducted on facial images simulated by noisy labels and real-time challenges. The training set consists of noisy labeled faces, and the testing set of labeled faces. Real-time challenges such as variations in poses, occlusion, illumination and facial expressions are also considered in analysing the effectiveness of the proposed face annotation method. Fig. 5 shows the topmost 5 matching similar faces with their annotation results for 2 sample faces from each dataset.

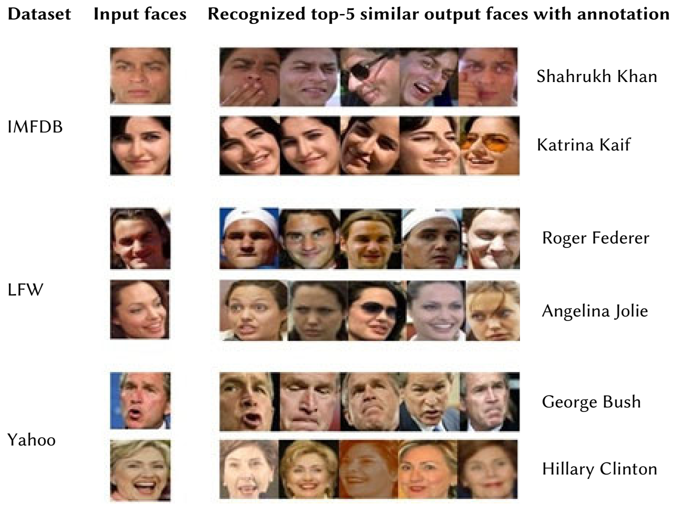


Fig.5. Sample of top-5 recognized similar images with annotation using the proposed method.

VI. Performance Analysis

A. Performance Metrics

The feasibility and effectiveness of the proposed face annotation method is analyzed using the performance metrics given in Equations (15)-(23). The precision, recall and F-score values are calculated using Equations (15), (16) and (17) respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{16}$$

$$\text{F} - \text{score} = 2.\frac{\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \tag{17}$$

The recognition rate is validated using Equation (18). The accuracy of the face annotation is evaluated using Equation (19).

$$\text{RecognitionRate} = \frac{\text{Number of correctlymatched images}}{\text{Total test images}} \times 100 \tag{18}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{19}$$

The true positive rate (TPR) determines the percentage of the face image that is correctly annotated, and is calculated using Equation (20).

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \tag{20}$$

The false positive rate (FPR) typically describes the possibility of falsely naming the input face image, and Equation (21) calculates it.

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{21}$$

The miss rate and error rate of the annotated results are calculated using Equations (22) and (23).

$$\text{Miss rate} = \frac{\text{FN}}{\text{FN+TP}} \tag{22}$$

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{Total}} \tag{23}$$

where TP is true positive, FP is the false positive, TN is the true negative, and FN is the false negative.

B. Fine-tuning the Normalization Parameter, Alpha-(α), for the Proposed SMNLR Face Annotation Method

The noise labels are refined, based on the fused similarity matrix. The fused similarity matrix generation approach uses the normalization parameter, alpha -(α), which is represented in Equation (8). The normalization parameter, α, that combines the two different similarity matrices is experimentally fixed using the three datasets of the LFW, IMFDB, and Yahoo. The impact of the normalization parameter, α, is evaluated in this experiment to find the optimum alpha value. The parameter, α, is set in the range {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}. Table III shows the experimental results.

Certain critical validations are to be drawn from Table III. When setting the value at α=0 and α =1, the performance of the proposed SMNLR method fluctuates, since the noise label refinement procedure becomes ineffective when α=0 and α=1 respectively. This is because, if α is set to 0, the similarity information from the cosine-based matrix can be avoided and the Mahalanobis-based matrix can be quietly updated to handle noise label refinement. At the same time, if α is set to 1, the similarity information from the Mahalanobis-based matrix can be avoided, and the first matrix can be gently updated to carry out noise label refinement. The process of fine-tuning the ambiguity of labeled faces is poorly performed, since the noisy nature of incorrectly labeled faces can be transmitted to correctly labeled faces through the similarity measurements of the fused matrix. Hence, it is clear that the values of 0 and 1 are not applicable to α. After fine-tuning α in

TABLE III. Finding the Optimal Alpha (A)-value for the Proposed Face Annotation Method

