Community as a Connector: Associating Faces with Celebrity Names in Web Videos

Zhineng Chen[1], Chong-Wah Ngo[1], Juan Cao[2], Wei Zhang[1]
{zhinchen, cscwngo}@cityu.edu.hk, caojuan@ict.ac.cn, wzhang34@student.cityu.edu.hk

[1] Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong

[2] Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China

ABSTRACT
Associating celebrity faces appearing in videos with their names is of increasingly importance with the popularity of both celebrity videos and related queries. However, the problem is not yet seriously studied in Web video domain. This paper proposes a *Community connected Celebrity Name-Face Association* approach (C-CNFA), where the community is regarded as an intermediate connector to facilitate the association. Specifically, with the names and faces extracted from Web videos, C-CNFA decomposes the association task into a three-step framework: community discovering, community matching and celebrity face tagging. To achieve the goal of efficient name-face association under this umbrella, algorithms such as the constrained density-based clustering and exemplar based voting are developed by leveraging different pieces of visual and contextual cues. The evaluation on 0.4 million faces and 144 celebrities shows the effectiveness of the proposed C-CNFA approach. Moreover, using the obtained associations, encouraging results are reported in celebrity video ranking.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Algorithms, Measurement, Experimentation

Keywords
Name-face association, celebrity videos, community analysis

1. INTRODUCTION
With the prosperity of video sharing websites, Web videos are being captured and populated at an accelerating rate. Among the huge deposit of videos and query logs in these websites, as reported in [7], many of them are indeed about celebrities. However, the user-provided tags are often incomplete or even noisy, and mostly locate at video level rather than segment or keyframe level. The browsing, searching and ranking of celebrity videos by keywords as performed by commercial search engines are by no means efficient. For example, ranking videos by celebrity names often results in less meaningful lists that occasionally rank videos without the appearance of the searched celebrity at the top of the list, given that there are a large number of videos tagged with celebrities.

MM'12, October 29–November 2, 2012, Nara, Japan.

To enable the content-based access of celebrity videos, a key technique is by tagging celebrity faces with their names. In other words, tags are labeled directly at the face level rather than video level. The task is generally referred to as name-face association [1, 2]. In the literature, there have been several researches undergone for this task in the domain of news videos [3, 4], episodes of the TV series [5], movies [6] and news images [1, 2]. Different from processing Web videos, where most techniques suffer from the sparse text problem, faces extracted from these domains are often rich of textual information. For instance, the transcripts of movies and news videos already provide a vivid text cue for face tagging, especially when the timeline information is available to synchronize the transcript and video segments. It is also worth to notice that name-face association in Web videos is generally a much harder task than other domains, due to the unconstrained video capturing environment and unpredictable number of celebrity names and faces.

In this paper, we investigate the association of celebrity names and faces in Web video domain. This domain often covers celebrities with a wide range of professions and different levels of interaction. Therefore, instead of directly associating faces with names as done by traditional approaches [1, 2, 6], we propose a *Community connected Celebrity Name-Face Association* approach (C-CNFA), where the name and face communities in Web video collections are exploited and used as intermediate connectors to facilitate the association. Specifically, with the extracted names and faces, C-CNFA decomposes the association task into a three-step framework: community discovering, community matching and celebrity face tagging. It associates faces with names following the path of face - face community - name community - name, as illustrated in Figure 1. To achieve efficient association, several novel algorithms are developed. In community discovering, name and face communities are discovered by leveraging the co-occurrence between names or face clusters. A face cluster, which is assumed to contain faces of only one person, is generated by first performing the constrained density-based clustering video-by-video and then applying the agglomerative clustering on all videos. In community matching, a "soft weighting" strategy is proposed to simultaneously match a face community to multiple name communities with the largest video co-occurrence, i.e., the percentage of common videos where the names and faces are extracted from. It thus avoids name-face mismatches as much as possible at community level. While in celebrity face tagging, an exemplar based voting method is employed. Video co-occurrence and celebrity faces on the Web are jointly considered to label faces in matched communities. We conduct experiments on 0.4 million faces and 144 celebrities, in which the effectiveness of the C-CNFA approach is demonstrated. In addition, we perform evaluation on celebrity video ranking, from which better ranking lists are generated by incorporating the obtained associations.

