Face Recognition in Videos by Label Propagation

Vijay Kumar, Anoop M. Namboodiri, C.V. Jawahar
International Institute of Information Technology, Hyderabad, India
{vijaykumar.r@research., anoop@, jawahar@}iiit.ac.in

***Abstract*—We consider the problem of automatic identification of faces in videos such as movies, given a dictionary of known faces from a public or an alternate database. This has applications in video indexing, content based search, surveillance, and real time recognition on wearable computers. We propose a two stage approach for this problem. First, we recognize the faces in a video using a sparse representation framework using l_1-minimization and select a few key-frames based on a robust confidence measure. We then use transductive learning to propagate the labels from the key-frames to the remaining frames by incorporating constraints simultaneously in temporal and feature spaces. This is in contrast to some of the previous approaches where every test frame/track is identified independently, ignoring the correlation between the faces in video tracks. Having a few key frames belonging to few subjects for label propagation rather than a large dictionary of actors reduces the amount of confusion. We evaluate the performance of our algorithm on Movie Trailer face dataset and five movie clips, and achieve a significant improvement in labeling accuracy compared to previous approaches.**

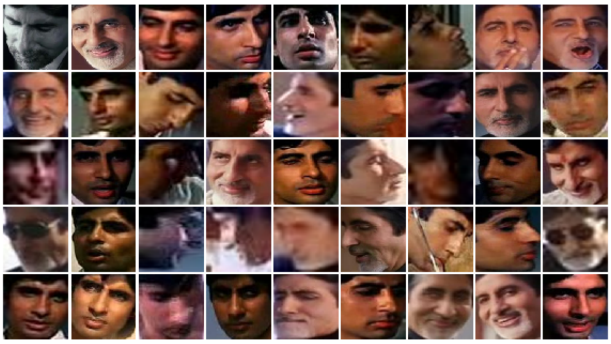


Fig. 1: Example faces from IMFDB of the Indian actor *Amitabh Bachchan* collected from different movies. Faces are detected and cropped manually from a large collection of movies selected at different stages of actor's career. Notice the large variations in expression, occlusion, pose and age.

I. Introduction

Face recognition in unconstrained settings still remains an unsolved problem even after decades of focused research. This is primarily due to large variability exhibited by faces in terms of occlusion, pose, expression, scale and sensors. After making a steady progress in the constrained settings with limited pose, expression, occlusion, and illumination variations, the research is currently focused on recognizing faces in unconstrained (popularly known as *in the wild*) settings. The interest is further strengthened by the introduction of new unconstrained datasets such as LFW [1], YouTube Celebrities [2], and PubFigs [3], which are harvested from the Internet. Ever since their introduction, there has been a significant progress in improving the verification and recognition performance.

The above mentioned advances have also triggered the interest for face recognition in unconstrained videos [4], [5]. In this work, we consider one such application; that of identifying actors in movies. Specifically, we are given a large dictionary of actors, and the objective is to label the actors in a specific movie. Identification of actors in movies can help in many applications such as video indexing, actor-specific scene retrieval, etc. However, the problem is very challenging due to large variations in appearance, pose, facial expressions, occlusions and camera motion. Some of the recent methods [4], [5] approach the problem by identifying individual faces and face tracks independently, while ignoring the large amount of unlabeled faces and the correlation that exists among the faces and tracks in a video.

We approach this problem with two important observations. The first is related to the appearance of actors and the second concerns the number of actors in a movie. The actor's appearance is usually consistent across most parts of a movie and there could be multiple face tracks where the actor has similar appearance. Moreover, the change of pose and expression in movie shots are not sudden but smooth. Hence, if the frames that are consistent with the dictionary are labeled correctly, remaining frames can be labeled through propagation. We also note that the number of main (lead) actors in a movie is typically low, usually less than 10. Hence if a small labeled seed set from the movie is available, labels can be propagated efficiently with fewer confusions than having a dictionary with large number of subjects.

