9-2014

Name-face association in web videos: A large-scale dataset, baselines, and open issues

Zhi-Neng CHEN

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Wei ZHANG

Juan CAO

Yu-Gang JIANG

Chen ZN, Ngo CW, Zhang W *et al.* Name-face association in Web videos: A large-scale dataset, baselines, and open issues. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(5): 785–798 Sept. 2014. DOI 10.1007/s11390-014-1468-z

Name-Face Association in Web Videos: A Large-Scale Dataset, Baselines, and Open Issues

Zhi-Neng Chen[1,2] (陈智能), Chong-Wah Ngo[2,*] (杨宗桦), *Member*, *IEEE*, Wei Zhang[2] (张　炜)
Juan Cao[3] (曹　娟), and Yu-Gang Jiang[4] (姜育刚)

[1] *Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

[2] *Department of Computer Science, City University of Hong Kong, Hong Kong, China*

[3] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*

[4] *School of Computer Science, Fudan University, Shanghai 200433, China*

E-mail: zhineng.chen@ia.ac.cn; cscwngo@cityu.edu.hk; wzhang34-c@my.cityu.edu.hk; caojuan@ict.ac.cn; ygj@fudan.edu.cn

Received February 24, 2014; revised July 3, 2014.

Abstract Associating faces appearing in Web videos with names presented in the surrounding context is an important task in many applications. However, the problem is not well investigated particularly under large-scale realistic scenario, mainly due to the scarcity of dataset constructed in such circumstance. In this paper, we introduce a Web video dataset of celebrities, named WebV-Cele, for name-face association. The dataset consists of 75 073 Internet videos of over 4 000 hours, covering 2 427 celebrities and 649 001 faces. This is, to our knowledge, the most comprehensive dataset for this problem. We describe the details of dataset construction, discuss several interesting findings by analyzing this dataset like celebrity community discovery, and provide experimental results of name-face association using five existing techniques. We also outline important and challenging research problems that could be investigated in the future.

Keywords Web video, celebrity, name-face association, dataset construction, community analysis

1 Introduction

With the prosperity of video sharing activities on the Web, videos are being captured, searched and browsed at an accelerating rate. Among the huge deposit of videos and query logs on these websites, as reported in [1-2], many of them are about celebrities. While techniques for video search and annotation are becoming increasingly important as a result of the video growth[3-6], there is relatively little work being conducted in the celebrity domain.

The search and browsing of celebrities by keywords from a large volume of Internet videos, as current mainstream video search engines do, is far beyond satisfaction for the following reasons. First, the majority of user-supplied tags are given at the video level rather than the segment or keyframe level. Locating a portion of a video where faces of celebrities appear remains a function unsupported by any commercial search engines. Second, user-supplied tags are often incomplete or even noisy[7]. A video with celebrities is not necessarily tagged with their names. Similarly, a video tagged with celebrities also does not imply the existence of the celebrities in the videos. As a result, the search of celebrities could very likely lead to a mixture of desirable and noisy results. Third, a good result ranking in celebrity video search is difficult to get, because there could be a huge number of videos tagged with the same celebrity, and pure name matching cannot well distinguish which video is more relevant. A naive way for ranking the videos could be simply made by counting the number of times as well as the size of a celebrity's face appearing in a video, which obviously does not sound to be an ideal solution.

To enable better content-based search and browsing of celebrity videos, a key technique is to tag the celebrities' faces with their names. Instead of tagging at the video level, tags labeled at the face level are much more accurate and helpful in related applications. This task is generally referred to as name-face association in the literature[8-9]. Over the years, many techniques and benchmark datasets have been developed for this task

Regular Paper

This work was supported by a research grant from City University of Hong Kong under Grant No. 7008178, and the National Natural Science Foundation of China under Grant Nos. 61228205, 61303175 and 61172153.

*Corresponding Author

in the domains of news videos[10-15], TV series[16-19], movies[20], news images[8-9,21-25] and Web videos[26-27]. However, the only dataset built on top of Web video is the YouTube Faces dataset[26]. The other studies were experimented against datasets like the TRECVID benchmark, popular TV series and movies, which are generally related to only a few celebrities or characters, although some of them[17] also have a large quantity of faces. While in [26], the duration of the videos is very short, and each person only associates with no more than 6 videos. Thus the relationships among people are largely underestimated. Some of the existing datasets pose the challenge of recognizing faces under "wild" conditions due to such as variations in pose, illumination and face expression. However, these datasets are relatively narrow in coverage and (or) small in size, which limits their applicability in characterizing the diversity of faces in large-scale realistic Web video repositories.

In this paper, we aim at constructing a Web video celebrity dataset that depicts the existence of celebrities and their relations as realistic as possible, supporting the research on large-scale name-face association. To this end, we describe the construction of a new dataset, named as WebV-Cele①, which consists of 75 073 videos and contains 2 427 celebrities and 649 001 faces. The dataset covers a large and diverse visual appearance of faces, and a wide range of celebrities with different professions. To our knowledge, WebV-Cele is the most diverse large-scale dataset in the literature for this problem.

We also perform a series of tasks on this dataset, including face detection, name verification, community analysis, and manual name-face association, and observe that celebrities are naturally grouped into meaningful communities closely correlated to hot topics. These are summarized in Section 3. In addition, we conduct experiments to evaluate several existing techniques, including weak association, support vector machine (SVM), multiple instance learning, graphic-based clustering and image matching, for name-face association using the WebV-Cele dataset in Section 4. The results can serve as baselines for further work developed on this dataset. We further outline several research issues that could be studied using this dataset in Section 5. Finally, Section 6 concludes this paper.

2 Related Work and Datasets

Face recognition has received intensive research attention over the past few decades. Recognizing faces under unconstrained capturing conditions nevertheless remains a highly difficult problem[28-29]. Name-face association, which utilizes surrounding metadata to assist recognition, is generally regarded as a feasible methodology for face recognition "in the wild". A common approach is to first weakly associate every name found in metadata with every face detected from images or videos, and then the refinement based on visual similarity and contextual clues is conducted to remove false matches. Existing studies in name-face association mostly differ in the way of how refinement is formulated, which largely depends on the domains and information available for use. Based on the refinement, we can broadly categorize related studies into classification-based[11-14,19,23], clustering-based[8,10,15-18,20-22] and knowledge-based[1,27,30] methods.

The classification-based methods learn discriminative models to predict the correction of the weakly associated name-face instances. For example, in [13], instead of learning models for each person, a unified SVM classifier was trained to determine the correction of each name-face instance based on multiple modalities extracted from the transcripts, optical character recognition (OCR) result and speech track of news videos. Due to the need of labeling a large number of name-face pairs for learning, the work was later extended to partial learning under multiple instance setting, i.e., MIL[14], where only partial label information is required for model training. Similar in spirit, the work by [23] also proposed MIL for learning metric of association in news photo domain, and the authors of [11-12] adopted semi-supervised learning to propagate name-face alignments from labeled to unlabeled examples in news video domain. Recently, a multi-class classification framework was proposed to identify persons in TV series[19], which jointly considers labeled and unlabeled data as well as the temporal relations between face tracks. Despite these efforts, supervised and semi-supervised learning methods are difficult to be scaled up to large datasets for requiring sufficient training samples to guarantee a high recognition rate.

Observing that celebrities or major characters usually appear recurrently in news videos, TV series and movies, the clustering-based methods investigate name-face association by focusing on mining visual similarities between faces (face tracks) and contextual information derived from video structure and (or) prior knowledge. These methods were found to be more practical and could also produce satisfactory performance as demonstrated by early work in the domain of news videos[10,15] and TV series[16]. For instance in [16], by exploiting the time-coded information from subtitles

①http://vireo.cs.cityu.edu.hk/WebV-Cele/, Aug. 2014.

and speaker identities from scripts, names are automatically aligned with speaking faces and then propagated to other faces. Similarly in [20], global name-face matching in movie domain is proposed to match affinity graphs of faces and names through aligning movie scripts and subtitles. Another influential work is the graph-based clustering methods developed by [21] in Web image domain, where the faces detected from an image collection are jointly modeled as a graph and the relationship between faces is determined based upon the similarity of facial features. With the assumption that the faces of a person should exhibit higher visual similarity and reside in a dense subgraph, the problem was converted to identifying densely connected subgraphs corresponding to the names. Nevertheless, clustering without strong contextual cues as in news and movie domains is difficult in general and is likely to generate noisy results.

In contrast, the knowledge-based methods tackle the problem of weak contextual clue by leveraging online information sources for learning face models[1] and identifying social networks[30]. In [1], based on the fact that there are plenty of celebrity photos freely available online, face models are learnt by automatically crawling training examples from the Web without human labeling. In [30], using user-supplied identity tags on photos (tags directly labeled on the faces) from Facebook, social context is exploited for face recognition through supervised structure prediction. The recent work by [27] proposed a 3-step pipeline that leverages name and face communities as connectors to map names to faces. The name and face communities are constructed based on the metadata of Web videos and the visual similarity of detected faces respectively.

