MSR-Asia at TREC-11 Video Track

Xian-Sheng Hua [1], *Pei Yin* [2+], *Huajian Wang* [3+], *Junfeng Chen* [1], *Lie Lu* [1]
Mingjing Li [1], *Hong-Jiang Zhang* [1]
[1] Media Computing Group, Microsoft Research Asia
[2] Department of Computer Science and Technology, Tsinghua University
[3] Department of Electrical Engineering, Tsinghua University

Abstract

The Media Computing Group of Microsoft Research Asia participated in all the three tasks of Video tracks of TREC-11, including automatic Shot Boundary Determination, Semantic Feature Extraction and Video Search. A robust shot detector was proposed. Systems for semantic feature extraction and video retrieval which integrated many recent research results of this group's are presented.

1. Introduction

Media Computing Group of Microsoft Research Asia (MSRA) took part in the TREC-11 Video Track in 2002, where runs for all three tasks are submitted, including automatic Shot Boundary Determination (SBD), Semantic Feature Extraction (SFE) and Video Search (VS). The systems, algorithms and experimental results of these three tasks are presented in Section 2, Section 3 and Section 4, respectively.

2. Shot Boundary Detection

The current SBD system of MSRA, Microsoft Smart Shot Detector 2002 (MSSD02), is developed based on the MSSD01 system [1] last year. MSSD02 is specially concentrated on improving the performance of gradual transition (GT) detection. In MSSD02, the main feature for SBD is frame difference, i.e., the total difference of the bin-wise histogram comparison between two consequent frames in R, G and B channels. Generally the shot boundaries are then determined according to the approaches presented in [1][2]. To improve the detection accuracy of GT, several heuristic rules are applied to MSSD02.

In MSSD01 system, histogram comparison is performed on gray frames, while color histograms which contain more information are employed in MSSD02. And the bins of the color histogram are quantified to reduce the effects of noises caused by digitalization and luminance variations. In addition, video frames are also down-sampled to reduce the negative effects that caused by camera or object motions, and to speed up SBD at the same time. The down-sampling factor is 4 (2 for height and 2 for width). With quantization and down-sampling, the system runs three times faster than real time.

2.1. System and algorithms

2.1.1. Shot detection by comparison of frame histogram

Many previous SBD systems, such as MSSD01 [1], are developed to detect CUT and GT based on frame comparison by the bi-Threshold scheme proposed in [2]. The frame

[+] This work was performed when the authors were visiting Microsoft Research Asia as interns.

comparison can be performed in many forms, such as pixel by pixel on video frames or bin by bin on the frame histograms. In MSSD02 system, frame difference is generated by pair-wise histogram comparison.

Algorithm 1 (Shot Detection)

Denote the difference of two consecutive frames F_{i-1} and F_i as D_i, and Th_1 and Th_2 as the high threshold and low threshold for the frame difference, respectively.

(1) If $D_i > Th_1$, it indicates that there is a CUT between F_{i-1} and F_i ;

(2) If a sequence of D_i at the length of m (D_{i-m+1}, … , D_i) satisfies

$$Th_2 \le \min_{i-m+1 \le j \le i} \{D_j\} \le \max_{i-m+1 \le j \le i} \{D_j\} \le Th_1 \quad (1)$$

$$D_{i+1} \le Th_2 \quad (2)$$

and m is large enough, it indicates that there is a GT from F_{i-m+1} to F_i .

Experiments proved that most of shot boundaries are successfully detected by this algorithm. However, it is observed that frequently there are many cases which violate these rules. For instance, at the shot boundaries, frame differences vary greatly. Some values are lower than Th_2, while some values are higher than Th_1. Based on this observation, the above-mentioned algorithm is improved by the following rule.

Rule 1 (Threshold Tolerance for GT Detection)

In GT detection, some tolerance for D_i sequence falling below Th_2 should be taken into consideration. A D_i sequence can be regarded as GT candidate, if there are always at least N frame differences above Th_2 in any sub sequence of D_{i-m+1}, … , D_{i-1}, D_i. In our system, $M = 5$ and $N = 4$.

In **Algorithm 1**, high frame differences are regarded as indicators of CUT and a series of medium frame differences are regarded as GT. Actually, detection performance is improved by integrating CUT detection and GT detection with the above-mentioned rules as follow.