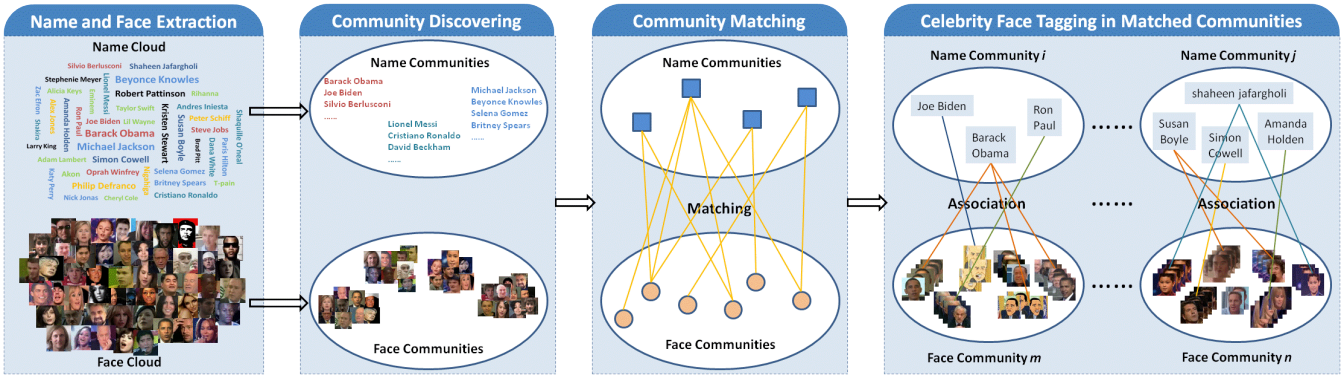

Figure 1: Illustrative flowchart of the proposed community connected celebrity name-face association approach.

2. THE C-CNFA APPROACH

2.1 Community Discovering

We first extract names from metadata (titles and tags) surrounding videos. The metadata often comprises words or phrases lacking a standard grammar structure. It is thus difficult to analyze it by traditional natural language processing techniques. Therefore, we propose a Wikipedia based name extraction method to identify celebrity names. In particular, names are extracted by stepwise testing whether a word or a succession of words in metadata could represent the name of a person, by applying a heuristic rule that a name always has a category tag representing the birth year of the person (e.g., 1945 births) in its Wikipedia page.

Generally speaking, the co-occurrence of names somewhat reflects the interaction of celebrities, especially when looking into a large number of videos. Therefore, we build a graph $G_{name} = \langle V_n, E_n, W_n \rangle$, where $V_n = \{n_1, n_2, \cdots, n_m\}$ represents m names, W_n is the weight matrix whose element $w_{i,j}$ is defined as

$$w_{i,j} = \frac{|S_i \cap S_j|}{|S_i|} + \frac{|S_i \cap S_j|}{|S_j|} \quad (1)$$

where S_i is the set of videos with name n_i found in metadata and $|S_i|$ denotes its cardinality. Thus, a sparse edge matrix E_n is constructed by connecting each name to five other names with the largest weights. Based on G_{name}, we further employ the Walktrap algorithm [9] to discover the inherent name communities by splitting the graph. Names in the same community are assumed to interact more frequently than in different communities.

We divide the task of discovering face community into a three-step scheme: within-video clustering, among-video clustering and face community discovering. The scheme aims at utilizing contextual and visual cues originated from different aspects to cope with the large variance of faces detected from Web videos.

The objective of within-video clustering is to produce one or more face clusters for each video, where one cluster contains only faces of one person. There are several cues could be exploited for the clustering. First, multiple faces that spatially appear together in a keyframe usually reside in different face clusters. These faces should not be assigned to the same name. Second, similar faces which are temporally close in the time axis (e.g., from multiple keyframes of a shot) are likely to belong to the same person. These two cues can be served as "cannot-link" and "must-link" constraints in the clustering. Third, a person who appears in a video is likely to have the same color of hair and to wear the same clothes. Thus, in addition to measuring the similarity of face appearance, head and upper bodies also provide strong contextual cues to reveal the identity of a person. Therefore the similarity $s_{i,j}$ between two faces f_i and f_i is defined as

$$s_{i,j} = \exp\left(-\frac{d_{i,j}}{\sigma}\right), \quad \text{where } d_{i,j} = \frac{d_{i,j}^{face} + d_{i,j}^{head} + \delta_i \cdot \delta_j \cdot d_{i,j}^{body}}{2 + \delta_i \cdot \delta_j} \quad (2)$$

where σ is a parameter balancing the similarities between faces. δ_i is a binary value denoting the appearance of upper body of face f_i in the keyframe, as the upper body may not appear together with the face. The superscripts in Eq. 2, namely *face*, *head* and *body*, represent different human parts. Particularly, two 166-dimensional color histogram features are respectively extracted from head and upper body. A 1937-dimensional feature extracted from 13 facial regions [5] is used as the face descriptor. Given the high dimensionality, we perform PCA to reduce the dimension of the face descriptor to 100. Based on these constraints and features, a constrained density-based clustering algorithm is employed to group faces belonging to the same person.