We present an approach for face recognition in movies given a large labeled dictionary of faces. The first step is to identify all the detected faces in a video with the best available face recognition techniques such as Sparse Representation based Classifier (SRC) [6]. We then retain the labels of those key-frames that are highly confident based on a robust Sparsity Concentration Index (SCI) [6]. In the second stage, we consider only the faces from the given video and treat the key-frames as labeled and remaining frames as unlabeled. We cast the problem as a transductive semi-supervised learning and propagate the labels from the labeled key-frames to remaining frames. We consider the similarities in feature as well as temporal space while propagation, thus effectively exploiting the correlation among faces within and between tracks. This is in contrast to a recently introduced methods such as [4], where the average face in a track is labeled independently. Our approach will be highly suitable for offline annotation of movies, trailers, etc.

A. Related Work:

There have been many attempts to identify faces in videos in the last few years. These methods can be roughly grouped into three categories: key-frame based, temporal model based and image-set based approaches (see [7] for a complete overview). Key-frame based approaches [8]–[10] rely on single image face recognition by treating video as a collection of images and performing recognition on all or a set of selected frames. Instead of recognizing every frame, only a set of key frames that are of *good* quality or suitable for recognition are selected based on several heuristics. Several approaches to select good quality frames were proposed including relative positions of eyes and nose [9], robust statistics to filter out noisy face images [10], etc. Once the key frames are identified for every track, majority voting schemes are used to finally label the entire video.

Temporal model based approaches [11] take into account the correlation between consecutive frames in a video. They model the face dynamics such as non-rigid facial expressions and rigid head movements to learn how individual faces vary through a video sequence. On the other hand, Image-set based approaches [12] consider the face tracks as image sets and model the distributions of face images in each set and compare the similarity of distributions for recognition. They do not consider the temporal coherence between the frames addressing the cases where temporal sequence of faces is not available due to limitations of detection, etc.

Our approach has some resemblance with key-frame and temporal based approaches. However, unlike existing key-frame based approaches that focuses on selecting good quality images, our approach selects the images that are labeled confidently. We do not model any face dynamics as opposed to temporal coherence based approaches but only consider the temporal proximity while propagating the labels. We also make a mention of few other attempts in identifying characters in sitcoms where additional clues such as clothing and audio [5], [13], and relations between characters [14] are exploited. [14] employs a semi-supervised scheme and uses the weakly labeled data by aligning the subtitles with speaking face track. Note that speaking detection itself is a challenging problem and may not be extended for cases without subtitles.

II. Building a Dictionary of Movie Actors

Inspired by LFW [1] and PubFigs [3], a new face database *Indian Movie Face Database (*IMFDB*)* is introduced recently to further promote the face recognition research in unconstrained settings. The database consists of $34,512$ face images of 100 Indian actors collected from approximately 103 video clips and movies. It includes 67 male and 33 female actors with at least 200 images for each actor. This movie face database could help in characterizing people, activities and retrieving shots by aligning the scripts as done in the past [15].

IMFDB[1] is created from a large and diverse collection of Indian movies to capture as much variations as possible. The faces in movies have rich variety of resolution, illumination, age, makeup, pose and expression. It is created through a careful selection and extraction of faces from the movies. First, movies were selected at different stages of every actor's career in order to account for age variations. They belong to five different Indian languages namely, Hindi, Telugu, Kannada, Malayalam and Bengali to have a diverse appearance of actors. Also, the movies were from a large time period ($1970 - 2012$), thereby including variations in image quality, recording medium, resolution, and illumination. From each video, faces are detected manually without relying on face detectors. The pose of actors will have extreme variations in movies corresponding to variety of shots in songs, action sequences, etc. Though the manual detection of faces incurred cost and significant human effort, it resulted in a very rich set of pose variations producing new and challenging poses, which may be difficult to detect for current state-of-the-art detectors. Once the frames are detected, only few diverse set of images were selected for each movie to avoid any similarity among images. Fig 1 shows a few images of actor *Amitabh Bachchan* from the database showing rich diversity in pose, expression, illumination, occlusion, age, resolution and quality. IMFDB provides detailed annotation for each actor in terms of pose, expression, makeup, age, occlusion and illumination.

