

Research Article

Large Scale Near-Duplicate Celebrity Web Images Retrieval Using Visual and Textual Features

Fengcai Qiao, Cheng Wang, Xin Zhang, and Hui Wang

College of Information Systems and Management, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Fengcai Qiao; qiaofengcail25@gmail.com

Received 10 July 2013; Accepted 2 August 2013

Academic Editors: J. T. Fernandez-Breis, S.-S. Liaw, and J. H. Sossa

Copyright © 2013 Fengcai Qiao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Near-duplicate image retrieval is a classical research problem in computer vision toward many applications such as image annotation and content-based image retrieval. On the web, near-duplication is more prevalent in queries for celebrities and historical figures which are of particular interest to the end users. Existing methods such as bag-of-visual-words (BoVW) solve this problem mainly by exploiting purely visual features. To overcome this limitation, this paper proposes a novel text-based data-driven reranking framework, which utilizes textual features and is combined with state-of-art BoVW schemes. Under this framework, the input of the retrieval procedure is still only a query image. To verify the proposed approach, a dataset of 2 million images of 1089 different celebrities together with their accompanying texts is constructed. In addition, we comprehensively analyze the different categories of near duplication observed in our constructed dataset. Experimental results on this dataset show that the proposed framework can achieve higher mean average precision (mAP) with an improvement of 21% on average in comparison with the approaches based only on visual features, while does not notably prolong the retrieval time.

1. Introduction

Near-duplicate images are defined as images of the same object or scene viewed under different imaging conditions [1]. Retrieval of near-duplicate images has boosted many real-world applications such as image annotation, image understanding [2], and content-based image retrieval (CBIR) [3]. It can improve traditional text-based image search by filtering duplicate query results, can be used as a bridge to relate two web pages in different languages, and can also provide similarity clues for recognizing visual events and searching news video clips. Among the tremendous amount of web images, celebrity images including posters, portraits, news photos, movie snapshots are of particular interest to the end users. Moreover, in agreement with the observation of [4], we find that near-duplication is also particularly prevalent in queries for celebrities and historical figures. For example, on January 8, 2013, the top 1000 thumbnails returned by the Google image search for the query “Kate Winslet” contained 51 images (see Figure 1 for some of them) about Kate’s arriving in London for Titanic 3D premiere. These images are exactly the same with each other or slightly different in scales,

cropping, and text insertion manners or captured in the same (3D) scene but with different cameras.

Most recent near-duplicate image retrieval systems [5–9] adopt bag-of-visual-words (BoVW) model, which is a popular and practical method for large-scale image datasets. A typical BoVW based method includes routines which builds visual vocabulary, performs SIFT descriptors [10] quantization and online near-duplicate image retrieval. Because of the application of textual indexing and retrieval schemes, large scale near-duplicate retrieval becomes simple and effective. Nevertheless, the deficiency of current BoVW methods lies in the approximation error incurred by quantization and the loss of geometric relations between local features caused by the histogram-based representation adopted. Both factors reduce the discriminative power of local image features. Although many improvements such as hamming embedding (HE), weak geometry consistency (WGC), and enhanced weak geometry consistency (E-WGC) have been presented [1, 6, 11] to overcome these problems to some extent, they all depend solely on visual clues. In addition, to our best knowledge, no near-duplicate image retrieval methods have been experimented totally on celebrity image datasets. However,



FIGURE 1: Examples of near-duplicate images contained in the top 1000 thumbnails returned by Google image search engine for the query “Kate Winslet” on January 8, 2013.

as well known, nearly all web images are surrounded with texts in their hosting webpage. By properly utilizing these surrounding texts we may get considerable improvements in the precision and recall of near-duplicate retrieval results. Hence, we propose in this paper a new text based data-driven reranking framework, which utilizes textual features and is combined with state-of-art BoVW schemes. Our main contributions to large scale near-duplicate celebrity web image retrieval are three pronged.

First, we construct a 2-million-celebrity-image-text dataset from the Internet about 1089 celebrities and divide near-duplicate celebrity web images into three categories in our ground-truth dataset according to our observation. All the images’ surrounding texts are also kept and properly organized for further use. The necessities are listed below. (i) No existing available dataset provides images together with their accompanying text. (ii) None of them has focused only on celebrities or persons. (iii) Our constructed dataset provides fine-grain ground-truth labels, that is, the specific category of near-duplication of each near-duplicate image.

Second, we thoroughly evaluate six representative existing local visual feature based methods based on our constructed dataset. According to our experiments, BoVW with weak geometry consistency outperforms other methods when utilizing only visual feature.

Finally, to leverage the useful cues embedded in the surrounding texts of web images and hence to improve the precision and recall, we propose a text-based data-driven

reranking framework. The framework improves the state-of-art local visual feature based BoVW methods through further exploiting the accompanying texts of images. It is worthwhile to note that the input of our framework is the same as BoVW schemes which means that we do not need to input the query image’s accompanying text although we utilize text information in the retrieval process. Experiments on our constructed dataset of 2 million celebrity web images show that the proposed method achieves 21% improvement on average compared with pure BoVW methods.

