Tsinghua University at TRECVID 2004:
Shot Boundary Detection and High-level Feature Extraction

Jinhui Yuan, Wujie Zheng, Le Chen, Dayong Ding, Dong Wang, Zijian Tong, Huiyi Wang, Jun Wu
Jianmin Li, Fuzong Lin, Bo Zhang
State Key Laboratory of Intelligent Technology and System
Department of Computer Science and Technology Tsinghua University
Beijing 100084, P. R. China

Abstract

Our shot boundary detection system mainly consists of three components: fade out/in(FOI)detector, cut(CUT)detector and gradual transition(GT)detector. The key technique of FOI detector is the recognition of monochrome frame. In CUT detector, second order derivative method is applied to get cut candidates,and then flashlight detector and other gradual transition filter module are incorporated to eliminate false positives. In GT detector, short GT and long GT are treated separately, more specifically,twin comparison algorithm for short GT and finite state automaton (FSA) model for long GT. Ten runs are submitted. And the evaluation result shows that our system is among the best. The characteristics of the runs are summarized as follows.

thuai02	Baseline system with default architecture and default parameter settings.
thuai05	Second order derivative scheme for short GT.
thuai07	Looser conditions for short GT detector.
thuai08	Without gradual transition filter module.
thuai10	Prolong the duration of post-cut modules,including flashlight detector and gradual transition detector.
thuai14	Baseline system Derived from thuai10 with higher coefficients for motion based self-adaptive threshold.
thuai15	Even higher coefficients for motion based self-adaptive threshold.
thuai16	Increase the higher threshold of FSA model.
thuai17	Further heighten the higher threshold of FSA model.
thuai19	Replace the second order derivative of CUT detector with first order derivative.

In Feature Extraction task, the visual modal detector forms our baseline system. And several kinds of other detectors based on text and timing cue are added. The system is applied to three concepts[1], namely: “Basket scored”, “Bill Clinton” and “Beach”. For “Basket scored”, the visual modal detector achieves 0.561 in AP. And the added text and timing information brings 20% increase in AP and arrives the best run for that concept. The submitted runs of “Basket scored” are:

B_thuai_10	Baseline system. AP-based Borda fusion of results generated by different visual models.
B_thuai_5	Baseline + commercial filter
B_thuai_1	Baseline + commercial filter + SRI operation.
B_thuai_7	Baseline + commercial filter + SRI operation + shot cluster operation.

1. Introduction

This year we participate in two tasks of shot boundary detection (SBD) and feature extraction (FE).

Various automatic shot boundary detection algorithms have been proposed(see [7],[11],[13]). Most of them seem to fail under roughly the same conditions such as flashlight or object/camera movement. To boost the performance, we design and integrate specific modules to eliminate various disturbance. The evaluation shows that our system exhibits excellent performance.

Compared with the traditional Content Based Image Retrieval (CBIR), Content Based Video Retrieval (CBVR) has much more information from multiple modalities and various information fusion methods to deal with the semantic gap between

[1] “Concept detection” and “feature extraction” are used interchangeably in this paper. But as the latter are often confused with the frequently used “low-level feature extraction”, the former is preferred.

the extracted low level features and high level user queries. Since FE is the preliminary steps towards CBVR, we implement different detectors based on visual, text, and time modalities and combine them for boosting the detection accuracy.

Following this brief introduction, SBD is treated in Section 2 and FE in Section 3. Conclusions and future directions are given in Section 4.

2. Shot Boundary Detection

2.1. System Overview

Most of existing algorithms are developed based on the observation that frames surrounding a boundary generally displays a significant change of the visual content while those within the same shot usually maintains consistency.However, the observation is neither a sufficient nor a necessary condition of shot boundaries. On the one hand, abrupt illumination change or object/camera's large movement will also lead to significant change of visual content. On the other hand, some statistical features, adopted to represent the visual content, are not expressive enough to reflect shot transition occurring in the same scene. Therefore, to achieve high performance, the system must try to reduce the disturbance of illumination and motion, and adopt expressive features.

As an attempt to overcome the aforementioned shortcomings, our system can be characterized as "two stages, three modules". Two stages refer to feature extraction stage and shot boundary detection stage. Firstly, various required features, which form a compact representation of each frame's visual content, are extracted from either compressed domain or uncompressed domain. Then shot boundary detectors, including FOI detector, CUT detector and GT detector, decide whether a shot transition occurs at a certain position via analyzing the feature variation. Figure 1. depicts the architecture of our system.