IMFDB differs from existing unconstrained databases such as LFW [1] and PubFigs [3] in the following aspects. While the images in these databases are collected from Internet sources such as Yahoo news, IMFDB is created from movies, thus including larger variations in pose, illumination, and resolution. Images of the celebrities collected from Internet will have similar public appearance giving a very small set of variations. Also, since the images are collected through a search query, retrieved results may not be diverse and have limited age variations. Public figures and celebrities often *retain* their identity (appearance, dress patterns, expressions) when they are in public leading to a constrained variations while actors in movie can have a large variety of expressions, pose and appearance. IMFDB also includes large age variations through a careful selection of movies. More details can be found in our earlier work [16].

III. Automatic Face Identification in Movies

Given a labeled dictionary of actors, we wish to identify the faces in a movie. The proposed method (Fig 2) consists of two stages. In the first stage, referred to as key-frame selection, we label all the faces (see Section IV-D for a discussion) in a movie using highly successful Sparse Representation based face recognition algorithm. We retain the labeling of a few highly confident key-frames measured through a Sparsity Concentration Index (SCI) [6]. In the second stage, we propagate the labels from the key frames to remaining frames incorporating constraints in the temporal and feature space. Our approach is based on the intuition that, certain faces in the movie may have similar appearance, pose, with the faces in the dictionary and are labeled with high degree of confidence. If one could identify such labellings, then one can propagate the labels from the selected key frames to the remaining frames effectively. This is particularly helpful for scenes involving zoom in or zoom out of actors face or in scenes where there is a gradual pose change (Fig 3).

Given a labeled dictionary $\mathbf{D} = [D_1, D_2, \ldots, D_c]$ with D_i containing training examples belonging to class $i \in \{1, 2, \ldots, c\}$, we need to annotate a set of N faces $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ present in a video. Throughout this paper, we assume that each of the columns of D and X have a unit l_2-norm.

[1] http://cvit.iiit.ac.in/projects/IMFDB/

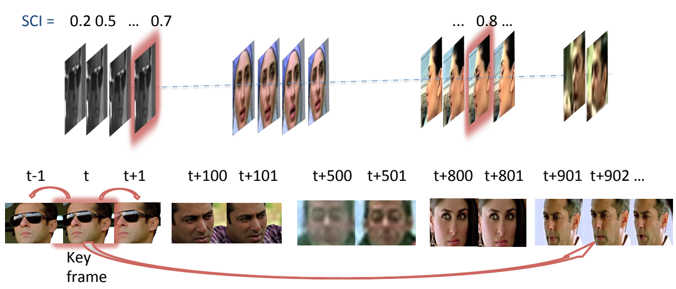


Fig. 2: Overview of our approach. We label all the frames independently in a movie using SRC and select only highly confident key frames based on SCI. The labels of the key frames are then propagated to remaining frames by imposing the constraints in time and feature space.

A. Key-Frame Selection:

We use the highly successful still-image face recognition algorithm, Sparse Representation based Classifier (SRC) [6] for initially labeling the video frames using the labeled dictionary. SRC represents each image as a *sparse* linear combination of dictionary faces. This is based on the notion that face images lie on a low-dimensional manifold and given a sufficient number of training examples, they span this low dimensional space [17]. Any face thus can be represented as a linear combination of training samples from its class. When there are large number of classes, this representation will be *sparse*. The sparse solution is obtained by solving the following l_1 minimization, which gives sparse solutions under certain conditions.

$$\hat{\alpha_i} = \arg\min_{\alpha_i} ||X_i - D\ \alpha_i||_2 + \lambda||\alpha_i||_1 \qquad (1)$$

where λ is a Lagrangian constant which controls the trade-off between reconstruction error and sparsity. The sample X_i is finally assigned the label of a class of training samples that best reconstructs the sample.

$$\text{label}(X_i) = \arg\min_{j} ||X_i - D_j\ \hat{\alpha_{ij}}||_2 \qquad (2)$$

where D_j are the training samples belonging to class j and α_{ij} are the corresponding weights.

