Research Article

Audio-Visual-Based Query by Example Video Retrieval

Sujuan Hou^{1,2} and Shangbo Zhou^{1,2}

¹ College of Computer Science, Chongqing University, Chongqing 400030, China

² Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400030, China

Correspondence should be addressed to Shangbo Zhou; shbzhou@cqu.edu.cn

Received 3 March 2013; Accepted 23 April 2013

Academic Editor: Hua Li

Copyright © 2013 S. Hou and S. Zhou. 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.

Query by example video retrieval aims at automatic retrieval of video samples which are similar to a user-provided example from video database. Considering that much of prior work on video analysis support retrieval using only visual features, in this paper, a two-step method for query by example is proposed, in which both audio and visual features are used. In the proposed method, a set of audio and visual features are, respectively, extracted from the shot level and key frame level. Among these features, audio features are employed to rough retrieval, while visual features are applied to refine retrieval. The experimental results demonstrate the good performance of the proposed approach.

1. Introduction

While query by example image retrieval is becoming a mature field [1–5], the query by example video retrieval (QEV) is still in its infancy. Video retrieval in particular has received much attention in the area of digital video processing. For videos, the management of ever growing databases is time consuming when done completely manually. This is why automatic systems are required to lighten the job. Usually, retrieving a video from a big video database requires knowledge on the players/actors, team/director name, or other textual information. These textual queries are efficient for the cases when users have quite precise knowledge on what they are searching for within manually indexed and structured databases. However, this classical approach cannot work when users need to search video clips in highly unstructured database such as the web or to discover video clips without information on players/actors or team/director name. Now a widely accepted query type for video is query by example [6]. Query by example video retrieval aims at automatic retrieving of samples from a database, which are similar to the example provided by the user.

As discussed by Hu et al. [7], there are two widely accepted query types for video retrieval: nonsemantic-based video query types and semantic-based video query types.

Nonsemantic video query types include query by example [8], by sketch [9] and by objects [10]. Semantic-based query types include query by key words and by natural language [11, 12].

An example video may be a video clip, composed of a set of shots describing a particular event. Although people often treat a video as a sequence of images, it is actually a compound medium, integrating diverse media such as realistic images, graphics, text, and audio. Multimedia data are represented by features from multiple media sources. Among these media, audio is a significant part for videos. In fact, some audio information in video plays an important role in video detection. For instance, the audio signal from sports video is capable of characterizing some short duration events, such as a goal or audience applause. Although it has attracted many attentions from the researchers, most of them focus on applying audio features in event detection or copy detection [13]. Little research has been conducted in applying audio signal to video retrieval. In this sense, query by example video retrieval can be a compromise solution which can fill the gap between low-level features and human perception. The search for videos that are similar to a given query example, from the point view of their content, has become a very important part of research.

In this paper, a two-step retrieval framework for QEV applying audio-visual features is presented to facilitate video

retrieval and access. We focus on the features from shot level and key frame level, respectively. For shot level, audio feature analysis is firstly conducted on each clip for rough retrieval; after that the scope of objective videos in database is narrowed. Then the fine similar measurement is conducted by applying visual features at key frame level. The system finally returns the similar videos according to the visual features. Based on our approach, existing video processing techniques and algorithms can be integrated to find similar videos. We use sports videos and other social videos as our test bed. Our performance study on video datasets indicates that the proposed approach offers good performance.

The rest part of this paper is organized as follows. Section 2 describes the two-step retrieval system architecture using audio and visual features, the specific extracted audio and visual features in system are elaborated in Section 3, and Section 4 proposes the similarity measure. Section 5 presents the experimental settings and performance evaluation. Section 6 concludes this paper.

2. Two-Step Retrieval and Management Architecture

The proposed two-step retrieval and management scheme is shown in Figure 1.

There are two stages in the proposed video retrieval scheme: offline database archiving and online video retrieval.

During the offline database archiving stage, in order to support shot-based video retrieval, video preprocessing is essential: segment video using shot cut technique [14] and extract audio-visual features. Audio features are stored in database for rough retrieval while visual features from key frames are used for fine retrieval.

The video preprocessing procedures for the video retrieval stage are the same as those for the database archiving stage. After feature extraction, a two-step retrieval method is conducted: the first step is rough retrieval based on audio features, the aim is to narrow the scope of target videos by excluding irrelevant videos; the second step is to refine retrieval, which is to confirm the final result videos using similarity measure method discussed in Section 4. Lastly, the system returns the result videos.