Normalization parameter (α)	Performance of noise label refinement					
	LFW		IMFDB		Yahoo	
	Accuracy (%)	Error rate (%)	Accuracy (%)	Error rate (%)	Accuracy (%)	Error rate (%)
α=0	72	23.8	69	24.7	74	22.6
α=0.1	78	19.5	75	20.7	80	18.1
α=0.2	83	15.3	81	17.5	84	14.7
α=0.3	87	10.4	85	14.4	89	10.6
α=0.4	93	6.8	90	9.3	94	5.2
α=0.5	98	1.3	96	2.2	97	2.1
α=0.6	94	5.6	92	7.1	93	4.7
α=0.7	90	8.3	89	10.5	91	8.3
α=0.8	86	11.2	83	15.8	87	10.7
α=0.9	82	17.5	78	19.4	80	14.5
α=1	74	21.7	70	23.3	78	20.8

TABLE IV. Testing the Performance of the Proposed Method for Varying Proportions of Noisy Labels on Different Datasets

Datasets	Performance Metrics	Proportions of noisy labels (%)									
		10	20	30	40	50	60	70	80	90	100
LFW	Recall	97	97	96	95	94	89	87	85	82	80
	Precision	96	95	95	93	92	90	89	86	85	82
	Accuracy	97	97	96	95	94	91	90	89	87	85
	Miss Rate	2.1	2.7	3.2	4.4	6.8	7.1	7.6	8.1	8.7	10.3
	Error Rate	1.9	2.3	2.9	3.6	5.1	6.4	8.3	9.2	10.5	12.4
IMFDB	Recall	96	94	92	92	90	88	86	83	82	79
	Precision	95	93	93	91	91	89	87	85	83	81
	Accuracy	97	95	92	95	94	93	90	89	86	84
	Miss Rate	2.5	2.9	3.6	4.7	5.9	6.4	8.3	9.6	10.4	11.3
	Error Rate	2.6	3.5	3.8	3.9	5.0	6.9	9.2	10.2	11.8	12.6
Yahoo	Recall	98	97	96	94	91	89	87	86	84	81
	Precision	96	95	93	91	93	92	90	88	85	83
	Accuracy	97	96	95	93	92	90	88	84	82	85
	Miss Rate	1.5	2.7	3.6	5.8	6.5	7.9	8.4	9.7	10.1	12.7
	Error Rate	2.0	2.9	3.1	6.5	7.0	7.3	9.2	10.7	11.8	11.6

the range {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}, it is noted that the proposed method achieves improved results when assigning α to 0.5 on the three datasets, and hence the parameter α is fixed at 0.5. Since the fused similarity matrix comprises several discriminative and prominent details of the cosine and Mahalanobis similarity matrices, it is most effective at exploring noise labels and face naming.

C. Testing the Performance of the Proposed Face Annotation Method Under Different Levels of Noise Labels

Noise label refinement is a difficult issue in face naming. In this experiment, the performance of the proposed face annotation method is tested under different levels of noise labels. Face images and their names are randomly selected from the LFW, IMFDB and Yahoo databases for the training and testing sets. The training dataset contains 12,000 noisy labels of 800 faces. For the purpose of evaluation, different noise levels are simulated in a range from 0% to 100% by updating the randomly-allocated noise labels of each subset in the training set. Here, each level of the noisy labeled faces of all the subsets is applied separately to the proposed method, and the experimental results are shown in Table IV.

Table IV shows that the proposed method refines all the noisy labeled faces perfectly when the noise level ranges from 10% to 30%. When the noise level ranges from 30% to 50%, the proposed SMNLR method reaches 94% accuracy and a lower error rate. When the noise percentage varies from 50% to 100%, it is seen that almost all the noise labels are refined, while still obtaining an accuracy of over 80%. This clearly illustrates the robustness of the proposed SMNLR method in dealing with noisy labeled faces. The SMNLR eliminates the noise labels and re-assigns the correct labels, based on the distance of the least similarity value of each instance. Table IV shows that the SMNLR outperforms different levels of noise, except when the ambiguity percentage is greater than 50%. Hence, the SMNLR achieves enhanced results at low- and middle-levels of noise and becomes vulnerable at high noise levels. The underlying reason for these results is that high ambiguity levels affect the least distance component of the label refining similarity matrix, with the possibility of co-occurrence at such high ambiguity levels.