Once all the frames are labeled using the above method, we select the key frames based on confidence score given by Sparsity Concentration Index (SCI) [6] defined as

$$SCI(i) = \frac{c \cdot \max_j ||\hat{\alpha_{ij}}||_1/||\hat{\alpha_i}||_1 - 1}{c-1} \qquad (3)$$

where c denote the number of classes. This score gives a measure of how well a query sample is represented using the samples from a particular class. SCI =1 indicates that the query is represented from only a single class while SCI = 0 indicate that weights are spread uniformly across all the classes. We use this score as a confidence measure to consider the labeling of a face. This is based on a intuition that faces that have similar appearance and pose as that of dictionary elements will have a high SCI while other faces that have different pose, appearance will have low SCI (Fig 3). By identifying such key frames, labels can be propagated to remaining frames effectively.

B. Propagating the Key-Frames

Once the key frames are selected, their labels are propagated to the entire video using label propagation framework [18]. This approach has following advantages. First, since a dictionary contains a large number of subjects, the chance for confusion is high. However, the number of actors in a movie is typically less. Hence, if we could label few frames belonging to set of subjects confidently, then these labels can be propagated through out the video effectively with less confusion. Moreover, the appearance of the actors in different frames of video is usually similar making propagation simpler compared to labeling them independently.

In this step, we consider the key frames selected in the first step as labeled set X_l and remaining frames as unlabeled X_u. Let $\mathbf{X} = [X^l\ X^u] \in \mathbb{R}^{d\times N}$ be the data matrix and $F \in \mathbb{R}^{N\times c}$ be a non-negative matrix with each row corresponding to a data point. F can be treated as a scoring function that indicate how likely a data point belong to particular class. The label of a face X_i, $\{i = 1, 2, .., N\}$ can be obtained from F as $y_i = \arg\max_j F_{ij}$, where $j = \{1, 2, \ldots, c\}$. Let $Y \in \mathbb{R}^{N\times c}$ denote the initial labeling matrix. For the labeled dictionary faces, we define $Y_{ij} = 1$ if $y_i = j$ and 0 otherwise. For unlabeled faces, we assign $Y_{ij} = 0\ \forall j$ where j ={1, 2,..., c}.

Given X, we construct an undirected graph $\langle V, E\rangle$ using both labeled and unlabeled points. Each node in the graph corresponds to a face and the edges E between them represent similarity. Larger the edge weight, higher the similarity between the faces. In this work, we consider two kinds of similarities - one in feature and other in time space. Faces that are closer in temporal space should belong to same class and hence have large weights. This is achieved as follows,

$$\sum_{i,j} V_{ij}||F_i - F_j||^2 \qquad (4)$$

where $V_{ij} = \exp(-(t_i - t_j)\gamma_{ij}/2\sigma^2)$. t_i and t_j denote the frame number of i-th and j-th frames and γ_{ij} denote absolute sum of difference of co-ordinate of i-th and j-th frame. Intuitively, above constraints indicate that neighboring frames with similar bounding boxes for faces should belong to similar class.

Similarly, the points similar in appearance (feature space) should have large weights. This can be achieved as,

$$\sum_{i,j} W_{ij}||F_i - F_j||^2 \qquad (5)$$

where W_{ij} denote the appearance weights. A common measure of appearance similarity is Gaussian function given as $W_{ij} = \exp(-||x_i - x_j||^2/2\sigma^2)$, where σ controls the spread of the Gaussian function. However, such a similarity measure based on Euclidean distance may not be accurate for face recognition as it is sensitive to illumination and expression variations. Instead, we obtain the weights using the nearest neighbor based sparse representation approach proposed in [19]. Here, weights are obtained by representing each face as a linear combination of its nearest neighbors, thereby preserving both sparsity [6] and locality [20]. This method encourages the creation of edges only with neighboring samples achieving effective propagation. We solve the below equation for constructing the appearance weights.

$$\hat{w}_i = \arg\min_{w_{ik}} ||x_i - \Sigma_{k:x_k \in \mathsf{N}(x_i)}\ x_k w_{ik}||_2 + \beta||w_i||_2$$
$$\text{s.t}\ \forall_k\ w_{ik} \geq 0 \quad (6)$$

where $\mathsf{N}(x_i)$ denote the k neighboring samples of x_i and β is a Lagrangian constant that controls the trade-off between two terms. Appearance weight matrix $W \in \mathbb{R}^{N \times N}$ is then constructed as:

$$W_{ij} = \begin{cases} \hat{w}_i(k), & \text{if } x_j \in \mathsf{N}(x_i) \\ 0, & \text{otherwise,} \end{cases}$$

$\hat{w}_i(k)$ denotes the k-th element of vector $\hat{w}_i$ corresponding to k-th neighbor. Weights obtained by this method may not be symmetric i.e $w_{ij} \neq w_{ji}$. We make the final weights symmetric with the operation: $w_{ij} = w_{ji} = (w_{ij}+w_{ji})/2$.