3. Features

When designing a QEVR system, we pay attention to both audio and visual features. With respect to the fusion of them, as mentioned in the literature [15], there are two types of approaches: some audio-video feature fusion approaches try to directly evaluate the interaction of the two features [16]; other methods map first the features into a subspace where this relationship is enhanced and can therefore be estimated. In this paper, we adopt the latter: the audio features are firstly mapped into a feature subspace for rough retrieval; after that the retrieval results are optimized by visual features.

3.1. Audio Features at Shot Level. There are many features that can be used to characterize audio signals. Generally they can

be separated into two categories: time domain and frequency domain. To reduce the sensitivity to noisy for single audio feature, several audio features are employed to make sure they are complementary.

To map the audio features into a feature subspace, the feature vector representation of audio clips can be obtained in a three-step way.

Step 1 (meta-feature extraction). For each frame, six features are calculated: energy entropy, short time energy, zero crossing rate, spectral roll-off, spectral centroid, and spectral flux. This step leads to six feature sequences for the whole audio signal. above basic audio feature are described as follows.

(a) Energy entropy: it reflects the energy in the time-frequency domain volatility and distribution. We define probability density function (PDF) as p_i [17],

$$p_i = \frac{\lambda_i}{\sum_{i=1}^M \lambda_i}, \quad i = 1, 2, \dots, M, \quad (1)$$

where λ_i is amplitude of the i th frame after fast Fourier Transform (FFT), M is the length of audio frame. Then, for the i th frame, the entropy can be described as H_i ,

$$H_i = -\sum_{k=1}^M p_k \log p_k, \quad i = 1, 2, \dots, M. \quad (2)$$

(b) Short-time energy: it is one of the short-time analysis functions reflecting slowly changing of audio signal over time. The short-time energy of the i th frame can be defined as E_i [18]:

$$E_i = \sum_{m=1}^N x_i^2, \quad (3)$$

where N is the length of the i th frame.

(c) Zero crossing rate (ZCR): it is a correlate of the spectral centroid. It is defined as the number of time-domain zero crossings within the processing window and can distinguish devoiced and voiced. The ZCR is high at high frequency band while relatively low at low frequency band. The ZCR of speech signal $x_n(m)$ can be defined as follows [18]:

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} |\text{sgn}[x_n(m)] - \text{sgn}[x_n(m-1)]| \quad (4)$$

$$\text{sgn}[x] = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases}$$

(d) Spectral features: they characterize a signal's distribution of energy and frequency, which are calculated based on the short-time Fourier transform and performed frame by frame along the time axis. Three features are employed.

(1) Spectral centroid: this measure is obtained by evaluating the "center of gravity" using the Fourier transform's frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency

FIGURE 1: Block diagram of the proposed two-step method for QEVR.

weighted by amplitudes, divided by the sum of the amplitudes, or

$$\text{Spectral centroid} = \frac{\sum_{k=1}^N kF[k]}{\sum_{k=1}^N F[k]}, \quad (5)$$

here, k is the amplitude of bin number, $F[k]$ is the center frequency of that bin.

(2) Spectral flux [19]: it is a feature that measures the degree of variation in the spectrum across time. The 2-norm of the frame-to-frame spectral amplitude difference vector is $\| |X_i| - |X_{i-1}| \|$.

(3) Spectral roll-off [19]: this measure distinguishes voiced from unvoiced speech. Unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, where most of the energy for unvoiced signal is contained in lower bands. This is a measure of the “skewness” of the spectral shape—the value is higher for right-skewed distributions.

Step 2 (statistical feature). In order to achieve computational simplicity and detection effectiveness, for each of the six feature sequences, we use the statistical characteristic value of above features to quantify them, that is: energy entropy standard deviation (Std), signal energy Std by mean (average)

ratio, zero crossing rate Std, spectral roll-off Std, spectral centroid Std, and spectral flux Std by mean ratio. This step leads to six single statistic values (one for each feature sequence). Those six values are the final feature values that characterize the audio clip using a six-dimensional feature vector, which is denoted as $A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$.

Step 3 (normalized statistical feature). Normalization can ensure that contributions of all audio feature elements are adequately represented, preventing one feature from dominating the whole feature vector.