D. Performance Evaluation of the Proposed Method by Varying Number of Suggested Labels with Respect to the Matching Score

A suggested list is created for each instance of the training set, using the matching score representation. The maximum number of possibilities of extra names for each instance is analysed, based on the matching score. Therefore, this experiment is conducted to find the best combination of matching score levels with size of the suggested labels list. The performance of various combinations of matching score levels with varying sizes of the suggested list is demonstrated in Table V. The matching score levels range from 10% to 50% and the suggested list size that includes 2, 3, 4, 5, 6, 7, 8, 9, and 10 are considered for this experiment, with Table V listing the results.

TABLE V. Performance Evaluation of the Proposed Method By Varying Size of Suggested Labels List and Level of Matching Score

Matching Score level (%)	Datasets	Annotation Accuracy (%)								
		Number of suggestion labels for each face								
		2	3	4	5	6	7	.8	9	10
10	LFW	86.2	90.6	91.7	92.3	93.8	84.7	81.9	80.3	79.2
	IMFDB	85.5	91.2	92.3	93.5	95.9	83.4	82.7	81.6	76.5
	Yahoo	84.4	90.3	91.5	92.1	96.7	86.2	81.5	80.2	77.6
	Average	85.3	90.7	91.8	92.6	95.4	84.7	82	80.7	77.7
20	LFW	90.4	93.4	95.1	95.8	96.6	89.3	86.2	82.3	80.7
	IMFDB	89.1	91.6	94.8	96.9	97.9	86.9	84.5	81.9	79.3
	Yahoo	90.3	90.9	96.5	97.7	97.5	87.6	85.4	83.8	81.5
	Average	89.9	91.9	95.4	96.8	97.3	87.9	85.3	82.6	80.5
30	LFW	84.7	87.2	89.8	94.5	92.8	83.6	78.5	79.3	75.8
	IMFDB	82.4	85.3	90.8	92.2	94.6	81.8	80.2	78.2	76.5
	Yahoo	81.6	84.9	91.4	93.3	93.9	80.5	79.6	77.5	74.2
	Average	82.9	85.8	90.6	93.3	93.7	81.9	79.4	78.3	75.5
40	LFW	79.6	80.2	82.3	84.5	86.8	81.7	76.5	73.3	69.5
	IMFDB	78.3	81.4	82.6	83.1	85.6	80.4	74.7	70.3	70.4
	Yahoo	76.1	80.7	83.1	82.2	83.8	79.3	71.2	72.1	68.6
	Average	78	80.7	82.6	83.2	85.4	80.4	74.1	71.9	69.5
50	LFW	74.7	78.6	80.5	82.4	84.9	72.8	70.4	68.6	66.4
	IMFDB	75.2	77.4	79.3	83.8	85.7	78.5	69.5	67.8	67.2
	Yahoo	74.7	79.2	81.4	81.6	83.9	74.3	71.5	69.3	64.8
	Average	74.8	78.4	80.4	82.6	84.8	75.2	70.4	68.5	66.1

Table V shows that the proposed method produces enhanced results only when the number of suggested labels is less than 6, and decreases progressively as the label size changes from 7 to 10. When the suggested list size ranges from 2 to 6 with a matching score of 10%, the maximum probability of suggested labels for each instance compensates for imbalances in labelling. Consequently, when the suggested list size ranges from 2 to 6 with a matching score of 20%, reliable extra labels for each instance are generated and the annotation performance improved, because of a high probability that. the list of suggested names belongs to the unknown test face. On the contrary, when the number of extra labels ranges from 7 to 10 for each instance, it degrades the performance of the proposed method, and the lower accuracy obtained as a result is noted in Table V. Thus, it is concluded that the number of suggested list sizes is set to a value of 6 with a matching score level of 20%. The suggested list creation procedure enhances the reliability of the SMNLR by building a number of extra labels for each instance, using the fused similarity matrix.

E. Importance of Label Refinement in Face Annotation

In real-life, name labels of faces are incorrect or imperfect, stemming from the manual annotation of online applications. Making face annotation much more reliable by using noise labels is a major issue for real-time face annotation systems. Hence it is essential to refine the label ambiguity of faces without the loss of original labels. To this end, this experiment is conducted to validate the label quality before and after the label refinement process using the TPR, FPR, and accuracy. The values of each are shown in Table VI.