The rest of this paper is organized as follows. Section 2 discusses the related work. The celebrity web image-text dataset is introduced in Section 3, followed by our proposed data-driven reranking retrieval framework in Section 4. Experiments are presented and analyzed in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

Given a query image, near-duplicate image retrieval aims to find all images that are identified as essentially the same by human subjects. Near-duplicate image retrieval has been intensively addressed and researched in the last decade, partly attributed to the emergence of TRECVID benchmark for multimedia search. Researchers in this community often conduct their work in different application perspectives, typically including copyright protection, CBIR [12], and lately object or pattern recognition [13]. Generally, existing near-duplicate

image retrieval methods make use of some global or local statistics of low-level features extracted from the query and database images. They often concatenate the (statistics of local) features into a single vector or matrix and use it as the image representation for similarity evaluation. Hence, we can largely categorize these methods into two groups, that is, the global feature based and the local feature based ones, which are overviewed separately as follows.

Global feature based algorithms have been researched for a long time. Wang et al. [14] proposed a fast and effective method in which each image is compactly represented by a K-bit hash code to detect all visually duplicate groups in large image collections. References [15, 16] extended the idea of Wang et al. [14] by adding the canny edge information to the hash code with the purpose of further improving the discriminative power of the image representation.

In local feature based methods, bag-of-visual word (BoVW) scheme has been widely adopted in different contexts such as visual object recognition/classification, CBIR, and semantic concept detection. All these researches adopt BoVW as an image representation which involves routines for visual vocabulary construction, vector quantization, and pair-wise descriptor matching. However, it is notable that BoVW often produces many false matches which are mainly introduced by the vector quantization error. To tackle this issue, geometric verifications, for example, Hamming embedding (HE) [1] and weak geometric constraint (WGC) [1], have been proposed to prune these false visual word matches. Different from BoVW, Zhong et al. [9] presented a novel scheme where image features are bundled into local groups. Each group of bundled features becomes much more discriminative than a single local feature, and within each group simple and robust geometric constraints can be efficiently enforced. They achieved a 77% precision improvement over the baseline BoVW based approach and a 24% improvement over full geometric verification alone.

Before closing this section, we would emphasize that our research in near-duplicate celebrity web images retrieval adopts a methodology similar to the search-based images annotation [15, 17–19] rather than the traditional face recognition based approaches. In the emerging area of *Internet Vision* [20], the search-based scheme has been adopted by many researchers. Besides, the textual keyword based approaches have been successful applied to image search, while there are, to the best of our knowledge, very few reported researches that have applied textual information to near-duplicate images retrieval. In this paper we proposed a novel framework which combines BoVW schemes with textual information with the purpose of improving duplicate celebrity web image retrieval.

3. The Celebrity Web Image-Text Dataset

3.1. Data Collection. To collect celebrity web images, from January 8, 2013, to May 3, 2013, we constructed a celebrity dictionary using the 1089 top searched celebrity names in 5 areas. We use these names as query keywords and performed image search using Google image search engine and six news

search engines including Google news, Yahoo news, Ifeng news, Sina news, Panguoso news, and Baidu news. The URL patterns for each page of the returned images are listed in Table 1. For Google image search engine, we downloaded the returned 1000 images (at best) and the main text (including title, content, and image captions) on their hosted web pages for each query name. Images with exactly the same URLs are removed. For each news search engine, we crawled the latest 5 pages of returned results for each query, saving the news's text and accompanied images. Similarly, news web pages are also saved without the same URLs. The crawling and collection process lasted for nearly four months. In the end, we get more than 2 million images and their related text totally (see Table 2 for details). These data are organized in unit of image-text pairs which means that each image is related to some text. Each image-text pair is assigned with a unique ID to index the image and its related text.

3.2. Types of Near-Duplication and Ground-Truth Labeling.

We inspect a substantial number of these celebrity images to identify instances and types of near-duplication. According to our observation, three main categories of near-duplicates exist in these celebrity images (see Figure 2 for some examples of each category). (i) *Major duplicates* refer to images being exactly the same or images which are identified exactly the same by human subjects but have (slightly) different scales, colors, file formats, luminance intensities, and so forth. (ii) *Partial duplicates* refer to images of which some parts are exactly or fully duplicate to (some parts of) others. (iii) *Scene-object duplicates* are images sharing the same 3D scene or the same object (with object-class variability like gesture variation) but captured by different cameras at different time. Images in each ground-truth group include one or more near-duplicate categories. The geometric and photometric transformations related to each category of near-duplicates are detailed in Table 3.

For the ground-truth labeling, we randomly select three celebrities from each area. Hence, 15 persons are selected in all. For each of them, we manually identify from his/her pictures collected the near-duplicate ones and their specific near-duplication types. In this way, we obtain the ground-truth composed of 57 groups of 2180 identified near-duplicate images. Each group contains images of all three categories of near-duplicates. The basic dataset contains near-duplicate images of our ground-truth images. As in [9], for evaluation purposes we identify and remove these duplicate images from the basic dataset by querying the database using every image from the ground-truth dataset and manually examining the returned images sharing any common parts with the query images. The number of ground-truth images in each type is listed in Table 4.

As listed in Table 4, major duplicates, partial duplicates, and scene-object duplicates account for almost the same in the ground-truth dataset, with 31.74%, 30.18%, and 38.08%, respectively. Many major duplicates exist because many websites just copy and publicize the original images without large modifications. Partial duplicates also occupy considerable large portion. It is noticeable that scene-object duplicates are very common for celebrity dataset. These images are usually

TABLE 1: URLs of the search engines used in the experiments.