Using the above similarity measures, we finally propagate the labels by solving the objective function as,

$$Q(F) = \arg\min_{F_i} \frac{\gamma_1}{2} \sum_{i,j}^{N} V_{ij}||F_i - F_j||^2 + \frac{\gamma_2}{2} \sum_{i,j}^{N} W_{ij}||F_i - F_j||^2$$
$$+ \frac{\gamma_3}{2} \sum_{i}^{N} ||F_i - Y_i||^2 \quad (7)$$

Third term in the above equation ensures that labeling of the labeled set is not changed much from its initial labeling. First and second terms are equivalent to $F^T L_1 F$ and $F^T L_2 F$ where $L_1 = D^1 - V$ and $L_2 = D^2 - W$ and $D^1_{ii} = \sum_j V_{ij}$ and $D^2_{ii} = \sum_j W_{ij}$. Here, L_i is the Laplacian of a graph.

Thus, our objective function Eqn 7. reduces to

$$Q(F) = \arg\min_{F} \gamma_1 F^T L_1 F + \gamma_2 F^T L_2 F + \gamma_3 ||F - Y||^2 \quad (8)$$

Differentiating the quadratic function $Q(F)$ with respect to F and equating to 0 we get,

$$\gamma_1 L_1 F + \gamma_2 L_2 F + \gamma_3 (F - Y) = 0$$
$$(\gamma_1 L_1 + \gamma_2 L_2) F + \gamma_3 F - \gamma_3 Y = 0$$
$$F^* = \gamma_3 (\gamma_1 L_1 + \gamma_2 L_2 + \gamma_3)^{-1} Y \quad (9)$$

The labels of unlabeled samples can then be predicted using $y_i = \arg max_j F^*_{ij}$.

C. Rejecting unknown faces:

The video frames may consist of unknown faces that are not present in the dictionary. The algorithm should be able to reject the labeling of any such faces based on a confidence measure. We consider the ratio of two largest labeling scores as a confidence measure to accept or reject the labeling of a face [19]. Intuitively, the scoring vector F_i for a sample should have high score corresponding to true class and very less score to remaining classes indicating the contribution of a single class in the reconstruction of a sample. Based on this intuition, we define the labeling dominance score (LDS) that accepts or rejects the labeling of a face as follows.

$$\text{LDS (i)} = \frac{F_{ij}}{\arg \max\limits_{k, k \neq j} F_{ik}} \text{ where } j = \arg\max_j F_{ij} \quad (10)$$

We consider the annotation of a face when the gap between two largest scores F_{ij} is high.

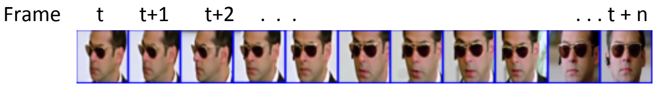


Fig. 3: A scene where there is a gradual change of pose of an actor. The dictionary may not contain all the poses of an actor. In such scenarios, even if few images in the pose transition that are similar to dictionary set are identified with high confidence, their labels can be propagated to remaining poses effectively.

IV. Experiments and Results

In this section, we initially analyze the complexity of IMFDB using existing state-of-the-art methods and then report the performance of our proposed approach on Movie Trailer dataset [4] and movie clips.

A. Supervised:

We evaluate the performance of k-nearest neighbor (KNN), Sparse representation based classifier (SRC) [6], Collaborative representation based classifier (CRC) [21] and Locality constrained linear coding (LLC) [20] on IMFDB. IMFDB consists of 34512 images belonging to 100 actors. Each actor has at least 200 images. For each actor, we selected 100 images for training and remaining images as testing. We resized all the images to 80×80. We extracted two kinds of features, dense SIFT and LBP with a block size of 20 and 8, respectively using VLFEAT library [22]. We further reduced the dimension of the features to 90 using PCA. The experiments are carried out 10 times with random training and test sets and average results are reported. Table I shows the performance of the state-of-the-art techniques on IMFDB. Notice that the performance of these methods on IMFDB is very low indicating the large variability of faces.

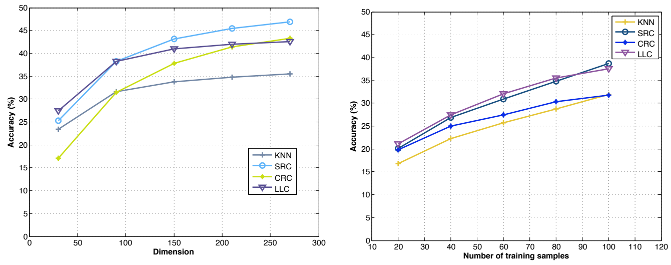
Fig 4 (a) and (b) shows the performance of various methods for various feature dimension and training examples using LBP features. It is clear that the performance is very low with less labeled examples and increases steadily as the training data is increased, indicating the inability of these methods to recognize samples with less training data.

TABLE I: Supervised recognition rates (mean ± std%) of various methods on IMFDB using LBP and SIFT.

Method	LBP	SIFT
KNN	31.70 ± 0.32	21.85 ± 0.16
SRC [6]	38.00 ± 0.57	23.82 ± 0.26
CRC [21]	31.30 ± 0.41	20.02 ± 0.24
LLC [20]	37.92 ± 0.34	26.05 ± 0.27

B. Semi-Supervised annotation of Isolated faces

Next, we show the effectiveness of our approach on IMFDB in a semi-supervised setting. Since the isolated faces in IMFDB do not have temporal information, we set γ_1 to 0. For this experiment, we consider a subset of 20 actors from IMFDB. We create a semi-supervised setting by selecting 10 randomly examples for each actor as labeled and remaining examples as unlabeled. We extract the LBP features with block size of 8 and reduce the dimension to 90 using PCA. We set the values of γ_2 and γ_3 to 0.7 and 0.3 respectively which gave best results. Table II shows the comparison of



(a) (b)

Fig. 4: Recognition rates of various methods for various (a) dimension of the feature space and (b) number of training examples for each actor using LBP features

our approach with the current state-of-the-art approaches. Our approach which considers the coherence within different tracks clearly achieves better performance even in the absence of temporal constraints. Table III shows the effect of various number of training examples using LBP features.

TABLE II: Semi-supervised recognition rates (mean ± std%) of various methods on IMFDB using LBP and SIFT.

Method	LBP	SIFT
KNN	20.45 ± 0.12	14.57 ± 0.15
SRC [6]	30.52 ± 0.22	20.16 ± 0.21
CRC [21]	24.54 ± 0.14	14.64 ± 0.36
LLC [20]	31.14 ± 0.41	20.61 ± 0.22
MSSRC [4]	30.52 ± 0.22	20.80 ± 0.21
Our approach	**37.34 ± 0.22**	**24.81 ± 0.31**

TABLE III: Semi-Supervised recognition rates [%] of various methods on IMFDB for different number of labeled examples using LBP.

Method	5 Train	10 Train	20 Train	30 Train	40 Train
KNN	12.1	20.45	32.01	38.80	40.12
SRC [6]	30.20	31.56	38.07	47.09	50.13
CRC [21]	28.65	24.54	30.43	41.24	43.28
LLC [20]	28.27	31.14	37.35	44.29	48.71
MSSRC [4]	30.20	31.56	38.07	47.09	50.13
Our approach	35.21	37.34	45.69	52.18	56.40

C. Annotation of faces in Videos

Movie Trailer Face Dataset: Movie Trailer Face Dataset [4] consists of features of 4485 face tracks from 101 movie trailers released in the year 2010. These trailers are collected from YouTube and contain the celebrities presented in the PubFig Dataset [3] along with additional 10 actors. The dataset contains only 35% of known actors from PubFigs+10 presenting a new scenario of rejecting unknown faces in videos. For details on detection and feature extraction, please refer to [4]. The labeled dictionary consists of 34522 images (PubFigs + 10 additional actors) with each actor having a maximum of 200 images.

Since the dataset does not contain bounding box and frame information, we set γ_1 to 0. We set $\gamma_2 = 0.3$ and $\gamma_3 = 0.7$ for which we obtained best results. For key-frame selection, we selected top 50% highly confident initial labels based on SCI. We use LDS explained in Section 3 for accepting/rejecting the final labeling of a face. As done in [4], we use SCI as a confidence measure for SRC, MSSRC and CRC algorithms, and distance for k-NN and SVM. Fig 5 shows the Precision-recall curves of various methods and accuracies are shown in Table IV. Our approach clearly achieves higher recognition accuracy and precision in rejecting unknown faces even in the absence of temporal information. We also conducted an experiment by considering only known actor tracks, results of which are shown in Table V. Our approach clearly outperforms previous approaches in identifying known actors by a large margin when a complete dictionary of actors is available.

TABLE IV: The performance of various methods (in %) in the presence of unknown actors.

Method	Accuracy	Average Precision
1-NN	23.60	9.53
SVM	54.68	50.06
CRC [21]	41.93	36.33
SRC [6]	47.78	54.33
MSSRC [4]	50.52	58.69
Our approach	**55.98**	**59.34**

TABLE V: Performance of various methods in identifying known actors

Method	Accuracy (%)
1-NN	23.60
SVM	54.68
CRC [21]	41.93
SRC [6]	47.78
MSSRC [4]	50.52
Our approach	**68.98**

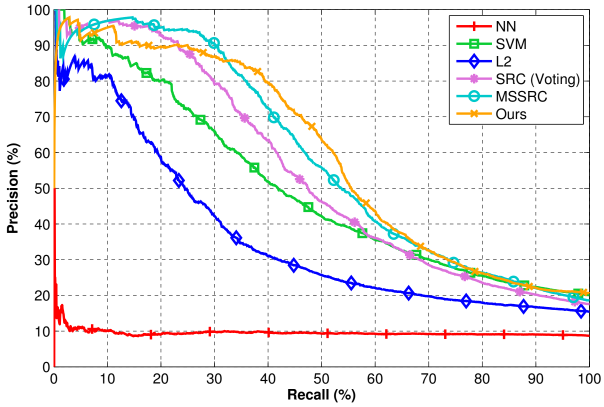


Fig. 5: Precision vs Recall of various methods on Movie Trailer Dataset. The performance of our approach in rejecting unknown actors is comparable to MSSRC.

Our test set: We created a test set of 5 movie clips of length ranging from 2 − 8 minutes from YouTube. Each movie clip has at least 2 actors from IMFDB. The faces are detected using Viola-Jones algorithm for frontal and profile views. Actors not in IMFDB are treated as unknowns.

We considered a subset of labeled dictionary with 20 actors from IMFDB. For each actor, we selected 50 images for training. To handle the large variations of faces, we extracted multiple features. Gabor features were extracted at two scales and four orientations. HOG and dense-SIFT features were extracted with a cell size of 8. Each of these features are then normalized and dimensionality reduced to 300 using PCA. For key-frame selection, we selected top 50% highly confident initial labels based on SCI. We selected the values of γ_1, γ_2 and γ_3 to 0.7, 1 and 0.3 respectively. For MSSRC [4], we created the tracks based on similarity of bounding box and LBP features. The performance of various approaches in identifying the known actors in the movie clips is shown in Table VI. Clearly, our proposed approach outperforms the existing methods including the recent state-of-the-art [4]. As can be seen in Figure 6, LDS is very effective in rejecting the unknown actors compared to other methods which use SCI. SCI is found not so effective for CRC and LLC which use l_2-minimization. Table VII presents our results on test movie clips in terms of average precision indicating superior results.

TABLE VI: Performance of various methods in identifying known actors

Method	Accuracy (%)
KNN	19.23
SRC [6]	34.20
CRC [21]	22.07
LLC [20]	33.22
MSSRC [4]	36.77
Our approach	**54.51**

TABLE VII: Our approach outperforms other methods in average precision by atleast 34%.

Method	Average Precision (%)
KNN	13.33
SRC [6]	39.60
CRC [21]	15.16
LLC [20]	43.85
MSSRC [4]	42.31
Our approach	**77.50**

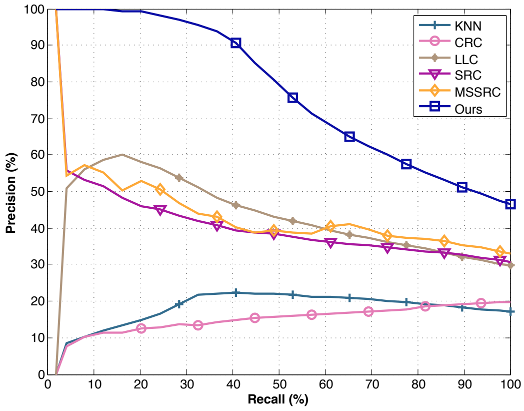


Fig. 6: Precision vs Recall of various methods. Our approach rejects the unknown actors better than other methods.

D. Discussion

During key-frame selection stage, labeling every frame and selecting few frames can be expensive. Instead, one can follow several heuristics such as labeling the average track as done in [4]. But the performance of such methods depend on face detectors and tracking. One can also stop the initial labeling after sufficient number of key-frames are available. Note that, unlike [4], we have not employed tracking of faces, but our temporal constraint makes sure that neighboring frames have similar labels achieving similar effect. This is more robust than assigning the label of average track face where an error in initial tracking may degrade the performance.

V. Conclusion

In this paper, we have presented an approach for identifying faces in movies given a labeled dictionary of actors. Our proposed approach involves two stages. In the first stage, all the frames in the movies are labeled using a sparse representation algorithm using a labeled dictionary of actors and only confident labellings are considered based on confidence measure. In the second stage, we propagate the labels from key-frames to remaining frames imposing constraints in the temporal and feature space. We finally showed that our method outperforms the recently proposed approaches on a movie clips in recognizing known actors at the same time achieving high precision in rejecting unknown actors.

Acknowledgment

This work is partly supported by the MCIT, New Delhi. Vijay Kumar is supported by TCS research fellowship.

References

[1] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. of Massachusetts, Tech. Rep., 2007.

[2] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley., "Face Tracking and Recognition with Visual constraints in real-world Videos," in *CVPR*, 2008.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.

[4] E. Ortiz, A. Wright, and M. Shah, "Face Recognition in Movie Trailers via Mean Sequence Sparse representation-based classification," in *CVPR*, 2013.

[5] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, ""knock! knock! who is it?" probabilistic person identification in tv-series," in *CVPR*, 2012.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *PAMI*, 2009.

[7] C. Shan, "Face recognition and retrieval in video," in *VSM*, 2010.

[8] M. Zhao, J. Yagnik, H. Adam, and D. Bau, "Large scale learning and recognition of faces in web videos," in *FG*, 2008.

[9] D. Gorodnichy, "On Importance of Nose for Face Tracking," in *FG*, 2002.

[10] Berrani S.A and Garcia C, "Enhancing face recognition from video sequences using robust statistics," in *AVSS*, 2005.

[11] G. Edwards, C. Taylor, and T. Cootes, "Improving identification performance by integrating evidence from sequences," in *CVPR*, 1999.

[12] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *ECCV*, 2002.

[13] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *IVC*, 2009.

[14] M. Buml, M. Tapaswi, and R. Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data." in *CVPR*, 2013.

[15] K. P. Sankar, C. V. Jawahar, and A. Zisserman, "Subtitle-free movie to script alignment." in *BMVC*, 2009.

[16] S. Shetty *et al.*, "Indian Movie Face Database: A Benchmark for Face Recognition under wide variations," in *NCVPRIPG*, 2013.

[17] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *PAMI*, 2001.

[18] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Tech. Rep., 2002.

[19] Vijay Kumar, Anoop. M. Namboodiri, and C. V. Jawahar, "Sparse Representation based Face Recognition with Limited Labeled Samples," in *ACPR*, 2013.

[20] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[21] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" *ICCV*, 2011.

[22] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2010.