Algorithm 2 (Integrated CUT and GT Detection)

(1) If $D_i > Th_2$, **Rule 1** is applied to count the length of the GT candidate sequence, without considering whether it is greater than Th_1 or not.

(2) If the length of the GT candidate sequence is long enough, it is regarded as a GT; otherwise, go to (3).

(3) If there is a D_j in the sequence valued higher than Th_1, there is a CUT between F_{j-1} and F_j. Otherwise, the GT candidate sequence is regarded as a false alarm which possibly caused by object or camera motions.

2.1.2. Luminance variation

Flash detection [3] only handles part of detection errors caused by luminance variation. In

MSSD02, edge information is taken into consideration to decrease false alarms caused by luminance variation, especially for detecting shot boundaries in older films. Abrupt transitions are detected by comparing frame differences as **Algorithm 1** or **2**, and verified by edge differences between two consecutive frames. This rule (**Rule 2**) is based on the observation that luminance variation will not bring large differences on the edge maps.

Rule 2 (Luminance Variation Resistance)

Denote edge map of F_i as E_i, while E_i is the response of Sobel edge detector. ED_i, i.e., the edge difference is obtained by the following equation.

$$ED_i = \frac{\sum_{m,n} \left| E_i(m,n) - E_{i-1}(m,n) \right|}{\sum_{m,n} E_i(m,n)} \quad (3)$$

In MSSD02, an abrupt shot transition should satisfy $ED_i > 0.1$.

2.1.3. GT suppression

It is observed that intensive motions often cause false alarms. A simple rule to decrease such false alarms is to set a higher threshold for the minimum shot duration. However, a number of boundaries of shorter shots will be missed when applying this rule. Based on the observation that the intensive motions usually affect a series of frame differences, it is more likely that these false alarms are detected as GTs. Thus setting a higher threshold for the temporal distance between two GTs will decrease this type of false alarms, while avoid missing short shot boundaries to the utmost at the same time.

2.2. Experimental results of shot boundary detection

2.2.1. Experiments configuration

Testing data: SBD evaluation is performed on clips randomly selected from both Video TREC 2002 data (28 clips, 3.5GB, 6 hours) and Video TREC 2001 data (4 clips, 1.2GB, 2.2 hours). Ground-truth of shot boundaries is manually annotated by three students majored in arts by a ground-truthing tool developed by this group.

System configuration:

Table 1. System Configuration for SBD Evaluation.

Configuration \ System ID	Sys01	Sys02-1	Sys02-2	Sys02-3	Sys02-4
System	MSSD01	MSSD02	MSSD02	MSSD02	MSSD02
Minimal Shot Duration	-	5 frames	25 frames	30 frames	30 frames
GT Suppression	-	x	x	x	√

Sys01: MSSD 2001
Sys02-1: MSSD 2002 with minimal shot duration of 5 frames
Sys02-2: MSSD 2002 with minimal shot duration of 25 frames
Sys02-3: MSSD 2002 with minimal shot duration of 30 frames
Sys02-4: MSSD 2002 with minimal shot duration of 30 frames plus “GT suppression” rule

Evaluation protocol

C1: VideoTREC02 standard evaluation scheme.

C2: C2 is derived from C1, while the only difference is that C2 also extends GT bi-directionally by 5 frames, which is more reasonable when there are many short gradual transitions.

2.2.2. SBD evaluation results

Table 2. Evaluation Results on 2002 Data Using Protocol C1.

2002 Data	Sys01	Sys02-1	Sys02-2	Sys02-3	Sys02-4
SB Detected	0.519	0.812	0.773	0.764	0.764
SB Inserted	0.773	0.633	0.410	0.379	0.314
SB Deleted	0.481	0.188	0.227	0.236	0.236
GT Precision	0.567	0.747	0.737	0.738	0.74
GT Recall	0.553	0.900	0.900	0.895	0.899

Table 3. Evaluation Results on 2001 Data Using Protocol C1.

