CeleLabel: An Interactive System for Annotating Celebrities in Web Videos

Zhineng Chen[1], Jinfeng Bai[1], Chong-Wah Ngo[2], Bailan Feng[1], Bo Xu[1]
{zhineng.chen, jinfeng.bai}@ia.ac.cn, cscwngo@cityu.edu.hk, {bailan.feng, xubo}@ia.ac.cn

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

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

ABSTRACT
Manual annotation of celebrities in Web videos is an essential task in many people-related Web services. The task, however, poses a significant challenge even to skillful annotators, mainly due to the large quantity of unfamiliar and greatly varied celebrities, and the lack of a customized system for it. This work develops CeleLabel, an interactive system for manually annotating celebrities in the Web video domain. The peculiarity of CeleLabel is to exploit and display multiple types of information that could assist the annotation, including video content, context surrounding and within a video, celebrity images on the Web, and human factors. Using the system, annotators can interactively switch between two views, i.e., merging similar faces and labeling faces with names, to approach the annotation. User studies show that the CeleLabel leads to a much better labeling efficiency and satisfaction.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

Keywords
Celebrity annotation; Web video; interactive system

1. INTRODUCTION
Manual annotation is a popular mean to create ground truth labels in many visual tasks. This work focuses on building an interactive system for annotating celebrities in the Web video domain, i.e., identifying the celebrities appearing in Web videos and labeling them with their corresponding names. The task becomes increasingly important with the explosion of people-related Web videos and services. However, the annotation of celebrities in Web videos poses a significant challenge even to skillful annotators, mainly due to the following three reasons:

- Web video repositories cover a large number of celebrities with a wide range of nations and professions, in which most celebrities are not familiar to annotators.
- Celebrities maybe with significant variations in visual appearance both within and among videos, such that it might be difficult to correctly distinguish them from others.
- There is no customized system for this problem so far, such that multiple clues for the annotation could not be leveraged.

MM'14, November 3–7, 2014, Orlando, Florida, USA.
ACM 978-1-4503-3063-3/14/11.
http://dx.doi.org/10.1145/2647868.2654879

In the past few years, a number of efforts have been devoted to building databases for challenging visual problems, such as the Trecvid collaborative annotation [1]. These efforts, despite making substantial progress, mainly focus on annotating common concepts or objects which are intuitive for humans. Celebrities, however, are much more special concepts whose inter-class differences are less obvious and not easily described. It would be quite difficult to annotate them if, in the simplest way, only snapshots containing people is presented to annotators. To ease the problem of annotation, it is essential to provide more information about the video and the celebrities to annotators.

Motivated by this observation, we develop a system named CeleLabel for supporting the manual annotation of celebrities in the Web video domain. CeleLabel exploits four types of information to facilitate the annotation: video content, context surrounding and within a video, celebrity images on the Web and human factors. Given a Web video, face tracks and candidate celebrity names are extracted. Based on this, relevant celebrity images are collected from the Web. Parallel, the face tracks are ranked based on several content based factors. At the labeling stage, the annotation task is proceed by either merging two similar faces or labeling the top ranked face track with a candidate celebrity name. Meanwhile, celebrity images most similar to the top ranked track are also displayed for reference. With the labeling continues, the ranking of remaining face tracks are adjusted after every labeling action, by incorporating additional contextual clues and human factors like human perception habits, face track co-occurrence, etc, ensuring that the face track suggested to annotators is always the worthiest. Since multiple clues that could be exploited to facilitate the annotation are leveraged, CeleLabel leads to a much better labeling efficiency and satisfaction as demonstrated in user studies.

2. TECHNOLOGIES
2.1 Preprocessing
Web videos are first preprocessed to extract candidate celebrity names and face tracks. In the implementation, we employ the Wikipedia-based name entity extraction method [2] to extract names from metadata (title and tags) surrounding the given video, and apply the detection-based tracking method [3] to extract face tracks from the video, respectively.

2.2 Merging Similar Faces
It is common that faces of a single person fall into different extracted face tracks. These tracks are visually similar to each other. Compared with determining their respective name correspondence, it is easier to determine whether they refer to the same individual. Therefore, we develop a merging view for annotating face tracks from the angle of merging two similar face tracks rather than separately assigning names to them. The similarity measurement between face track f_i and f_j, denoted as $sim(f_i, f_j)$, is defined as:

$$sim(f_i, f_j) = e^{-\Delta t_{i,j}/t_\theta} \cdot (1 - CO_{i,j}) \cdot vs(f_i, f_j) \quad (1)$$

In the formula, $e^{-\Delta t_{i,j}/t_\theta}$ is a time-decay function, in which $\Delta t_{i,j}$ is the appearing time difference between track f_i and f_j, t_θ is a time slot threshold controlling the influence of time difference. The function favors "nearby" tracks rather than "far away" tracks, thus the time consistency is preserved to some extent. $CO_{i,j}$ is the co-occurrence status of the two tracks. If they are time-overlapped, $CO_{i,j}$ equals to 1, otherwise it is 0. By this way, time-overlapped face tracks would not appear in the merge view even their faces are similar, as two faces co-occurred hardly refer to the same identity. $vs(f_i, f_j)$ is the visual similarity defined as the minimum distance between faces of the two tracks:

$$vs(f_i, f_j) = \min_{\boldsymbol{f}_m^i \in f_i, \boldsymbol{f}_n^j \in f_j, i \neq j} \|\boldsymbol{f}_m^i - \boldsymbol{f}_n^j\| \quad (2)$$

where $\boldsymbol{f}_m^i$ is facial feature of the m-th face of track f_i. In the merging view, face track pairs are suggested to annotators according to their $sim(f_i, f_j)$.