Search engine	URL pattern
Google image	http://www.google.com.hk/search?tbm=isch&q=&k&start=&
Google news	http://www.google.com.hk/search?q=&k&tbm=nws&start=&
Yahoo news	http://news.yahoo.cn/s?q=&k&page=&
Ifeng news	http://search.ifeng.com/sofeng/search.action?q=&k&p=&
Sina news	http://search.sina.com.cn/?q=&k&c=news&page=&
Panguso news	http://news.panguso.com/newssearch.htm?q=&k&p=&
Baidu news	http://news.baidu.com/ns?tn=news&rn=20&word=&k&p=&

TABLE 2: Number of the crawled celebrity images.

Category	# Names	# Images (Google)	# Images (news)	# Images (total)
Politicians	245	219030	171255	390285
Actors	312	423072	218398	641470
Singers	212	281112	211788	492900
IT leaders	189	142254	157657	299911
Sports stars	131	130519	114930	245449
Total	1089	1195987	874028	2070015

taken on the same day and in the same event and hence often look like a series of snapshots of a period of video. From the subjective point of view, they also should be near-duplicate images.

4. Data-Driven Reranking Retrieval Framework

4.1. Framework. Near-duplicate celebrity web images retrieval is a process whose goal is to locate all major duplicate, partial-duplicate or scene-object images in a large corpus of web images given a query image I_q :

$$I_q \xrightarrow{\text{near-duplicate images retrieval}} \{I_i \mid I_j \in D, \delta(I_q, I_j) = 0, j = 1, 2, \dots, n\}. \quad (1)$$

D is the basic dataset. n is the number of near-duplicate images retrieved. $\delta(I_q, I_j)$ is the near-duplicate estimation function which is defined as

$$\delta(I_q, I_j) = \begin{cases} 0, & \text{Sim}(I_q, I_j) \geq \theta, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

$\text{Sim}(I_q, I_j)$ is the similarity of I_q and I_j . θ is the selected threshold got from heuristic manner or learning. The key of $\text{Sim}(I_q, I_j)$ lies in image representation method and distance calculation. In the BoVW based near-duplicate image retrieval methods, an image is represented as a vector which is based on local features. Distances are calculated using cosine similarity and can be efficiently conducted by exploiting the inverted file index [21] (Section 4.2).

Denote the text context of top k images closest to query image I_q retrieved by BoVW based methods to be Y_q .

Intuitively, Y_q is most likely to contain related text information about the people appearing in the query image I_q . From our observation, a celebrity's name, which affirmatively appears at least once in webpage title, content, or celebrity image captions, is a metric of vital importance to the celebrity images. Based on this observation, a two-stage solution for data-driven reranking retrieval framework is proposed:

- (i) *BoVW-based search*, retrieving near-duplicate images of I_q .
- (ii) *text-based reranking*, mining candidate celebrity names from Y_q , computing confidence scores, and updating the similarity of each I_j .

Figure 3 illustrates the proposed data-driven reranking retrieval framework. Two major components in the framework are BoVW-based near-duplicate images retrieval and text-based reranking. The former step conducts near-duplicate images retrieval using BoVW-based methods which is only dependent on visual features. Then the later step reranks the retrieval results according to each image's text context feature. Finally, we get the retrieval results using both visual and text features.

4.2. BoVW-Based Near-Duplicate Images Retrieval. In this stage, we conduct near-duplicate images retrieval using BoVW methods. All the query images and images in the basic dataset are preprocessed firstly. SIFT descriptors are extracted from the images and clustering is carried out to quantize the descriptors into a visual vocabulary. Each descriptor in the image is then mapped to the nearest visual word in the vocabulary (vector quantization) and this forms a bag of visual words for each image. As the vocabulary is very large in this paper (1 M), to efficiently quantize without exhaustive search of the nearest visual words, we adopt a three-level vocabulary [1].

A bag of visual words is represented in the form of histogram, each bin of which accumulates the number of visual words found in the image. In practice, each bin of the histogram has its weight deriving from tf-idf (term frequency-inverse document frequency) method. Denote the histogram of an image I_j to be

$$I_j = (t_1, \dots, t_i, \dots, t_m)^T. \quad (3)$$

TABLE 3: The Image alterations related to each category of near-duplicates.

Near-duplicate category	Alterations
Major duplicate	Typical image alterations include adding logos or inserting texts, changes of pixel color vectors, color space, file format, or pixel luminance intensities, and some geometric transforms such as scaling, mirroring, and rotation.
Partial duplicate	Alterations undergone commonly include cropping, pasting an image to overlay a part of another, zooming, and adding border.
Scene-object duplicates	Sharing the same scene or undergone object-class variability. Variations between near duplicate images of this category are often caused by alterations such as viewpoint changes, partial occlusion, and differences in (intrinsic/extrinsic) camera parameters.



FIGURE 2: Examples of three categories of near-duplicate celebrity images.

TABLE 4: Number of ground-truth images in each category.