2001 Data	Sys01	Sys02-1	Sys02-2	Sys02-3	Sys02-4
SB Detect	0.660	0.683	0.675	0.666	0.666
SB Insert	0.332	0.419	0.355	0.329	0.317
SB Delete	0.340	0.317	0.325	0.334	0.334
GT Precision	0.751	0.618	0.646	0.630	0.634
GT Recall	0.690	0.833	0.822	0.841	0.850

Table 4. Evaluation Results on 2002 Data Using Protocol C2.

2002 Data	Sys01	Sys02-1	Sys02-2	Sys02-3	Sys02-4
SB Detect	0.599	0.823	0.786	0.777	0.777
SB Insert	0.693	0.623	0.397	0.367	0.300
SB Delete	0.401	0.177	0.214	0.223	0.223
GT Precision	0.654	0.757	0.743	0.745	0.746
GT Recall	0.551	0.908	0.907	0.904	0.907

Table 5. Evaluation Results on 2001 Data Using Protocol C2.

2001 Data	Sys01	Sys02-1	Sys02-2	Sys02-3	Sys02-4
SB Detect	0.879	0.917	0.909	0.898	0.898
SB Insert	0.112	0.114	0.121	0.096	0.085
SB Delete	0.121	0.083	0.091	0.101	0.102
GT Precision	0.778	0.661	0.678	0.669	0.658
GT Recall	0.748	0.908	0.884	0.908	0.910

3. Feature Extraction

Semantic shot features are extracted from each shot based on the “common shot boundaries”.

3.1. Visual features

Firstly, multiple key-frames are extracted from each shot. The number of the key-frames in one shot depends on the duration of the shot and the histogram variation within the shot. All

visual features, including indoor, outdoor, cityscape, landscape, face and text overlay are extracted on these key-frames.

Indoor/Outdoor/Cityscape/Landscape

Each key-frame is segmented into 5×5 macro blocks. Color Moment (CM) and Edge Direction Histogram (EDH) are extracted from each block [4]. According to these features, the key-frames are then classified into indoor/outdoor classes and cityscape/landscape classes by an AdaBoost-based classifier. Different trained models are employed for classifying indoor/outdoor and cityscape/landscape. Final outputs of these four features are determined by linearly combining the classification confidences over all the key-frames of each shot.

Face

Face areas are detected in the shot key-frames by the approach based on DAM (Direct Appearance Model) [5] and FloatBoost [6]. Face Ratio, defined by the following equation, is regarded as the possibility that a shot contains face.

$$R = \frac{F}{N} \tag{1}$$

where F is the number of the key-frames in which at least one face is detected, and N is the total number of the key frames in the corresponding shot. This detector is tested on 749 images, which are got from the search test set, and precision of 60.2% is obtained.

Text overlay

The text detector is similar to the approach described in paper [7], but improved by applying the heuristic rules proposed in [8] and [9], which help to get more accurate text bounding boxes and decrease false alarms. Then we verify the detection result by Optical Character Recognition.

3.2. Audio features

The following basic features are employed to obtain semantic audio features, including High Zero-Crossing Rate Ratio (HZCRR), Low Short-Time Energy Ratio (LSTER), Spectrum Flux (SF), Linear Spectral Pairs (LSP) distance, Band Periodicity (BP) and Noise Frame Ratio (NFR). Based on these features, a SVM-based classifier is applied to classify the audio stream into four classes, which include speech, music, environmental sound and silence [10].

Dialog/Monolog classification is based on speaker change detection [11]. In the shots that contain speech, if speaker change is detected, then the shot is classified as dialog, otherwise it is monolog.

4. Retrieval

4.1. System overview

MSRA video retrieval system, Q-Video, is consisted of the following components: shot boundary detection module (SBD), feature extraction module (FE), video management module (VM), query construction module (QS) and feedback analyzer module (FA), as illustrated by **Figure 1** (SBD is integrated in *Feature Extractor*, while VM is integrated in *Feature Database*).

SBD module is the same as the system presented in Section 2, but the actual shot boundaries used in Q-Video are the "common shot boundaries" due to the requirements of TREC-11 Video Track. The search task is also based on these common shot boundaries.

