Near-Duplicate Video Clustering Using Multiple Complementary Video Signatures

Jun-Tae Lee, Kyung-Rae Kim, Won-Dong Jang, and Chang-Su Kim
Korea University, Seoul, Korea
E-mail: {jtlee,krkim,wdjang}@mcl.korea.ac.kr; changsukim@korea.ac.kr
Tel/Fax: +86-2-3290-3806

***Abstract*—A near-duplicate video clustering algorithm based on multiple complementary video signatures is proposed in this work. We use three kinds of frame descriptors: RGB histogram, color name histogram, and ternary pattern. Then, we convert each kind of frame descriptors for a video into a video signature based on the bag-of-visual-words scheme. Consequently, we have three signatures to represent the video. These signatures are complementary to one another, since they are robust to different near-duplication types. Also, we develop a clustering technique to refine pairwise matching results and categorize near-duplicate videos. Experimental results on an extensive video dataset show that the proposed algorithm detects near-duplicate videos more effectively than conventional algorithms.**

I. Introduction

With the growth of online video sharing and searching services, users conveniently search, download, and even re-upload videos after editing. Due to the vigorous video sharing, there are a huge number of near-duplicate videos (NDVs) in the Internet. Redundant videos occupy up to 93% of search results to user queries [1]. These NDVs take users a longer time than necessary to find desired videos. Also, redundant storage is required for near-duplicate or identical videos. It is hence essential to eliminate or at least categorize NDVs. However, it is hard to detect NDVs because of various types of video modification, including logo and subtitle insertion, resizing, cropping, and photometric change. A lot of attempts have been made to detect NDVs in the last decade [2]–[13].

An NDV detector performs three steps in general: frame descriptor extraction, video signature generation, and near-duplicate detection. First, given a query video, features for each frame are extracted to yield a frame descriptor. Second, the frame descriptors for all or key frames are converted into a video signature. An approach to this conversion is the indexing, which converts frame descriptors into votes, corresponding to indices of a video signature vector. For example, Jégou *et al.* [2] and Hu [3] employ the bag-of-visual-words scheme and the local sensitive hashing, respectively, for the indexing. Another approach is to concatenate frame descriptors to yield a video signature [4], [5]. Third, NDVs are retrieved by comparing the signature of the query video with those of database videos. Note that the clustering of NDVs in a database can improve the efficiency of the NDV retrieval and the database management. Some NDV detectors [6], [7] categorize NDVs in an ad-hoc fashion, but little work has been done to develop efficient clustering techniques for NDVs.

Let us categorize conventional NDV detectors into three groups according to their frame descriptor types: global, local, and hybrid of global and local. Global descriptors, which extract summarized features of an entire frame, are more robust to logo and subtitle insertion than local descriptors are. Global descriptors also require less computations, but are less discriminative since they do not consider spatial information. To obtain global features, invariant to photometric variations, Grana *et al.* [8] extract a color histogram in the HSV color space. In addition to color histograms, Zheng *et al.* [9] adopt spatial derivative filters to extract gradient features, and Shafeian and Bhanu [10] employ the mean and variance of a color distribution. Shang *et al.* [11] describe each frame with a binary pattern, and combine the patterns of successive frames to obtain the visual shingle of them.

Local descriptors extract features around interest points. A typical interest point detector finds points that are tolerant of geometric and photometric transform [14]. Then, around each interest point, local features are extracted and then combined into a frame-level descriptor. Yang *et al.* [15] divide a local window around each interest point into 3×3 patches, and compute the difference between the average gray-levels of each pair of patches. They employ the local sensitive hashing to aggregate the difference patterns. Wei *et al.* [16] extract the SIFT feature [17] for each interest point and exploit the bag-of-visual-word scheme to combine those features. To reduce the computational complexity of SIFT, Liu *et al.* [18] employ the ranks of gradient magnitudes within a local window. However, inserted logos or subtitles generate numerous interest points around them. These irrelevant interest points disrupt local descriptors easily.

To overcome the disadvantages of the global and the local descriptors, some NDV detectors use a hybrid of global and local frame features. Wu *et al.* [12] first select NDV candidates using a global color histogram with a high recall rate, and filter out only distinctly different videos. Then, they identify the remaining videos using local descriptors. Song *et al.* [13] extract a global color histogram and local binary patterns, and then concatenate them. They also apply a hashing technique to detect NDVs efficiently. However, the blending of different features may weaken the specific strength of each individual

This work was supported partly by the National Research Foundation of Korea(NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (No. 2009-0083495), and partly by Samsung Electronics Co., Ltd.

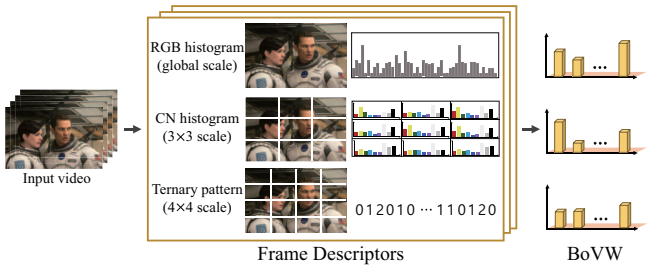
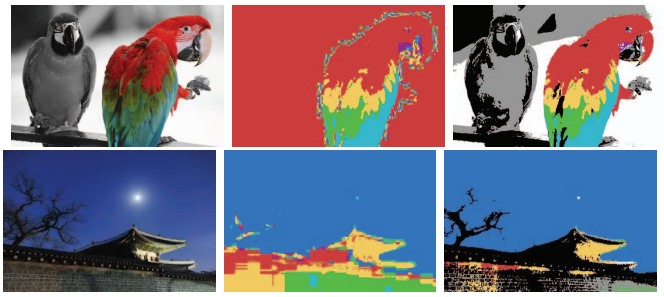


Fig. 1. An overview of the proposed video signature extraction process. Multiple frame features are extracted and combined into the corresponding video signatures. Each video signature is a bag-of-visual-words (BoVW) descriptor.



(a) (b) (c)

Fig. 2. Color naming: (a) input images, (b) uniformly quantized hue images, and (c) quantized images by the proposed color naming scheme. Whereas achromatic and dark regions in (a) are converted into colors that are uncorrelated to the human perception in (b), perceptually better quantization results are obtained by the proposed scheme in (c).

feature, degrading the NDV detection performance.

In this work, we propose a novel hybrid NDV detector using multiple complementary frame descriptors. Specifically, we use a global RGB histogram, a local color name histogram, and a global ternary pattern to describe a frame. To preserve the specific strength of each descriptor, we process the descriptors in a multi-modal manner. We convert the three kinds of frame descriptors into the corresponding video signatures using the bag-of-visual-words scheme separately. The resultant three signatures work collaboratively, since they are robust to different near-duplication types. Moreover, we cluster NDVs in a video database and refine pairwise matching results to discover undetected NDVs and eliminate false matches. Experimental results demonstrate that the proposed NDV detector significantly outperforms the conventional detectors in [11], [13].

The rest of the paper is organized as follows. Section 2 describes the proposed multiple complementary signatures, and Section 3 develops the NDV clustering algorithm. Section 4 discusses experimental results. Finally, Section 5 concludes this work.

II. Complementary Video Signatures

Video signature extraction is important for robust NDV detection. However, it is hard to represent the characteristics of a video with a single signature, since there are various near-duplication types. To describe videos effectively, we extract multiple kinds of signatures, which are complementary to one another.

We first extract frame descriptors and then combine them into video signatures using the bag-of-visual-words scheme. Our main contributions are (1) the extraction of frame descriptors in different scales and different color spaces, which are discriminative as well as robust to various near-duplication types, and (2) the multi-modal process that makes the signatures complement one another. Fig. 1 illustrates the extraction of video signatures.

A. Frame Descriptors

We use three frame descriptors, which are RGB histogram, color name histogram, and ternary pattern. Since each descriptor is robust to different near-duplication types, they work collaboratively for reliable NDV detection. To reduce the complexity, we uniformly sample frames at the rate of 2 frames/second. We extract the three kinds of frame descriptors for the sampled frames $I_1, \ldots, I_N$.

RGB histogram: We extract RGB color histograms, denoted by $f_1^{\mathrm{RGB}}, \ldots, f_N^{\mathrm{RGB}}$, which describe the whole frames globally. We quantize each color channel into 16 levels and construct a histogram. Then, we cascade the histograms for the three channels into a single 48-dimensional histogram. This global RGB histogram is robust to cropping and trimming, but sensitive to photometric modifications.

Color name histogram: It is necessary to develop a descriptor that is invariant to photometric modifications, such as color variations, white balance variations, and tone mapping. When colors are adjusted by such modifications, they vary extremely in the RGB color space. While the human visual system recognizes original and varied colors as the same one, their RGB coordinates are significantly different. Thus, it is hard to produce a photometrically invariant frame descriptor in the RGB space. In contrast to the RGB space, hue in the HSV space separates the chromatic components of colors. Thus, the hue-based color quantization, called color naming, is relatively robust to photometric discrepancies [19]. However, the conventional color naming is unreliable in the cases of low brightness or low saturation conditions. Fig. 2(b) exemplifies the problem of the hue-based color naming. To overcome this issue, we develop a hierarchical color naming scheme. We first partition the HSV space into a chromatic subspace and an achromatic subspace. Then, we quantize each subspace into representative colors, as illustrated in Fig. 2(c). The proposed scheme provides more correlated results to the human perception than the conventional uniform quantization.

When a color has a saturation value lower than a threshold, its chromaticity is hard to perceive. Also, the threshold varies with brightness conditions. Thus, we divide the HSV color space into a chromatic subspace and an achromatic subspace by employing a brightness-based saturation threshold. Specifically, for each brightness V, we empirically determine the saturation threshold $\tau(V)$ by

$$\tau(V) = S_{\max} - \sum_{i=1}^{n} \alpha_i u(V - V_i) \qquad (1)$$

where $u(\cdot)$ is the unit step function, V_i and α_i are the locations and heights of steps, respectively. In this work, the number n of steps equals 3, and the values of α_i and V_i are $(90, 8, 2)$ and $(20, 55, 85)$, respectively. Also, $S_{\max}$ is the maximum saturation, *i.e.* $S_{\max} = 100$. Note that $0 \leq S \leq 100$. When a color has a saturation value $S < \tau(V)$, it is declared to belong to the achromatic subspace. Otherwise, it belongs to the chromatic subspace. Notice that we employ the step functions to model abrupt changes in the color perception of the human visual system according to the brightness conditions.

Next, we determine representative colors for each subspace. For the achromatic space, we quantize a color into one of three levels, *black*, *gray*, or *white*, when its brightness satisfies $V < V_1$, $V_1 \leq V < V_2$, or $V \geq V_2$, respectively. On the other hand, colors in the chromatic subspace are mapped to six levels, *red*, *yellow*, *green*, *cyan*, *blue*, and *violet*, by quantizing their hue values uniformly.

Based on the color naming scheme, we describe the input frames $I_1, \ldots, I_N$ by their color name histograms $f_1^{\text{CN}}, \ldots, f_N^{\text{CN}}$. To preserve spatial information partially, we partition a frame into 3×3 patches and then extract the 9-dimensional color name histogram for each patch, as shown in Fig. 1. By cascading the patch-level histograms, we generate the 81-dimensional frame descriptor.

Ternary pattern: We also exploit ordinal relations of patches to describe a frame, which are robust against photometric editing [11]. We first divide a frame into 4×4 patches, as shown in Fig. 1, and compute the average luminance of each patch. Then, for a pair of patches (i, j), we describe the ordinal relation in a ternary digit by

$$T_{ij} = \begin{cases} 0 & \text{if } L_i - L_j > \eta, \\ 1 & \text{if } |L_i - L_j| \leq \eta, \\ 2 & \text{if } L_j - L_i > \eta, \end{cases} \tag{2}$$

where L_i and L_j are the luminance levels of patches i and j, respectively, and η is a threshold that is fixed to 20. We examine all possible pairs of patches. Consequently, we obtain the 120-dimensional frame descriptors $f_1^{\text{TPN}}, \ldots, f_N^{\text{TPN}}$, where $120 = \binom{16}{2}$.

B. Video Signatures

For video signature generation, we should integrate the three kinds of frame descriptors. There are two possible approaches: early fusion and late fusion. In the early fusion, for each frame, the three descriptors are first combined into a single unified frame descriptor. Then, a video signature is generated using the unified descriptors of the frames. The early fusion, however, may weaken the discriminative capability of each individual descriptor. On the other hand, in the late fusion, for each kind of frame descriptors, we generate a separate video signature without the descriptor unification. In this work, we adopt this multi-modal approach of the late fusion to preserve the merits of different frame descriptors.

We adopt the bag-of-visual-words scheme [20]–[22] to convert the three kinds of frame descriptors into the corresponding video signatures. In this work, visual words correspond to quantized frame descriptors, and a visual signature is generated by constructing the histogram of the visual words. To learn the visual words, we extract frame descriptors from a training set, which is composed of 78,400 frames and is not used as test videos. Then, the frame descriptors are partitioned into K clusters, whose centroids become the visual words. We employ the K-means clustering technique, and set K to 500, 900, and 500 for the RGB histograms, the color name histograms, and the ternary patterns, respectively. Consequently, given an input video, we convert the frame descriptors $\{f_1^{\text{RGB}}, \ldots, f_N^{\text{RGB}}\}$, $\{f_1^{\text{CN}}, \ldots, f_N^{\text{CN}}\}$, and $\{f_1^{\text{TPN}}, \ldots, f_N^{\text{TPN}}\}$ into the three video signatures $\mathbf{h}^{\text{RGB}}$, $\mathbf{h}^{\text{CN}}$, and $\mathbf{h}^{\text{TPN}}$, respectively.

III. Clustering Near-Duplicate Videos

We develop an NDV clustering algorithm for managing a video database. First, we conduct pairwise video matching. Let us describe the matching of video signatures based on the RGB histograms only, since the matching based on the color name histograms and the ternary patterns are performed similarly. We measure the dissimilarity between two video signatures $\mathbf{h}_i^{\text{RGB}}$ and $\mathbf{h}_j^{\text{RGB}}$ using the Hellinger distance [23], given by

$$d(\mathbf{h}_i^{\text{RGB}}, \mathbf{h}_j^{\text{RGB}}) = \sqrt{1 - \frac{1}{K^{\text{RGB}} \sqrt{\bar{h}_i^{\text{RGB}} \bar{h}_j^{\text{RGB}}}} \sum_{k=1}^{K^{\text{RGB}}} \sqrt{\mathbf{h}_i^{\text{RGB}}(k) \mathbf{h}_j^{\text{RGB}}(k)}} \tag{3}$$

where $\bar{h}_i^{\text{RGB}}$ and $\bar{h}_j^{\text{RGB}}$ are the element means of the vectors $\mathbf{h}_i^{\text{RGB}}$ and $\mathbf{h}_j^{\text{RGB}}$, respectively. Also, K^{RGB} is the dimensionality of the RGB histogram signatures, *i.e.*, $K^{\text{RGB}} = 500$. Then, we construct the binary matching matrix $\mathbf{A}^{\text{RGB}} \in \mathbb{R}^{M \times M}$, where M is the number of videos in the database. If the Hellinger distance $d(\mathbf{h}_i^{\text{RGB}}, \mathbf{h}_j^{\text{RGB}})$ is smaller than a threshold, we match the corresponding two videos and assign bit '1' to the (i, j)th element in $\mathbf{A}^{\text{RGB}}$. Specifically,

$$a_{ij}^{\text{RGB}} = \begin{cases} 1 & \text{if } d(\mathbf{h}_i^{\text{RGB}}, \mathbf{h}_j^{\text{RGB}}) < \rho^{\text{RGB}}, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where ρ^{RGB} denotes the matching threshold. We generate the matching matrices $\mathbf{A}^{\text{CN}}$ and $\mathbf{A}^{\text{TPN}}$ similarly. Then, we construct the unified matching matrix $\mathbf{A} = [a_{ij}]$ by

$$a_{ij} = a_{ij}^{\text{RGB}} \vee a_{ij}^{\text{CN}} \vee a_{ij}^{\text{TPN}} \tag{5}$$

where $\vee$ denotes the binary addition. The matching thresholds ρ^{RGB}, ρ^{CN}, and ρ^{TPN} are tightly assigned with small numbers 0.4, 0.4, and 0.5, respectively. This is because the three kinds of video signatures are robust to complementary near-duplication types, and a pair of NDVs are matched via (5) as long as at least one kind of their video signatures are similar to each other.

Despite this complementary matching rule, some NDVs may be undetected and some different videos may be falsely detected. Therefore, we refine the pairwise matching results by

TABLE I
NEAR-DUPLICATION TYPES IN THE MCL-ONEVID DATASET. A VIDEO MAY EXPERIENCE ONE OR MORE TYPES OF NEAR-DUPLICATION.

Near-duplication types	Frequency
Identical	10 %
Resolution change	80 %
Cropping	30 %
Intensity adjustment	40 %
Color change	20 %
Logo/Subtitle insertion	40 %
Trimming	15 %
Contents modification	5 %
Unclassified	5 %

exploiting the overlap ratio γ_{ij} between the NDVs of videos i and j, defined as

$$\gamma_{ij} = \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{\max(|\mathcal{M}_i|, |\mathcal{M}_j|)} \tag{6}$$

where $|\cdot|$ denotes the number of elements in a set. $\mathcal{M}_i$ and $\mathcal{M}_j$ are the set of NDVs of videos i and j, which are detected by the pairwise matching, respectively. Next, we refine the matching by

$$\hat{a}_{ij} = \begin{cases} 1 & \text{if } \gamma_{ij} > \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\delta = 0.5$ is a threshold. With this refinement, some new matches are inserted and some old matches are eliminated. Finally, using the refined matching matrix $\hat{\mathbf{A}} = [\hat{a}_{ij}]$, we construct a graph whose nodes and edges represent videos and near-duplicate matches, respectively. To determine NDV clusters, we find connected components in the graph by employing the depth-first search [24].

IV. EXPERIMENTAL RESULTS

A. Dataset: MCL-ONEVID

To simulate online video searching, we have collected NDVs from video-sharing websites YouTube [25], Vimeo [26], and Todou [27]. We have made queries to select popular videos. Videos, which have been retrieved with the same query, include many NDVs. The resultant video dataset, called MCL online near-duplicate video dataset (MCL-ONEVID), is composed of 20,000 videos whose total duration is about 667 hours. Table I shows the frequencies of near-duplication types in MCL-ONEVID.

B. Experimental Setting

We compare the proposed algorithm with two conventional algorithms STF_LBP [11] and MFH [13]. STF_LBP employs spatiotemporal features, and generates video signatures using the inverted file method. MFH describes frames with a hybrid of global and local features, and generates signatures based on learning-based hashing.

As a preprocessing step, we eliminate borders within each video. Since borders modify video signatures drastically, we detect positions, luminance intensities of which are consistent throughout all frames in a video, and eliminate them. Note that the border elimination is applied to the conventional algorithms as well.

TABLE II
COMPARISON OF THE PRECISION, RECALL, AND F-MEASURE SCORES. THE TOP THREE ROWS ARE THE PAIRWISE MATCHING RESULTS, WHILE THE BOTTOM ROW LISTS THE SCORES USING THE NDV CLUSTERING ALGORITHM IN SECTION 3. THE HIGHEST SCORE IN EACH METRIC IS HIGHLIGHTED IN BOLDFACE.

		Precision	Recall	F-measure
Pairwise detection	STF_LBP [11]	0.02	0.30	0.02
	MFH [13]	0.59	0.80	0.33
	Proposed	**0.94**	0.81	0.87
Clustering	Proposed	0.93	**0.86**	**0.89**

We assess the NDV detection performance using three quality metrics: precision, recall, and F-measure. The precision and recall scores are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. A true positive is a real NDV pair that is correctly matched. A false positive is an irrelevant video pair that is mistaken for near-duplicates. A false negative occurs when a real NDV pair is undetected. The F-measure is defined as a weighted harmonic mean of the precision and recall scores,

$$\text{F-measure} = \frac{(1+\beta) \cdot \text{Precision} \cdot \text{Recall}}{\beta \cdot \text{Precision} + \text{Recall}}, \tag{9}$$

where β is set to 0.3 as generally done.

C. Near-Duplicate Video Retrieval Performance

Table II compares the precision, recall, and F-measure scores. We observe that the proposed pairwise NDV matching outperforms both STF_LBP and MFH in all measures. Moreover, the proposed NDV clustering further improves the recall and F-measure scores at the cost of a small reduction in the precision score. Especially, the clustering increases the recall rate significantly by discovering unmatched pairs in the pairwise matching.

STF_LBP yields inferior results, since it employs ordinal features only and cannot handle various near-duplication types properly. Moreover, its temporally connected features are redundant, since consecutive frames are similar in most videos. MFH provides better results than STF_LBP by exploiting multiple features, but its early fusion strategy weakens individual features. Also, its features are not specialized for the NDV detection. In contrast, the proposed algorithm employs three complementary features, which are designed to be robust against different types of NDV modification. Consequently, the proposed algorithm provides significantly higher scores than STF_LBP and MFH. Fig. 3 shows examples of NDVs in MCL-ONEVID, detected by the proposed algorithm.

V. CONCLUSIONS

In this work, we proposed an effective NDV detector, which employed three complementary video signatures to achieve

(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. Examples of NDV detection results. The videos in the same row are clustered into the same near-duplicate group by the proposed algorithm. Original contents in (a) and (e) are edited with various near-duplication types, such as logos and subtitles (b, c, d, f), cropping (c, g), and photometric variations (b, c, d, g, h).

high discriminative capability as well as robustness to various near-duplication types. We used three frame descriptors: RGB histogram, color name histogram, and ternary pattern. Then, we adopted the bag-of-visual-words scheme to convert each kind of frame descriptors into a video signature. We exploited the three kinds of video signatures collaboratively to detect NDV pairs. Moreover, we developed the clustering technique to refine pairwise matching results and group NDVs. Experimental results demonstrated that the proposed NDV detector significantly outperforms the conventional detectors [11], [13] on the extensive MCL-ONEVID dataset.

REFERENCES

[1] X. Wu, C.-W. Ngo, H. A. G. Hauptmann, and H.-K. Tan, "Real-time near-duplicate elimination for web video search with content and context," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 196–207, Feb. 2009.

[2] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large sacle image search," *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 316–336, Feb. 2010.

[3] S. Hu, "Efficient video retrieval by locality sensitive hashing," in *Proc. IEEE ICASSP*, Mar. 2005, pp. 449–452.

[4] X.-S. Hua, X. Chen, H.-J. Zhang, "Robust video signature based on ordinal measure," in *Proc. IEEE ICIP*, Oct. 2004, pp. 685–688.

[5] J. Revaud, M. Douze, C. Schmid, and H. Jegou, "Event retrieval in large video collections with circulant temporal encoding," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2459–2466.

[6] S. S. Cheung and A. Zakhor, "Fast similarity search and clustering of video sequences on the world-wide-web," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 524–537, Jun. 2005.

[7] T.-Y. Hung, C. Zhu, G. Yang, and Y.-P. Tan, "Video organization: Near-duplicate video clustering," in *Proc. IEEE ISCAS*, May 2012, pp. 1879–1882.

[8] C. Grana, R. Vezzani, and R. Cucchiara, "Enhancing HSV histograms with achromatic points detection in video retrieval," in *Proc. ACM CIVR*, Jul. 2007, pp. 302–308.

[9] L. Zheng, G. Qiu, J. Huang, and H. Fu, "Salient covariance for near-duplicate image and video detection," in *Proc. IEEE ICIP*, Sep. 2011, pp. 2537–2540.

[10] H. Shafeian and B. Bhanu, "Integrated personalized video summarization and retrieval," in *Proc. IEEE ICPR*, Nov. 2012, pp. 996–999.

[11] L. Shang, L. Yang, F. Wang, K.-P. Chan, and X.-S. Hua, "Real-time large scale near-duplicate web video retrieval," in *Proc. ACM Multimedia*, Nov. 2011, pp. 531–540.

[12] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. ACM Multimedia*, Sep. 2007, pp. 218–227.

[13] J. Song, Y. Yang, and Z. Huang, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *Proc. ACM Multimedia*, Nov. 2011, pp. 423–432.

[14] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[15] X. Yang, Q. Zhu, and K.-T. Cheng, "Near-duplicate detection for images and videos," in *Proc. the First ACM workshop on Large-scale multimedia retrival and mining*, May 2009, pp. 73–80.

[16] S. Wei, Y. Zhao, C. Zhe, C. Xu, and Z. Zhu, "Frame fusion for video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 1, pp. 15–28, Jan. 2011.

[17] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE ICCV*, 1999, vol. 2, pp. 1150–1157.

[18] H. Liu, H. Lu, Z. Wen, and X. Xue, "Gradient ordinal signature and fixed-point embedding for efficient near-duplicate video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 555–566, Apr. 2012.

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[20] S. Battiato, G. M. Farinella, G. Gallo, and D. Ravi, "Scene categorization using bag of textons on spatial hierarchy," in *Proc. IEEE ICIP*, Oct. 2008, pp. 2536–2539.

[21] X. Li and A. Godil, "Exploring the bag-of-words method for 3D shape retrieval," in *Proc. IEEE ICIP*, Nov. 2009, pp. 437–440.

[22] K. Kersorn and S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 211–222, Feb. 2012.

[23] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[24] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT Press, 2001.

[25] YouTube. [Online]. Available: http://www.youtube.com/.

[26] Vimeo. [Online]. Available: https://vimeo.com/.

[27] Todou. [Online]. Available: http://www.todou.com/.