TABLE VI. An Evaluation of the Noise Label Refinement Capability for the Proposed SMNLR Face Annotation Method

Datasets	Without label refinement			With label refinement		
	TPR (%)	FPR (%)	Accuracy (%)	TPR (%)	FPR (%)	Accuracy (%)
LFW	0.60	0.50	0.58	0.97	0.1	0.96
IMFDB	0.52	0.45	0.52	0.96	0.03	0.96
Yahoo News	0.63	0.62	0.54	0.98	0.05	0.97
WDB	0.46	0.55	0.45	0.95	0.1	0.93
Average	0.55	0.53	0.52	0.96	0.07	0.95

The labeling accuracy between noisy labeled faces and refined labeled faces is compared using the TPR, FPR, and accuracy, which reveals contrary results. Table VI proves that the annotated faces with noise label refinement have a high TPR, accuracy value and a low FPR. Further, it clearly reveals that the proposed face annotation method is most reliable and robust.

F. Performance Analysis of the Proposed Face Annotation Method for Different Real-time Challenges

Real-time challenges in face images are commonly a challenge for face annotation. Annotating challenging face images is a difficult task in computer vision, and considerably affects classification and labeling performance. Hence the effectiveness of the proposed SMNLR method is analysed by performing this experiment on expression, occlusion, illumination and pose challenges, using the LFW, IMFDB and Yahoo databases. Table VII displays the performance for SMNLR face annotation against different real life challenging faces.

Table VII clearly shows that the SMNLR method has produced better results for real-time challenges. This is because more than one convolutional filter in the CNN can generate more useful and essential features from the significant facial components such as spatial local contrast, frequency descriptions and orientation properties. In addition to that, the convolution filters use the edges, gradients, directions and corner extraction techniques to obtain more complex features of face image and it overcomes the real-time challenges. However, when compared to normal face recognition, the recognition rate for challenging faces is slightly reduced in terms of expression and occlusion. Since the intrinsic feature information between pixels is not fully extracted from faces, and consequently produces a lower recognition rate.

TABLE VII. A performance Evaluation of the Proposed Face Annotation Method for Real-time Challenges

Datasets	Real-time challenges	Performance Metrics		
		Precision (%)	Recall (%)	Accuracy (%)
LFW	Normal	98.3	96.4	94.4
	Expression	90.5	91.9	86.6
	Illumination	93.6	94.2	90.4
	Occlusion	86.6	84.5	79.7
IMFDB	Normal	97.3	94.4	96.4
	Expression	91.5	90.8	82
	Illumination	95.6	93.3	89.4
	Occlusion	83.7	80.5	79.7
Yahoo	Normal	97.5	96.3	96.8
	Expression	90.5	89.7	81.4
	Illumination	92.7	91.4	88.1
	Occlusion	98.7	97.4	96.8

G. Performance Comparison of the Proposed Face Annotation Method With Existing Methods

To compare and evaluate the effectiveness of the proposed face annotation method with other state-of-the-art-methods, recall and error rate results are displayed in Table VIII. LFW and Yahoo are the most commonly used universal datasets in the face naming community, and are considered for a comparison with all other methods. In all, 4000 samples for training and 3000 samples for testing are taken from LFW, while 5500 samples for training and 4650 samples for testing are taken from the Yahoo dataset. Table VIII shows the experimental results for both the LFW and Yahoo datasets.

TABLE VIII. A comparison of the Proposed Face Annotation Method With State-of-the-art Methods

Face Annotation Methods	LFW		Yahoo	
	Recall (%)	Error rate (%)	Recall (%)	Error rate (%)
Chen's method [3]	78.5	19.4	71.3	21.7
Zeng's method [31]	65.7	23.2	64.1	27.4
Cour's method [32]	74.3	22.5	78.5	23.1
Liu's method [33]	88.1	9.1	90.2	8.4
Su's method [1]	83.4	14.6	80.6	12.9
Kumar's method [35]	90.6	9.2	89.8	10.3
Proposed SMNLR method	97.2	1.9	96.9	2.4