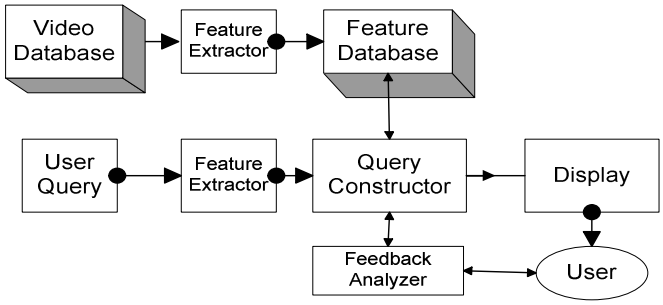

Figure 1. MSRA Video Retrieval System Overview.

4.2. Feature extraction module:

Except for the semantic features introduced in Section 3, nine other low-level feature sets are used in Q-Video, including Color Moment (CM), Dominant Color (DC) [12], HSV Histogram (HSVH), Color Layout (CL), Edge Histogram (EH), Color Texture Moment (CTM), Kirsh Direction Density (KDD), Wavelet Feature (WF) and Motion Texture (MT) [13]. Different distance metrics are employed for different feature sets. The first eight features are extracted from shot key-frames, while Motion Texture is extracted from the motion vector field of each entire shot.

4.3. Video management module:

All the extracted features are stored in a feature database and indexed by video ID and shot ID. The management module allows users browse the video freely and submit query request. **Figure 2** illustrates the interface of the system.

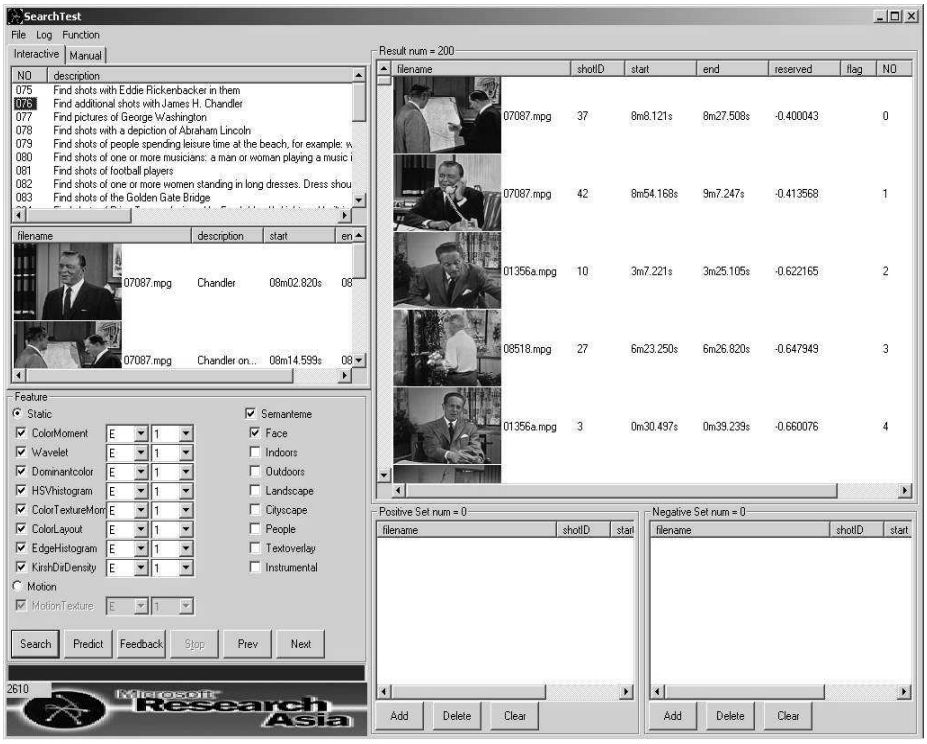

Figure 2. Interface of Q-Video.

4.4. Query construction module:

The search task of TREC-11 Video Track competition is QBE (Query by Example) based video shot retrieval. For each search topic, one or more video clips or images representing the content of that topic are provided by NIST. The Q-Video system analyzes these data and extracts the all the features from them.

Nine similarities are obtained by calculating similarity on the nine different feature sets. The overall similarity is a linear combination of the nine similarities, wherein the weights are adjusted manually according to the retrieval topics by the users.

The semantic features are applied as filters to filtrate particular shots which do not meet the requirements of the retrieval tasks. For example, Topic 79 is to find shots in which there are people spending leisure time at the beach. The shots which are relative to this topic must have the property of "Outdoor". Whether using these filters or not are determined manually according to the retrieval topics.