In this paper, each audio feature is computed over each shot audio clip (30 ms, duration). For $A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$, suppose $a_{\max}$ and $a_{\min}$ are the maximum and minimum value of feature components, respectively, which can be defined as follows:

$$a_{\max} = \max_{1 \leq i \leq 6} |a_i|, \quad a_{\min} = \min_{1 \leq i \leq 6} |a_i|. \quad (6)$$

A normalized audio feature vector A_N consisting of six components is constructed, denoted as (7).

$$A_N = \left\{ \frac{a_1 - a_{\min}}{a_{\max} - a_{\min}}, \frac{a_2 - a_{\min}}{a_{\max} - a_{\min}}, \dots, \frac{a_6 - a_{\min}}{a_{\max} - a_{\min}} \right\}. \quad (7)$$

FIGURE 2: Audio segmentation scheme.

3.2. Visual Features at Key Frame Level. We segment video clips into shots and extract representative frames from each shot. The key frames of a shot reflect the characteristics of the video to some extent. Traditional image retrieval techniques can be applied to key frames to achieve video retrieval. In the field of image index, image representation became color oriented, since most of the images of interests are in colors. Many of the previous researches used the color composition of an image [20]. The advantages of color histogram are that it is invariant for translation and rotation of the viewing axis. However, using a single image attribute for retrieval may lack sufficient discriminatory information, such as color feature cannot convey the space information of image. Texture feature describes the spatial correlation between pixels, compensating for this shortcoming.

For the color feature, we defined it in terms of a histogram in the quantized hue-saturation-value (HSV) color space, since the HSV space conforms more to human perceptual similarity of colors, this property is used in quantizing the color space into a small number of colors. The color basic properties are emphasized and then the color histogram bins are chosen as 180. For the texture, we defined it in terms of GLCM (gray level cooccurrence matrix) [21, 22]. Here, we adopted eight indicators extracted from GLCM, namely, the mean and standard deviation of energy, entropy, and contrast and inverse difference moment. Energy is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form. Entropy measures the disorder of an image and it achieves its largest value when all elements are in GLCM. Contrast is a difference moment of the GLCM, and it measures the amount of local variations in an image. Inverse difference moment measures image homogeneity. This parameter achieves its largest value when most of the occurrences in GLCM are concentrated near the main diagonal. Users can choose other feature representation according to their own requirements.

Shot S is represented as a weighted matrix by the following:

$$V_s = [w_c \times V_c \quad w_t \times V_t] \quad (8)$$

$$\begin{aligned} V_c &= [c_1, c_2, \dots, c_{180}] \\ V_t &= [t_1, t_2, \dots, t_8], \end{aligned} \quad (9)$$

where w_c and w_t represent the weight of color, and texture, respectively, when describing a representative frame of shot S . Considering that cognition on shot can be quite subjective therefore, different weight assignments may reflect different

user requirements and preferences. The symbols V_c and V_t represent normalized feature vector of color and texture, respectively, t_i ($1 \leq i \leq 180$) and t_j ($1 \leq j \leq 8$) represent the components of HSV histogram and GLCM, respectively. The Normalization method of visual feature vector is the same as that of audio feature vector as elaborated in Section 3.1.

4. Similarity Measure

The overall similarity matching contains two levels: rough retrieval based on audio features and refined retrieval based on visual features.

4.1. Rough Retrieval Based on Audio Feature. Firstly, the audio signals were segmented at 30 ms/frame with Hamming window, which is the basic unit for feature extraction as shown in Figure 2.

And then, the audio features elaborated in Section 3.1 are used to construct feature vector for audio frame. The audio samples were collected with 44.1 kHz sample rate, stereo channels, and 16 bits per sample.

Feature analysis is conducted on each clip. And the aim of rough retrieval is to narrow the scope of objective videos in database utilizing the audio feature vector mentioned above.

A normalized Euclidean distance $d(a_i, a_j)$ is used to measure the similarity between audio clips from query video shot i and database video shot j . a_i and a_j are the normalized feature vectors depicting the characteristics of audio clip i and j .

$$d(a_i, a_j) = \sqrt{\sum_{k=1}^6 (a_{ik} - a_{jk})^2}. \quad (10)$$

We say i and j are dissimilar if and only if $d(a_i, a_j) > \varepsilon_{\text{audio}}$ (a distance threshold). And the greater d , the greater dissimilarity between a_i and a_j .