2.3 Labeling Faces with Names

Parallel to the merging view, we also develop a labeling view to let annotators directly label faces with names. To determine which face track is the most worthy suggested to annotators, video content, context clues of the video and human factors are jointly investigated. The suggestion score of face track f_i is defined as:

$$Sug_i = (1 - P_i) \cdot (\overline{Sai}_i + \overline{JS}_i) \quad (3)$$

In the formula, P_i is a function representing whether track f_i has been skipped by annotators at the labeling stage. P_i equals to 1 if it is, otherwise it is 0. $\overline{Sai}_i$ and $\overline{JS}_i$ are the normalized salience Sai_i and joint similarity JS_i of face track f_i. The face tracks are ranked and suggested to annotators according to their suggestion scores. The salience Sai_i is computed by:

$$Sai_i = e^{-s_\theta/s_i} + e^{-d_\theta/d_i} \quad (4)$$

In the formula, variable s_i and d_i are the average size and duration of track f_i, respectively. s_θ and d_θ are two thresholds set empirically to control the influence of the two variables. The salience measurement favors to suggest large and long face tracks.

On the other hand, JS_i suggests face tracks by considering both human factors and the similarity in Eq. (1). It is defined as:

$$JS_i = \sum_{j=1}^{K} L_j \cdot sim(f_i, f_j) \quad (5)$$

where K is the number of face tracks in the video, L_j is a function indicating the manual labeling status of track f_j. It equals to 1 if already labeled, and 0 if not labeled yet. JS_i gives higher score to tracks which are similar to previously labeled ones in general.

According to above formulas, the top ranked face tracks trend to have the following properties: salient in the video, visually similar, and nearby but no overlapped to previous annotated tracks. All the properties are basically in accordance with human perception habits. At the labeling stage, annotators can choose either the merging view or the labeling view to conduct the annotation.

2.4 Celebrity Images

Because of unfamiliar with the celebrities, it is common that annotators are unsure about who refers to the presented face track. In such cases, consulting external sources about relevant celebrities seems to be a wise choice. Based on this observation, we collect and present relevant celebrity images as follows.

Firstly, we use extracted celebrity names one-by-one to issue Google Image Search, and crawl the top ranked Web images for every celebrity. Secondly, the celebrities are ranked, according to the average similarity between images of the celebrity and the face track currently displayed in the labeling view, to decide which celebrity should be shown first. Thirdly, images of the shown celebrity are also ranked according to their similarity to the displayed face track, to determine which images to be presented.

2.5 Interface

The interface of CeleLabel consists of three building blocks, i.e., annotating area, reference area, and history. The annotating area provides the two annotation views: the merging view displays two most similar tracks, annotators can click either "same" to merge them, or "next" to load the next two most similar tracks; the labeling view displays a representative faces of the top ranked track and the celebrity names. Annotators can select a celebrity name, "unknown", or "Not Face" to label the track as a celebrity, an unknown person or a false positive track. The reference area presents a celebrity name and six of its images, helping annotators to compare and distinguish who refers to the track. Annotators can click "next" or "prev" to load the next or previous similar celebrity name and its images for reference. The history displays already labeled face tracks and their names, serving as a summary of the current annotation. Content of the three areas are updated dynamically with every labeling action. An illustrative demo of the system is available at www.youtube.com/watch?v=q_jFVIyMHek.

3. USER STUDY

To evaluate the effectiveness of CeleLabel, we recruit 10 assessors to experience the annotation using both the CeleLabel and a baseline system, which shows only face snapshots of a video and its surrounding title and tags. Our objective is to evaluate whether the annotation of celebrities in Web videos could be boosted by using the proposed CeleLabel.

In the study, 20 preprocessed Web videos are given. The assessors are divided into two groups each with 5 people. The first group is asked to annotate the first 10 videos using CeleLabel and the other 10 videos using the baseline system. On the contrary, the other group is asked to annotate the first 10 videos using the baseline and the other 10 videos using CeleLabel. They are asked to manually annotate all extracted face tracks. The time spent on annotation is automatically recorded by both systems. Besides, the scores (1-5) are given by each assessors to evaluate the user experience. The higher score indicates the higher satisfaction. The results show that compared with the baselines (326 seconds and 2.3 on average), CeleLabel achieves much better performance (147 seconds and 4.4 on average). Besides, all assessors agree on that the display of relevant celebrity images is quite helpful. Some assessors also point out that the show of annotation history is useful. The user study clearly shows, by offering more clues about the video and the celebrities to annotators, CeleLabel provides a much better labeling efficiency and satisfaction.

4. ACKNOWLEDGMENTS

The work described in this paper was supported by the National Natural Science Foundation of China (#61303175, #61228205).

5. REFERENCES

[1] P. Over et al. Trecvid 2011: an overview of the goals, tasks, data, evaluation mechanisms and metrics. *TREC Video Retrieval Evaluation*, 2011.

[2] Z. Chen, J. Cao, T. Xia, Y. Song, et al. Web video retagging. *Multimed. Tools Appl.*. 55(1): pp. 53-82, 2011.

[3] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Image Vision Comput.*. 27(5): pp. 545-559, 2009.