Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 8, August 2018, pp. 1824-1832
DOI: 10.23940/ijpe.18.08.p19.18241832

Video Indexing and Retrieval based on Key Frame Extraction

Wenshi Wang[a,b], Zhangqin Huang[a,b,*], Weidong Wang[a,b], Shuo Zhang[a,b], and Rui Tian[a,b]

[a]*Beijing Advanced Innovation Center for Future Internet Technology, Beijing University of Technology, Beijing, 100124, China*
[b]*Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing, 100124, China*

Abstract

With the ever growing amount of digital video data becoming available, people are gradually challenged to come up with methods that facilitate video indexing and retrieval. This paper presents a key frame based method that employs shot boundary detection and "bag-of-visual-words (BoW)" based on local keypoints for key frame extraction and semantic concept detection. The performance of BoW features is optimized by choosing appropriate representation choices. Once video frames are represented by BoW features, we can adopt a spectral clustering algorithm for the generation of key frames in each shot, and then we can classify these key frames using support vector machines for video indexing. Finally, this paper performs a query by concept search for video retrieval. The experimental results demonstrate that the proposed approach is capable of retrieving videos. Compared with the existing related method, the proposed method yields better results for key frame extraction and yields a mean average precision (MAP) of 0.68 for the video retrieval model.

Keywords: frame entropy; bag-of-visual-words; spectral clustering; key frame extraction; semantic concept detection; video retrieval

(Submitted on May 1, 2018; Revised on June 13, 2018; Accepted on July 20, 2018)

1. Introduction

In the wake of growing multimedia technologies and Internet application technologies, video sequences are readily acquired by users and data distribution is very easy. Meanwhile, the size of the video data we can acquire is increasing; multimedia information, including images, sounds, and videos, can also be uploaded by unprofessional users daily. This enables people to process a large number of video information, but it is hard to access the desired video sequences in spite of the technical progress in indexing and retrieval [1]. Currently, the problem we face is the lack of efficient methods of retrieval of video sequences rather than the lack of the video resources. Thus, efficient video processing demands, such as classification, indexing, retrieval, storage, and summarization, are a precondition of efficient acquirement of desired video sequences, especially in the situation of dramatic data increase.

To address the problem of textual data retrieval, a common method involves searching keywords by accurate accouplement in the database. This method is effective for most text information retrievals [2]. Compared with textual information retrieval, there exists a different problem to be taken into account for the retrieval of video data because of the difference of structured data and semantic data. Semantic concept detection technology has become an effective approach in video indexing and retrieval, and it is essentially a classification process that classifies images and predicts the labels of images [3]. In video retrieval, a detector that indicates the semantic concepts of query is selected for solving a given query by keyword or example, and the desired video sequences are then ranked and presented to users.

This work presents an approach for key frame based indexing and retrieval of videos, including shot segmentation, representation of BoW features, key frame extraction, and video indexing and retrieval. The proposed approach first detects the shot boundaries based on frame entropy and local keypoint matching. Then, the performance of BoW features is optimized by choosing appropriate representation choices due to keypoint-based image representation. After video frames have been expressed by BoW features, the proposed method employs a spectral clustering algorithm to select key frames in

* Corresponding author.
E-mail address: zhuang@bjut.edu.cn

video shots, which are used for semantic concept detection. Finally, this paper considers an automatic query by concept finding for video retrieval, in which the user query is selected from used concepts.

The remaining sections of this work are shown below. Section 2 surveys existing works. Section 3 describes the details of our work, including shot segmentation, feature representation, key frame extraction, and semantic concept detection. Experimental evaluation of the proposed key frame selection and video indexing and retrieval is discussed in section 4. Section 5 described conclusions and future works.

2. Related Work

Video retrieval permits people to retrieve desired shot sequences in a video database. Related works on video retrieval methods are vast, and it is not our goal to cover them here completely. Souza et al. [4] present a method for content-based indexing and retrieval of TV video sequences. The main steps of this method include video boundary detection, key frame selection, and content-based video retrieval, where hashing and k-d tree methods are employed to accomplish video indexing and retrieval. Mallick et al. [5] present a key frame based retrieval method. They exploit a clip detection method for the extraction of key frames. In addition, they employ a color correlation histogram to represent the key frame. This method can perform the query by example, where key frames selected from the query video are matched with those in the video database using the similarity measure. Saravanan [6] introduces valid information retrieval methods based on key frame extraction. This paper presents the query by image example, in which a frame would be mapped with labeled key frames by means of similarity measures. O'Connor et al. [7] apply a few video indexing tools, including shot segmentation, key frames selection, shots clustering, and news story segmentation, to automatically index video data for succedent misalignment browsing. Chen et al. [8] propose a novel multimedia retrieval approach based on key frames of videos. This method compares video spatial-temporal feature curves to determine the similarity between the query video and the videos that the user wants to search.