Along with these techniques, there have been several benchmark datasets developed for name-face association, e.g., Yahoo! News[18], Fan-Large[24], LFW[25], and YouTube Faces[26]. Each of the former three datasets contains thousands of names and hundreds (tens) of thousands of images, which were all designed for supporting research on image-based name-face association. The YouTube Faces dataset contains 3 425 videos of 1 595 persons, which was primarily constructed for studying the problem of face pair matching, i.e., deciding whether two faces represent the same individual. The duration of these videos is thus very short, with an average length of only about 8 seconds. Moreover, each person only associates with no more than six videos. Because of this dataset bias, the actual relations of people can hardly be reflected. For most of the name-face association techniques developed for videos, they were experimented against datasets like TV series[12,16-19], movies[20], and news videos[11,13-14,22]. The total duration of these videos ranges from dozens to hundreds of hours. Although a large number of faces could also be extracted[17], most of these datasets contain only a few celebrities or characters.

Compared with the existing datasets, our constructed dataset WebV-Cele is more comprehensive in terms of the number of videos, celebrity subjects and faces. More importantly, our dataset is drawn from nearly 250 000 YouTube videos. This results in a more representative and less subjective dataset that naturally covers a wide range of faces and names, which may also reflect the true distributions of faces and names. Compared with [26], our final dataset contains more than 4 000 hours of videos with an average video duration of 219 seconds. In addition, the faces are extracted from keyframes rather than from consecutive frames as in [15, 26], containing much less near-duplicates and thus showing more diversified facial appearances. These unique characteristics of our dataset make it an ideal benchmark for future research in this area.

3 Dataset WebV-Cele

The WebV-Cele dataset is created on top of the MCG-WEBV — a real-world Web video dataset released a few years earlier[31]. The dataset includes two parts: CoreData and ExpandedData. The former contains 14 473 "Most Viewed" videos of "this month" crawled from the 15 predefined YouTube channels from December 2008 to November 2009. ExpandedData is composed of additional 234 414 "Related Videos" of the videos in CoreData. The videos in MCG-WebV have been decomposed into shots, and more than five millions of keyframes were extracted to represent these shots.

3.1 Face and Name Extraction

We employ commercial software developed by the IS'vision company② for frontal face detection. In total, there are 1 556 265 detected faces, with size equal to or larger than 40×40 pixels. The number of videos and keyframes containing at least one face are 62.0% and 26.9% respectively, clearly indicating that face is a major entity in Web videos. Many of the detected faces appeared in videos were captured "in the wild" with severe variations in head pose, expression, illumination and degree of motion blur, as illustrated in Fig.1, where the faces of former U.S. president George W. Bush are given as an example. Fig.2 also gives the distribution of faces according to resolution. Only 12.7% of faces are in close-up view, i.e., with size 150×150 pixels or

②http://www.isvision.com/cn/index, July 2014.

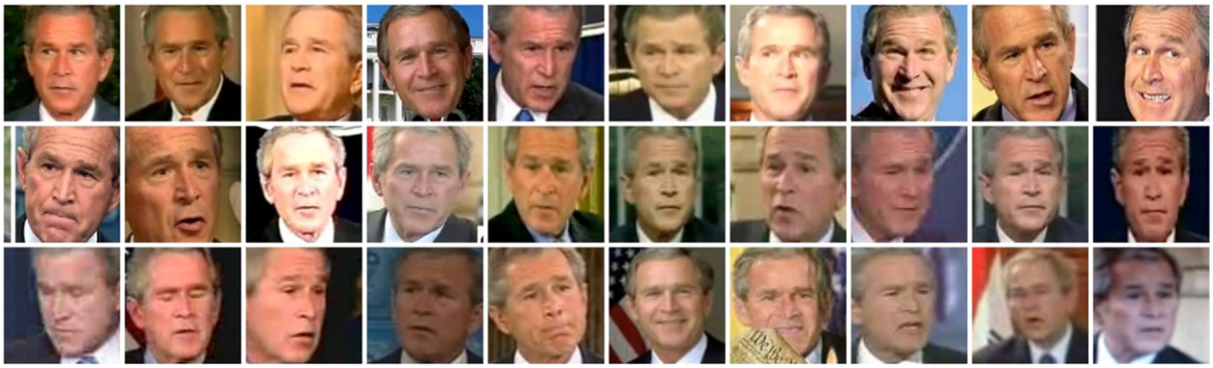


Fig.1. Faces of former U.S. president George W. Bush extracted from Web videos. Large variations in head pose, expression, illumination, background and motion blur can be observed.

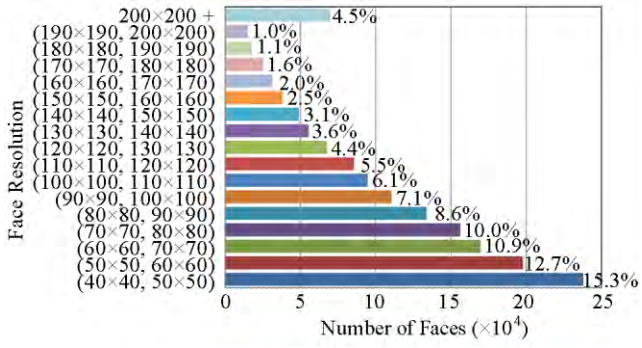


Fig.2. Distribution of the detected faces with respect to face size.

even larger, and nearly half of the faces are smaller than 80×80 pixels. The prevalence of the low-resolution faces is one of the major challenges for name-face association in Web videos.

The name entities are extracted from metadata (titles and tags) surrounding videos. We implement a Wikipedia-based extraction algorithm for this task. First, candidate names are extracted by stepwisely testing whether a word or a succession of words in metadata could represent the name of a person. The candidate names are searched and verified against the categories provided by Wikipedia. To confirm whether a word or a succession of words refers to a person, we use a heuristic that the person's birth year should appear in the category description on Wikipedia. Fig.3 briefly illustrates the flow of name entities extraction. Throughout the process, we only keep the longest name found from the

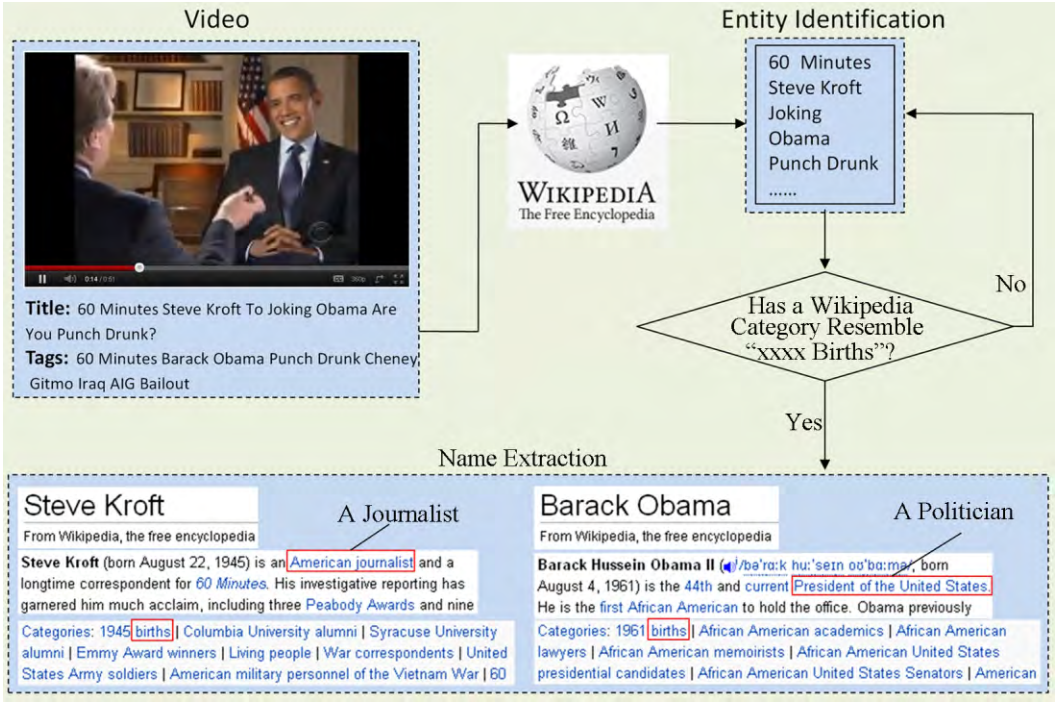


Fig.3. Framework of name extraction. Candidate names are extracted from metadata and verified against the Wikipedia categories.

successive words in the metadata, e.g., we keep "Barack Obama" rather than "Barack" and "Obama" individually if the two words appear successively. Once a candidate is identified, we also extract the occupation along with the full name.

Based on this method, a total of 209 001 name occurrences are extracted. These occurrences can be further grouped into 21 847 different names, and by Wikipedia, they correspond to 17 552 unique persons. An interesting observation is that 20% of persons contribute to 85% of name occurrences, whereas 65% of persons with names appeared less than three times. Fig.4(a) shows the distribution of name frequency in the dataset. The distribution can be well modeled by power-law distribution[32] with $\alpha = -0.83$. In other words, the percentage of names appearing at least x times is proportional to $x^{-0.83}$. We also show the distributions of the number of detected faces and names per video in Fig.4(b) and Fig.4(c) respectively. The two distributions are also modeled by power-law distribution with $\alpha = -0.40$ and $\alpha = -0.27$, implying the percentage of videos containing at least x faces or names is proportional to $x^{-0.40}$ or $x^{-0.27}$, respectively. For videos containing at least one face (name), there are 13.9 faces (2.5 names) per video on average.

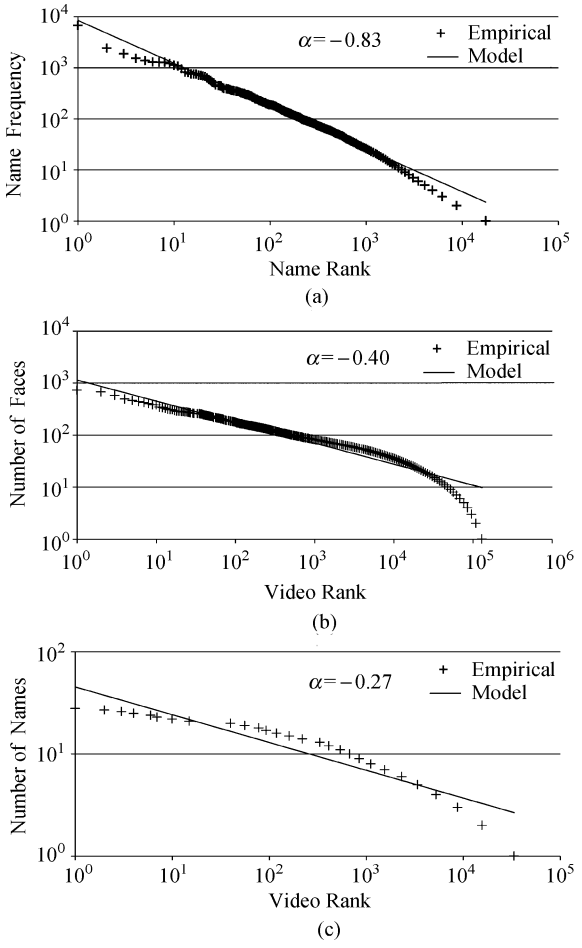


Fig.4. (a) Distributions of the name frequency in the dataset. (b) Number of faces per video in the dataset. (c) Number of names per video in the dataset. The name (video) rank refers to the list generated by ranking the names (videos) in descending order, according to the name frequency (the number of faces or names).

With these faces and names, interestingly, names and faces do not always co-occur in videos. Among the detected faces, only 42.2% of them are from videos containing names. On the other hand, there are 17.3% of videos containing names in the metadata but no faces are detected. The statistics also highlight the challenge of name-face association in realistic scenario. Not only should the within video association of names and faces be investigated, but also labeling faces from videos with no name presence and omitting videos with no face detected should be studied, for the sake of better exploiting the celebrity videos.

3.2 Celebrity Mining

We further extract a subset of videos with 2 427 celebrities for a more detailed analysis. The names of these celebrities appear at least ten times in the whole dataset. Fig.5 lists the top-20 popular persons and their name frequencies in the dataset. As videos in the dataset are uploaded to YouTube during the years 2008 and 2009, these person names are highly correlated to hot news events in this period. For instance: most videos tagged with *Barack Obama* are about "U.S. Presidential Election 2008"; videos tagged with *Michael Jackson* are likely linked to "the death of Michael Jackson"; clips tagged with *Robert Pattinson* and *Kristen Stewart* are excerpt from the movie "Twilight"; and videos tagged with *Susan Boyle* refers to her unexceptional performance in "Britain's Got Talent".

Among the 2 427 celebrities, 144 ones are brought to our extra attention as they also have at least ten occurrences in CoreData. Based on the professions extracted from Wikipedia descriptions, we can roughly group the 144 celebrities into five categories:

- Internet Star (25): vlogger and Internet celebrity;
- Artist (63): singer, actor, actress and model;
- Politician (21): politician, business man and religious leader;
- Sportsman (21): athlete and coach;
- Journalist (22): journalist, broadcaster, TV host, judge, writer and director.

Obviously, artists and Internet stars are the major groups of celebrities in Web video repositories. This statistics aligns well with a 2009's report from YouTube③, where 61% of people voted for "to be enter-

③http://youtubereport2009.com/, Feb. 2014.

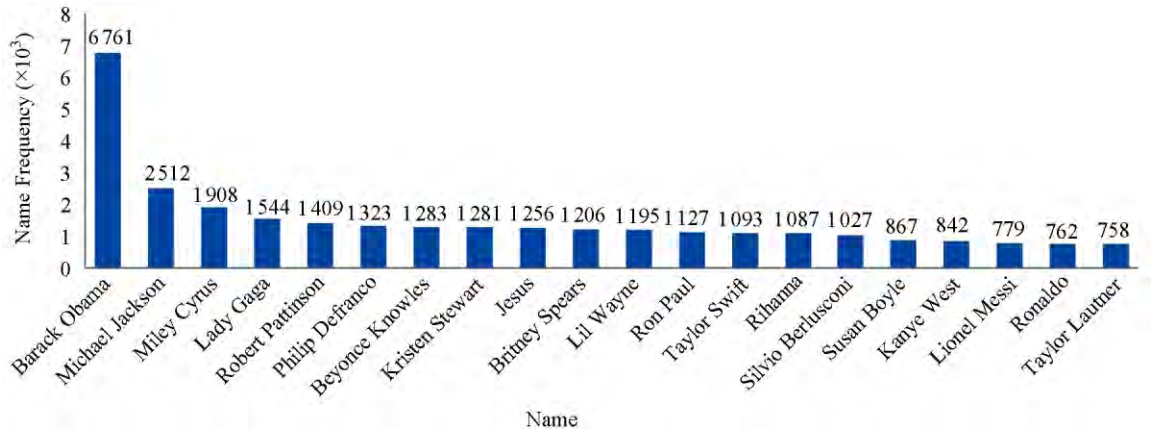


Fig.5. 20 most popular persons and their name occurrences in the dataset. The number on top of each bar is the occurrence of the corresponding celebrity.

tained" as the main reason for browsing YouTube. Note that the mapping between professions and celebrities is not necessarily one-to-one. For example, *Arnold Schwarzenegger* is known as both Artist and Politician. Fig.6 shows the celebrities highlighted with their professions.

3.3 Relationship Mining

Celebrities could be grouped by their social networks. We attempt to discover communities based on the 144 most popular celebrities in CoreData. The mining starts by quantifying the pairwise relationships between the celebrities. We employ the asymmetric co-occurrence measure proposed in [33] for this purpose. Denote $link(n_i, n_j)$ as the relationship between celebrities n_i and n_j, the measure is defined as

$$link(n_i, n_j) = \frac{|n_i \cap n_j|}{|n_i|} + \frac{|n_i \cap n_j|}{|n_j|},$$

where $|n_i|$ is the cardinality of n_i, namely the number of videos with name n_i found in metadata. In this way,

Fig.6. 144 most popular celebrities in the CoreData, ranked in descending order of name frequency from left to right and top to bottom. The bounding box indicates professions, blue: Internet Star, green: Artist, red: Politician, gray: Sportsman, dark red: Journalist.

a sparse graph is constructed by connecting each name to five other closest names, i.e., with the largest $link(n_i, n_j)$. We further employed the Walktrap algorithm[34] to discover the communities by partitioning the graph. There are 12 found communities, of size as small as 4 persons to as large as 26 persons, depicted in Fig.7.

From our analysis, these communities can be linked to both hot topics and celebrities' professions. For example, as shown in Fig.7, the community highlighted by a black dotted circle is a set of famous football stars, while the celebrities in communities highlighted by red and blue dotted circles are about the judges and contestants in "Britain's Got Talent", and the actors and the original author of the movie "Twilight" respectively. Denote $Cele_{x,y}$ as the celebrities with profession identity x and community identity y, C_y as the community with identity y, the community level celebrities co-occurrence $CO_{i,j}$ between professions i and j can be computed by

$$CO_{i,j} = \sum_{Cele_{x=i,y}} \frac{|Cele_{x=i,y}|}{|c_y|}. \quad (1)$$

Using (1), we can get the closeness level of celebrities of one community to celebrities of other communities. Table 1 presents the result, where the relations among five professions are shown as a co-occurrence matrix. The statistics give insights about how communities are formed. For example, celebrities with the same professions (Internet Star, Artist, Sportsman) are likely to be grouped into the same communities. On the other hand, celebrities belonging to closely related professional groups such as Politician and Journalist, Artist and Journalist, frequently appear together in the same communities.

Table 1. Community Level Celebrities' Co-Occurrence Across Different Professions

Profession	I	A	P	S	J
I	**53.1**	27.7	11.9	1.9	5.4
A	11.0	**68.5**	3.1	5.1	12.3
P	13.3	10.1	**37.2**	5.4	34.1
S	2.2	15.3	5.4	**73.6**	3.5
J	6.3	**33.9**	31.6	3.1	25.1

Note: I stands for "Internet Star", A for "Artist", P for "Politician", S for "Sportsman", and J for "Journalist".

3.4 Ground-Truth and Visual Features

To obtain accurate labels for a part of the dataset, we recruited two assessors for manual annotation. The

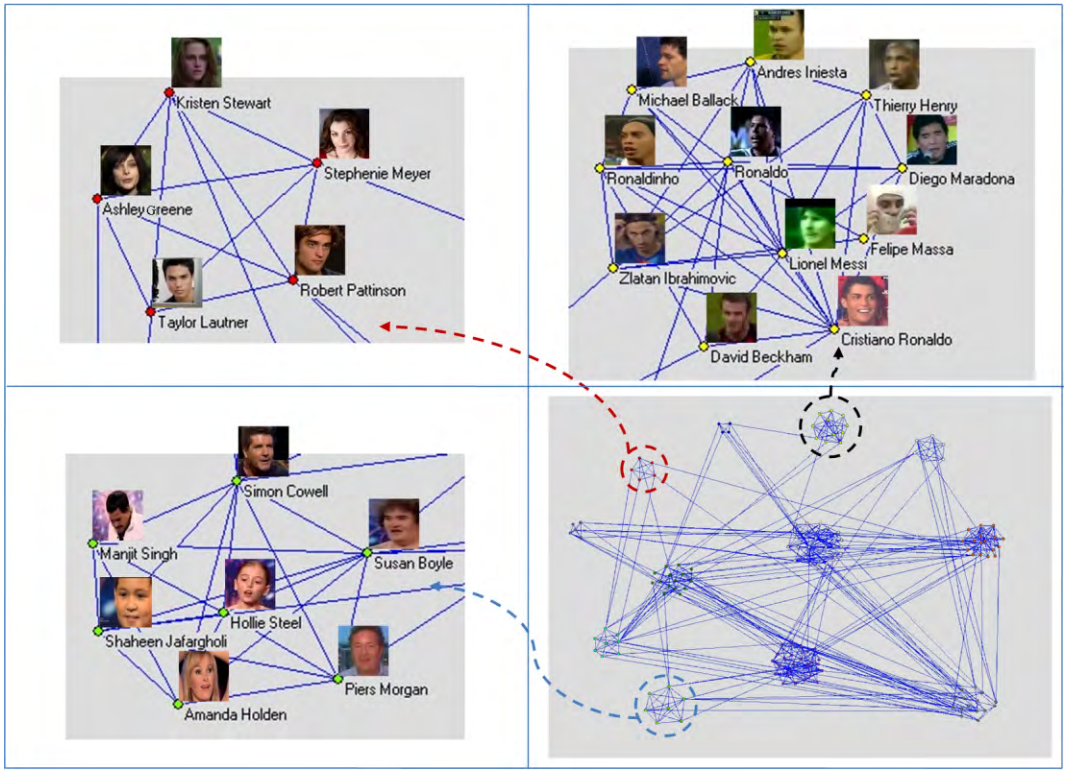


Fig.7. 12 communities discovered by graph partitioning (indicated at the bottom-right) and the zoom-in view of three communities with visual examples. Vertices represent celebrities and edges between them denote their relationships. The colors on vertices represent communities.

procedure was as following. We first collected 3 194 videos from CoreData. These videos were all tagged with at least one of the 144 celebrity names. We then brute-forcely associated every name with all faces found in a video, generating a total of 75 817 name-face pairs to be judged. The assessors were asked to label each pair as "correct", "incorrect", or "hard-to-determine", independently. The two assessors then compared their labeling and resolved inconsistency through discussion. The ground-truth was eventually formed by having 19 216 correct pairs, 55 977 incorrect pairs, and there were 624 pairs labeled as "hard-to-determine". Most "hard-to-determine" pairs were from heavily blurred faces, which are too hard to be recognized.

To facilitate research studies on this dataset, we also split the ground-truth into two subsets, named as *Even* and *Odd* respectively. The partition is based on the month when a video was uploaded. The *Even* (*Odd*) subset contains the videos uploaded during the even (odd) months. Partitioning dataset in this way basically leads to a better separation of videos for the *Even* and *Odd* subsets, as closely related videos (e.g., videos about the same breaking news) are often uploaded within a short period, i.e., a few days. Thus they are less likely to be partitioned into different subsets. The *Even* (*Odd*) subset contains 17 828 (24 290) faces and 28 740 (47 077) name-face pairs.

We also release two sets of visual features for each face. The first set contains: 1) 1937-dimensional (1937-D) pixel-wised signature, and 2) 1664-D SIFT signature, both extracted from 13 facial regions such as eyes and mouths[16]. The second set is composed of six visual properties to represent the head and upper body of a person. These features are: 1) 166-D color histogram, 2) 166-D color correlogram, 3) 225-D color moments, 4) 96-D co-occurrence texture, 5) 108-D wavelet texture gird, and 6) 320-D edge histogram. Readers are referred to [31] for techniques employed for feature extraction.

In summary, the final WebV-Cele dataset contains 2 427 popular celebrities found from 75 073 videos, and 649 011 faces from 56 905 (out of the 75 073 totally, some of which do not contain detected faces) videos along with their visual features. A subset of the dataset, which contains 3 194 videos, is manually labeled to name a total of 42 118 faces against 144 celebrity names.

4 Name-Face Association Baselines

We conduct experiments to evaluate several existing techniques[1,13-14,21], ranging from supervised (unsupervised) learning to knowledge-based inference, for name-face association in the WebV-Cele dataset. Note that name-face association has the following three properties: 1) a face can only associate with a name appearing in the metadata, or *null* if no corresponding name is found (null assignment); 2) a face can be associated with at most one name (uniqueness constraint); and 3) a name can be associated with multiple faces sequentially appearing in a video, but only one face at most in one scene (temporal compatibility). Notably, there are a few heavily edited videos conflicting with the temporal compatibility, i.e., multiple faces of the same individual are found in the same frame because of post-editing. However, we omit these cases here as they are very rare according to our observation.

We adopt five metrics to evaluate the performance of name-face association. The first three are Face Accuracy (FA), Face Precision (FP) and Face Recall (FR), evaluating the performance at the face level. FA is the fraction of correctly associated faces (including null assignments) over all the detected faces. FP is the same as FA, except that null assigned faces are not included for evaluation. The two metrics have both been used in previous work[9,24]. FR calculates the fraction of correctly associated faces over all the labeled celebrity faces. On the other side, Celebrity Precision (CP) and Celebrity Recall (CR) evaluate the performance at the celebrity name level, where CP of name A is the fraction of detected name-face associations referring to A that are correct, and CR of name A is the fraction of faces correctly associated with A over all the labeled faces of A. In all the experiments, we omit faces that are labeled as "hard-to-determine". The baseline methods are briefly described as following:

Weak Association (*WA*). This is a baseline run that brute-forcely associates every face detected in a video to all the names in the metadata, including null assignment. Note that WA does not meet the uniqueness constraint.

SVM Classification (*SVM*)[13]. A one-against-all SVM classifier is trained for each name in the dataset. A face is classified to the name whose classifier outputs the largest score. "Null" is assigned to a face if the largest score does not exceed an empirically set threshold, i.e., a threshold deciding whether the classification is confident enough or not.

Multiple Instance Learning (*MIL*)[14]. Instead of using manually labeled name-face pairs as training examples, MIL uses the positive and negative bags for learning classifiers of person names. A positive bag refers to a video with at least one correct name-face instance. Compared with SVM, MIL requires much less effort in manual labeling since the annotation happens at the video level rather than the face level. The mi-SVM with RBF kernel (an MIL method)[14] is used in this experiment. A face is classified to a name or set as *null*, similar to the aforementioned SVM-based settings.

Graph-Based Clustering (*GC*). A face graph is constructed with faces as nodes and face similarities as edge weights. Name-face association is carried out in an unsupervised manner, by mining the dense subgraphs from the graph, where each subgraph corresponds to faces of a particular name. In the implementation, the iterative approach proposed in [21] is employed to determine the solution, i.e., the subgraphs and their respective names. Different from [21] that associates all faces in a subgraph to the corresponding name, which is likely to mistakenly name the faces of unknown people, we introduce an additional step to classify those dissimilar faces to null assignments. Specifically, the average similarity score between faces within the subgraph is calculated face-by-face. Null assignment is declared if the similarity score of a face is below an empirically set threshold.

Image Matching (*IM*). Since face photos of popular celebrities can be easily searched from the Web, similar to the idea introduced by [1], this method matches a face to the Web images of celebrities. In the implementation, for each name, we crawled the top-10 images from Google Image Search for matching. The *K*NN classifier is used (where $K = 1$) for name-face association. Null assignment is activated if the similarity of a face to its nearest neighbor is below an empirically set threshold.

In the experiment, we ensure that there are at least five positive faces per celebrity for learning and testing, i.e., both the *Even* and *Odd* subsets have at least five positive samples. As a result, only 81 out of 144 celebrities are considered. For SVM and MIL, we use *Even* and *Odd* as training and testing sets alternatively. The reported results are the average of this 2-round testing. To keep consistency, the reported results are also the average on both the *Even* and *Odd* for WC, GC and IM. The 1937-D pixel-wised signature described in Subsection 3.4 is used in SVM, MIL, GC and IM.

Fig.8 shows the FA-FR and FP-FR curves of the five baseline methods. The curves are obtained by calculating FA, FP and FR based on differently set null thresholds. Comparing the two supervised methods, SVM consistently gives better results. This implies the strength of SVM is over MIL, where the learning is conducted at the face level which is more precise than at the bag level. For the unsupervised GC and knowledge-based IM, GC generally has better FP but worse FA. The reasons are: on one hand, cases of associating faces to wrong celebrity names, and associating unknown faces (i.e., faces not labeled by any celebrity name) to celebrity names, are relatively few for GC, leading to higher FP when compared with IM. On the other hand, GC marks many celebrity faces as null assignment, which reduces its FA value with respect to IM. For comparison between supervised and unsupervised methods, MIL performs slightly better than GC and IM in general. From our analysis, on one hand, faces in videos suffer from the effect of low resolution, motion blur, etc. Therefore the visual based face clustering is not robust. It is difficult to name all faces with respect to the corresponding celebrity names except a few highly similar ones. On the other hand, matching faces from Web images, which are usually sharply focused and in high resolution, to poorer quality faces in videos, is also quite difficult.

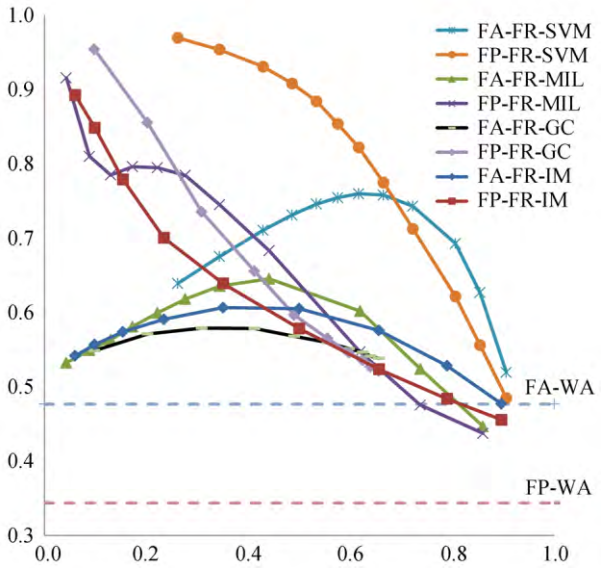


Fig.8. FA-FR and FP-FR curves of the five baseline methods on name-face association. The eight solid lines correspond to the performance of the four visual-based methods, respectively, while the two horizontal dotted lines give FA and FP of WA at FR 1.0.

We manually adjust the null thresholds such that the four visual based methods all have an FR of around 0.5. Table 2 gives the corresponding FAs and FPs. The four visual based methods perform significantly better than the text-based WA. For example, by using SVM, MIL, GC and IM, the improvements are 54.6%, 33.5%, 18.9% and 26.9% in terms of FA, and 161.8%, 81.7%, 72.0% and 68.4% in terms of FP, respectively. The improvements clearly validate the effectiveness of analyzing visual features of faces.

Table 2. FA and FP of the Baseline Methods at an FR of Approximately 0.5

	WA	SVM	MIL	GC	IM
FA	0.476 7	0.737 0	0.636 2	0.566 9	0.605 1
FP	0.343 5	0.899 4	0.624 3	0.590 8	0.578 3

We then look into the CP and the CR on individuals to further evaluate the performance of name-face association at celebrity name level. Table 3 lists the de-

Table 3. CP and CR of the Five Baseline Methods Tested Against 81 Celebrities

Name	Celebrity Precision (CP)					Celebrity Recall (CR)				
	WA	SVM	MIL	GC	IM	WA	SVM	MIL	GC	IM
Adam Lambert	0.052 1	0.043 4	0.250 0	0.000 0	0.151 1	1.000 0	0.062 5	0.031 2	0.000 0	0.406 2
Adolf Hitler	0.141 2	0.783 3	0.144 8	0.201 0	0.410 6	1.000 0	0.777 7	0.986 1	0.388 8	0.416 6
Akon	0.111 2	0.000 0	0.143 5	0.108 3	0.000 0	1.000 0	0.000 0	0.593 7	0.062 5	0.000 0
Alex Jones	0.358 5	0.166 6	0.213 7	0.000 0	0.431 1	1.000 0	0.004 6	0.855 1	0.000 0	0.663 5
Alicia Keys	0.237 9	0.250 0	0.236 1	0.000 0	0.45 10	1.000 0	0.06 12	0.346 9	0.000 0	0.265 3
Amanda Holden	0.076 3	0.713 8	0.674 1	0.000 0	0.270 6	1.000 0	0.611 1	0.734 5	0.000 0	0.518 5
Ana Kasparian	0.15 25	0.777 7	0.174 4	0.000 0	0.337 9	1.000 0	0.518 5	1.000 0	0.000 0	0.333 3
Ashley Greene	0.060 9	0.000 0	0.069 4	0.000 0	0.205 3	1.000 0	0.000 0	0.357 1	0.000 0	0.642 8
Barack Obama	0.221 1	0.932 1	0.809 5	0.326 0	0.420 7	1.000 0	0.420 0	0.641 6	0.667 2	0.447 2
Beppe Grillo	0.315 1	0.829 3	0.834 7	0.741 1	0.552 3	1.000 0	0.416 1	0.705 2	0.771 6	0.450 8
Beyonce Knowles	0.1983	0.9824	0.7155	0.4512	0.4599	1.000 0	0.3958	0.4861	0.4722	0.5763
Bill Reilly	0.208 8	0.750 0	0.105 8	0.312 6	0.289 6	1.000 0	0.239 4	0.718 3	0.464 7	0.704 2
Bubbi Morthens	0.958 6	0.961 5	0.000 0	0.991 2	1.000 0	1.000 0	0.165 4	0.000 0	0.942 4	0.014 3
Cenk Uygur	0.530 8	0.987 0	0.500 0	0.813 8	0.718 8	1.000 0	0.7753	0.3405	0.2536	0.7536
Charlie Mcdonnell	0.699 2	0.960 4	0.500 0	0.718 8	0.741 2	1.000 0	0.263 4	0.166 6	0.962 3	0.688 1
Cheryl Cole	0.055 3	0.171 4	0.541 6	0.000 0	0.133 9	1.000 0	0.095 2	0.761 9	0.000 0	0.714 2
Christian Bale	0.564 9	0.964 2	0.837 6	0.671 4	0.819 3	1.000 0	0.218 3	0.781 6	0.597 7	0.528 7
Christine Gambito	0.607 8	0.986 8	1.000 0	1.000 0	0.929 1	1.000 0	0.669 3	0.282 2	0.387 0	0.475 8
Cory Williams	0.647 6	1.000 0	0.000 0	0.826 3	0.875 0	1.000 0	0.441 1	0.000 0	0.970 5	0.338 2
Cristiano Ronaldo	0.364 3	0.000 0	0.000 0	0.540 3	0.650 0	1.000 0	0.000 0	0.000 0	0.872 3	0.659 5
Dana White	0.409 8	0.500 0	0.000 0	0.369 4	0.407 5	1.000 0	0.026 6	0.000 0	0.933 3	0.506 6
Dan Brown	0.713 9	0.996 0	0.500 0	0.879 8	0.815 5	1.000 0	0.624 5	0.180 5	0.848 3	0.393 5
Dan Johnson	0.659 6	0.722 2	0.000 0	0.611 9	0.437 5	1.000 0	0.419 3	0.000 0	1.000 0	0.451 6
David Guetta	0.557 3	0.500 0	0.000 0	0.000 0	0.500 0	1.000 0	0.054 7	0.000 0	0.000 0	0.753 4
David Letterman	0.448 9	0.936 8	0.875 0	0.492 2	0.395 4	1.000 0	0.704 6	0.894 5	0.776 3	0.037 9
Demi Lovato	0.331 2	0.814 2	0.944 4	1.000 0	0.684 6	1.000 0	0.374 4	0.270 1	0.061 6	0.511 8
Ellen Degeneres	0.143 9	0.373 5	0.561 6	0.000 0	0.294 0	1.000 0	0.358 9	0.641 0	0.000 0	0.666 6
Eminem	0.336 2	0.000 0	0.000 0	0.000 0	0.461 7	1.000 0	0.000 0	0.000 0	0.000 0	0.371 3
Flo Rida	0.487 6	0.000 0	0.000 0	0.500 0	0.500 0	1.000 0	0.000 0	0.000 0	0.796 6	0.135 5
George W. Bush	0.169 2	0.811 3	0.835 2	0.620 1	0.554 1	1.000 0	0.622 9	0.557 3	0.696 7	0.827 8
Glenn Beck	0.270 7	0.829 3	0.867 1	0.741 8	0.697 3	1.000 0	0.147 5	0.267 7	0.469 9	0.431 6
Gucci Mane	0.162 0	0.000 0	0.052 2	0.000 0	0.000 0	1.000 0	0.000 0	0.304 3	0.000 0	0.000 0
Hillary Rodham Clinton	0.033 5	0.159 0	0.031 2	0.000 0	0.144 1	1.000 0	0.235 2	0.294 1	0.000 0	0.529 4
Hollie Steel	0.0549	0.6699	0.5782	0.000 0	0.4583	1.000 0	0.5662	0.6024	0.000 0	0.6385
Iman Crosson	0.207 8	0.187 5	0.252 5	0.000 0	0.416 6	1.000 0	0.093 7	1.000 0	0.000 0	0.156 2
James Cameron	0.146 7	0.000 0	0.000 0	0.061 6	0.083 3	1.000 0	0.000 0	0.000 0	0.272 7	0.090 9
James Rolfe	0.326 8	0.882 3	0.629 4	0.616 8	0.650 0	1.000 0	0.313 4	1.000 0	0.641 7	0.164 1
Jeremy Clarkson	0.236 9	0.200 0	0.338 5	0.170 3	0.289 6	1.000 0	0.025 9	0.792 2	0.402 5	0.285 7
John Green	0.839 4	0.962 2	0.500 0	0.894 9	0.970 5	1.000 0	0.733 1	0.331 5	0.897 5	0.247 9
Justine Ezarik	0.549 2	0.941 8	0.991 9	0.651 8	0.763 0	1.000 0	0.592 9	0.433 6	0.902 6	0.422 5
Justin Bieber	0.483 5	0.745 3	0.860 0	0.661 6	0.665 3	1.000 0	0.3257	0.174 2	0.295 4	0.318 1
Katy Perry	0.127 8	0.192 3	0.162 4	0.000 0	0.344 1	1.000 0	0.098 0	0.803 9	0.000 0	0.176 4
Kevjumba	0.483 9	0.982 2	1.000 0	0.860 8	0.808 2	1.000 0	0.765 4	0.415 9	0.712 3	0.787 6
Kobe Bryant	0.212 6	0.500 0	0.075 0	0.327 9	0.433 3	1.000 0	0.037 0	0.444 4	0.851 8	0.481 4
Kristen Stewart	0.137 1	0.490 4	0.391 0	0.000 0	0.243 5	1.000 0	0.12 75	0.536 9	0.000 0	0.583 8
Lady Gaga	0.172 6	0.786 7	0.6676	0.278 5	0.421 2	1.000 0	0.379 4	0.694 5	0.386 6	0.568 0
Lauren Luke	0.975 7	0.998 7	1.000 0	0.986 5	0.995 5	1.000 0	0.706 5	0.177 2	0.698 7	0.527 5
Lil Wayne	0.242 8	0.770 8	0.351 0	0.208 5	0.663 3	1.000 0	0.138 6	0.712 8	0.198 0	0.277 2
Lindsay Lohan	0.258 1	0.500 0	0.000 0	0.000 0	0.462 1	1.000 0	0.017 8	0.000 0	0.000 0	0.785 7
Lisa Donovan	0.275 5	0.500 0	0.306 8	0.330 4	0.364 5	1.000 0	0.044 0	0.290 7	0.603 5	0.568 2
Mahmoud Ahmadinejad	0.222 2	0.500 0	0.125 0	0.270 5	0.385 1	1.000 0	0.100 0	0.250 0	0.900 0	0.550 0
Manny Pacquiao	0.229 1	0.820 5	0.703 9	0.232 8	0.305 8	1.000 0	0.242 9	0.411 2	0.504 6	0.542 0
Marco Travaglio	0.925 9	0.998 1	0.998 1	0.000 0	0.985 8	1.000 0	0.729 2	0.579 3	0.000 0	0.717 6

(to be continued)

Table 3. (continued)

Name	Celebrity Precision (CP)					Celebrity Recall (CR)				
	WA	SVM	MIL	GC	IM	WA	SVM	MIL	GC	IM
Mariah Carey	0.232 2	0.879 7	0.176 0	0.000 0	0.641 0	1.000 0	0.512 3	0.727 2	0.000 0	0.198 3
Marina Orlova	0.694 4	0.993 4	1.000 0	0.890 9	0.935 4	1.000 0	0.511 2	0.458 8	0.885 7	0.514 9
Megan Fox	0.068 4	0.458 3	0.436 9	0.000 0	0.228 3	1.000 0	0.090 9	0.727 2	0.000 0	0.545 4
Michael Buckley	0.616 6	0.920 0	0.880 9	0.000 0	0.501 7	1.000 0	0.767 8	0.873 1	0.000 0	0.113 2
Michael Jackson	0.174 7	0.621 0	0.506 0	0.215 7	0.481 4	1.000 0	0.208 0	0.463 6	0.538 8	0.340 8
Miley Cyrus	0.156 0	0.222 2	0.582 5	0.246 9	0.301 2	1.000 0	0.056 4	0.254 1	0.489 4	0.552 9
Natalie Tran	0.882 1	0.994 1	0.998 9	0.976 2	0.972 9	1.000 0	0.789 9	0.580 6	0.940 3	0.504 2
Nigahiga	0.529 6	0.945 0	0.987 1	0.875 8	0.542 2	1.000 0	0.383 0	0.345 7	0.325 4	0.366 1
Oprah Winfrey	0.047 3	0.400 0	0.250 0	0.000 0	0.230 5	1.000 0	0.166 6	0.125 0	0.000 0	0.833 3
Perez Hilton	0.181 0	0.416 6	0.044 1	0.000 0	0.313 8	1.000 0	0.175 4	0.105 2	0.000 0	0.333 3
Peter Schiff	0.102 7	0.166 6	0.064 8	0.000 0	0.151 7	1.000 0	0.043 4	0.304 3	0.000 0	0.391 3
Philip Defranco	0.801 3	0.997 4	0.999 0	0.910 2	0.958 2	1.000 0	0.742 4	0.547 0	0.777 6	0.701 3
Piers Morgan	0.067 5	0.884 0	0.793 7	0.000 0	0.150 1	1.000 0	0.550 3	0.664 4	0.000 0	0.308 7
Richard Hammond	0.234 0	0.878 2	0.340 3	0.000 0	0.380 8	1.000 0	0.220 7	0.857 1	0.000 0	0.181 8
Ricky Hatton	0.097 4	0.000 0	0.018 9	0.000 0	0.400 0	1.000 0	0.000 0	0.263 1	0.000 0	0.631 5
Rihanna	0.158 8	0.875 0	0.946 1	1.000 0	0.434 2	1.000 0	0.035 5	0.514 7	0.094 6	0.325 4
Robert Pattinson	0.1104	0.6156	0.3448	0.1181	0.2035	1.000 0	0.3407	0.4592	0.0888	0.4666
Ron Paul	0.161 4	0.878 4	0.746 6	0.704 5	0.491 1	1.000 0	0.648 9	0.706 1	0.044 8	0.742 8
Rush Limbaugh	0.119 7	0.541 6	0.625 0	0.000 0	0.773 8	1.000 0	0.147 0	0.058 8	0.000 0	0.441 1
Sarah Palin	0.081 2	0.588 2	0.030 0	0.007 5	0.131 6	1.000 0	0.464 2	0.214 2	0.035 7	0.750 0
Sean Hannity	0.092 4	0.244 7	0.194 0	0.000 0	0.220 4	1.000 0	0.413 7	0.620 6	0.000 0	0.517 2
Selena Gomez	0.176 7	0.496 5	0.534 9	0.390 8	0.379 7	1.000 0	0.072 7	0.490 9	0.436 3	0.672 7
Shaheen Jafargholi	0.139 8	0.672 6	0.746 4	0.000 0	0.448 0	1.000 0	0.449 7	0.373 2	0.000 0	0.861 2
Simon Cowell	0.129 8	0.893 4	0.864 4	0.000 0	0.517 2	1.000 0	0.620 3	0.786 0	0.000 0	0.243 3
Soulja Boy	0.180 4	0.142 8	0.184 3	0.100 0	0.375 0	1.000 0	0.057 1	0.942 8	0.028 5	0.057 1
Susan Boyle	0.257 6	0.961 2	0.986 2	0.587 6	0.707 1	1.000 0	0.579 9	0.606 5	0.470 9	0.748 1
Taylor Lautner	0.089 3	0.884 6	0.226 9	0.000 0	0.215 5	1.000 0	0.185 7	0.342 8	0.000 0	0.400 0
Taylor Swift	0.201 2	0.523 3	0.556 8	0.269 9	0.299 5	1.000 0	0.237 3	0.540 4	0.474 7	0.363 6
Average	0.171 8	0.899 4	0.624 3	0.590 8	0.578 3	1.000 0	0.503 7	0.504 3	0.501 5	0.499 9

tailed CPs and CRs of the five methods tested against the 81 celebrities. The performance varies greatly among different methods as well as celebrities. For example, by using SVM and MIL, the CP for *Taylor Lautner* is 0.884 6 and 0.226 9 respectively, while the corresponding CR is 0.185 7 and 0.342 8, respectively. The results from different methods appear to be complementary, where in this case, the fusion of the results is likely to boost the performance. Variations in face appearance also result in distinctions in performance across different celebrities. Some faces are easy to name (e.g., *Natalie Tran*) compared with others (e.g., *Hillary Rodham Clinton*) where all the methods show low performance.

To further analyze the problem and understand the pros and cons of each method, Fig.9 details the results with respect to the celebrities' profession. The performance also varies across different profession groups. For example, the CP of Internet Star is much higher than that of the other four profession groups in general. The results could roughly explain the performance differences among celebrities. Videos with Internet stars (e.g., *Natalie Tran*) are typically vloggings, with a celebrity in front of a camera for delivering show. Using WA only can already achieve a CP of more than 60% on average. In contrast, celebrities of professional groups such as Political and Sportsman, whose faces are often small and exhibit large variations in visual appearance, are often captured "in the wild". Finally, the visual feature used in these baseline methods is probably not sufficient for resolving this challenging problem. More sophisticated method considering more advanced features is expected to improve the results.

5 Future Research Directions

While we have presented the experimental results of several baseline techniques on this dataset, there are still plenty of challenges ahead. We envision three problems that can be studied on the WebV-Cele dataset.

5.1 Cross-Video Face Tagging

Our experiments considered only within video association. In real scenarios, there could exist faces in a video that have no name in its metadata, and vice versa.

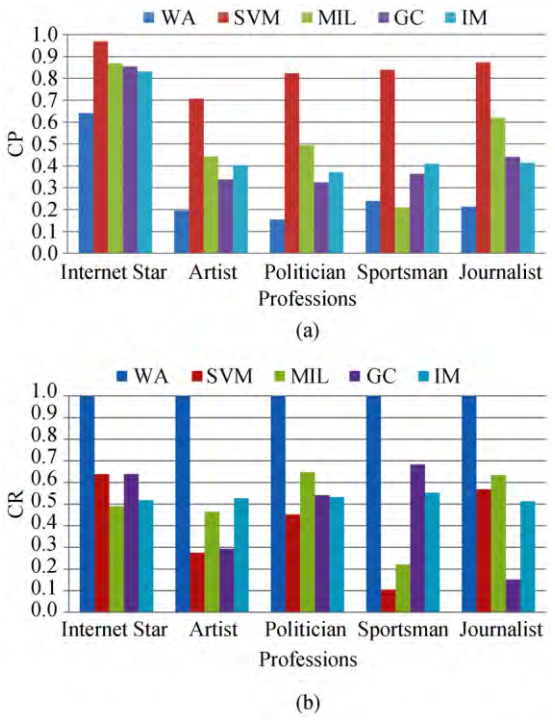


Fig.9. Celebrity precision and celebrity recall of the five baseline methods tested against 81 celebrities with respect to the celebrities' professions. (a) CP. (b) CR.

Thus, how to label faces with missing names and how to omit names with no face presence become important problems. A potential solution is by modeling the proximity of faces detected from different videos as a graph, for example, and then propagating the tagging results from individual videos to similar faces through inference. The scalability issue, however, could be a challenging problem when there is a large number of faces to be associated, e.g., in the WebV-Cele with 0.6 million of faces.

5.2 Context-Aware Labeling

Rich context information exists in Web videos, though some of which could be noisy. Here, we briefly summarize three dimensions of context: temporal, social and visual background, which are worth further exploration. In videos, the appearance of faces could change over time, that is, multiple faces may appear together in one scene or separately in different scenes. The simple fact that two faces which happen in a scene should not be labeled as the same name already gives constraints such as how false labels could be removed when tagging along the temporal dimension[11,21-22]. There are also various social clues in video sharing websites that could be leveraged for tagging[30-31,35-36]. These clues include the view count, author, channel, geolocation, and "related videos" to a video. By analyzing these social signals, the task of name-face association could be modeled as a probabilistic framework. For example, discovering the interest of an author through his or her uploading history could probably help predict the celebrities in the next video he or she will upload, and relating videos of high view counts to hot news at the uploading time could also reveal groups of celebrities that will appear in the videos. Lastly, visual background context like the clothes of celebrities or scenes and objects that often appear together with some celebrities also gives valuable clues for face tagging.

5.3 Community Discovery for Name-Face Association

As analyzed in Subsection 3.3, the groups of celebrities that appear together in videos are tightly correlated with their underlying communities. This presents two interesting dimensions of the problem: how to leverage the knowledge of communities such as celebrities' professions to assist the association; and how to apply results of name-face association for social network analysis and discovery. Both dimensions of problems can be iteratively solved, by leveraging a known community structure to assist name-face association, and by using the results of association to refine the community structure and discover new relationship.

6 Conclusions

This paper has presented and released the WebV-Cele dataset, a benchmark that is expected to stimulate research on large-scale name-face association in the Web video domain. The dataset consists of 2 427 celebrities covering a wide range of subjects and professions, and has totally 649 001 faces with diverse visual appearances. To our knowledge, this is the largest video dataset with the most diverse content for this problem.

We have described the construction of dataset and discussed some statistics regarding the distributions and social networks of celebrities. Experiments were conducted on a subset of 3 194 videos containing 42 118 faces and 144 celebrities, using five baseline approaches on name-face association including supervised, unsupervised and knowledge-based methods. The results are encouraging but also reveal several challenges of the problem. In future, we are interested in developing a user-friendly tool for the fast and accurate annotation of celebrity videos, and incorporating advanced semi-supervised methods in order to exploit the large and comprehensive dataset.

The research on name-face association in Web videos is still in its infancy stage. There are many issues deserving investigations, among which we have discussed

three interesting and important ones that could be explored using this dataset in the future.

References

[1] Zhao M, Yagnik J, Adam H *et al.* Large scale learning and recognition of faces in Web videos. In *Proc. the 8th IEEE FGR*, Sept. 2008, pp.1-7.

[2] Zhang X, Zhang L, Wang X J, Shum H Y. Finding celebrities in billions of Web images. *IEEE Trans. Multimedia*, 2012, 14(4): 995-1007.

[3] Xie H, Zhang Y, Tan J, Guo L, Li J. Contextual query expansion for image retrieval. *IEEE Trans. Multimedia*, 2014, 16(4): 1104-1114.

[4] Yao T, Ngo C W, Mei T. Circular reranking for visual search. *IEEE Trans. Image Processing*, 2013, 22(4): 1644-1655.

[5] Liu J, Huang Z, Cai H, Shen H T, Ngo C W, Wang W. Near-duplicate video retrieval: Current research and future trends. *ACM Computing Surveys*, 2013, 45(4): Article No.44.

[6] Zhang L, Zhang Y, Gu X, Tang J, Tian Q. Scalable similarity search with topology preserving hashing. *IEEE Trans. Image Processing*, 2014, 23(7): 3025-3039.

[7] Chen Z, Cao J, Xia T *et al.* Web video retagging. *Multimedia Tools and Application*, 2011, 55(1): 53-82.

[8] Berg T L, Berg A C, Edwards J *et al.* Names and faces in the news. In *Proc. the 2004 IEEE CVPR*, Jun. 2004, 2: 848-854.

[9] Bu J, Xu B, Wu C *et al.* Unsupervised face-name association via commute distance. In *Proc. the 20th ACM Multimedia*, Oct. 29-Nov. 2, 2012, pp.219-228.

[10] Satoh S, Nakamura Y, Kanade T. Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 1999, 6(1): 22-35.

[11] Pham P T, Tuytelaars T, Moens M F. Naming people in news videos with label propagation. *IEEE MultiMedia*, 2011, 18(3): 44-55.

[12] Pham P T, Deschacht K, Tuytelaars T, Moens M F. Naming persons in video: Using the weak supervision of textual stories. *J. Visual Communication and Image Representation*, 2013, 24(7): 944-955

[13] Yang J, Hauptmann A G. Naming every individual in news video monologues. In *Proc. the 12th Annual ACM Multimedia*, Oct. 2004, pp.580-587.

[14] Yang J, Yan R, Hauptmann A G. Multiple instance learning for labeling faces in broadcasting news video. In *Proc. the 13th Annual ACM Multimedia*, Oct. 2005, pp.31-40.

[15] Duygulu P, Hauptmann A. What's news, what's not? Associating news videos with words. In *Proc. the 3th CIVR*, Jul. 2004, pp.132-140.

[16] Everingham M, Sivic J, Zisserman A. Hello! My name is ... buffy — Automatic naming of characters in TV video. In *Proc. the 17th BMVC*, Sept. 2006, pp.889-908.

[17] Ramanan D, Baker S, Kakade S. Leveraging archival video for building face datasets. In *Proc. the 11th ICCV*, Oct. 2007, pp.1-8.

[18] Cinbis R G, Verbeek J, Schmid C. Unsupervised metric learning for face identification in TV video. In *Proc. the 13th ICCV*, Nov. 2011, pp.1559-1566.

[19] Bäuml M, Tapaswi M, Stiefelhagen R. Semi-supervised learning with constraints for person identification in multimedia data. In *Proc. the 26th IEEE CVPR*, Jun. 2013, pp.3602-3609

[20] Zhang Y F, Xu C, Lu H *et al.* Character identification in feature-length films using global face-name matching. *IEEE Trans. Multimedia*, 2009, 11(7): 1276-1288.

[21] Guillaumin M, Mensink T, Verbeek J, Schmid C. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 2012, 96(1): 64-82.

[22] Ozkan D, Duygulu P. Interesting faces: A graph-based approach for finding people in news. *Pattern Recognition*, 2010, 43(5): 1717-1735.

[23] Guillaumin M, Verbeek J, Schmid C. Multiple instance metric learning from automatically labeled bags of faces. In *Proc. the 11th ECCV*, Sept. 2010, pp.634-647.

[24] Ozcan M, Jie L, Ferrari V *et al.* A large-scale database of images and captions for automatic face naming. In *Proc. the 22nd BMVC*, Aug. 29-Sept. 2, 2011, Article No. 29.

[25] Huang G B, Ramesh M, Berg T *et al.* Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[26] Wolf L, Hassner T, Maoz I. Face recognition in unconstrained videos with matched background similarity. In *Proc. the 2011 IEEE CVPR*, Jun. 2011, pp.529-534.

[27] Chen Z, Ngo C W, Cao J, Zhang W. Community as a connector: Associating faces with celebrity names in Web videos. In *Proc. the 20th ACM Multimedia*, Oct. 29-Nov. 2, 2012, pp.809-812.

[28] Ruiz-del-Solar J, Verschae R, Correa M. Recognition of faces in unconstrained environments: A comparative study. *EURASIP Journal on Advances in Signal Processing*, 2009, pp.1-19

[29] Wang D, Hoi S C H, He Y, Zhu J. Mining weakly labeled Web facial images for search-based face annotation. *IEEE Trans. Knowledge and Data Engineering*, Jan. 2014, 26(1): 166-179.

[30] Stone Z, Zickler T, Darrell T. Toward large-scale face recognition using social network context. *Proceedings of the IEEE*, 2010, 98(8): 1408-1415.

[31] Cao J, Zhang Y D, Song Y C *et al.* MCG-WEBV: A benchmark dataset for Web video analysis. Technical Report, Institute of Computing Technology, CAS, May 2009.

[32] Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data. *SIAM Review*, 2009, 51(4): 661-703.

[33] Sigurbjornsson B, Zwol R V. Flickr Tag recommendation based on collective knowledge. In *Proc. the 17th Int. Conf. World Wide Web*, Apr. 2008, pp.327-336.

[34] Pons P, Latapy M. Computing communities in large networks using random walks. In *Proc. the 20th ISCIS*, Oct. 2005, pp.284-293.

[35] Song Y C, Zhang Y D, Cao J, Xia T, Li J T. Web video geolocation by geotagged social resources. *IEEE Trans. Multimedia*, 2012, 14(2): 456-470.

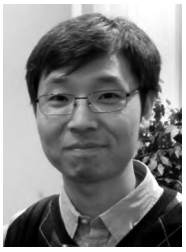
[36] Wu X, Ngo C W, Hauptmann A G, Tan H K. Real-time near-duplicate elimination for Web video search with content and context. *IEEE Trans. Multimedia*, 2009, 11(2): 196-207.

Zhi-Neng Chen received his B.S. and M.S. degrees in computer science from the College of Information Engineering, Xiangtan University, China, in 2004 and 2007, respectively, and Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2012. He is currently an assistant professor of the Institute of Automation, Chinese Academy of Sciences, Beijing. He was a senior research associate with the Department of Computer Science, City University of Hong Kong, in 2012. His current research interests include large-scale multimedia information retrieval and video processing.

Chong-Wah Ngo received his B.S. and M.S. degrees in computer engineering from Nanyang Technological University, Singapore, and Ph.D. degree in computer science from the Hong Kong University of Science and Technology. He was a postdoctoral scholar with the Beckman Institute, University of Illinois at Urbana-Champaign, Champaign. He was a visiting researcher with Microsoft Research Asia. He is currently an associate professor with the Department of Computer Science, City University of Hong Kong. His current research interests include large-scale multimedia information retrieval, video computing, and multimedia mining. Dr. Ngo is currently an associate editor of the IEEE Transactions on Multimedia. He is the conference co-chair of the ACM International Conference on Multimedia Retrieval 2015 and Pacific Rim Conference on Multimedia 2014, the program co-chair of the ACM Multimedia Modeling Conference 2012, and the ACM International Conference on Multimedia Retrieval 2012, and the area chair of the ACM Multimedia 2012. He was the chairman of the ACM (Hong Kong Chapter) from 2008 to 2009.

Wei Zhang received his B.E. degree in computer science from School of Computer Software, M.E. degree in computer science from the School of Computer Science and Technology in Tianjin University, China, in 2008 and 2011, respectively. He is now a Ph.D. candidate in computer science with the Department of Computer Science in City University of Hong Kong, and a visiting student of the Department of Electrical Engineering, Columbia University, New York, in 2014. His research interests include large-scale video retrieval and digital forensic analysis.

Juan Cao received her Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2008. She is currently an associate professor of the Institute of Computing Technology, Chinese Academy of Sciences. She was a senior research associate of the Department of Computer Science, City University of Hong Kong, in 2009, and a visiting scholar of the Department of Electrical Engineering, Columbia University, New York, in 2010. Her current research interests focus on large-scale Web video processing and multimedia retrieval.



Yu-Gang Jiang received his Ph.D. degree in computer science from the City University of Hong Kong, in 2009. During 2008∼2011, he was with the Department of Electrical Engineering, Columbia University, New York. He is currently an associate professor of computer science with Fudan University, Shanghai. His research interests include multimedia retrieval and computer vision. Dr. Jiang is an active participant of the Annual U.S. NIST TRECVID Evaluation and has designed a few top-performing video analytic systems over the years. His work has led to a Best Demo Award from ACM Hong Kong, the second prize of ACM Multimedia Grand Challenge 2011, and a recognition by IBM T. J. Watson Research Center as one of ten "Emerging Leaders in Multimedia" in 2009. He has served on the program committees of many international conferences and is a guest editor of a special issue on Socio-Mobile Media Analysis and Retrieval, IEEE Transactions on Multimedia.