4.2. Refined Retrieval Based on Visual Features. The scope of objective videos is narrowed after rough retrieval. In this section, we will address how the temporal similarity of two videos can be exactly measured properly.

We use key frames to represent a shot in video, and each key frame is represented by (8) using the visual features elaborated in Section 3.2. Then, a large number of key frames are extracted from query video as the baseline for comparison. Alternative temporal sampling is also possible, while our current comparison is in key frame features only.

The similarity between the shots in above subclassification and the query video shot can be estimated with the Euclidean distance. Firstly, let us denote shots set of above subclassification and query video by $S_c = \{s_1, s_2, \dots, s_i, \dots, s_N\}$ and s_q , respectively. The shots similar to query video s_q are defined as (11).

$$\text{SimS}(s_q) = \{s_i \mid d(s_i, s_q) \leq \varepsilon_{\text{visual}}, 1 \leq i \leq N\}, \quad (11)$$

where $d(s_i, s_q)$ is the Euclidean distance function to measure the similarity of s_i and s_q . The parameter N represents

FIGURE 3: Shot matching between query video and database video belongs to subclassification.

the number of shots in S . We say s_i and s_q are similar if $(d(s_i, s_q) < \epsilon_{\text{visual}}, s_q)$, where ϵ_{visual} is a small positive value called shots similarity threshold.

Now, considering that many of current videos are the results of synthesis and editing in the later stage, let's denote the query video by $V_q = \{s_1, s_2, \dots, s_q, \dots, s_L\}$ and database video belonging to above sub-classification by $V_d = \{s_1^1, s_1^2, \dots, s_v^m, \dots\}$, where L is the number of shots in query video, m and v represent serial number of shot and video, respectively. The shot matching result between V_q and V_d can be seen as in Figure 3.

Based on the above observations, the database video which is similar to query video V_q is defined by (12).

$$\text{SimV}(V_q) = \left\{ V_i \mid i = \arg \max_x \left\{ \sum_x \sum_{j=1}^m \sum_{q=1}^l 1 \{ \text{SimS}(s_x^j) \} \right\} \right\}, \quad (12)$$

where $\text{SimS}(s)$ denotes the shot which is similar to shot s .

5. Experimental Results/Case Study

5.1. Evaluation Criteria for Experiment. For analyzing the effectiveness of our proposed approach, two major criteria in retrieval system, namely precision and recall, are used to measure the related experimental evaluations. The precision and recall are defined as follows [5]:

$$\begin{aligned} \text{precision} &= \frac{|\text{correct}|}{|\text{retrieved}|} \times 100\%, \\ \text{recall} &= \frac{|\text{correct}|}{|\text{relevant}|} \times 100\%, \end{aligned} \quad (13)$$

where correct is the number of relationally retrieved videos, retrieved is the number of all retrieved videos by the proposed approach, and relevant is the ground truth representing the number of all relational videos in the database. The criterion precision delivers the ability for hunting the desired videos in user's mind and the recall represents the ability for finding the accumulated positive videos in a query session.

5.2. Experimental Results

5.2.1. Parameter Settings. Before evaluating the proposed approach, the appropriate parameter settings need to be

TABLE 1: Precision performance on four types of video by varying ϵ_{audio} .

Precision	$\epsilon_{\text{audio}} = 0.2$	0.4	0.5	0.6	0.8
Badminton	23.41%	22.4%	22.4%	22.4%	22.4%
Wrestling	7.89%	6.38%	8.51%	8.51%	8.51%
Social	41.67%	21.28%	21.28%	21.28%	21.28%
Basketball	32.26%	21.28%	21.28%	21.28%	21.28%

TABLE 2: Recall performance on four types of video by varying ϵ_{audio} .

Recall	$\epsilon_{\text{audio}} = 0.2$	0.4	0.5	0.6	0.8
Badminton	85.84%	85.84%	85.84%	85.84%	85.84%
Wrestling	75%	75%	100%	100%	100%
Social	50%	100%	100%	100%	100%
Basketball	67	100%	100%	100%	100%

elicited. During our related experiments, two group parameters need to be assigned: the first one is the distances threshold, denoted as ϵ_{audio} and ϵ_{visual} in (10) and (11), which measure the similarity between query video and database videos from the perspective of audio features and visual features, respectively.