4.5. Feedback analyzer module

A SVM-based learning procedure is applied in the interactive Q-Video system. Firstly, for each topic, retrieval result consists of a series of video shots ranked by their similarity to the retrieval topic is obtained by the manual retrieval system with default parameters (so no manual efforts involved indeed). Then the users browse these shots and label the ones fitting the topic and also the ones not relative to the topic at all. The labeled fitting shots are regarded as positive feedback, while the ones labeled as irrelative are considered negative feedback. These feedbacks are the inputs of the SVM-based learning procedure. This process will stop when the users get satisfactory retrieval result.

4.6. Search with the result of ASR:

ASR results are also integrated in the Q-Video system. Searching with ASR is implemented by matching key words which are extracted from the ASR output.

5. Acknowledgement

The authors would like to thank Lei Zhang, Yu-Fei Ma, Xinli Zou and Dong Zhang for providing parts of system components.

6. References

[1] Yu-Fei Ma, Jia Sheng, Yuan Chen, Hong-Jiang Zhang, "MSR-Asia at TREC-10 Video Track: Shot Boundary Detection Task", *Video TREC 2001*.

[2] HongJiang Zhang, Atreyi Kankanhalli, Stephen W. Smoliar, "Automatic Partition of Full Motion Video", *Multimedia System1*, pp.10-28, 1993.

[3] Dong Zhang, Wei Qi, Hong-Jiang Zhang, "A New Shot Detection Algorithm," *2nd IEEE Pacific-Rim Conf on Multimedia (PCM2001)*, pp. 63-70, Beijing, China, October 2001.

[4] Lei Zhang, Mingjing Li, Hong-Jiang Zhang, "Boosting Image Orientation Detection with Indoor vs. Outdoor Classification", *IEEE Workshop on Applications of Computer Vision*, December 3-4, 2002, The Orlando World Center Marriott, Orlando, FL USA.

[5] Stan Z. Li, S.C. Yan, H.J. Zhang, Q.S. Cheng, "Multi-View Face Alignment Using Direct Appearance Models". *In Proceedings of The 5th International Conference on Automatic Face and Gesture Recognition*. Washington, DC, USA. 20-21 May, 2002.

[6] Stan Z. Li, Z.Q. Zhang, Harry Shum, H.J. Zhang, "FloatBoost Learning for Classification". In

Proceedings of The 16-th Annual Conference on Neural Information Processing Systems (*NIPS*), Vancouver, Canada, December 9-14, 2002.

[7] Xiang-Rong Chen, Hong-Jiang Zhang, "Text Area Detection from Video Frames," *2nd IEEE Pacific-Rim Con. on Multimedia (PCM2001)*, pp. 222-228, Beijing, China, October, 2001.

[8] Xian-Sheng Hua, et al., "Automatic Location of Text in Video Frames," *Proceeding of ACM Multimedia 2001 Workshops: Multimedia Information Retrieval (MIR2001)*, pp. 24-27, Ottawa, Canada, October 5, 2001.

[9] Xian-Sheng Hua, Pei Yin, Hong-Jiang Zhang, "Efficient Video Text Recognition Using Multiple Frame Integration," *2002 International Conference on Image Processing (ICIP2002)*, Rochester, New York, Sept 22-25, 2002.

[10] Lie Lu, Stan Z. Li, Hong-Jiang Zhang, "Content-Based Audio Segmentation Using Support Vector Machines," *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 956-959, 2001.

[11] Lie Lu, Hong-Jiang Zhang, "Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis", *Accepted by 10th ACM International Conference on Multimedia 2002.*

[12] Tong Ling and Hong-Jiang Zhang, "Integrating Color And Spatial Features for Content-Based Video Retrieval," *Invited Paper, Proceedings of 2001 International Conference on Image Processing*, Thessaloniki, Greece. October 7-10, 2001.

[13] Yu-Fei Ma, Hong-Jiang Zhang, "Motion Pattern based Video Classification using Support Vector Machines", *2002 IEEE International Symposium on Circuits and Systems, Theme: Circuits and Systems for Ubiquitous Computing(ISCAS2002)* , Scottsdale, Arizona, USA, May 2002.