The first step will produce one to several face clusters for a video. As a person name often appears in many videos, it is common that faces of the person are scattered across different face clusters come from different videos. Since the face appearance of a person seldom changes, faces from two clusters, both of the same person, often show large similarities. However, it is also not surprised that a portion of faces are not tightly grouped in the feature space, due to the variance of facial expression, capture environment, et al. Therefore, similar to [6], a bi-threshold method is proposed to merge face clusters across videos. Two face clusters $fc_i = \{f_1, f_2, \cdots f_m\}$ and $fc_j = \{f_1, f_2, \cdots f_n\}$ are merged into a new cluster if the minimum and average distance between faces in the two clusters are both below two predefined thresholds T_{min} and T_{ave}. Note that only the 100-dimensional face descriptor is employed for this clustering, as a person is not necessarily to wear the same clothes and have unchanged hairstyle. The clustering is repeatedly performed until all face clusters in the video collection meeting the two thresholds are merged. Note that a merged face cluster may contain faces from both tagged and untagged videos.

Given a number of merged face clusters each assumed to contain the faces of a specified person from different videos, the interaction between two face clusters can be measured based on their video co-occurrence, e.g., by Eq. 1. Therefore, the third step, namely face community discovering, is performed by firstly constructing a face cluster graph G_{face} and then splitting the graph to face communities following the same procedure as in discovering name communities.

2.2 Community Matching

The matching between the discovered name and face communities is also measured by video co-occurrence. Since a face can

only be assigned to at most one name, intuitively, the matching from face to name should be one-to-one. Nevertheless, as community discovery is not always perfect, imposing the constraint could run into risk that the face tagging (Section 2.3) will completely fail if the matching is incorrect. For robustness, we instead consider one-to-many matching by allowing a face community simultaneously match with multiple name communities, which we name this strategy "soft weighting". By doing so, the matching between communities is less sensitive to video co-occurrence. In our implementation, each face community basically matches to *K* name communities having the largest video co-occurrence. The sensitivity of *K* will be analyzed in Section 3.2.2.

2.3 Celebrity Face Tagging

With the matched communities, the task becomes tagging faces in a face community using the names in the matched name communities, i.e., celebrity face tagging.

Celebrity face tagging is a task by no means easy, as the interaction between names and faces is relatively sparse at the name or face cluster level, i.e., many names and face clusters come from totally non-overlapped video sets. Therefore their interaction is immeasurable. This is defined as the sparse interaction problem. We propose to use external cues from the Web to address the sparse interaction problem. Specifically, an exemplar based voting method is proposed to utilize celebrity faces on the Web to tag faces. Since celebrity photos can be easily searched from Web, we crawl the top 64 images of every celebrity from Google Image Search. Face detection is performed again on the images and those with only one face response are kept as exemplars.

Denote $FC = \{f_1, f_2, \cdots, f_n\}$ as a face cluster in the face community, $N = \{n_1, n_2, \cdots, n_m\}$ as one of the matched name communities, $E_j = \{e_1, e_2, \cdots, e_p\}$ as the p exemplars of name $n_j \in N$. The association score between FC and n_j is defined as

$$s(n_j) = \frac{1}{n}\sum_{i=0}^{n} \delta\left(p(f_i), n_j\right) * c(f_i) \quad (3)$$

where $p(f_i)$ is the label of face f_i determined by a *5-NN* classifier, and $c(f_i)$ is the confidence of assigning the label measured based on majority voting of the five nearest neighbors. The function $\delta(x, y)$ is an indicator function that equals to 1 if $x = y$, and 0 otherwise. By Eq. 3, the association scores between FC and all the candidate names can be measured, and the name with the largest score is assigned as the label of the FC, provided that the score is larger than a given threshold. Otherwise, FC is classified as an unknown face cluster.

3. PERFORMANCE EVALUATION

3.1 Dataset

We construct a Web video celebrity dataset named Cele-WebV on top of the MCG-WEBV core videos [8]. The dataset contains 14,473 representative videos evenly crawled from the 15 predefined YouTube channels during Dec. 2008 to Nov. 2009. The videos have been decomposed into shots, from which 409,900 keyframes containing 133,997 frontal faces (20.0% are close-up views, i.e., resolution larger than 150*150) are provided. Note that 66.8% of videos contain faces, and 43.5% of faces appear together with upper bodies. There are 13.9 faces per video on average.

By employing the proposed Wikipedia based name extraction method, 3,621 names corresponding to 3,231 different persons are obtained. By merging names of the same person, 144 names appeared at least ten times are extracted. These names construct the celebrity names of Cele-WebV. Note that 22.1% of videos have celebrity names, and these videos contain 1.6 celebrity names on average.

To label the dataset, we brute-forcely associate every celebrity name and face found in a video, generating a total of 74,540 name-face pairs to be judged. Two assessors are recruited to label the pairs one-by-one independently. An extra human judge is introduced if the two assessors provide inconsistent labels. The ground-truth is eventually formed by having 19,240 name-face associations, which is used both in evaluating the performance of the proposed C-CNFA approach and celebrity video ranking. Note that for celebrity video ranking, a video is defined as a positive sample of celebrity *j* if and only if it contains at least one face of *j*.

3.2 Evaluation

3.2.1 Community discovering

We first evaluate the performance of the two face clustering methods. Specifically, *p-accuracy* is proposed as the evaluation metric, which is defined as the percentage of a celebrity's faces in a cluster if it contains at least one face of the celebrity, and the largest percentage is chosen when multiple celebrities' faces simultaneously appeared. By averaging the p-accuracy over all face clusters, 0.808 and 0.624 are reported for the within-video and among-video clustering, respectively. It is worth to notice that the p-accuracy only partially measures the performance of face clustering, as face clusters without celebrity faces are ignored in the evaluation. The results indicate that the obtained celebrity clusters are also mixed with a portion of faces of other people, showing the challenges of clustering faces detected from Web videos.

By applying the proposed community discovering method, 12 name communities and 859 face communities are obtained from the 144 celebrities and 10,382 (merged) face clusters, respectively. Since quantitatively analyzing the quality of discovered communities is somewhat subjective, we propose to use YouTube channel labels to estimate it. Our assumption is that the more elements (i.e., names or faces) of a community come from the same video channel, the better the community is. Thus, we calculate the distribution of channels for each community. The most dominate channel is picked out and its percentage is averaged over all communities. By this way, the percentages of 0.423 and 0.720 are reported for name and face communities, respectively, showing that the C-CNFA approach successfully produces communities distributing unevenly over evenly crawled dataset. A large portion of elements in the same community are come from the same video channel.

3.2.2 Soft weighting

We then evaluate the effectiveness of the soft weighting strategy using the accuracy and recall of the proposed C-CNFA approach. Given a celebrity, the accuracy is defined as the fraction

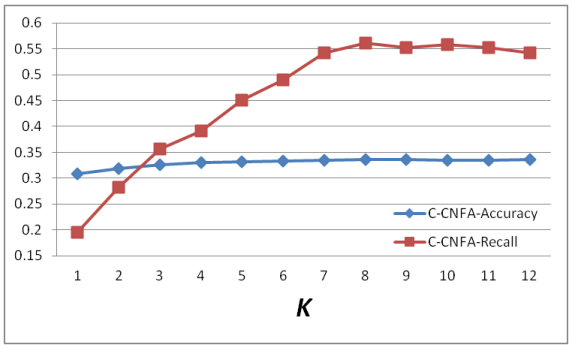

Figure 2: The performance of the C-CNFA varying with *K*.

of face clusters that are correctly tagged. A face cluster is regarded as correctly tagged if it contains at least one face of the celebrity. On the other hand, recall is defined as the fraction of the celebrity's faces that are correctly tagged.

Figure 2 gives the performance of the C-CNFA varying with *K*. Basically, increasing *K* leads to better performance in both accuracy and recall. When *K* reaches 8, C-CNFA performs best, which is significantly better than the performance of the baseline matching strategy, i.e., *K*=1, with improvements of 7.7% and 186.4% in terms of accuracy and recall, respectively. The substantial improvement on recall clearly shows the effectiveness of incorporating more name communities as potential successful matches.

3.2.3 Name-face association

To evaluate the performance of the proposed C-CNFA approach, three other name-face association approaches are experimented for comparison as follows:

1) **Brute-Force Association (BFA)**: BFA is a baseline that simply associates every celebrity name with every face in a video.

2) **Direct Association (DA)**: DA directly associates every face cluster with at most one name based on video co-occurrence. Note that this approach is similar to the state-of-the-art approach [2] in spirit.

3) **C-CNFA without Exemplar Voting (CA-EV)**: CA-EV differs from C-CNFA in celebrity face tagging step, where celebrity faces on the Web are not exploited. To settle the sparse interaction problem, video co-occurrences between communities, as well as between name and face clusters, are jointly considered in CA-EV. We empirically set *K*=6 for CA-EV.

Table 1 lists the accuracy and recall of the four approaches test against the 144 celebrities on Cele-WebV. BFA is the textual baseline and thus has the lowest accuracy and the perfect recall, because the evaluation only conducts on labeled celebrity faces. For the other three approaches, the recall of both CA-EV and C-CNFA are better than DA, showing the effectiveness of exploiting communities. Moreover, C-CNFA performs better than CA-EV in both accuracy and recall, indicating exemplar based voting plays an essential role in effective association. It is also observed that CA-EV and C-CNFA are moderately worse than DA in accuracy. These results highlight the importance of discovering communities consistently between name and face domains.

Table 1. Accuracy and recall of the four different approaches.

	BFA	***DA***	***CA-EV***	***C-CNFA***
Accuracy	0.2251	0.3411	0.3182	0.3361
Recall	1	0.465	0.5542	0.5604

4. CELEBRITY VIDEO RANKING

Given a set of videos tagged with celebrity names, celebrity video ranking aims at ranking the videos according to the visual appearance of celebrities. Obviously, this is an interesting function not offered by mainstream commercial search engines.

In this section, we propose an *Association Ranking* (AR) method to use the association results to rank celebrity videos. Given a query of celebrity j, AP produces the ranking score of video v_i by

$$r_{AR}(v_i) = p_i\big(face(j)\big) + p_i\big(closeup(j)\big) \quad (4)$$

where $v_i \in S_j$ (S_j is similarly defined in Eq. 1), $p_i\big(face(j)\big)$ and $p_i\big(closeup(j)\big)$ are the percentages of keyframes containing faces and close-up views of celebrity j in v_i, respectively.

Two methods, *Views Ranking* (VR) and *Face Ranking* (FR), are employed as baselines. For the same query, the two methods produce ranking lists based on viewed count or face statistics of videos. In FR, the ranking score for a video $v_i \in \mathcal{V}_j$ is defined as

$$r_{FR}(v_i) = p_i(face) + p_i(closeup) \quad (5)$$

where $p_i(face)$ and $p_i(closeup)$ are the percentages of keyframes containing faces and close-up views in v_i, respectively.

Table 2 shows the performance of the three methods on MCG-WEBV core videos. AR performs better than the two baselines, with improvements of 16.1% and 7.5% in terms of MAP over 144 celebrities, respectively. This is not surprise as the top results of VR not necessarily contain detectable frontal faces, and FR often ranks undesired videos at the top of ranking lists. For example, the videos produced by YouTube personalities who stand in front of camera to talk about other celebrities. On the contrary, AR only assigns large ranking scores to videos truly containing faces of the celebrity, thus achieves the best performance.

Table 2. Performance of 144 celebrities for VR, FR and AR.

	VR	***FR***	***AR***
MAP	0.5000	0.5400	0.5807

5. CONCLUSIONS

We have presented the C-CNFA approach for associating faces with celebrity names in Web videos. The experiments conducted on name-face association and celebrity video ranking basically validate our proposal. Performance improvements are also observed when compared to the BFA, DA and CA-EV. The quality of name and face community discovery, though, is limited by the fact that names and faces in Web videos is of high diversity. Thus, future work includes the incorporation of external knowledge to discover more accurate name and face communities. We are also interested in testing the C-CNFA approach on an even larger dataset containing thousands of names and millions of faces.

6. ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7008178).

7. REFERENCES

[1] T. Berg, A.C. Berg, J. Edwards, et al. Names and Faces in the News. *In CVPR*, 2004.

[2] M. Guillaumin, T. Mensink, J. Verbeek. Face Recognition from Caption-Based Supervision. *In IJCV, 2012*.

[3] J. Yang and A.G. Hauptmann. Naming Every Individual in News Video Monologues. *In ACM Multimedia*, 2004.

[4] J. Yang, et al. Multiple Instance Learning for Labeling Faces in Broadcasting News Video. *In ACM Multimedia*, 2005.

[5] M. Everingham, et al. Hello! My name is . . . Buffy, Automatic Naming of Characters in TV Video. *In BMVC*, 2006.

[6] Y.F. Zhang, C.S. Xu, H.Q. Lu, et al. Character Identification in Feature-Length Films using Global Face-Name Matching. *IEEE Trans. on Multimedia*, 2009.

[7] M. Zhao, J. Yagnik, H. Adam, et al. Large scale learning and recognition of faces in web videos. *In IEEE FG*, 2008.

[8] J. Cao, Y.D. Zhang, et al. MCG-WEBV: A Benchmark Dataset for Web Video Analysis. *Technical Report*, 2009.

[9] P. Pons, and M. Latapy. Computing Communities in Large Networks using Random Walks. *In ISCIS*, 2005.