The second group is the feature weight. For audio feature vector described in Section 3.1, consider that audio features are just used for filtering some irrelevant candidate videos and to keep the balance between them, we equally treat the weight of six audio features during the stage of rough retrieval. In the same way, in addition, 0.5 is adopted as our default setting both for w_c and w_t .

The appropriate parameters setting about distance threshold ϵ_{audio} was decided by experiments on four types of videos, which is shown as in Tables 1 and 2 as follows.

From the above two tables, we can see that the best setting for ϵ_{audio} is 0.5, where the scope of target video begins to stabilize. It probably leave out some relevant videos if the value of ϵ_{audio} is too small while increase the amount of videos need computed in the second step if the value is too big.

Applying the similar approach, we find that the appropriate parameters setting for ϵ_{visual} is also 0.5.

5.2.2. Experiments. The proposed approach is tested on real-world experimental data that was acquired from the collection of videos containing sport videos and other social science videos. In our experiment, the sport videos are composed of 830 shots while social videos composed of 227 shots. Three kinds of sport videos are included: badminton, wrestling and basketball, respectively.

In order to specify the retrieval procedure based on proposed two-step retrieval mechanism, a schematic example is given in Figures 4 and 5 as follows.

In online retrieval, the query video including 27 shots is introduced to the system, and then both audio and visual features are computed for each shot and key frames respectively. The two-step retrieval procedure is implemented subsequently: the first step is rough retrieval based on audio features, for 61 raw videos with 1057 shots in database,

FIGURE 4: (a) Part sequences of query video. (b) Part raw target video sequences in database.

TABLE 3: The comparison with experimental results with audio-visual based and visual based.

	NS	Retrieval results with audio-visual based		Retrieval results with visual based	
		Precision	Recall	Precision	Recall
Badminton	20	80.52%	88.61%	62.02%	58.64%
Wrestling	21	57.70%	100%	48.39%	100%
Social	28	100%	100%	87.5%	100%
Basketball	41	99.18%	67.31%	72.47%	50.25%
Average	—	84.35%	88.98%	67.595%	77.223%

and after rough retrieval there are 18 target videos that are retained; the second step is to refine retrieval based on visual features, after which 10 target videos are reversed as the final target videos.

Figure 4 represents part sequences of query video and part videos in database, where part (a) shows the query video sequences while part (b) lists raw target videos in database.

Figure 5 demonstrates the retrieval results, in which part (a) is rough retrieval based on audio features and part (b) is refined retrieval with visual features.

To evaluate the performance of proposed two-step retrieval mechanism in retrieving video in terms of effectiveness and efficiency, we have identified four groups of experiments (badminton, wrestling, basketball, and social science, resp.).

Table 3 shows the comparison results between proposed audio-visual-based retrieval method and audio-based method, in which NS denotes the number of shots in query video.

We also consider two other kinds of situations: one is retrieval with rough retrieval with visual features while refine retrieval with audio features, and the other is retrieval with only audio features. The results are shown in Table 4.

From Table 4 we can find that generally, audio features work worse than visual-based features. However, these two kinds of features are very complementary. When we just simply combined audio and visual features, the average precision and recall can be improved by 16.755% and 11.757%, respectively compared with only visual based, improved by 66.135% and 20.2%, respectively, compared with only audio based.

FIGURE 5: (a) Rough retrieval based on audio features. (b) Refine retrieval based on visual features.

TABLE 4: The comparison with experimental results with visual-audio and only audio features.

	NS	Retrieval results with visual-audio based		Retrieval results with audio based	
		Precision	Recall	Precision	Recall
Badminton	20	30%	90%	22.41%	85.84%
Wrestling	21	11.11%	25%	7.89%	75%
Social	28	66.67%	20%	21.28%	14.29
Basketball	41	31.25%	100%	21.28%	100%
Average	—	34.78%	58.75%	18.215%	68.78%

TABLE 5: Comparison between this paper's approach and Kong's approach.

	This paper's approach		Kong's approach [12]	
	Precision	Recall	Precision	Recall
Average	84.35%	88.98%	62.025%	85.325%

We also compare this paper's approach with Kong's approach [12] in Table 5, which also support the shot-based retrieval from the perspective of precision and recall.