Table VIII clearly demonstrates that the proposed SMNLR method has produced significant results, when compared to state-of-the-art methods. This is because the fused similarity matrix obtains efficient similarity measures between faces with associated noise labels, and eliminates noise labels significantly when resolving the label refinement task. The error rate of the proposed method is also much lower than all other methods. The recall values of the methods of Chen et al. and Cour et al. indicate that their face naming performance is slightly worse than all other methods. The method advanced by Zeng et al. provides a lower recall value and higher error rate because their procedure fails to effectively handle noise labels and other irrelevant information, which impacts annotation results. The methods recommended by Liu et al. and Su et al. achieve recall rates of up to 83.4% and 90.2% respectively. In the methods above, most label ambiguity issues are resolved, and improved results are achieved by comparing them to the methods of Chen et al., Cour et al. and Zeng et al. That's Kumar et al. The method propounded by Kumar et al. produces slightly better results than all other methods, because they employed Convolutional neural networks for feature extraction. Table VIII proves that the proposed SMNLR method outperforms other related state-of-the-art-methods.

VII. Conclusion

In this paper, a new method named as Similarity Matrix based Noise Label Refinement (SMNLR) is proposed for face annotation. Two different similarity matrices can be acquired from first and second similarity matrix learning schemes respectively. In addition, these two matrices are fused to distinguish the uniqueness of faces. Generally, noise labels are refined by using cluster based approaches. On the contrary to existing methods, the proposed SMNLR method effectively exploits the noise label refinement approach for resolving the ambiguity of labels. Since SMNLR is proficient of exploiting the essential minimum distance value representation of faces, it is effective to identify variations within faces. It is noted that the proposed method produced significant results under different level of noisy labeled facial images. It is also observed that the CNNs deep feature offers improved results for annotation. Further, it makes the suggested labels list to overcome the problem of labeling the face that is not occurred in training set. The extensive experiments have been conducted to validate the proposed method using the three databases, such as IMFDB, LFW and Yahoo. The noise labels are synthesized on these three datasets. Moreover, the proposed SMNLR method outperforms various state-of-the-art methods. Finally, it is concluded that the similarity measurements based label refinement approaches can effectively handle the ambiguously labeled facial images for face annotation.

References

[1] X. Su, J. Peng, X. Feng, and J. Wu, "Labeling faces with names based on the name semantic network," *Multimedia Tools and Applications*, vol. 75, no. 11, pp. 6445-6462, 2016.

[2] D. Wang, S. C. Hoi, Y. He, J. Zhu, T. Mei, and J. Luo, "Retrieval-based face annotation by weak label regularized local coordinate coding," *IEEE transactions on pattern analysis and machine intelligence,* vol. 36, no. 3, pp. 550-563, 2014.

[3] C. H. Chen, V. M. Patel, and R. Chellappa, "Learning from ambiguously labeled face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp.1653-1667, 2017.

[4] S. C. Huang, M. K. Jiau, and Y. H. Jian, "Optimisation of automatic face annotation system used within a collaborative framework for online social networks," *IET Computer Vision,* vol. 10, no. 5, pp. 351-360, 2016.

[5] D. Wang, S. S. Hoi, and Y. He, "A unified learning framework for auto face annotation by mining web facial images," *Proceedings of the 21st ACM international conference on Information and knowledge management,* 2012, pp. 1392-1401.

[6] D. Wang, S. C. Hoi, Y. He, and J. Zhu, "Mining weakly labeled web facial images for search-based face annotation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 1, pp.166-179, 2012.

[7] G. Gao, M. Xu, J. Shen, H. Ma, and S. Yan, "Cast2face: assigning character names onto faces in movie with actor-character correspondence," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 26, no. 12, pp.2299-2312, 2015.

[8] M. Zhu, N. Wang, X. Gao,J. Li J, and Z. Li, "Face Photo-Sketch Synthesis via Knowledge Transfer," *In Proceedings of the Twenty-Eight International Joint Conference on Articial Intelligence (IJCAI),* 2019, pp. 1048-1054.

[9] B. Shikha, P. Gitanjali, and D. Pawan Kumar, "An Extreme Learning Machine-Relevance Feedback Framework for Enhancing the Accuracy of a Hybrid Image Retrieval System," *International Journal of Interactive Multimedia & Artificial Intelligence,* vol. 6, no. 2, pp. 15- 27, 2020, doi: 10.9781/ijimai.2020.01.002.

[10] H. Ito, and H. Koshimizu, "Face image retrieval and annotation based on two latent semantic spaces in fiars," *In Eighth IEEE International Symposium on Multimedia (ISM'06),* 2006, pp. 831-836.