Memar et al. [9] propose a unified semantic-based method for resemblance computation to improve the retrieval performance in video indexing and retrieval. Bartolini et al. [10] introduce a complete video retrieval system. It is based on the automatic semantic label of video offering descriptions. Then, video shots can be searched using tags and visual features. Aly et al. [11] address a searching method based only on a noisy concept detector output. They search the unobservable binary events in videos using a probabilistic ranking framework that models the possibility of relevance. It can be implemented by the absence and presence of concepts. Wei et al. [3] present the usage of a detector for large-scale video retrieval. It considers the fusion of semantics, reliability, observability, and variety of concept detectors for query answering in video domains. Jeong et al. [12] propose an approach for automatic video annotation. It uses noumena to enhance the performance of information retrieval and sharing procedures. Zarchi et al. [13] present the semantic-based framework for data search, which detects the objects and recognizes upper-level concepts specifying existing co-occurring targets by mean of two conceptual layers, respectively. Toharia et al. [14] study the effects of semantic tests on video searches by artificially modifying the performance of detection of semantic features. Experimental results demonstrate that the effect of varying the capability of semantic detection is not reflected in search capability.

Lu et al. [1] address the problem of wholesale information search based on spatial indexing and querying of Field-Of-View (FOVs). They present the novel indexing method, which harnesses FOVs' camera positions, directions, and view-distances. Lee et al. [15] present a data cube model for high dimension indexing and searching of monitoring videos. This approach offers a high dimension process for target objects of videos. It relies on the chronological view, events, and position of the model. Yoshitaka et al. [2] adopt the Gaze detection method to find the focus of interest in videos. This contributes to enhanced accuracy of information search and video abstraction.

3. The Proposed Method

This section presents our method for key frame based video indexing and search. It consists of four main steps. The first step consists of implementing the segmentation of videos into shots using the shot segmentation method. It detects the shot boundary using frame entropy analysis as well as SIFT local features matching. The second step consists of optimizing the performance of "bag-of-visual-word" (BoW) feature representation of images by selecting proper representation choices in each shot. The representation choices greatly affect the performance of semantic concepts detection. The third step consists of summarizing the video shots by extracting key frames, which is based on the BoW features and a spectral clustering algorithm. One or more key frames will be selected in each shot by analyzing the content complexity. A succinct representation of the video data decreases the number of videos needed in video analysis and contributes to efficient indexing and retrieval operations. In the fourth step, the proposed method performs the semantic concept detection by

adopting support vector machines (SVM) and retrieves the videos of interest by concept search. The above mentioned four steps are illustrated in Figure 1.