Near-duplicate category	# Images	Percentage
Major duplicate	692	31.74%
Partial duplicate	658	30.18%
Scene-object duplicate	830	38.08%
Total	2180	100%

In (3), m is the vocabulary size and

$$t_i = \text{tf}(i) \times \text{idf}(i) = \frac{n_{ij}}{n_j} \times \ln \frac{N}{1 + N_i}. \quad (4)$$

In (4), n_{ij} is the number of the i th visual word appearing in I_j , n_j is the total number of visual words in I_j , N is the number of images in the basic dataset, N_i is the number of images in which the i th visual word appears. Denote $t_i(I_q)$ and $t_i(I_j)$ to be the weight of i th bin in image I_q and I_j , respectively. The closeness between I_q and I_j is calculated via cosine similarity:

$$\text{Sim}(I_q, I_j) = \frac{\sum_{i=1}^m t_i(I_q) \times t_i(I_j)}{\sqrt{\sum_{i=1}^m t_i(I_q)^2 \sum_{i=1}^m t_i(I_j)^2}}. \quad (5)$$

$\text{Sim}(I_q, I_j)$ can be efficiently computed using inverted file indexing because the size of vocabulary is usually very large,

and therefore the histogram is normally very sparse. In the inverted file indexing, each row stores the relationship of visual word and image. Local features like x , y , scale, orientation, and hamming code can be also stored in index cells dependent on the BoVW method used. Therefore, cosine similarity is evaluated only on a subset of images in the basic dataset (nonzero entries in the images).

In the retrieval process, to further speed up the online retrieval time we use hamming embedding (HE) [1] scheme. HE maintains a binary signature for each SIFT descriptor. The binary signature is stored in the inverted file index cells. During retrieval, any two matched visual words can be pruned if the hamming distance between their signatures is large than a threshold. At the same time, weak geometry consistency (WGC) [1] or enhanced weak geometry consistency (E-WGC) [11] can also be conducted and embedded into the inverted file indexing to fast reranking the images according to their geometric consistency with the query image. We name the BoVW method without HE and geometric consistency checking as traditional BoVW. As a result, there are six BoVW variations in all traditional BoVW, BoVW + HE, BoVW + HE + WGC, BoVW + HE + E-WGC, BoVW + WGC, and BoVW + E-WGC.

4.3. Text-Based Reranking. As mentioned above, celebrity's name plays a vital role in the retrieval of near-duplicate

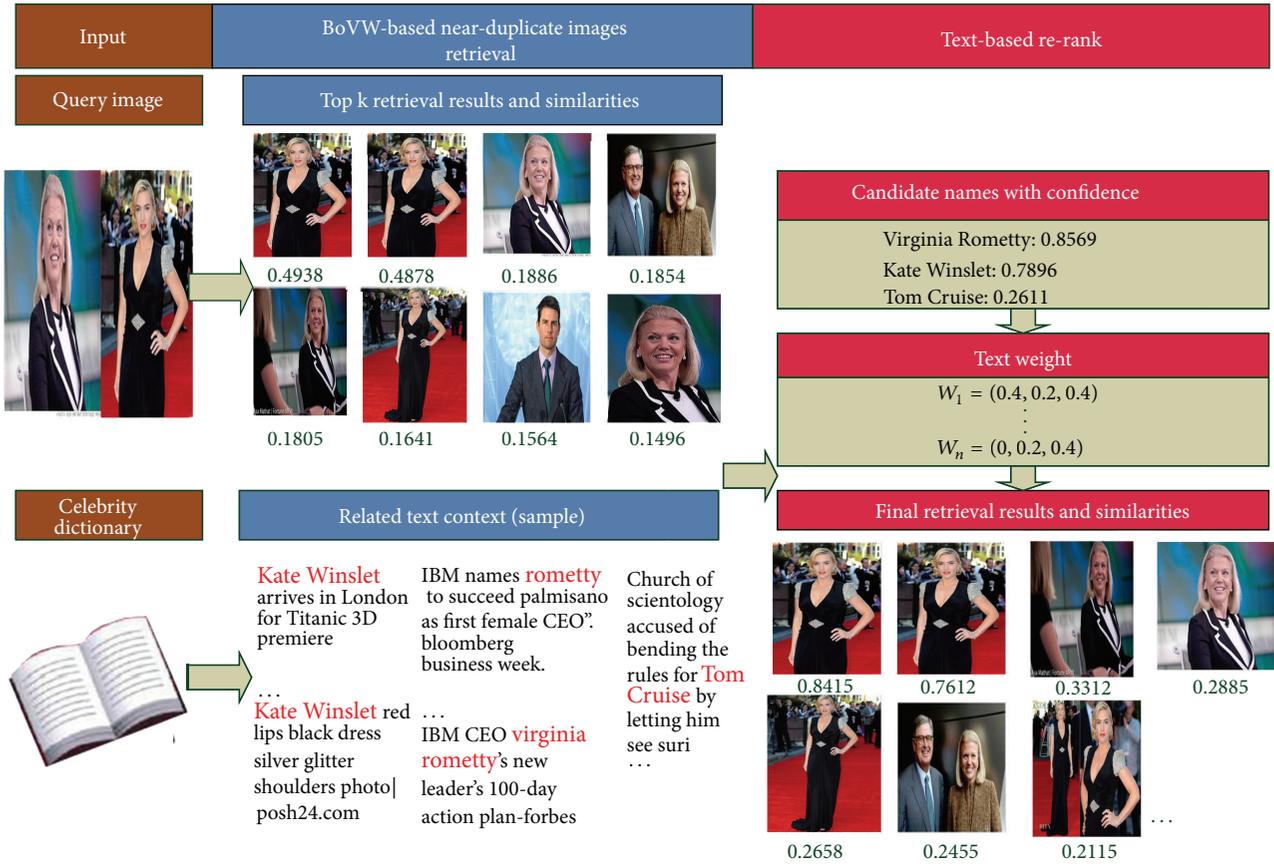


FIGURE 3: Framework for data-driven reranking retrieval.

celebrity images. Our text-based reranking scheme is dependent on the identification of celebrity names from the text context of the top k images retrieved by BoVW method. The top k images are most likely to contain the same person(s) appearing in the query image. Then, with the identified name or names, text retrieval can be efficiently performed using inverted file indexing. For images with text containing identified names, intuitively, the similarities of which should be larger than original values to some extent. In other words, there exist many false positive images whose similarities are bigger than some true positives when retrieval is conducted under large scale celebrity images dataset. By properly importing celebrity's name, we can filter many false positives which impossibly share any common parts with the query image, thus reranking true positive images to anterior order.

4.3.1. Candidate Name Identification and Confidence Computing. Firstly, we mine candidate names from Y_q and compute each candidate name's confidence score. The method used in this step is mainly inspired by [15], but we adopt the idea with a very different purpose, that is, to discover the possible celebrities contained in the given image. Given the name dictionary, all the possible names are defined as

$$V = \{v_l \mid l = 1, \dots, K\}, \quad (6)$$

where K is the number of celebrities in our name dictionary. With the name dictionary and text context Y_q of the query image, one or more candidate names can be identified. Denote candidate names extracted from Y_q to be V_q ($V_q \subset V$). Given a name $v_l^{(q)} \in V_q$, denote $p(v_l^{(q)} | Y_q)$ to be the confidence score of $v_l^{(q)}$. Denote the text features for $v_l^{(q)}$ to be $T(v_l^{(q)})$. $T(v_l^{(q)})$ considers the following three aspects.

- (i) *Type of names:* We consider a term in context Y_q as a candidate name in two ways: (i) full name, for example, "Steve Jobs;" (ii) partial name, for example, "Steve" or "Jobs;" we observed that full name match was a very reliable indicator for the occurrences of celebrity names. Partial match can increase the confidence score because people often use partial names for simplicity.
- (ii) *Type of text:* in our basic dataset, three types of text related to a celebrity image are extracted and stored, that is, caption, page title, and content. Different types of text have different importance. For instance, information in page title or caption is more predictive of the identity of the celebrity than information in content.
- (iii) *Frequency:* frequency f corresponds to the number of times that $v_l^{(q)}$ occurs in Y_q . Intuitively, $v_l^{(q)}$ is more

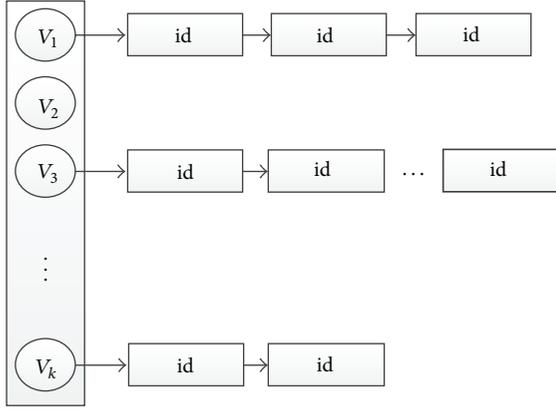


FIGURE 4: Structure of inverted file indexing used in related text retrieval.

likely to appear in the query image if its frequency is high. In our experiment, $v_l^{(q)}$ is under consideration only if $f \geq 2$.

The confidence score of $v_l^{(q)}$ is estimated by

$$P(v_l^{(q)} | Y_q) = \frac{1}{1 + e^{-(W^T T(v_l^{(q)}) + b)}}, \quad (7)$$

where parameters W and b are learnt by logistic regression.

4.3.2. Related Text Voting and Updating the Similarities. In this step we first perform text retrieval using the names identified in (1) as query keywords to find text containing specified name. This procedure can be efficiently conducted using inverted file indexing, whose structure is illustrated in Figure 4. As each text-image pair has been assigned a unique ID in our basic dataset (Section 3.1), the index stores the name-ID relationship. There are K rows in all, each of which corresponds to a celebrity name and ids to the list of texts which contain the name. As a consequence, given a celebrity name, the texts are hashed into the index and the IDs of text containing this name are acquired immediately.

We divide the text of an image into three types: page title, content, and image caption. As each type has different importance to identify the celebrity image, the text information of image I_j can be represented by a vector

$$T_j = (W_{\text{title}}, W_{\text{content}}, W_{\text{caption}}). \quad (8)$$

We assign the weights of W_{title} , W_{content} , and W_{caption} with 0.4, 0.2, and 0.4, respectively, according to experiential observation. To make clear, if the queried celebrity name appears in the title, then we assign W_{title} with value 0.4, otherwise 0. Same procedure is applied to content (title) and caption. As a result, we construct three inverted file indexing structures: page title index, content index, and caption index. For each query name $v_l^{(q)} \in V_q$, fast retrieval is conducted in

the three indexing structures at the same time. Denote $S(v_l^{(q)})$ to be the retrieved text set for $v_l^{(q)}$:

$$S(v_l^{(q)}) = \{T_j | T_j \subset T, v_l^{(q)} \in T_j\}. \quad (9)$$

Each element T_j in $S(v_l^{(q)})$ represents the text information of image I_j . Denote w_j to be the score of text information T_j . w_j is simply estimated by

$$w_j = W_{\text{title}} + W_{\text{content}} + W_{\text{caption}}. \quad (10)$$

Finally, for each image I_j whose related texts are in $S(v_l^{(q)})$, we update its similarity to the query image according to both visual features and text features. The data-driven reranking framework adjusts the similarity computed in (5) by

$$\text{Sim}_{\text{final}}(I_q, I_j) = 2^{\gamma(p+w_j)} \times \text{Sim}(I_q, I_j). \quad (11)$$

p is the confidence score of $v_l^{(q)}$ estimated by (7). w_j is the score of text information T_j . The factor γ is a parameter used to control the influence of text. Thus, The original visual-only-based similarity $\text{Sim}(I_q, I_j)$ is boosted, by a factor corresponding to the text information of image, for images whose related text information is in $S(v_l^{(q)})$ and $v_l^{(q)} \in V_q$. For images whose text does not contain any name in V_q , the similarities of which stay unchanged.

5. Experimental Results

As pointed out in [1], near-duplicate is a kind of relationship for image pairs. Thus two natural measures are the precision and recall of duplicate image pairs. A ‘‘correct’’ image pair means it belongs to the intersection of a detected group and a ground-truth group. Then, the image pair precision (IPP) and image pair recall (IPR) are defined, respectively, as

$$\text{IPP} = \frac{\# \text{correct pairs}}{\# \text{detected pairs}} * 100\%, \quad (12)$$

$$\text{IPR} = \frac{\# \text{correct pairs}}{\# \text{ground-truth pairs}} * 100\%.$$

As mentioned in Section 3.2, the ground-truth dataset is composed of 57 groups of 2180 identified near-duplicate images. In our evaluation, we randomly select two images from each ground-truth group as query images and hence obtain 114 queries in all. Following [9] we use mean average precision (mAP) as our evaluation metric. For each query image we do the retrieval work and compute its precision-recall curve, from which we obtain its average precision and then take the mean value over all queries as the mAP. To evaluate the performance with respect to the size of the dataset, we also build four smaller datasets (50 K, 500 K, 1000 K, and 1500 K) by sampling the basic dataset.

5.1. Evaluation

5.1.1. Baseline Methods. As mentioned in Section 4.2, we use six BoVW based methods in all. They are traditional BoVW,

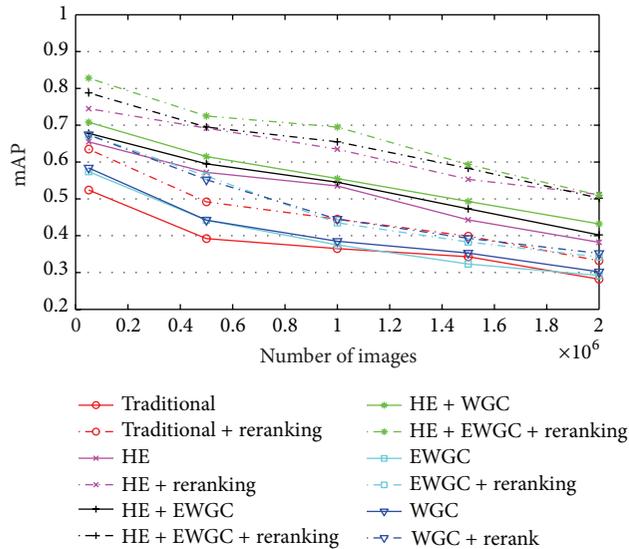


FIGURE 5: Comparison of different methods using mAP: the six BoVW methods without text-based reranking are in solid lines. Dashed lines are the BoVW methods combined with our data-driven reranking framework.

BoVW + HE, BoVW + HE + WGC, BoVW + HE + E-WGC, BoVW + WGC, and BoVW + E-WGC, making up of our baseline methods. To ensure the coverage of vocabulary, we choose a visual vocabulary of 1M visual words for the two-million-basic dataset.

5.1.2. mAP Comparisons. Each of the six BoVW variations can be combined with our data-driven reranking framework. As a result, 12 methods are compared with each other. In our implementation, the hamming code in HE is 32 and the threshold for hamming distance is set at 15. The parameter γ in (11) is set to be 0.8.

Figure 5 compares the above 12 methods using mAP, leading to three observations. First, our data-driven reranking framework boosts all the six BoVW methods which are dependent only on visual features. On the 2-million dataset, the improvements of mAP for traditional BoVW, BoVW + HE, BoVW + HE + WGC, BoVW + HE + E-WGC, BoVW + WGC, and BoVW + E-WGC are 17%, 34%, 25%, 18%, 16%, and 17%, respectively, 21% on average. Second, as shown by the curves labeled “HE + WGC” and “HE + WGC+ rerank”, BoVW method outperforms other BoVW methods with and without our text-based reranking stage. Finally, the hamming embedding is of vital importance to the near-duplicate celebrity images retrieval. On the 2-million dataset, when combining HE with our reranking stage, BoVW + HE, BoVW + HE + WGC and BoVW + HE + E-WGC all achieve almost the same high mAP than others, with highest mAP 0.508 of BoVW + HE + WGC (79% improvement over the traditional BoVW without HE and our reranking).