[11] A. Kasthuri, and A. Suruliandi, "A survey on face annotation techniques," *In 4th International Conference on Advanced Computing and Communication Systems (ICACCS),* IEEE, 2017, pp. 1-9.

[12] Y. Yang, Y. Liu, and J. Liu, "Automatic face image annotation based on a single template with constrained warping deformation," IET Computer Vision, vol. 7, no. 1, pp.20-28, 2013.

[13] J. Zhu, S.C. Hoi, and M. R. Lyu, "Face annotation using transductive kernel fisher discriminant," *IEEE Transactions on Multimedia,* vol. 10, no. 1, pp.86-96, 2007.

[14] H. Zitouni, M. F. Bulut, and P. Duygulu, "Recognizing faces in news photographs on the web," *In 2009 24th International Symposium on Computer and Information Sciences, IEEE,* 2009, pp. 50-55.

[15] G. B. De Souza, D. F. da Silva Santos, R.G. Pires, A. N. Marana, and J. P. Papa, "Deep texture features for robust face spoofing detection," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 64, no. 12, pp.1397-1401, 2017.

[16] K. Tang, X. Hou, Z. Shao, and L. Ma, "Deep feature selection and

projection for cross-age face retrieval," *In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2017, pp. 1-7.

[17] A. Kasthuri, A. Suruliandi, and S. P. Raja, "Gabor-oriented local order feature-based deep learning for face annotation," *International Journal of Wavelets, Multiresolution and Information Processing, vol.* 17, no. 05, p. 1950032, 2019, doi.org/10.1142/S0219691319500322.

[18] K. Anburajan, S. Andavar, and P. Elango, "An Empirical Evaluation of Name Semantic Network for Face Annotation," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 13, no. 4, pp.557-571, 2020.

[19] V. Rihani, A. Bhandari, and C. P. Singh, "Face Recognition Using Convolution Filters and Neural Networks," *In IC-AI,* 2006, pp. 185-190.

[20] N. Venkatesh, M. Subhransu, and R. Manmatha, "Automatic image annotation using deep learning representations", *In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM* , 2015.

[21] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang, "Large scale automatic image annotation based on convolutional neural network," *Journal of Visual Communication and Image Representation*, vol. 49, pp.213-224, 2017.

[22] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," *In 2016 IEEE winter conference on applications of computer vision (WACV)*, 2016, pp. 1-8.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp.2884-2896, 2018.

[24] M. Khari, A. K. Garg, R. G. Crespo, and E. Verdú, "Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 7, p. 22, 2019, doi: 10.9781/ijimai.2019.09.002.

[25] M. S. Maheshan, B. S. Harish, and N. Nagadarshan, "A Convolution Neural Network Engine for Sclera Recognition," *International Journal of Interactive Multimedia & Artificial Intelligence,* vol. 6, no. 1, pp. 78-83, 2020, doi: 10.9781/ijimai.2019.03.006.

[26] L. Celona, S. Bianco, and R. Schettini, "Fine-grained face annotation using deep multi-task CNN," *Sensors,* vol. 18, no. 8, p. 2666, 2018.

[27] W. Jiang, and W. Wang, "Face detection and recognition for home service robots with end-to-end deep neural networks," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 2232-2236

[28] X. Sun, P. Wu, and S.C Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing,* vol. 299, pp.42-50, 2018.

[29] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3767-3780, 2019.

[30] O. C. Kurban, T. Yildirim, and A. Bilgiç, "A multi-biometric recognition system based on deep features of face and gesture energy image," In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA,)* 2017, pp. 361-364

[31] Z. Zeng, S. Xiao, K. Jia, T. H. Chan, S. Gao, D. Xu, and Y. Ma, "Learning by associating ambiguously labeled images," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2013, pp. 708-715.

[32] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels", *The Journal of Machine Learning Research*, vol. 12, pp. 1501-1536, 2011.

[33] X. Liu, S. Li, M. Kan, S. Shan, and X. Chen, "Self-error-correcting convolutional neural network for learning with noisy labels," *In 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* 2017, pp. 111-117.

[34] X. Su, J. Peng, X. Feng, J. Wu, J. Fan, and L. Cui, "Cross-modality based celebrity face naming for news image collections," *Multimedia Tools and Applications,* vol. 73, no. 3, pp.1643-1661, 2014.