In summary, we have achieved retrieval results with the proposed two-step hierarchical retrieval mechanism base on audio-visual feature, together with the visual-based versus audio-based versus visual-audio-based versus retrieval method in other literature. The comparative results indicates that the proposed method offers better performance.

6. Conclusion

Considering that audio is a significant part of videos, we present a novel and efficient method jointing audio features for video retrieval query by example. In our system, the audio features are firstly used for rough retrieval to narrow the scope of objective videos in database. Then, the visual features are applied to refine retrieval. Finally, the system returns the similar videos which are similar to a user-provided example. Experimental results indicate that the proposed approach owns better performance in retrieving video when query by example compared with other types of retrieval mechanism.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grants nos. 61103114 and 61004112).

References

- [1] H.-C. Hsin, T.-Y. Sung, Y.-S. Shieh, and C. Cattani, "Adaptive binary arithmetic coder-based image feature and segmentation in the compressed domain," *Mathematical Problems in Engineering*, vol. 2012, Article ID 490840, 14 pages, 2012.
- [2] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Fast wavelet-based image characterization for highly adaptive image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1613–1332, 2012.
- [3] Y.-K. Chan, M.-H. Tsai, J.-S. Wang, S.-W. Guo, and J.-L. Wu, "Color-texture-based image retrieval system using Gaussian Markov random field model," *Mathematical Problems in Engineering*, vol. 2009, Article ID 410243, 17 pages, 2009.
- [4] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang, "Intentsearch: capturing user intention for one-click internet image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1342–1353, 2012.
- [5] J.-H. Su, W.-J. Huang, P. S. Yu, and V. S. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 360–372, 2011.
- [6] S. Hou, S. Zhou, and M. A. Siddique, "Query by example video based on fuzzy c-means initialized by fixed clustering center," *Optical Engineering*, vol. 51, no. 4, pp. 047405-1–04740511, 2012.
- [7] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, no. 6, pp. 797–819, 2011.
- [8] A. K. Jain, A. Vailaya, and X. Wei, "Query by video clip," *Multimedia Systems*, vol. 7, no. 5, pp. 369–384, 1999.
- [9] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.
- [10] J. Sivic and A. Zisserman, "Video Google: efficient visual search of videos," in *Toward Category-Level Object Recognition*, pp. 127–144, Springer, New York, NY, USA, 2006.
- [11] Y. Aytar, M. Shah, and J. Luo, "Utilizing semantic word similarity measures for video retrieval," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.
- [12] J. Kong and C. Han, "Content-based video retrieval system research," in *Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT '10)*, pp. 701–704, July 2010.
- [13] P. Natarajan, V. Manohar, S. Wu et al., "Bbn viser trecvid 2011 multimedia event detection system," in *Proceedings of the NIST TRECVID Workshop*, 2011.
- [14] Y. Yusoff, W. Christmas, and J. Kittler, "Video shot cut detection using adaptive thresholding," in *Proceedings of the British Machine Video Conference*, pp. 362–372, Bristol, UK, 2000.
- [15] P. Besson, V. Popovici, J. M. Vesin, J. P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 63–73, 2008.
- [16] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Image and Video Retrieval*, pp. 488–499, Springer, New York, NY, USA, 2003.
- [17] S. Jiong-ning, F. De-sheng, and X. Yong-hua, "Speech detection algorithm based on energy-entropy," *Computer Engineering and Design*, vol. 12, p. 084, 2005.
- [18] Q. Cai, S.-N. Lu, and S.-T. Yang, "Video content detection method based on audio-visual features," *Computer Engineering*, vol. 22, p. 083, 2007.
- [19] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pp. 1331–1334, April 1997.
- [20] J. Han and K.-K. Ma, "Fuzzy color histogram and its use in color image retrieval," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 944–952, 2002.
- [21] M. Partio, B. Cramariuc, M. Gabbouj, and A. Visa, "Rock texture retrieval using gray level co-occurrence matrix," in *Proceedings of the 5th Nordic Signal Processing Symposium*, October 2002.
- [22] A. K. Mikkilineni, P. J. Chiang, G. N. Ali, G. T. C. Chiu, J. P. Allebach, and E. J. Delp, "Printer identification based on graylevel co-occurrence features for security and forensic applications," in *Proceedings of the 7th SPIE-IS and T Electronic Imaging-Security, Steganography, and Watermarking of Multimedia Contents VII*, vol. 5681, pp. 430–440, January 2005.