Keypoint Based Keyframe Selection

Genliang Guan, Zhiyong Wang, *Member, IEEE,* Shiyang Lu, Jeremiah Da Deng, *Member, IEEE,* David Dagan Feng *Fellow, IEEE*

***Abstract*—Keyframe selection has been crucial for effective and efficient video content analysis. While most of existing approaches represent individual frames with global features, we for the first time propose a keypoint based framework to address the keyframe selection problem so that local features can be employed in selecting keyframes. In general, the selected keyframes should be both representative of video content and containing minimum redundancy. Therefore, we introduce two criteria, Coverage and Redundancy, based on keypoint matching in the selection process. Comprehensive experiments demonstrate that our approach outperforms the state-of-the-art.**

***Index Terms*—Keyframe selection, keypoint, interest point, local features, video representation, video summarization**

I. Introduction

THE proliferation of video acquisition devices and the mounting interest of consumers in the access to video repositories have significantly boosted the demand for effective and efficient methods in retrieving and managing such multimedia data. A video is structurally composed of a number of stories, each story is depicted with a number of video shots, and each shot is essentially a sequence of images (i.e. frames) [1]. Due to the inherent temporal continuity of the consecutive frames within a video shot, there exists a great deal of redundant information among those frames. Therefore, selecting a set of frames to represent a video shot has been crucial for effective and efficient video content analysis.

Clustering based approaches [2] are proposed to group all frames in a video shot and identify cluster centres as keyframes. This is intuitive as the selected keyframes represent the prominent visual appearances and variations within a shot. Li *et al.* [3] turned the task of keyframe selection into a MINMAX rate distortion optimization problem for video summarization and Ngo *et al.* [4] tackled the clustering problem with the normalized cut algorithm. Recently, Panagiotakis *et al.* [5] proposed a novel keyframe selection algorithm based on three iso-content principles (Iso-Content Distance, Iso-Content Error and Iso-Content Distortion). According to the specific principle, the selected keyframes are equidistant in the video content curve and the most appropriate number of key frames is automatically estimated in supervised or unsupervised manners.

Genliang Guan, Zhiyong Wang, Shiyang Lu, and David Dagan Feng are with the School of Information Technologies, The University of Sydney, NSW 2006, Australia. e-mail: ({genliang.guan, zhiyong.wang, shiyang.lu, dagan.feng}@sydney.edu.au).

Jeremiah Da Deng is with Department of Information Science, University of Otago, P.O. Box 56, Dunedin, New Zealand. e-mail:(jeremiah.deng@otago.ac.nz).

Most of these approaches are however mainly subject to the following two limitations. First, they highly rely on global features such as color, texture and motion information, though adapting them to local features may be possible. As a result, local details of frames will be neglected, which makes the selected keyframes less representative, though global features coarsely represent visual characteristics of an image. Second, it is difficult to decide how many keyframes should be selected. For example, it is always challenging to set an appropriate threshold when two adjacent frames are compared. For the clustering based approaches, it is generally an open issue to set a reasonable number of clusters without prior knowledge.

Recently, local features such as the scale-invariant feature transform (SIFT) descriptor [6] have played a significant role in many application domains of visual content analysis such as object recognition and image classification due to their distinctive representation capacity. Hence, it would be beneficial to characterize each frame with local visual descriptors derived from the keypoints within the frame, and keyframe selection is to identify a number of frames whose keypoints are representative for the scene.

In light of the above observations, we propose a keypoint based keyframe selection framework summarized as follows. Firstly, keypoints are identified from each frame and descriptors are extracted for each keypoint. Secondly, a global pool of unique keypoints is formed to represent the whole video shot through keypoint matching. Finally, representative frames which best cover the global keypoint pool are chosen as keyframes. Two criteria, namely Coverage and Redundancy [7], are devised to ensure that each keyframe is selected to maximize the coverage of the keypoint pool and to minimize introducing redundant keypoints.

II. Keypoint Based Video Shot Representation

A. Keypoint Matching

Lowe's SIFT descriptor [6] is utilized for keypoint extraction and representation, though many other local features [8] are also applicable to our approach. The SIFT descriptor of each keypoint is a 128-dimension feature vector (a 4×4 array of orientation histograms with 8 orientation bins in each).

Straightforward keypoint matching based on SIFT descriptors will result in many false matches. Lowe proposed to improve matching robustness by imposing ratio test criterion (i.e. the ratio of the nearest neighbor distance to the second nearest neighbor distance is greater than a given threshold) [6]. However, there still exist two challenging problems.

Firstly, the cost of keypoint matching between two target frames is high. To exhaustively match keypoints, we have to calculate the distance between every pair of keypoints in

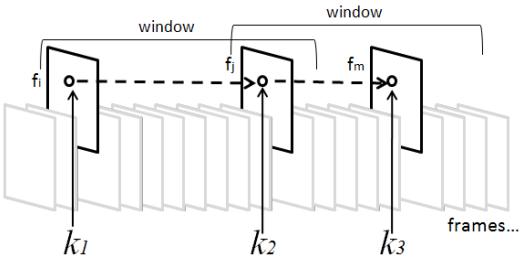

Fig. 1. Illustration of Inter-window keypoint chaining with overlapped windows, where k_1, k_2, and k_3 are matched keypoints.

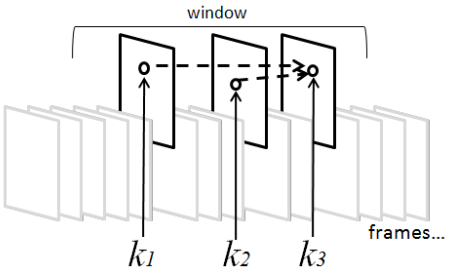

Fig. 2. Illustration of Intra-window keypoint chaining within one window, where k_1, k_2, and k_3 are matched keypoints and merged into one chain.

both frames, which is computationally expensive. In order to relieve this problem and take advantage of the continuity among adjacent frames, we adopt a matching strategy that considers only those candidate keypoints within a certain radius R of the target keypoint. Meanwhile, false matching can also be reduced with such a constraint. Secondly, there are a number of false-positive matches, and as a result, the global pool of keypoints K would contain noisy keypoints. To filter these false matches, the RANdom Sample Consensus algorithm (RANSAC) [9] is iteratively invoked to detect sets of geometrically consistent keypoint matches. This process is repeated until no further large set of matches (e.g. five matches in a group) can be found.

B. Keypoint Pool Construction

In order to build a global pool K from all keypoints k_x in each frame f_i to represent the content of a video shot, ideally every two frames f_i and f_j (a pair) within the shot should go through keypoint matching. However, such a strategy is very costly. Utilizing the inherent nature of visual continuity among consecutive video frames, we propose an Inter-window Keypoint Chaining scheme to constrain the pairing within a temporal window of size W without losing the discriminative power of keypoint matching, as illustrated in Fig. 1. Hence, keypoints are only matched within a window and chained across multiple windows. When a keypoint k_1 in frame f_i is matched with another keypoint k_2 in frame f_j, and the same keypoint k_2 is matched with a third keypoint k_3 in frame f_m, satisfying $|i-j| <= W$ and $|m-j| <= W$, we link these matches into a chain, which would finally contribute to the same unique keypoint in the global pool K without matching keypoints between f_i and f_m. As shown in our recent study [10], the window size can be adaptively determined by calculating visual variations between consecutive frames in terms of distribution correlation.

On the other hand, it may occur that true keypoint matches are dropped during matching. In order to make the matching more reliable, we also propose Intra-Window Keypoint Chaining. As shown in Fig. 2, k_1 is matched with k_3 but not k_2, and k_2 is matched with k_3. In this case, k_1, k_2 and k_3 will also be linked by a single chain, which could ease the problem of missed matching (e.g. k_1 should be a true match with k_2).

After the keypoint chaining on frames, each keypoint either belongs to a chain of matched keypoints or becomes an singleton. All singleton keypoints, which are very likely to be noisy keypoints, are discarded. Each chain is represented by its HEAD keypoint and the number of keypoints on that chain, denoted by (k_x, N_x). The global keypoint pool K is then formed by aggregating all (k_x, N_x). In order to reduce noisy chains, we further filter less important/unstable global keypoints by setting a threshold T for N_x.

III. KEYFRAME SELECTION

The goal of keyframe selection is to best represent a video shot with a minimal number of frames. That is, the keyframes are able to best represent the video shot while minimizing redundancy among them. In our case, to ensure the best representation, the keypoints of those keyframes should cover the global keypoint pool as much as possible. Since this can be formulated as a variation of the well-known Set Cover Problem which has been proven to be NP-complete [11], we adopt a greedy algorithm to approximately tackle this issue. At first, we choose the frame with the highest number of keypoints against the keypoint pool. Then, at each iteration, a frame is chosen as a keyframe if it best helps improve the coverage while minimizing redundancy. Therefore, we devise two metrics, namely *Coverage* and *Redundancy*, to guide the selection process.

In the selection process, the pool is separated into two sets, $K_{covered}$ and $K_{uncovered}$. At the beginning of the process, $K_{uncovered}$ contains all keypoints in K and $K_{covered}$ is empty. For frame f_i, denote its keypoint set as FP_i, then the *Coverage* of the frame to the pool can be defined as the cardinality of the intersection between FP_i and the uncovered set:

$$C(f_i) = |FP_i \cap K_{uncovered}|. \tag{1}$$

Likewise, *Redundancy* is defined as how many keypoints it contains in $K_{covered}$, reflecting how redundant it is based on the covered content in the shot:

$$R(f_i) = |FP_i \cap K_{covered}|. \tag{2}$$

The influence of frame f_i at an iteration is calculated in (3) as a balance of $C(f_i)$ and $R(f_i)$ controlled by α.

$$Influence(f_i) = C(f_i) - \alpha R(f_i) \tag{3}$$

A simplified illustration of the calculation is presented in Fig. 3. In this example, f_1 has higher coverage but also higher redundancy than f_2, so f_2 will be favored during the selection.

At the end of each iteration, the frame with the highest influence value and positive coverage will be selected as a keyframe, and $K_{covered}$ and $K_{uncovered}$ will be updated based

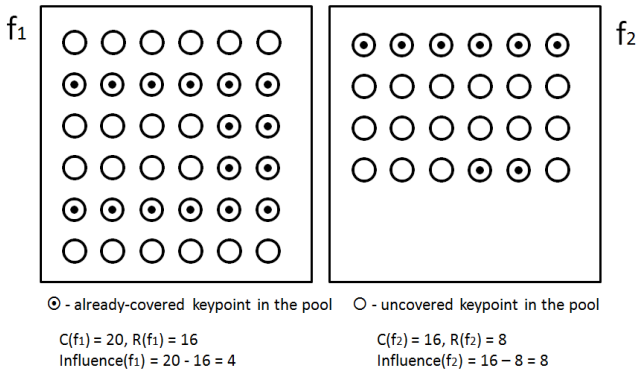

Fig. 3. A toy sample of calculating of the influence of frames, where f_2 is selected because of its higher influence.

TABLE I
THE TESTING VIDEOS FROM THE OPEN VIDEO PROJECT

Video Name	From Frame	To Frame	# of Frames
v25 A New Horizon, segment 02	664	900	237
v28 A New Horizon, segment 05	3223	3440	218
v33 Take Pride in America, segment 03	540	650	111
v39 Senses And Sensitivity, Introduction to Lecture 4 presenter	1838	1934	97
v40 Exotic Terrane, segment 01	1790	1989	200
v49 America's New Frontier, segment 07	150	500	351
v57 Oceanfloor Legacy, segment 04	1600	1800	201
v58 Oceanfloor Legacy, segment 08	540	633	94
v63 Hurricane Force - A Coastal Perspective, segment 03	867	1012	146
v66 Drift Ice as a Geologic Agent, segment 05	766	977	212

on the keypoints of the selected keyframe. The iteration repeats until all the keypoints are covered or a predefined coverage threshold of the pool K is reached.

IV. EXPERIMENTS AND DISCUSSIONS

A. Experimental Settings

We conduct experiments with two datasets. The first dataset is for case studies, consisting of 4 videos including the widely used Foreman and Costguard videos and two TV news shots (Tennis video and Zooming video). The second dataset is constructed from the Open Video Project (http://www.open-video.org) for quantitative evaluation. As described in Table I, it consists of 10 video shots across several genres (e.g. documentary, education, and history).

In our experiments, the results generally are not affected when the matching radius R is set above 100 and the window size W above 5. Hence we set the radius R to 100 (i.e. 100 pixels around a target keypoint) to reduce matching search space without sacrificing matching accuracy even in fast-motion scenes, and W to 5 so as to balance the computational cost and chaining accuracy. The threshold T to filter the unstable global keypoint affects the size of the keypoint pool and thus the granularity of details it captures. Empirically we have tried different settings in our experiments, but results shown in the following section is based on $T = 5$ to reduce noisy keypoints without losing noticeable details.

Our approach (denoted as KBKS in the figures) is compared against three state-of-the-art approaches, Iso-content distance [5], Iso-content distortion [5] and Clustering [2]. For the first two approaches we use the same Color Layout Descriptor as adopted in the original paper. For the clustering based method, we adopt the CEDD feature [12], which is a histogram representing color and texture features.

B. Case Studies

The sample frames for the four shots in discussion is presented in Fig. 5. The results for the Foreman video are displayed in Fig. 6. It is observed that our approach can capture different details when different coverage threshold values are specified. For example, the two frames under 73% coverage capture the key content, the foreman and the building. When the coverage is increased to 95%, different stages of the smiling face are captured. However, such details are missing in the results of the other methods, since they rely on global features. Meanwhile, it is also noticed that our approach misses the keyframe on the tower and sky. There are two reasons. One is that the transition is very short and some keypoint chains are discarded. The other is that there are not many keypoints due to a large portion of the uniform region and the influence score of those frames have been affected. In order to remedy this issue, we take global features into account by replacing (3) with the following equation:

$$InfluencNew(f_i) = \frac{C(f_i) - \alpha R(f_i)}{GolbalSim(f_i)}, \tag{4}$$

where

$$GolbalSim(f_i) = \sum_j Similarity(f_i, f_j). \tag{5}$$

That is, the influence of a frame f_i will be increased if it shares low similarity (i.e. small $GolbalSim(f_i)$) with other frames in terms of color and edge histogram. As shown in Fig. 4, such a simple strategy is able to effectively resolve the "missing sky" problem, though not being used in our other experiments.

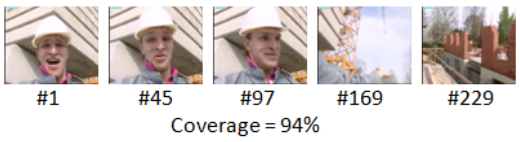

Fig. 4. New Keyframe selection results for the Foreman video.

For the Coastguard video (See Fig. 7) capturing that one boat overtakes the other, our approach selects not only the frames with both boats, but also more frames to get a higher coverage of keypoints as the background of the boat (e.g. the building and trees) keeps changing. The other two methods do capture both boats, but do not reflect the background change very well. In addition, from our selected keyframes, it seems easier for audience to understand the overtaking process.

The Tennis video contains two actions of the player with a very short panning and fading transition in between. Our selection algorithm clearly identifies these two action frames with a high keypoint coverage of 97%. The clustering-based method achieves the similar result with the help of predefined the number clusters (i.e., 2), and the Equidistance method selects the first and last frames.

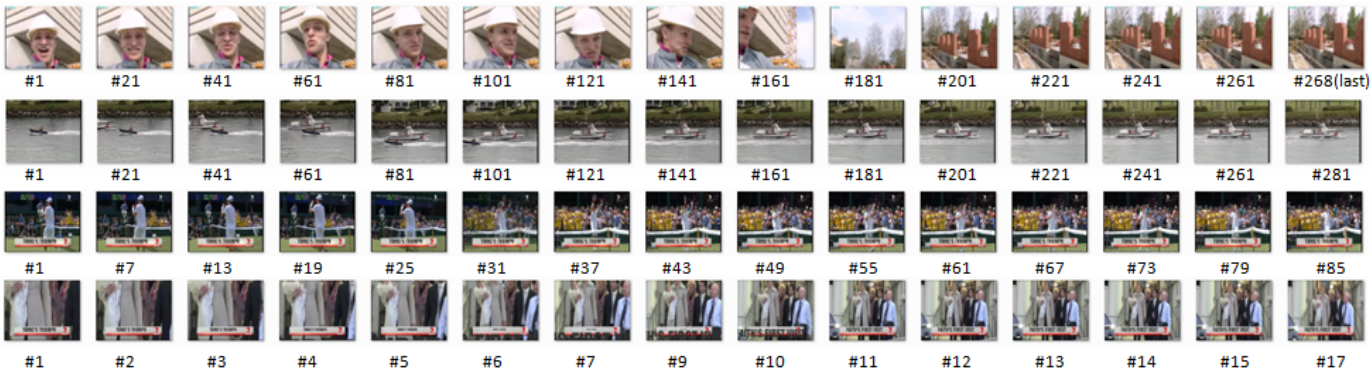

Fig. 5. Sample frames of the Foreman, Coastguard, Tennis, and Zooming videos (from top to down).

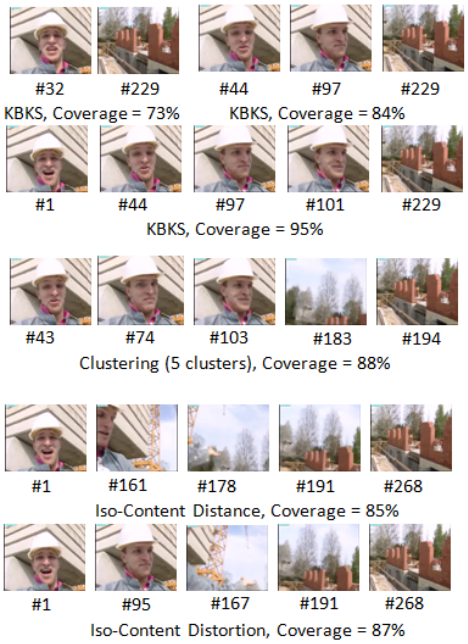

Fig. 6. Keyframe selection results for the Foreman video.

The last video is a short zoom-out footage. Our approach selects one keyframe near the end of the shot with a high coverage of 86%, since the frames at the beginning are part of such a keyframe. For the clustering-based method, if the number of cluster is set to 1, we get the keyframe with the middle frame of the shot. That is, clustering based approaches generally take the frame with average information as representative frames. For the Equidistance method, it has the limitation of selecting both the first and the last frames as a starting point, which is not necessary for many cases such as zooming.

C. Quantitative Evaluation

The ground-truth keyframes of the videos described in Table I are manually selected by three students with video processing background. When calculating the metrics, we average the results among the three ground-truth sets of keyframes. The number of target keyframes is set to five. As for our approach, we try different values of coverage starting from 50% until five keyframes are generated. The following metrics are chosen: Precision, Recall, F-score, and Dissimilarity.

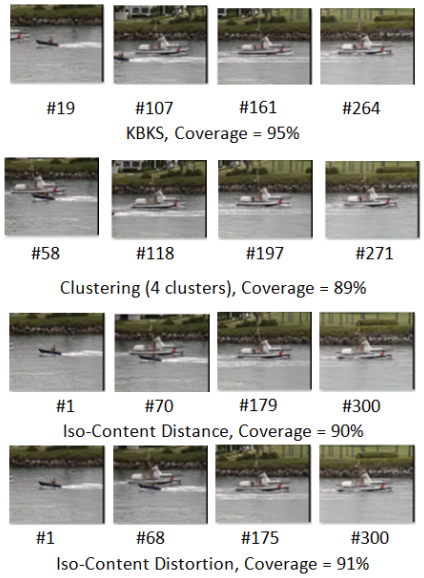

Fig. 7. Keyframe selection results for the Coastguard video.

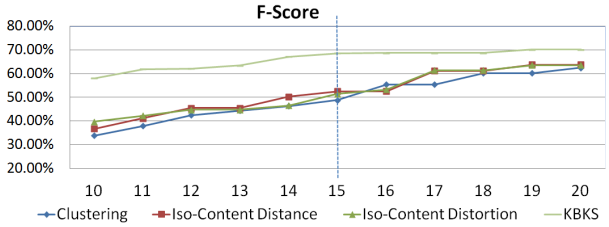

Fig. 8. Influence of X on the F-score.

A candidate keyframe is considered matched if being no more than X frames apart from a ground truth keyframe. A ground-truth keyframe will be matched with at most one candidate keyframes. F-score is a combination of both the precision and recall indicating the overall quality. Dissimilarity measures the difference between the candidate keyframes and the ground-truth keyframes. It is defined as:

$$Dissimilarity = \sum_{f_c} min_{f_t} d(f_c, f_t), \tag{6}$$

where f_c is a candidate keyframe and f_t is a ground-truth keyframe, and $d(f_c, f_t)$ is a distance measure of two keyframes, which is the difference of their frame indices.

In order to explore the influence of X, various experiments

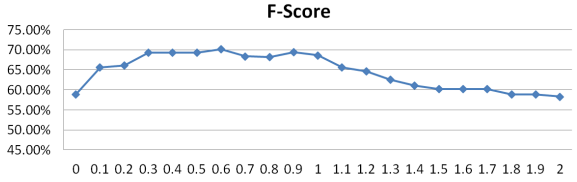

Fig. 9. Influence of α on the F-score.

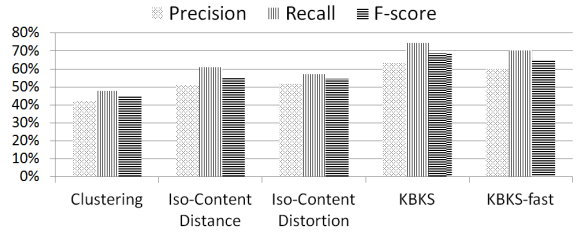

Fig. 10. Quantitative Evaluation on the second dataset in terms of Precision, Recall and F-score

were conducted by varying X from 10 to 20 while fixing α to 1 and T to 5. As shown in Fig. 8, the F-score of every method increases and stabilizes. While setting a high value for X does not reflect a true match, we set X to 15 in the following experiments. Similarly, experiments were conducted to explore the influence of α in (3) by setting X to 15 and T to 5, and varying α from 0 to 2. As shown in Fig. 9, α does influence the selection result, however, not in a significant way. F-Score grows when α increases from 0 to 0.3, and stabilizes between 0.3 and 1.2. This could be explained that a frame with a higher coverage introduces more new visual content and is more likely to introduce less redundancy. For the sake of simplicity, we set α to 1 in the following experiments.

As illustrated in Fig. 10, our approach achieves better performance in regards to precision, recall and F-score. The dissimilarity result shown in Table II also indicates that the results of our approach are more similar to the ground truth compared to other methods.

D. Computational Complexity

In our experiment, the frame size of Foreman and Coast-guard is 352 x 288, and frame size of the videos in the Open Video project is 352 x 240. With a standard 3.0GHz Dual core desktop computer, for a video shot of 300 frames (i.e. 10 seconds), the total time needed is roughly 150 seconds broken down into 150 seconds for the first step (Section II.A) and the second step (Section II.B) and less than 1 second for the third step (Section II.C) and the fourth step (Section III).

The computational cost of our approach is largely affected by the efficiency of Keypoint Extraction and Matching. As for Keypoint Extraction, it costs about 0.02 second to process one frame. Regarding Keypoint Matching, it takes about 0.1 second to process one frame-pair. Therefore, the time cost of keyframe selection on a video shot with N frame is roughly $N * 0.02 + W * N * 0.1 + 1$, and complexity is $O(N)$. When $N = 300$ and $W = 5$, the time cost is about 150 seconds.

In order to reduce the computational cost, we utilized the randomized kd-tree forest based matching algorithm [13] within the window. The matching speed is about ten times faster than the conventional matching algorithm. That is, the computational cost of the fast matching algorithm is about 15 seconds for 300 frames. As shown in the rightmost column of Fig. 10 and Table II, the performance of the fast algorithm (namely KBKS-fast) is still comparable to the original scheme, though approximated matching is employed in [13].

TABLE II
QUANTITATIVE EVALUATION ON THE SECOND DATASET: DISSIMILARITY

Clustering	Iso-Content Distance	Iso-Content Distortion	KBKS	KBKS-fast
35.3	29.72	30.72	27.5	28.1

V. CONCLUSION

In this paper we present a keyframe selection framework based on discriminative keypoints. A video shot is firstly represented by a global pool of keypoints through keypoint chaining. Secondly, a greedy algorithm is developed to select suitable keyframes based on the two intuitive metrics of Coverage and Redundancy. The experimental results on both case studies and quantitative evaluation demonstrate that our proposed approach is very promising. We will further apply this approach to video summarization in the future.

ACKNOWLEDGMENT

The work presented in this paper is partially supported by ARC grants. The authors would like to thank the editors and anonymous reviewers for their insightful and constructive comments.

REFERENCES

[1] D. Feng, W. Siu, and H. J. Zhang, Eds., *Multimedia Information Retrieval and Management*. Springer, 2003.
[2] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *IEEE International Conference on Image Processing*, 1998.
[3] Z. Li, G. M. Schuster, and A. K. Katsaggelos, "MINMAX optimal video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1245–1256, 2005.
[4] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 296–305, 2005.
[5] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on Iso-content principles," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, pp. 447–451, 2009.
[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.
[7] C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, and Y. Theodoridis, "Segmentation and sampling of moving object trajectories based on representativeness," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 7, pp. 1328–1343, Jul 2012.
[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 27, pp. 1615–1630, 2005.
[9] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.
[10] S. Lu, Z. Wang, M. Wang, M. Ott, and D. Feng, "Adaptive reference frame selection for near-duplicate video shot detection," in *IEEE International Conference on Image Processing*, 2010.
[11] R. M. Karp, "Reducibility among combinatorial problems," *Complexity of Computer Computations*, vol. 40, pp. 85–103, 1972.
[12] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *International Conference on Computer Vision Systems*, 2008.
[13] C. Silpa-Anan and R. Hartley, "Optimised kd-trees for fast image descriptor matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.