5.1.3. Impact of Factor γ . The factor γ in (11) is a parameter used to control the influence weight of text information. We

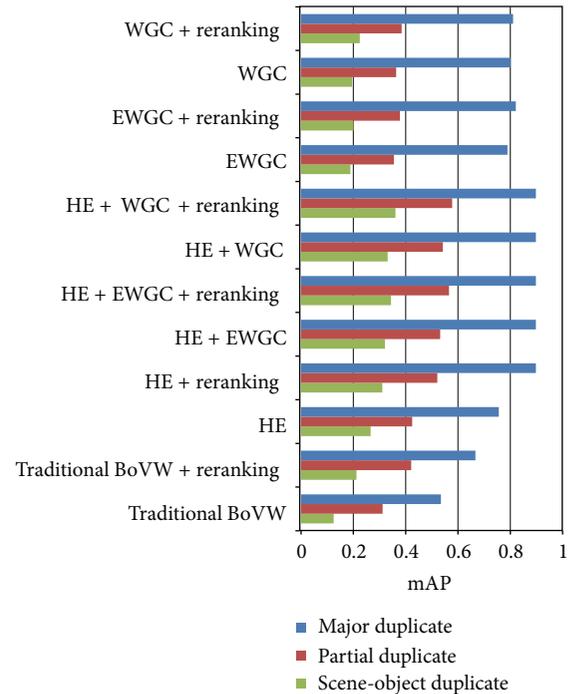


FIGURE 6: mAP with respect to each near-duplicate category on the 2-million dataset.

TABLE 5: Comparing the performance using different values of γ .

γ	0.2	0.4	0.6	0.8	1
mAP	0.488	0.493	0.501	0.508	0.499

TABLE 6: Average query time (not including feature extraction time).

BoVW + HE + WGC	BoVW + HE + WGC + re-ranking
2.9 s	3.7 s

test the performance of our data-driven reranking framework using different γ values on the 2-million dataset. As Table 5 shows, the mAP achieves highest value when γ is 0.8.

5.1.4. Runtime. We perform our experiments on a personal computer equipped with a single Core (TM) i5-2320 3.0 GHz CPU and 16 Gigabytes of RAM. Table 6 shows the average query time of BoVW + HE + WGC method and BoVW + HE + WGC plus our text-based reranking on the 2-million dataset. As can be seen, our reranking stage only introduces a modest time penalty (0.8 seconds). That is because we use inverted file indexing for fast text retrieval.

5.1.5. mAP of Each Near-Duplicate Category. In our ground-truth, each group contains images of all three categories of near duplicates. As a result, we also compute the mAP with respect to each near duplicate category with the goal of further investigating and comparing the capability of our method to identify different categories of near duplicates. We conduct the comparison on the 2-million dataset. As

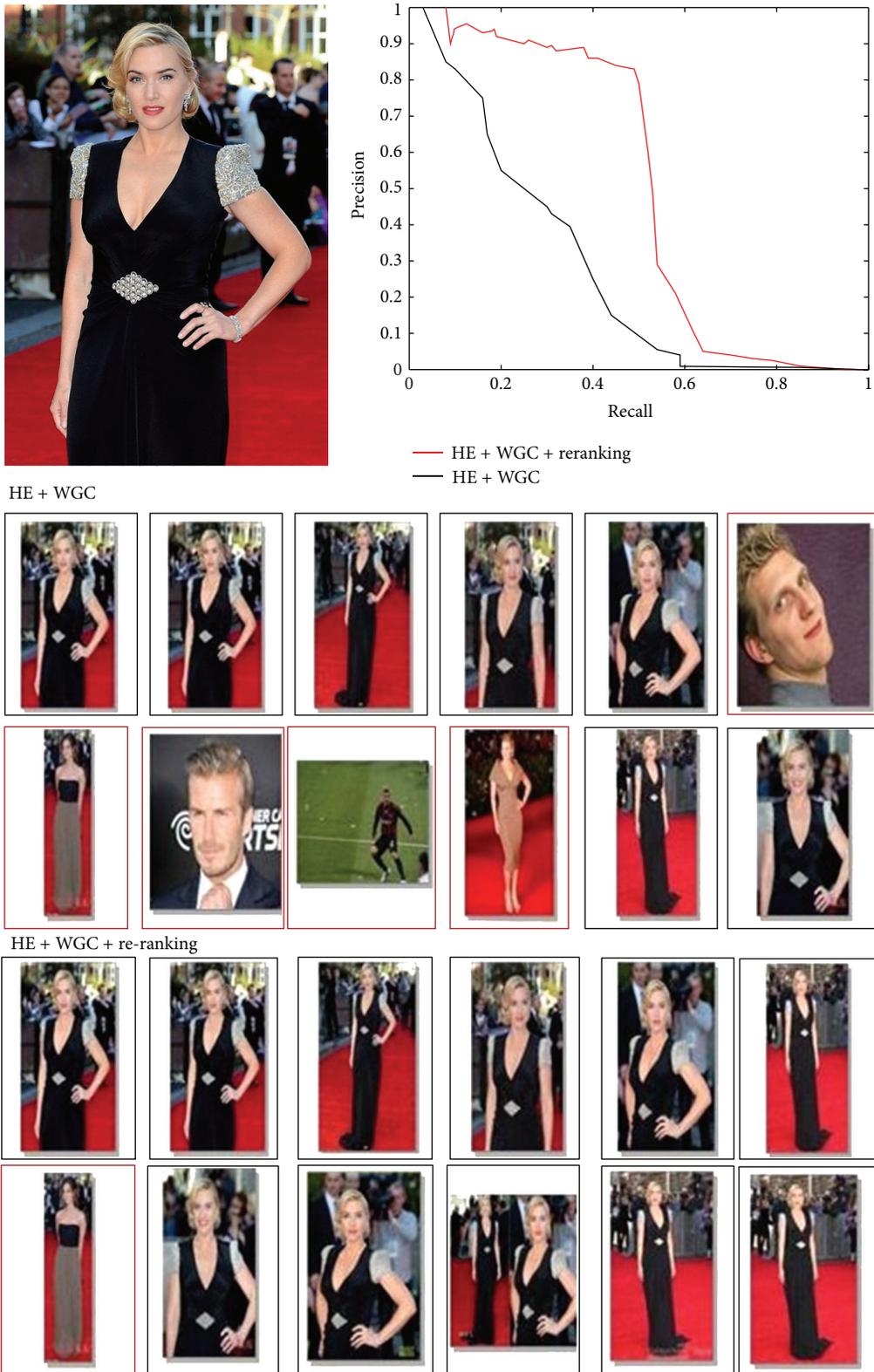


FIGURE 7: Top 12 images returned by BoVW with HE and WGC and BoVW with HE and WGC plus our text-based reranking. False positives are marked by red rectangles.

Figure 6 shows, major duplicate is easy to identify for all methods, while scene-object duplicate is hardest and partial duplicate in the middle. The mAPs of all categories are improved after our data-driven reranking stage. Obviously, “HE + WGC + re-rank” outperforms all the other methods especially for scene-object duplicate and partial duplicate.

5.2. Sample Results. Figure 7 gives example for our results on the 2-million dataset. We show the top 12 images returned by BoVW with HE and WGC and BoVW with HE and WGC plus our text based reranking. False positives are marked by red rectangles. The image at top left is the query image and top-right the precision-recall curves for the above two methods. It is obvious that our text-based reranking method filters many false positive images such as Beckham though they contain many visual local patches similar to those in the query image. For this query, we improve the mAP from 0.321 to 0.52, a 62% improvement.

6. Conclusions

We have proposed a text-based data-driven reranking framework. The framework is combined with six state-of-art BoVW schemes and improves the near-duplicate celebrity images retrieval results a lot. We also construct a 2-million-celebrity-image-text dataset with 1089 celebrities’ name query and divide near-duplicate celebrity web images into three categories in our ground-truth dataset according to our observation.

In the future work, we plan to improve our text-based reranking method by exploiting more text information not only celebrity names. After all, the text processing methods used are to some extent quite basic and nearly naïve. There is room for improvement in the future in this area. For example, finding events or topic from the text may also be more benefit for near-duplicate celebrity images retrieval. At the same time, we also want to attempt to research the use of text in general near-duplicate images retrieval not only celebrity images.

Acknowledgment

This work was supported by the National Science Foundation of China (NSFC) under Grant number 60902091.

References

- [1] H. Jégou, M. Douze, and C. Schmid, “Improving bag-of-features for large scale image search,” *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.
- [2] Z. Zhu, Y. Wang, and G. Jiang, “Statistical image modeling for semantic segmentation,” *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 777–782, 2010.
- [3] S. N. Sulaiman and N. A. M. Isa, “Adaptive fuzzy-K-means clustering algorithm for image segmentation,” *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2661–2668, 2010.
- [4] J. J. Foo, R. Sinha, and J. Zobel, “SICO: a system for detection of near-duplicate images during search,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME ’07)*, pp. 595–598, July 2007.
- [5] O. Chum, J. Philbin, M. Isard, and A. Zisserman, “Scalable near identical image and shot detection,” in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR ’07)*, pp. 549–556, Amsterdam, The Netherlands, July 2007.
- [6] H. Jegou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *Proceedings of the 10th European Conference on Computer Vision: Part I*, pp. 304–317, Marseille, France, 2008.
- [7] D. Nistér and H. Stewénius, “Scalable recognition with a vocabulary tree,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06)*, pp. 2161–2168, June 2006.
- [8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’07)*, June 2007.
- [9] W. Zhong, K. Qifa, M. Isard, and S. Jian, “Bundling features for large scale partial-duplicate web image search,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops ’09)*, pp. 25–32, June 2009.
- [10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [11] W. L. Zhao, X. Wu, and C. W. Ngo, “On the annotation of web videos by efficient near-duplicate search,” *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 448–461, 2010.
- [12] S. A. Berrani, L. Amsaleg, and P. Gros, “Robust content-based image searches for copyright protection,” in *Proceedings of the 1st ACM International Workshop on Multimedia Databases (MMDB ’03)*, pp. 70–77, New Orleans, La, USA, November 2003.
- [13] A. Qamra, Y. Meng, and E. Y. Chang, “Enhanced perceptual distance functions and indexing for image replica recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 379–391, 2005.
- [14] B. Wang, Z. Li, M. Li, and W. Y. Ma, “Large-scale duplicate detection for web image search,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME ’06)*, pp. 353–356, July 2006.
- [15] X. Zhang, L. Zhang, X. J. Wang, and H. Y. Shum, “Finding celebrities in billions of web images,” *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.
- [16] X. J. Wang et al., “Duplicate-search-based image annotation using web-scale data,” *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2705–2721, 2012.
- [17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: a database and web-based tool for image annotation,” *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 157–173, 2008.
- [18] X. Rui, M. Li, Z. Li, W. Y. Ma, and N. Yu, “Bipartite graph reinforcement model for web image annotation,” in *Proceedings of the 15th ACM International Conference on Multimedia (MM ’07)*, pp. 585–594, Augsburg, Germany, September 2007.
- [19] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma, “AnnoSearch: image auto-annotation by search,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06)*, pp. 1483–1490, June 2006.

- [20] S. Avidan, S. Baker, and Y. Shan, "Internet vision," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1367–1369, 2010.
- [21] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1470–1477, October 2003.




Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