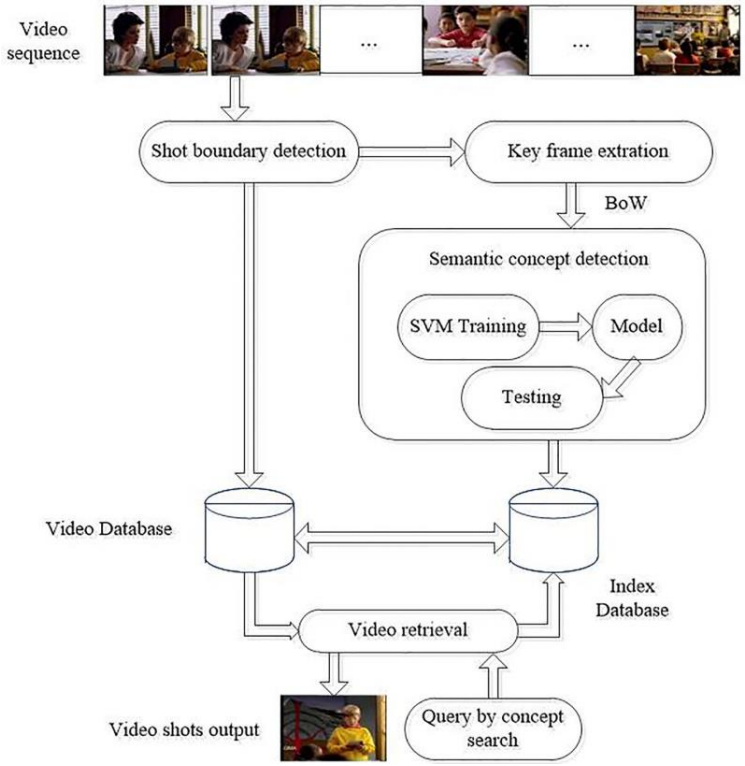

Figure 1. Overview of the proposed video retrieval system

3.1. Shot Segmentation

In our scheme, the segmentation of videos into shots will be first implemented using the shot boundary detection method proposed in [16]. Fade boundaries and candidate abrupt boundaries are determined by frame entropy analysis. Abrupt boundaries are determined by matching SURF's keypoints between the candidate boundary and the contiguous frame. Moreover, in the abrupt shot boundary detection, we use SIFT features instead of SURF features [17].

Define 1 Let a video $V=\{k_1, k_2, \cdots, k_n\}$ of length n be represented by entropy measurements $V_e=\{k_{e1}, k_{e2}, \cdots, k_{en}\}$, where k_{ei} represents the entropy of the i^{th} frame in the sequence. Then, the fade detectors are $F=\{f_1, f_2, \cdots, f_{n-1}\}$, where f_i is defined as

$$f_0 = 0$$
$$f_i = \begin{cases} 1, & k_{e_i} > k_{e_i+1} \\ 0, & k_{e_i} < k_{e_i+1} \\ b_{i-1}, & k_{e_i} = k_{e_i+1} \end{cases} \tag{1}$$

Define 2 Suppose $C=\{c_1, c_2, \cdots, c_k\}$ represents candidate boundaries. Each c_i is a label of a frame, and $Q=\{q_i\}$, $R=\{r_j\}$ are two sets of SIFT keypoints extracted from frame f_{ci} and f_{ci+1}, respectively. Let s be the Euclidean distance measurement, and β=0.6. We define point pair (q_i, r_j) as the match if the following equations hold:

$$s(q_i, r_j) = \min_{r_k \in R} s(q_i, r_k) \tag{2}$$

$$s(q_i, r_j) < \min_{r_l \in R, l \neq j} s(q_i, r_k) \times \beta \tag{3}$$

The main steps of the method are described next.

// *fad boundary detection*

(1) Apply the run length encoding (RLE) on F. Then, compute triples $g=\{g_1, g_2, \cdots, g_l\}$ by enhancement of RLE with the minimum entropy of a run. $g_j=(\Delta_j^1, \Delta_j^2, \Delta_j^3)$ and Δ_j^1 is a binary value. Δ_j^2 is the extent of the run of Δ_j^1. Δ_j^3 is the minimum entropy of the Δ_j^2 frame.

(2) for j=1, j++, j<=l
If $\Delta_j^2 > T_f$ and $0 \leqslant \Delta_j^3 \leqslant T_e$
{ If Δ_j^1=0, then the run is a fade-in;
else if Δ_j^1=1, then the run is a fade-out;
}; end

(3) After all fades have been detected, the fade boundaries can be detected by the cumulative sum of all frequencies $\sum_{i=1}^{j} \Delta_i^2$.

// *abrupt shot boundary detection*

(4) Assume frame f_{ci} is a candidate abrupt boundary, if the difference in entropy between frame f_{ci} and f_{ci+1} exceeds a threshold T_A, where f_{ci+1} is the adjacent frame of f_{ci}.

(5) Match SIFT features extracted from f_{ci} and f_{ci+1} using nearest neighbor search.

(6) Calculate the matching score γ between frame f_{ci} and f_{ci+1}

$$\gamma(f_{ci}, f_{ci+1}) = \frac{|M|}{|Q|} \tag{4}$$

(7) Calculate standard deviations of locations σ_x (x-direction), σ_y (y-direction) of interest points to overcome the problem of weak matches. Shifts in interest point positions from frame to frame are below T_σ.

(8) If $\gamma(f_{ci}, f_{ci+1}) < T_c$
f_{ci} is an abrupt boundary;
else if $\gamma(f_{ci}, f_{ci+1}) < T_c$ and $\sigma_{x,y}(f_{ci+1}) - \sigma_{x,y}(f_{ci}) > T_\sigma$
f_{ci} is an abrupt boundary;
end

(9) Finally, combine abrupt boundaries A and fade-out boundary set B-out as the final shot boundaries.

In our method, the parameter T_f is computed by using the number of frames per second (*fps*) and kept as *fps*×0.5, and T_e=2. The thresholds T_A, T_c, and T_σ in the abrupt boundary detection are empirically chosen as 0.01, 0.1, and 0.09, respectively.

3.2. Representation of BoW Feature

This section presents the BoW image representation based on SIFT keypoints [18] for key frame extraction and semantic concept detection. Feature representation choices have a great effect on the capability of video indexing, where the number of visual words can be adjusted by the number of clusters. In our method, we experiment with vocabularies of 100 visual words. In addition, visual word weighting has a great effect on video image searches. The soft weighting method [18] is applied to illustrate the importance of various words. For each interest point, it performs a nearest neighbor search. These top-N nearest words are selected. We employ the 100-dimensional vector $\mathbf{w}=[w_1, \cdots, w_t, \cdots, w_{100}]$, where w_t represents the weight of word t, defined as

$$w_t = \sum_{k=1}^{R} \sum_{l=1}^{S_k} \frac{1}{2^{k-1}} D_{sim}(l,t) \tag{5}$$

Where S_k is the number of interest points and the kth nearest neighbor is t. $D_{sim}(l,t)$ represents the cosine similarity between interest point l and t. Moreover, it is important for classifying the images to integrate the spatial location information of keypoints in an image. It can be implemented by three steps. First, frames are partitioned into equal-sized rectangular regions. Second, these visual word features are computed from each region. Third, features of regions are concatenated into the final vector. We experiment with 3×3 means having nine regions.

3.3. Key Frame Extraction

To perform key frame extraction, the frames of a video are represented by using the "bag-of-visual-words" and the local keypoints described by the SIFT descriptor (see Section II-B). Moreover, the similarity matrices for the pairs of frames in each shot is computed and employed as input to an improved spectral clustering method. It partitions video frames into groups, and the frame with the highest mean similar degree is selected as the key frame.

The improved spectral clustering algorithm [19] is shown as follows. Let $S=\{s_1, s_2, \cdots, s_N\}$ be a shot of length N, where s_k is the k^{th} frame in the shot and frames of the shot will be clustered into K clusters.

(1) Calculate the $N\times N$ similar degree matrix M for the pairs of frames in the shot S.

(2) Compute the diagonal matrix D, where $D(i,i)$ is equal to the summation of the elements of M's i^{th} row. Let L be equal to $I\text{-}D^{-1/2}MD^{-1/2}$.

(3) Build matrix $X=[x_1, x_2, \cdots, x_K]\in R^{N\times K}$ by computing the principal eigenvectors of L.

(4) Compute $Y=\{y_{ij}\}_{N\times K}$, where

$$y_{ij} = x_{ij} \Big/ \left(\sum_j x_{ij}^2 \right)^{1/2} \tag{6}$$

(5) Partition Y's row into K clusters based on fast global k means.

(6) Distribute frame s_i into group j when Y's i^{th} row is distributed to group j.

The similar degree measure we consider is the chi-square measure due to its effectiveness. Given two frame feature vectors x, y, the similarity measure $Sim(x,y)$ for the pairs of frames of each shot is computed as follows:

$$Sim(x, y) = 1 - \sum_i \frac{(x_i - y_i)^2}{0.5\times(x_i + y_i)} \tag{7}$$

The number of clusters cannot be predetermined. This attributes to the fact that content variation may be different in different shots. In our method, the number of the clusters is estimated by using a bootstrap method [20], in which the corresponding estimated clustering instability is minimized.

3.4. Video Indexing and Retrieval

The video search process includes semantic concepts detection, mapping, and video retrieval. In our method, after frames have been represented by BoW features, semantic concept detection can be performed using the method adopted in [18], where LibSVM package [21] is employed for training from labeled images as well as prediction of the labels of other images. The decision function of the test sample x is computed as follows:

$$f(x) = \sum_i \alpha_i y_i \kappa(x_i, x) - b \tag{8}$$

Where x_i represents the training sample. y_i and α_i are the class label of x_i and the learned weight of x_i, respectively. The χ^2 RBF kernel function $\kappa(x_i,x)$ has the following form:

$$\kappa(x, y) = e^{-\rho d(x,y)} \tag{9}$$

Where $d(\boldsymbol{x}, \boldsymbol{y})$ represents the χ^2 test function of two input vectors $\boldsymbol{x}$, $\boldsymbol{y}$, i.e.

$$d_{\chi^2}(x, y) = \sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)} \tag{10}$$

The training and testing data sets consist of 2000 key frames and the rest of the key frames respectively, which are extracted by the proposed method. We employ 20 semantic concepts to represent the training data and use the training model to detect the semantic concepts of the testing data. The set is a multiple label set. A key frame could have multiple concepts or none of the concept. Moreover, suppose there are l testing instances and d labels. Let y^i be the true label vector of the i^{th} instance and $\hat{y}^i$ be the predicted label vector. Evaluation criteria for multi-label classification problems, including exact match ratio, macro-average F-measure, and micro-average F-measure, can be computed using (11-13), respectively.

$$\frac{1}{l}\sum_{i=1}^{l} I\left[\hat{y}^i = y^i\right] \tag{11}$$

$$\frac{1}{d}\sum_{j=1}^{d} \frac{2\sum_{i=1}^{l} \hat{y}_j^{\,i} y_j^i}{\sum_{i=1}^{l} \hat{y}_j^{\,i} + \sum_{i=1}^{l} y_j^i} \tag{12}$$

$$\frac{2\sum_{j=1}^{d}\sum_{i=1}^{l} \hat{y}_j^{\,i} y_j^i}{\sum_{j=1}^{d}\sum_{i=1}^{l} \hat{y}_j^{\,i} + \sum_{j=1}^{d}\sum_{i=1}^{l} y_j^i} \tag{13}$$

Where I represents the binary indicator function.

In the video retrieval process, semantic concepts stored in video databases are used to map with the query by concepts specified in the video database. The retrieved video shots contributing to the selected query are ranked according to the number of key frames retrieved from the same shot, and the greater the number of key frames, the higher the priority of the retrieved shot. Using our method, the video retrieval system returns 50 shots for each query task. The mean average precision (MAP), which is the mean AP over all search queries, will be employed to evaluate the retrieval method.

4. Result and Analysis

This section presents the evaluation of the proposed method, which performs two groups of experiments: evaluation of key frame selection and video indexing and retrieval. We first describe the experimental platform and data sets that are used for our experiments. Then, we present the results and demonstrate the advantage that can be obtained by performing the proposed method for a few scientific video data sets.

The experiment was executed on ThinkCentre M6200t with Intel 3.20GHz and implemented in Matlab. Our experiments were performed on the video datasets selected from the Open Video Project, and the details of the videos tested are shown in Table 1.

Table 1. Details of the test video datasets

Video sequences	Duration (seconds)	Genre	Number of frames	Number of shots
Total	7763	Educational	232686	2458

4.1. Key Frame Extraction

The evaluation of a video summary produced by the key frame extraction method has many difficulties, because of the lack of the objective ground-truth. In our experiment, the compression ratio and fidelity measure are used to evaluate the

performance of key frame extraction. For a video sequence *V*, the fidelity measure value of the key frame set *keys* is computed as follows:

$$fidelity(V, keys) = 1 - \frac{\max_i \left\{ \min_j \left\{ D(i, keys_j) \right\} \right\}}{\max_D} \tag{14}$$

Where *D* represents the measure between the i^{th} image in video *V* and the *j*th image in *keys*. max_D is the largest possible value of measure *D*. The color histogram intersection method is employed to compute the distance measure *D*. The compression ratio can be calculated by

$$compression \text{ ratio} = \frac{K}{T} \tag{15}$$

Where *T* represents the number of images of videos and *K* the number of key frames extracted from the video.

Table 2 lists the comparative results of the proposed algorithm and related algorithm based on the color histogram [19]. A high fidelity value usually represents a good summarization produced by the key frame selection method. It can be seen that the proposed method has a better solution.

Table 2. Comparative results using compression ratio and fidelity measure

Methods	Number of key frames	Compression ratio	Fidelity
Proposed method	4109	98.2%	0.83
Color histogram [19]	5044	97.8%	0.79

4.2. Video Indexing and Retrieval

In our experiment, these key frames (2000 out of 4109) represented by 20 semantic concepts are employed to train the SVM classifier model. It is employed to detect the semantic concepts of the rest of the key frames. These concepts include: 1. adult, 2. children, 3. classroom, 4. screen, 5. computer, 6. chart, 7. animal, 8. outdoor, 9. sky, 10. underwater, 11. river, 12. road, 13. vehicle, 14. demonstrate, 15. device, 16. drawing, 17. rock, 18. interview, 19. water, and 20. crowd. Figure 2 shows the key frames corresponding to these concepts.

Figure 2. Key frames corresponding to 20 concepts.

For the multi-label problem, Table 3 lists the evaluation results of semantic concept detection for one key frame with maximum frame entropy in a shot and one or more key frames extracted by the proposed method. We can see that one or more key frames yield better results than the method based on one key frame. Figure 3 presents the average precision (AP) of each query by concept search (at top 30 shots) for the proposed method and the method based on one key frame. Table 4 lists the MAP of the video retrieval system, and the proposed method yields a MAP of 0.68. From Table 3, Figure 3, and Table 4, it can be seen that retrieval performance based on one or more key frames extracted by the proposed method is higher than that based on only one key frame in each shot. This is because of the fact that the content of a shot may not be effectively represented by using one key frame in each shot, and a concept can be well detected by using one or more key frames in each shot.

Table 3. Evaluation results of semantic concept detection

Key frame extraction method	Macro-average F-measures	Micro-average F-measures	Exact match ratio
One key frames only for each shot	0.712	0.436	0.5236
One or more key frames (the proposed method)	0.847	0.826	0.6813

Table 4. Performance results of video retrieval system using MAP measure

Measure	One key frame	One or more key frames
Mean average precision (MAP)	0.43	0.68

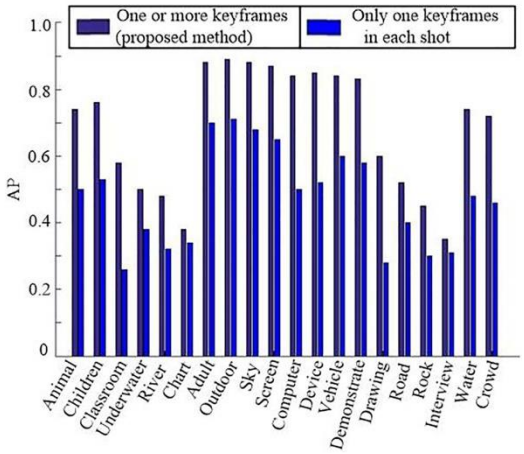

Figure 3. Performance results of each concept for one or more key frames and one key frame in the shot

5. Conclusions

The work proposes a unified approach for the indexing and retrieval of videos that can effectively retrieve video shots from a database with a reduced number of frames. This paper describes four steps: shot boundary detection, representation of BoW feature, key frame selection, and video indexing and retrieval. The experimental results demonstrate the effectiveness of the retrieval approach on educational videos. The proposed method improves the average fidelity of key frame extraction to 0.83 compared with the related method and yields a mean average precision (MAP) of 0.68 for the indexing and retrieval of videos. In future work, we will test this method on larger datasets and consider the fusion method of different aspects of concept classifiers that detect semantic concepts in videos.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (No. 61502018).

References

1. Y. Lu, C. Shahabi, and S. H. Kim, "Efficient Indexing and Retrieval of Large-scale Geo-tagged Video Databases," *Geoinformatica*, Vol. 20, No. 4, pp. 829-857, October 2016
2. A. Yoshitaka, "Image Video Indexing Retrieval and Summarization based on Eye Movement," in *Proceedings of 4th International Conference on Computing and Informatics* (ICOCI 2013), Sarawak, pp. 28-30, Malaysia, August 2013
3. X. Y. Wei and C. W. Ngo, "Fusing Semantics, Observability, Reliability and Diversity of Concept Detectors for Video Search," in *Proceedings of the 16th International Conference on Multimedia*, pp. 26-31, Vancouver, Canada, October 2008
4. C. L. D. Souza, F. L. C. Pádua, C. F. G. Nunes, G.T. Assis and G. D. Silva, "A Unified Approach to Content-based Indexing and Retrieval of Digital Videos," *Journal of Artificial Intelligence Research*, Vol. 3, No. 3, pp. 49-61, 2014
5. A. K. Mallick and S. Maheshkar, "Video Retrieval based on Color Correlation Histogram Scheme of Clip Segmented Key Frames," in *Proceedings of International Conference on Parallel, Distributed and Grid Computing* (PDGC), pp. 213-218, Waknaghat, India, December 2016
6. D. Saravanan, "Effective Video Data Retrieval using Image Key Frame Selection," in *Proceedings of International Conference on Computational Intelligence and Informatics* (ICCII), Hyderabad, India, May 2016
7. N. E. O'Connor, S. Marlow, N. Murphy, A. F. Smeaton, P. Browne, S. Deasy, H. Lee, and K. McDonald, "Fischlar: An Online System for Indexing and Browsing Broadcast Television Content," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, May 2001

8. X. Chen, K. Jia, and Z. Deng, "A Video Retrieval Algorithm based on Spatio-temporal Feature Curves and Key Frames," in *Proceedings of Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 287-290, Three Gorges, China, December 2008
9. S. Memar, L. S. Affendey, N. Mustapha, S. C. Doraisamy, and M. Ektefa, "An Integrated Semantic-based Approach in Concept Based Video Retrieval," *Multimedia Tools and Applications*, Vol.64, No. 1, pp. 77-95, May 2013
10. I. Bartolini, M. Patella, and C. Romani, "SHIATSU: Tagging and Retrieving Videos Without Worries," *Multimedia Tools and Applications*, Vol. 63, No. 2, pp. 357-385, March 2013
11. R. Aly, D. Hiemstra, A. D. Vries, and F. D. Jong, "A Probabilistic Ranking Framework using Unobservable Binary Events for Video Search," in *Proceedings of International Conference on Content-based Image and Video Retrieval*, pp. 349-358, Niagara Falls, Canada, July 2008
12. J. W. Jeong, H. K. Hong, and D. H. Lee, "Ontology-based Automatic Video Annotation Technique in Smart TV Environment," *IEEE Transactions on Consumer Electronics*, Vol. 57, No. 4, pp. 1830-1836, 2011
13. M. S. Zarchi, A. Monadjemi, and K. Jamshidi, "A Concept-based Model for Image Retrieval Systems," *Computers and Electrical Engineering*, Vol. 46, pp. 303-313, August 2015
14. P. Toharia, O. D. Robles, A. F. Smeaton, and A. Rodriguez, "Measuring the Influence of Concept Detection on Video Retrieval," in *Proceedings of International Conference on Computer Analysis of Images and Patterns*, Munster, Germany, September 2009
15. H. Lee, S. Park, and J. H. Yoo, "A Data Cube Model for Surveillance Video Indexing and Retrieval," in *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications*, Reykjavik, Iceland, July 2013
16. J. Baber, N. Afzulpurkar, and S. Satoh, "A Framework for Video Segmentation using Global and Local Features," *International Journal of Pattern Recognition and Artificial intelligence*, Vol. 27, No. 5, August 2013
17. J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo, "Evaluating Bag-of-visual-words Representations in Scene Classification," in *Proceedings of the International Workshop on Multimedia Information Retrieval*, MIR '07, pp. 197-206, ACM, New York, NY, USA, 2007
18. Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann, "Representations of Keypoint-based Semantic Concept Detection: A Comprehensive Study," *IEEE Transactions on Multimedia*, Vol. 12, No. 1, pp. 42-53, January 2010
19. V. T. Chasanis, A. C. Likas, and N. P. Galatsanos, "Scene Detection in Videos using Shot Clustering and Sequence Alignment," *IEEE Transactions on Multimedia*, Vol. 11, No. 1, pp. 89-100, January 2009
20. Y. X. Fang and J. H. Wang, "Selection of the Number of Clusters Via the Bootstrap Method," *Computational Statistics and Data Analysis*, Vol. 56, No. 3, pp. 468-477, March 2012
21. C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," (http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001)