[35] V. Kumar, A. Namboodiri, and C.V Jawahar, "Semi-supervised annotation of faces in image collection," *Signal, Image and Video Processing,* vol. 12, no. 1, pp.141-149, 2018.

[36] V. Maihami, and F. Yaghmaee, "Automatic image annotation using community detection in neighbor images," *Physica A: Statistical Mechanics and its Applications, 507,* 2018, pp.123-132.

[37] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," *In Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3574-3580.

[38] M. Zhu, J. Li, N. Wang and X. Gao, "A Deep Collaborative Framework for Face Photo–Sketch Synthesis," In *IEEE Transactions on Neural Networks and Learning Systems* vol. 30, no. 10, pp. 3096-3108, 2019.

[39] X. Wang, S. Wang, J. Wang, H. Shi, and T. Mei, "Co-mining: Deep face recognition with noisy labels," *In Proceedings of the IEEE international conference on computer vision,* 2019, pp. 9358-9367.

[40] S.C. Hoi, D. Wang, I.Y. Cheng, E.W. Lin, J. Zhu, Y. He, and C. Miao, "Fans: face annotation by searching large-scale web facial images," *In Proceedings of the 22nd international conference on World Wide Web, ACM*, 2013, pp. 317-320.

[41] S. C. Huang, M. K. Jiau and C. A. Hsu, "A High-Efficiency and High-Accuracy Fully Automatic Collaborative Face Annotation System for Distributed Online Social Networks," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 24, no. 10, pp. 1800-1813, 2014.

[42] J. Han, J. Hu, and W. Deng, "Constrained Spectral Clustering on Face Annotation System," *In: Chinese Conference on Pattern Recognition,* Springer, Singapore*,* 2016, pp. 3-12.

[43] S. Xiao, D. Xu, and J. Wu, "Automatic face naming by learning discriminative affinity matrices from weakly labeled images," *IEEE transactions on neural networks and learning systems,* vol. 26, no. 10, pp.2440-2452, 2015.

[44] D. Wang, and A. K. Jain, "Face retriever: Pre-filtering the gallery via deep neural net," *In International Conference on Biometrics (ICB), IEEE,* 2015, pp. 473-480.

[45] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," *In Proceedings of the IEEE International Conference on Computer Vision,* 2013, pp. 2408-2415.

[46] H. V. Nguyen, and L. Bai, "Cosine similarity metric learning for face verification," *In: Asian conference on computer vision, Springer, Berlin, Heidelberg*, 2010, pp. 709-720.

[47] Y. Ji, T. Lin, and H. Zha, "Mahalanobis distance based non-negative sparse representation for face recognition," *In International Conference on Machine Learning and Applications, IEEE,* 2009, pp. 41-46.

[48] E. Mostafa, A. M. Ali, and A. A. Farag, "Learning a non-linear combination of Mahalanobis distances using statistical inference for similarity measure," *IET Computer Vision*, vol. 9, no. 4, pp.541-548 2015

[49] M. Wang, Z. Wang, and J. Li, "Deep convolutional neural network applies to face recognition in small and medium databases," *In 4th International Conference on Systems and Informatics (ICSAI), IEEE*, 2017, pp. 1368-1372.

A. Suruliandi

A. Suruliandi completed his B.E. in Electronics & Communication Engineering in the year 1987 from Coimbatore Institute of Technology, Coimbatore. He completed his M.E. in Computer Science & Engineering in the year 2000 from Government College of Engineering, Tirunelveli. He obtained his Ph.D. in the year 2009 from Manonmaniam Sundaranar University, Tirunelveli. He is working as a professor in the Department of Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli. He is having more than 29 years of teaching experience. He published 50 papers in International Journals, 23 in IEEE Xplore publications, 33 in National conferences and 13 in International conferences. His research areas are remote sensing, image processing and pattern recognition.

A. Kasthuri

A. Kasthuri received her M.Sc Degree in Computer Science & Information Technology from Kamaraj University, Tamilnadu in 2012. She received her M.Phil Degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli in 2015. She is currently pursuing Ph.D Degree in Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu. Her Research interest includes Image Processing, Face Recognition, Person re-identification, Pattern Recognition.

S. P. Raja

S. P. Raja was born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He is having more than 14 years of teaching experience in engineering colleges. Currently he is working as an Associate Professor in the department of Computer Science and Engineering in Vellore Institute of Technology, Vellore. He published 38 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence.