SIViP
DOI 10.1007/s11760-017-1140-5

ORIGINAL PAPER

# Semi-supervised annotation of faces in image collection

Vijay Kumar[1] · Anoop Namboodiri[1] · C. V. Jawahar[1]

Received: 25 February 2017 / Revised: 26 May 2017 / Accepted: 26 June 2017

**Abstract** The objective of this work is to correctly detect and recognize faces in an image collection using a database of known faces. This has applications in photo-tagging, video indexing, surveillance and recognition in wearable computers. We propose a two-stage approach for both detection and recognition tasks. In the first stage, we generate a seed set from the given image collection using off-the-shelf face detection and recognition algorithms. In the second stage, the obtained seed set is used to improve the performance of these algorithms by adapting them to the domain at hand. We propose an exemplar-based semi-supervised framework for improving the detections. For recognition of images, we use sparse representation classifier and generate seed images based on a confidence measure. The labels of the seed set are then propagated to other faces using label propagation framework by imposing appropriate constraints. Unlike traditional approaches, our approach exploits the similarities among the faces in collection to obtain improved performance. We conduct extensive experiments on two real-world photo-album and video collections. Our approach consistently provides an improvement of ∼4% for detection and 5−9% for recognition on all these datasets.

**Keywords** Face annotation · Semi-supervised framework · Image collection · Label propagation

## 1 Introduction

Human faces are the most important entities in images and videos, consequently their detection and recognition are important for many commercial and consumer applications. A lot of progress has been made independently for these tasks. The current-day algorithms [10,12,17,19] already achieve impressive performance on various detection and recognition benchmarks. These algorithms are mostly aimed at general problem of detection and recognition of individual target face instances. However, there are certain consumer applications such as face tagging in albums and video indexing that require annotation of “collection of faces” appearing in multiple images. In such scenarios, it is beneficial to develop algorithms that leverage the presence of multiple collection instances to achieve superior performance.

In this work, we consider the problem of automatic annotation of faces in an image collection given a large dictionary of subjects. By an image collection, we mean a set of images with multiple instances of a limited number of subjects. For instance, a family photo-album containing multiple photographs of small number of people taken at different events. Our approach detects and recognizes faces in image collection in two stages as shown in Fig. 1. In the first stage, a set of seed instances are identified from the image collection using off-the-shelf face detection and recognition algorithms. In the second stage, the seed images obtained from the target domain are used to improve the performance of these algorithms in a semi-supervised framework. Our approach exploits the similarities among instances and is suitable for offline applications.

For face detection, we use exemplar detector [10,22] due to its superior performance and flexibility to retrain through semi-supervised learning. We select a few highly confident and diverse examples from the initial detection and cluster-

✉ Vijay Kumar
vijaykumar.r@research.iiit.ac.in

[1] International Institute of Information Technology, Hyderabad, India

Published online: 08 July 2017

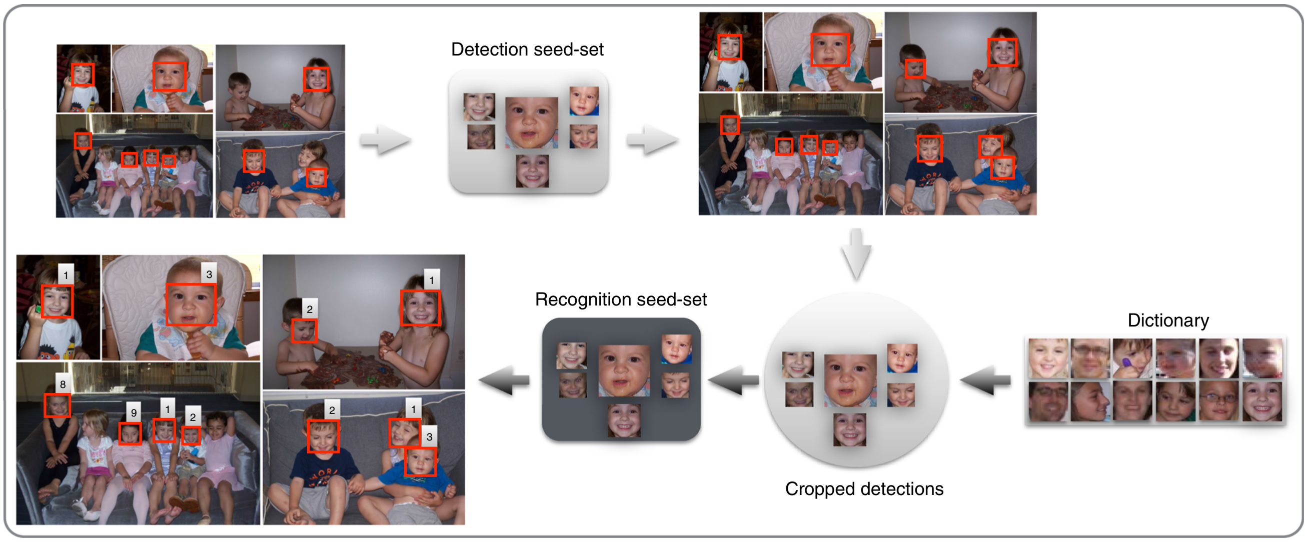


**Fig. 1** Overview of our face annotation approach. We detect and recognize faces in image collection in two stages. In the first stage, confident seed images are generated from the collection using off-the-shelf detection and recognition algorithms. In the second stage, the obtained seed images are used to further improve the performance of both detection and recognition using the semi-supervised approaches

ing. The exemplar detector is then retrained on these seed examples to adapt to the new domain. To recognize the detected faces, we follow a novel two-stage approach. We initially recognize the faces using the training dictionary using off-the-shelf recognition algorithm and retain the labels of highly confident examples based on a confidence measure. We then cast the recognition problem in a transductive semi-supervised framework treating the confident samples from the collection as the labeled set and rest as unlabeled set. We impose two constraints during propagation based on appearance and temporal similarities to exploit the relation among faces in different feature spaces.

Our approach achieves superior performance due to the following reasons. The appearance of the subjects that appear in multiple images of a collection is usually consistent. Our approach exploits this correlation among faces to propagate the labels from seed images to hard examples that are otherwise difficult to recognize. Also, unlike single-stage recognition approaches [15,23] that use dictionaries with large number of subjects, our dictionaries in the second stage contain only the subjects present in the collection, thereby reducing the confusion during multi-class classification.

The approach is related to our previous work [9] which focuses on semi-supervised recognition of faces in videos. This paper proposes a generic framework for annotating image collections with an end-to-end annotation pipeline including both detection and recognition stages. We conduct extensive experiments using deep convolutional features on several real-world collections of family albums and videos and demonstrate the improvements over baseline algorithms.

## 2 Related work

**Face detection** approaches can be roughly grouped into four categories, namely cascade AdaBoost, deformable parts, exemplar and deep network-based models. AdaBoost framework consists of a cascade ensemble of weak classifiers that detect faces in a coarse-to-fine strategy. Several features and convergence criteria are explored in this framework, starting from Viola–Jones (VJ) [25] to the latest works of [14]. Deformable part model (DPM)-based techniques [14,33] model the appearance of object parts and their relations. A vanilla DPM trained on a large database achieves impressive results [14]. Exemplar detectors [10,22] do not learn a global model that generalizes to the training examples. Instead, each exemplar generate voting maps in a retrieval framework, which are then aggregated to detect the faces. Recently, end-to-end trained convolutional neural networks (CNNs) are achieving superior performance for face detection [12,13].

**Semi-supervised face detection** schemes are typically employed when the number of labeled examples are limited, or to adapt the detectors to the target domain. Lie *et al.* [11] trains an SVM classifier with top positives and negatives obtained with VJ. Similarly in [7], less confident detections in each target image are re-scored through a regression process trained on top positives and negatives. This approach is not practical for large image collections as each target image requires separate training. Sebe et al. [20] trains a detector using few labeled examples along with a large number of unlabeled samples. The approach may be less relevant today due to the availability of large-scale face datasets.

**Face recognition** is another popular problem in which approaches that seek invariance to pose and illumination variations [3], hand-designed feature descriptors for improved discrimination [28], and low-dimensional feature representations [4,24] are proposed. One of the highly cited approaches in recent years is sparse representation classifier (SRC) [26] that represents each target image as a *sparse* linear combination of the training instances. For the related topic of face verification, metric learning approaches [6] learn embeddings that bring similar faces closer and dissimilar faces farther apart. More recently, CNN-based approaches [17,19] that learn a hierarchy of low-level to mid-level features through an end-to-end training are yielding excellent results for various face recognition and verification tasks.

**Semi-supervised face recognition** approaches are proposed when there are a few labeled training samples. In [18, 31], mean face templates for each subject are computed using either PCA and LDA. The unlabeled samples that are nearest to each of these class templates are augmented to labeled set to repeat the procedure. In [27], labeled examples are used as a constraint to obtain label constrained low-rank representations for face recognition.

**Face recognition in videos** is another related topic which contains several works in three categories—key frame, temporal model, and image-set-based approaches. A complete review of these approaches is given in [21]. A few works exploited additional cues in sitcom videos such as clothing and audio [23] and relationship between subjects [2].

## 3 Semi-supervised face detection

Given an image collection, our objective is to detect as many faces as possible by adapting off-the-shelf detector to the instances in the collection. The overview of our detection approach is shown in Fig. 2. We obtain initial detections from the collection using a generic detector trained on large database of images. We then select a few highly confident examples from the initial detections that are diverse and dissimilar through clustering. The detector is then retrained on these seed examples to adapt it to the new domain.

We choose the exemplar-based detector [10,22] due its high performance, simplicity, and flexibility it offers for adaption. The approach aligns well with our objective to improve performance using instances from the target domain. The exemplar detector uses retrieval framework for detection. A large database of training exemplars that cover better facial variations is initially constructed, and local features (dense SIFT) are extracted and quantized using a vocabulary. During testing, each exemplar generates multiple voting maps which are aggregated to locate the faces in the target image.

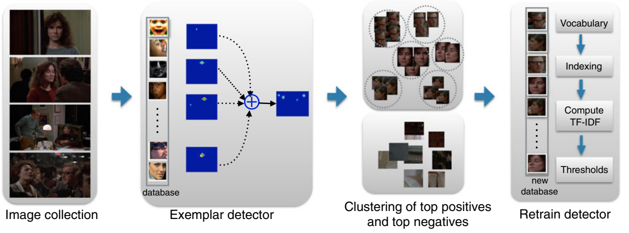


**Fig. 2** Overview of our detection approach. We apply off-the-shelf exemplar detector to get the initial detections. Top scoring positive and negative samples are selected and clustered to obtain a set of seed images. The exemplar detector is retrained by augmenting the seed images to exemplar database and re-computing the detector parameters

To achieve high recall, we initially apply the exemplar detector with very low threshold. Let $d_i$ be the set of initial detections and $s_i$ their corresponding scores. The scores $s_i$ are used to obtain a set of confident positive and negative instances. To achieve this, we define two thresholds $\omega_1$ and $\omega_2$ and consider those detections with scores $s_i$ greater than $\omega_1$ as top positives $P$ and less than $\omega_2$ as top negatives $N$.

$$\begin{aligned} P &= \{\, d_i, s_i > \omega_1,\ i = 1, 2, \ldots, n\} \\ N &= \{\, d_i, s_i < \omega_2,\ i = 1, 2, \ldots, n\} \end{aligned} \tag{1}$$

Once we obtain the sets $P$ and $N$, we cluster the images using $k$-means and select $\tau$ detections as seed images. This ensures that the obtained seed images are diverse which is essential for video collections containing near identical instances. If the diversity is not maintained, the most occurring exemplar instances dominate the voting process, severely degrading the performance. For each seed image, features are extracted and quantized, and augmented to the exemplar database. To retrain the exemplar detector, it is enough to update the inverse document frequencies of the visual words, unlike appearance-based detectors [12,14,25] which require retraining of the models from scratch.

## 4 Semi-supervised face recognition in image collection

We next recognize the detected faces given a labeled dictionary of subjects using a two-stage pipeline as shown in Fig. 3. In the first stage referred as *seed-set selection*, the detections are recognized using off-the-shelf face recognition algorithm. We then identify the key seed images that are confidently recognized using a robust confidence measure. In the second stage referred as *propagation*, we propagate the labels from the seed images to the remaining unlabeled collection images incorporating various constraints.

Our recognition approach follows the intuition that certain faces in the collection may have similar appearance, pose to the faces in the given dictionary. If such faces from the collection are recognized correctly, it is possible to propagate

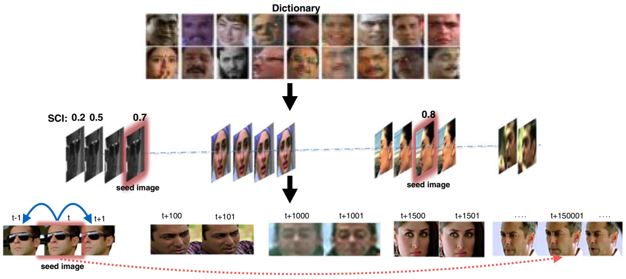


**Fig. 3** Overview of our proposed semi-supervised face recognition approach. In the first stage, we label all the images in a collection using SRC and select only highly confident seed images based on SCI. The labels of the seed images are then propagated to remaining images in the collection by imposing the constraints in time and feature space

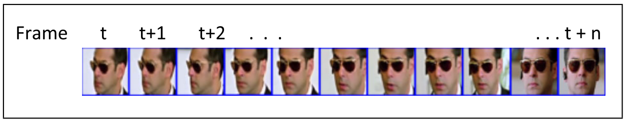


**Fig. 4** A possible scenario that depict the effectiveness of our approach on a video collection with a head pose change. If a few images in such shots are recognized correctly in the first stage, their labels can be effectively propagated to the remaining images in the second stage

their labels to the remaining unlabeled images as they belong to the same target domain. Our approach is applicable to scenarios that have images with high correlation among them. For instance, a video shot of zoom-in or zoom-out of a face in a fixed pose or video shot where there is a gradual change of pose of the subject as shown in Fig. 4. If one or few images in such shots are recognized confidently in the first stage, the labels can be propagated to the remaining faces effectively through propagation.

### 4.1 Seed-set selection

We use Sparse Representation Classifier (SRC) [26] as a off-the-shelf face recognition algorithm for initial labeling as it is robust to noise and occlusion and provides strong confidence measure based on reconstruction weights. Let $\mathbf{D} = [D_1, D_2, \ldots, D_M]$ be the dictionary of labeled examples and $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ is the collection of face instances. SRC represents the target image as a *sparse* linear combination of dictionary faces.

$$\hat{\alpha_i} = \arg\min_{\alpha_i} ||X_i - D\,\alpha_i||_2 + \lambda ||\alpha_i||_1, \tag{2}$$

where $\alpha_i$ is the representation of the sample $X_i$ and $\lambda$ is a Lagrangian constant which controls the trade-off between reconstruction error and sparsity. The label of $X_i$ is obtained using the minimum reconstruction error criteria as

$$\text{label}(X_i) = \arg\min_{j} ||X_i - D_j\,\hat{\alpha_{ij}}||_2, \tag{3}$$

where $D_j$ are the training samples belonging to class $j$ and $\alpha_{ij}$ are the corresponding weights. Once the images in the collection are labeled, we select a few confident images using Sparsity Concentration Index (SCI) [26] defined as

$$\text{SCI}(i) = \frac{c \cdot \max_j ||\hat{\alpha_{ij}}||_1 / ||\hat{\alpha_i}||_1 - 1}{c - 1}, \tag{4}$$

where $c$ denote the number of classes. This score indicates how well a target image is represented using the dictionary elements from a particular class. It is noted in [26] that $l_1$-based minimization results in nonzero coefficients that are concentrated with training examples from the correct class, even in the presence of noise and occlusions. Thus, the SCI score defined on the $l_1$-coefficients is robust and serves as a strong indicator for recognition confidence. An high score indicates that the target image is represented mostly from the single class and a very low score close to zero indicates contribution from all the classes. We consider this SCI score as a confidence measure to retain the label of a face.

### 4.2 Propagation of seed images

Once confident seed images are selected, their labels are propagated to the remaining unlabeled images in the collection through label propagation framework [32]. Since the collection contain correlated images, an image can be recognized using its appearance and its similarity with other images. This can be achieved in a graph-based framework where the image similarities are used to predict their labels. Recognition using seed images in the second stage reduces the domain mismatch between the dictionary and collection, and also the confusion rate when the dictionary has large number of subjects compared to the collection.

We reconsider $\mathbf{X} = [X_1, X_2, \ldots, X_N] = [X^l\ X^u]$ without loss of generality. Here, $X_l$ denote the labeled seed images from the first stage and $X_u$ denote the remaining unlabeled images in the collection. Let, $F \in \mathbb{R}^{N \times c}$ denote a nonnegative labeling matrix where the $i$th row of $F$ indicate the predicted class scores of $X_i$. Let $Y \in \mathbb{R}^{N \times c}$ be the initial labeling matrix. For the seed image $i$ belonging to class $j$, we define $Y_{ij} = 1$, and $0$ otherwise. For all the remaining unlabeled images, we assign a zero vector, i.e., $Y_{ij} = 0, \forall j$.

Given $X$, we construct an undirected graph $\langle V, E \rangle$ using both labeled and unlabeled images. Each node in the graph represents an image and the edges $E$ represent the similarities between images. Larger the edge weight, greater is the similarity between images. For image collections, we consider appearance similarity and consider both appearance and temporal similarities for video collections.

#### 4.2.1 Temporal Similarity

Let $t_i$ and $t_j$ be the absolute numbers that denote $i$th and $j$th frames, respectively. Also let $\upsilon_i = (\upsilon_i^{x1}, \upsilon_i^{y1}, \upsilon_i^{x2}, \upsilon_i^{y2})$ and $\upsilon_j = (\upsilon_j^{x1}, \upsilon_j^{y1}, \upsilon_j^{x2}, \upsilon_j^{y2})$ denote the $x$ and $y$ co-ordinates of top-left and bottom-right corners of the rectangles representing $i$th and $j$th face in the collection, respectively. We define temporal similarity [9] as follows,

$$W^t_{ij} = \exp\left(\frac{-(t_i - t_j)\chi_{ij}}{2\sigma_t^2}\right), \tag{5}$$

where $\sigma_t$ controls the spread of Gaussian function. $\chi_{ij}$ is defined as the absolute sum of the differences of the $i$th and $j$th image co-ordinates.

$$\chi_{ij} = \sum |\upsilon_i - \upsilon_j|. \tag{6}$$

A large value is assigned for $W^t_{ij}$ for pair of faces if they appear in subsequent frames and at similar locations. With above similarity, we define the temporal constraint as

$$\sum_{i,j} W^t_{ij}||F_i - F_j||^2. \tag{7}$$

During propagation, this constraint ensures that the images that are closer in temporal space have similar labels.

#### 4.2.2 Appearance Similarity

Images that are similar in appearance space (SIFT or CNN) should have large edge weights between them. While weights based on k-nearest neighbor and Gaussian function are commonly used in the literature, they are not robust to facial illumination and expression variations. For this reason, we employ a new scheme to compute the edge weights. We represent each image as a linear combination of its nearest neighbors to preserve the locality and impose nonzero constraints on the weights [9]. Our approach encourages the creation of edges only with similar samples as required for better performance of graph-based frameworks.

$$\hat{w}_i = \arg\min_{w_{ik}} ||X_i - \Sigma_{k:X_k \in \mathcal{N}(X_i)} X_k w_{ik}||_2 + \beta||w_i||_2$$
$$\text{s.t} \ \forall_k, \ w_{ik} \geq 0, \tag{8}$$

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

$$W^a_{ij} = \begin{cases} \hat{w}_i(k), & \text{if } X_j \in \mathcal{N}(X_i) \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

$\hat{w}_i(k)$ 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, $W^a_{ij} \neq W^a_{ji}$. To make the weights symmetric, we perform the below operation

$$W^a_{ij} = W^a_{ji} = \frac{W^a_{ij} + W^a_{ji}}{2}. \tag{10}$$

The appearance constraint is finally incorporated into the label propagation framework as

$$\sum_{i,j} W^a_{ij}||F_i - F_j||^2. \tag{11}$$

This constraint ensures that images that are highly similar in appearance space should belong to the same class.

#### 4.2.3 Propagation

We finally incorporate appearance (Eq. 11) and temporal (Eq. 7) constraints into label propagation formulation [32] to propagate the labels from labeled seed images to the unlabeled images as follows.

$$\begin{aligned} Q(F) = \arg\min_F \frac{\gamma_1}{2} \sum_{i,j}^N W^t_{ij}||F_i - F_j||^2 \\ +\frac{\gamma_2}{2} \sum_{i,j}^N W^a_{ij}||F_i - F_j||^2 \\ +\gamma_3 \sum_i^N ||F_i - Y_i||^2. \end{aligned} \tag{12}$$

If the first and second terms ensure that the images that are similar based on appearance and temporal weights have similar labels, third term retains the labels of the confident seed images. $\gamma_i$ are the parameters that control the trade-off between these three terms.

Let, $D^t_{ii} = \sum_j W^t_{ij}$ and $D^a_{ii} = \sum_j W^a_{ij}$ be the diagonal and symmetric matrices whose entries are row sums of $W^t_{ij}$ and $W^a_{ij}$, respectively. $L^t = D^t - W^t$ and $L^a = D^a - W^a$ are the Laplacian matrices that are symmetric and positive semi-definite, defined on temporal and appearance similarities, respectively.

Following [1], we can rewrite the first term in Eq. 12 as

$$\begin{aligned} \sum_{i,j}^N W^t_{ij}||F_i - F_j||^2 &= \sum_{i,j}^N (F_i^2 + F_j^2 - 2F_iF_j)W^t_{ij} \\ &= \sum_i^N F_i^2 D^t_{ii} + \sum_j^N F_j^2 D^t_{jj} \\ &\quad -2\sum_{i,j}^N F_iF_jW^t_{ij} = 2\bar{F}L^tF \end{aligned} \tag{13}$$

where $\bar{F}$ is the matrix transpose of $F$. Similarly, second term is equivalent to $2\bar{F}L^aF$. Thus, Eq. 12 can be rewritten as

$$Q(F) = \arg\min_F \gamma_1 \bar{F}L^tF + \gamma_2 \bar{F}L^aF + \gamma_3 ||F - Y||^2 \quad (14)$$

Differentiating $Q(F)$ with respect to $F$ and equating to 0,

$$2\gamma_1 L^t F + 2\gamma_2 L^a F + 2\gamma_3 (F - Y) = 0 \quad (15)$$

$$F^* = \gamma_3(\gamma_1 L^t + \gamma_2 L^a + \gamma_3)^{-1} Y \quad (16)$$

The final identity $y_i$ of the image $X_i$ can be obtained from $F^*$ as $y_i = \arg\max_j F^*_{ij}$, where $j = \{1, 2, \ldots, c\}$.

### 4.3 Rejection of unknown faces

The candidate boxes obtained from the detection stage may contain false positives or unknown faces that are not present in the dictionary. This is particularly common in video collections. The algorithm should be able to accept or reject the labels of such candidates after propagation. For this purpose, we define a confidence measure known as *label dominance score* (LDS) that measures the contribution of each class during the reconstruction of the sample. This is simply defined as the ratio of two largest class scores of the sample.

$$\text{LDS (i)} = \frac{F_{ij}}{\arg \max\limits_{k,\, k \neq j} F_{ik}} \quad \text{where } j = \arg\max_j F_{ij} \quad (17)$$

Intuitively, when an image has large edge weights with images belonging to a particular class, the scoring vector $F_i$ will have a high score for that particular class. When there is such clear dominance of one class, LDS will be high and we consider the labeling as confident and retain its final label.

## 5 Experiments and results

### 5.1 Datasets and setup

We evaluate our approach on several album and video collections whose instances exhibit pose, illumination, viewpoint variations that are seen in the real world. All these collections contain multiple images in which a small number of subjects appear in different images. Album collections contain multiple photographs of family members taken at various indoor and outdoor events, and video collections contain several frames of actors appearing in one or more videos. We show few images from these collections in Figs. 7 and 8.

**G-album** [5] consists of 589 photographs containing 931 faces belonging to 32 subjects. For each subject, we consider a maximum of 10 images to create a labeled dictionary and the rest as unlabeled collection.

**People in photo-albums (PIPA)** [30] is a large collection of 1438 user-uploaded albums collected from Flickr. The dataset consists of 37,107 photographs with 63,188 head instances belonging to 2356 identities. Similar to G-album setting, we consider a maximum of 10 images to create a labeled dictionary and the rest as unlabeled collection.

**Movie trailer face dataset** [15] consists 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 [8] along with additional 10 actors. The labeled dictionary consists of 34,522 images (PubFigs + 10 additional actors) with each actor having a maximum of 200 images. Since the dataset provides the raw descriptors based on the combination of LBP, HOG, and Gabor features), we show only the recognition results.

**Hannah movie dataset** [16] consists of face annotations for the entire movie *Hannah and Her Sisters*. The dataset has 153,833 frames with 202,178 face bounding boxes and 254 different labels of 41 named, 186 unknown characters, and 15 miscellaneous crowd regions. We create the labeled dictionary from IMDB photographs of actor's profile for each named character. We manually annotate the face bounding boxes for the dictionary images. Of the 41 named characters, only 26 prominent actors had profiles in IMDB. The labeled dictionary consists of 2385 images belonging to 26 prominent actors appeared in the movie.

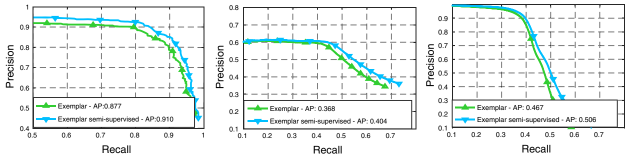
### 5.2 Features

We follow the same feature extraction procedure [10] based on dense SIFT for detection. For recognition, we extract deep CNN features using VGG model [17] trained on 2M faces of 2622 identities using `VGG-16` architecture. The features are extracted from seventh fully connected layer whose dimension is 4096. We apply PCA to reduce the feature dimension to 300.

### 5.3 Results

We use the publicly available implementation of [10] for exemplar face detection. We set the upper and lower thresholds $\omega_1$ and $\omega_2$ to 80th and 20th percentile of the detection scores to select the top 20% positives and negatives, respectively. For video collections, we further cluster the detections to select 300 diverse images. The obtained detections are resized and followed through the feature indexing pipeline as the original algorithm. The term frequencies and inverse document frequencies are updated based on seed images. The detection results in Fig. 5 show that performance improvement obtained using our approach.

For generating the seed images during recognition, we set the error tolerance $\lambda = 0.05$ and retain the labels of top 25% of images based on SCI. We set various parameters using



**(a)** G-album **(b)** PIPA **(c)** Hannah

**Fig. 5** Precision–recall curves of baseline and adapted exemplar detectors on various image collections. We notice a considerable improvement in the performance with our proposed approach that involves a simple augmentation of seed images into the exemplar database

**Table 1** Recognition performance (%) of various methods in terms of accuracy and average precision (AP) on image collections

| Method | G-album | | PIPA | |
|---|---|---|---|---|
| | Accuracy | AP | Accuracy | AP |
| 1-NN | 77.26 | 91.69 | 56.02 | 78.49 |
| SVM | 80.82 | 97.57 | 38.15 | 93.21 |
| CRC [29] | 75.33 | 94.58 | 51.27 | 76.79 |
| SRC [26] | 79.56 | 97.41 | 58.84 | 93.70 |
| MSSRC [15] | 79.56 | 97.41 | 58.84 | 93.70 |
| Our approach | 84.52 | 97.72 | 63.80 | 94.59 |

**Table 2** Recognition performance (%) of various methods in terms of accuracy and average precision (AP) on video collections

| Method | Hannah | | Movie Trailer | |
|---|---|---|---|---|
| | Accuracy | AP | Accuracy | AP |
| 1-NN | 32.71 | 15.87 | 23.60 | 9.53 |
| SVM | 44.82 | 83.42 | 54.68 | 50.06 |
| CRC [29] | 37.49 | 37.92 | 41.93 | 36.33 |
| SRC [26] | 39.76 | 49.97 | 47.78 | 54.33 |
| MSSRC [15] | 43.74 | 38.62 | 50.52 | 58.69 |
| Our approach | 54.19 | 84.79 | 55.98 | 59.34 |

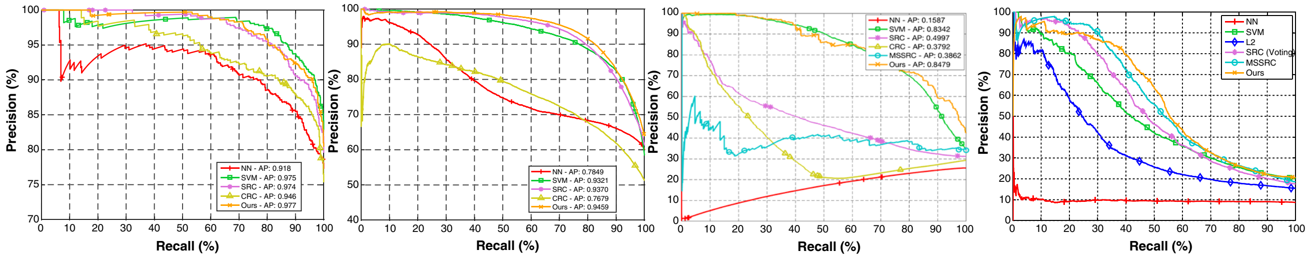
cross-validation on held-out set. We set the number of neighbors $k$ for computing appearance weights to 120, $\gamma_2 = 0.3$ and $\gamma_3 = 0.7$. For the collections with temporal information (e.g. G-album and PIPA), we set $\gamma_1 = 1$ and similarly for collections without temporal information (e.g. Hannah and Movie Trailer), we set $\gamma_1 = 0$.

We show the recognition performance of various approaches on image and video collections in Tables 1 and 2, respectively. Our approach which exploits the correlation among the instances outperforms other approaches that recognize instances independently. Note that, for image collections without tracks, MSSRC is same as SRC. We also show the average precision which measures the ability to reject unknown instances. We use SCI as a confidence measure for SRC, CRC, MSSRC and LLC algorithms, L2 Euclidean distance for k-NN and probability scores for SVM. We show the precision–recall (PR) curves of various methods in Fig. 6. It is clear from table that our confidence measure based on LDS is robust in rejecting unknown instances. The recognition rates of various approaches for different number of labeled training samples are shown in Table 3 for G-album dataset. The performance improvement is larger when there are limited labeled examples, and it becomes closer to other approaches as the labeled examples are increased. We show few qualitative results of our detection and recognition approach in Figs. 7 and 8, respectively.

**Ablation study** We show the effectiveness of different stages and constraints during recognition in Table 4. Clearly, we obtain an improvement with two-stage approach when appearance similarities in the collection are exploited. Additionally, when time information is used for videos, it brings further performance improvement. Finally, we show the performance of our approach for varying percentage of seed

**Table 3** Recognition rates [%] of various methods on G-album for different number of labeled examples

| Method | 1 Train | 2 Train | 5 Train | 10 Train | 20 Train |
|---|---|---|---|---|---|
| KNN | 56.12 | 65.85 | 74.15 | 77.26 | 80.43 |
| SVM | 57.45 | 64.40 | 78.09 | 80.82 | 85.42 |
| SRC [26] | 56.45 | 68.39 | 77.86 | 79.56 | 86.82 |
| CRC [29] | 56.27 | 68.28 | 77.62 | 75.33 | 85.87 |
| Our approach | 60.25 | 71.27 | 82.52 | 84.52 | 87.18 |



**(a)** G-album **(b)** PIPA **(c)** Hannah **(d)** MovieTrailer

**Fig. 6** Precision–recall curves of *recognition* algorithms on various image collections. Our confidence score based on LDS is more robust and performs better than all the approaches in rejecting unknown samples

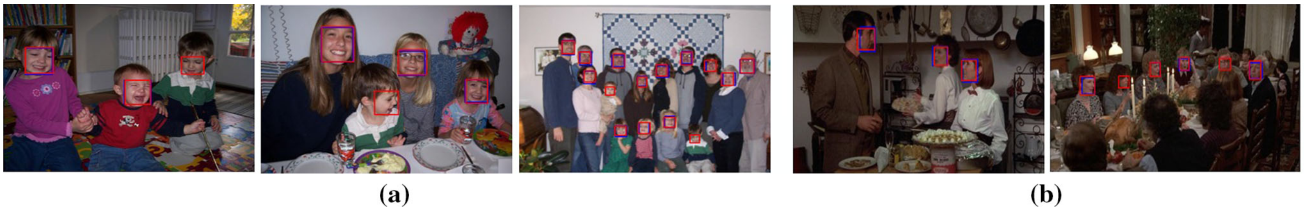


**Fig. 7** Performance improvement after adapting the exemplar detectors to the image collection from **a** Galbum and **b** Hannah datasets. The original detections are shown in blue, and new detections are shown in red. Notice how the faces that were by missed the original detector are detected after adaption using images from the seed set

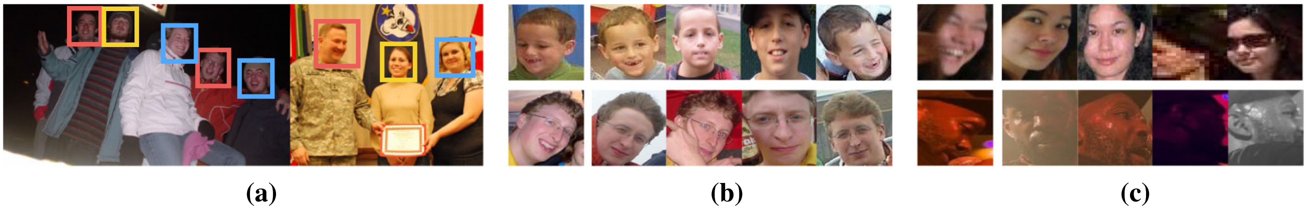


**Fig. 8** Qualitative results on PIPA: **a** the comparisons of our approach with single-stage approach of SRC. *Red boxes* show the success case of our approach and failure case of SRC. *Blue* (and *yellow*) boxes indicate instances that are (not) correctly predicted by both approaches. **b**, **c** The success and failure cases of our approach, using the test instance and its top 4 instances with nonzero appearance weights $w^a$

**Table 4** Ablation study: Recognition performance of our approach in the first stage (SRC), second stage based on appearance and temporal similarities alone, and overall performance with both the similarities

| Method | G-album | PIPA | Hannah | Trailer |
|---|---|---|---|---|
| Stage 1 (SRC) | 79.56 | 58.84 | 39.76 | 47.78 |
| Appearance | 84.52 | 63.80 | 47.83 | 51.13 |
| Temporal | – | – | 44.49 | 50.24 |
| Overall | 84.52 | 63.80 | 54.19 | 55.98 |

images retained in Fig. 9. The performance improvement is significant when we retain 20−30% of seed images and becomes closer to the performance of first stage with increasing proportion of seed images.

**Computational complexity** Our approach brings substantial performance improvements with minimal additional cost due to its two-stage process. With the feature extraction common across stages, the additional complexity is due to voting map generation for detection and $<W^t, W^a, F>$ computation for recognition in the second stage. The detection and recognition steps take ∼1.5 and ∼2 times more than the single-stage approach, respectively.

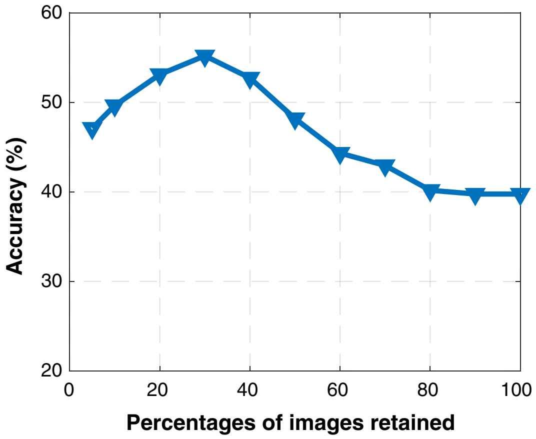


**Fig. 9** Recognition performance for varying percentage of seed images selected in the first stage using Hannah dataset. The improvement is significant when 20−30% of seed images are selected, and becomes less effective when large proportion of seed images are retained

## 6 Conclusion

In this paper, we present an approach for automatic annotation of faces in an image collection. We demonstrate a two-stage approach for both detection and recognition steps that exploits the similarities among collection instances to improve the performance. We first generate a set of seed images using off-the-shelf detection and recognition algorithms, which are then are used to improve the performance by adapting them to the target image collection. We propose an exemplar-based semi-supervised approach for improving the detections and label propagation-based framework for improving the recognition. Experiments on album and video collections show that our method obtains an improvement of ∼4% for detection and 5−9% for recognition tasks.

**Acknowledgements** This work is partly supported by the MCIT, New Delhi. Vijay Kumar is supported by TCS PhD fellowship 2012.

## References

1. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**, 1373–1396 (2003)
2. Buml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
3. De Marsico, M., Nappi, M., Riccio, D.: Face authentication with undercontrolled pose and illumination. Signal Image Video Process. **5**, 401–413 (2011)
4. Ding, C., Bao, T., Karmoshi, S., Zhu, M.: Single sample per person face recognition with KPCANet and a weighted voting scheme. Signal Image Video Process. (2017). doi:10.1007/s11760-017-1077-8
5. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
6. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: International Conference on Computer Vision (2009)
7. Jain, V., Learned-Miller, E.: Online domain adaptation of a pretrained cascade of classifiers. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
8. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE International Conference on Computer Vision (2009)
9. Kumar, V., Namboodiri, A., Jawahar, C.: Face recognition in videos by label propagation. In: International Conference on Pattern Recognition (2014)
10. Kumar, V., Namboodiri, A., Jawahar, C.: Visual phrases for exemplar face detection. In: IEEE International Conference on Computer Vision (2015)
11. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic part model for unsupervised face detector adaptation. In: IEEE International Conference on Computer Vision (2013)
12. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
13. Li, Y., Sun, B., Wu, T., Wang, Y.: Face detection with end-to-end integration of a CONvNet and a 3D model. In: European Conference on Computer Vision (2016)
14. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: IEEE European Conference on Computer Vision (2014)
15. Ortiz, E., Wright, A., Shah, M.: Face recognition in movie trailers via mean sequence sparse representation-based classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
16. Ozerov, A., Vigouroux, J.R., Chevallier, L., Perez, P.: Clothing cosegmentation for recognizing people. In: International Conference on Image Processing (2013)
17. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
18. Roli, F., Marcialis, G.L.: Semi-supervised PCA-based face recognition using self training. In: Joint International Association Of Pattern Recognition Workshop (2006)
19. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
20. Sebe, N., Cohen, I., Huang, T.S., Gevers, T.: Semi-supervised face detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2005)
21. Shan, C.: Face recognition and retrieval in video. In: Video Search and Mining. Springer, pp 235–260. doi:10.1007/978-3-642-12900-1_9 (2010)
22. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Detecting and aligning faces by image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
23. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: "Knock! Knock! Who is it?" Probabilistic person identification in TV-series. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
24. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**, 71–86 (1991)
25. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (2001)
26. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210–227 (2009)
27. Yang, C., Ye, M., Tang, S., Xiang, T., Liu, Z.: Semi-supervised low-rank representation for image classification. Signal Image Video Process. **11**, 73–80 (2017)
28. Yu, W., Gan, L., Yang, S., Ding, Y., Jiang, P., Wang, J., Li, S.: An improved LBP algorithm for texture and face classification. Signal Image Video Process. **8**, 155–161 (2014)
29. Zhang, L., Yang, M., Feng, X.: Sparse representation or Collaborative representation: Which helps face recognition?. In: IEEE International Conference on Computer Vision (2011)
30. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: improving person recognition using multiple cues. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
31. Zhao, X., Evans, N.W., Dugelay, J.L.: Semi-supervised face recognition with LDA self-training. In: International Conference on Image Processing (2011)
32. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical Report, CMU (2002)
33. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)