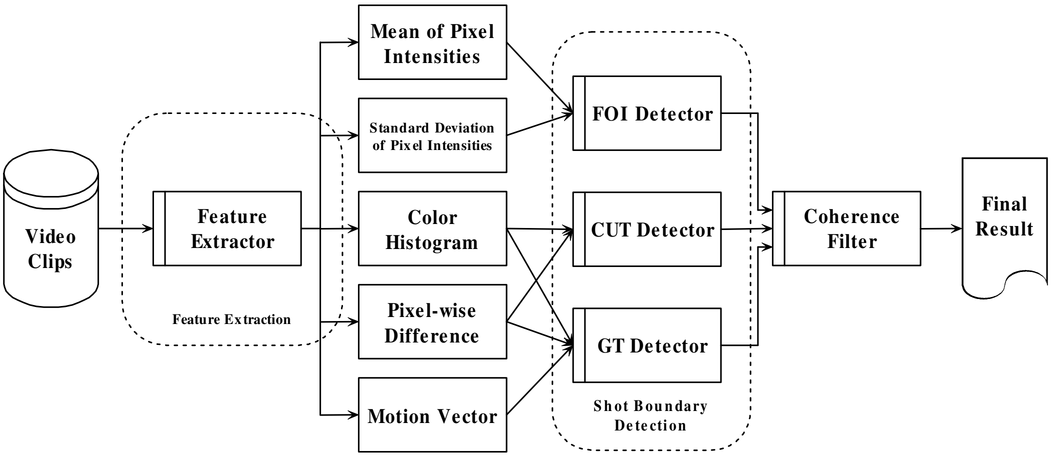


Figure 1: System overview of shot boundary detection task.

Dividing the system into two stages provides several advantages. Once various features are extracted, they can be utilized for good. Therefore,we can avoid repeated feature extraction job, which is computational intensive. In addition, obtaining the features of the whole video, the system can adjust the various global thresholds according to the distributions of features.

2.2. Representation of Visual Content

Various features have been exploited to represent the visual content of each frame[5]. As to shot boundary detection task, a trade-off of efficiency and expressiveness has to be considered. So, here we select five simple but complementary features, trying to reflect visual content from different aspects.

In our system, color histogram of 48 bins in RGB space, 16 bins for each channel, is utilized and histogram intersection method is adopted to calculate the dissimilarity of two histograms. Furthermore, pixel-wise difference feature, as a supplement to color histogram, introduces spatial information. To detect flashlight effect and monochrome frame, the mean value and standard deviation of each frame's pixel intensities are also calculated. Abrupt change of illumination can be detected by tracking the variation of mean gray value. Moreover, stable intensities, a prominent characteristic of monochrome frame, can be reflected by small standard deviation feature. Besides features from uncompressed domain, motion vectors from compressed domain are also extracted and synthesized to reflect the global motion of a frame.

2.3. System Components

In order to integrate multiple cues to detect a shot boundary,either CUT detector or GT detector in our system consists of two steps: getting candidates and sifting candidates. In this section, algorithms for detection different type transition are discussed.

2.3.1 FOI Detector

The key problem of FOI detection is the recognition of monochrome frame, since there is at least one monochrome frame within the FOI transition but monochrome frame seldom appears elsewhere. One dominant characteristic of monochrome frame is its low standard deviation of pixel intensities. Thus standard deviation feature is utilized in FOI detection process. FOI detection process is described as follows.

Step 1. Detect monochrome frame. A monochrome frame is declared once its standard deviation feature is below a given threshold T_s, which is heuristically determined.
Step 2. Judge the type of entering transition,abrupt or gradual.If it is gradual, search the fade out boundary of the previous shot.
Step 3. Judge the type of exiting transition,abrupt or gradual.If it is gradual, track the fade in boundary of the next shot.

2.3.2 Cut Detector

[10] considers that identification of cuts has been somehow successfully tackled. In fact, a simple threshold scheme can achieve relatively good performance. Therefore, here CUT detector firstly employs simple threshold method to get candidates, then incorporates several post-processing modules to sift the candidates. Different from traditional threshold method, a threshold method, called second order derivative, is proposed to boost the precision of cut candidates.

Second Order Derivative Method

The traditional threshold method, in which feature variation between adjacent frames is directly compared with a global threshold T_c, usually works well for video sequences illustrated in Figure 2(a). In Figure 2(a), most feature variations at cut transition exceed T_c, while variations within the same shot usually are smaller than T_c. However, due to illumination change or object/camera motion, a lot of video sequences as Figure 2(c) demonstrats exist. In such video sequences, even feature variations within the same shot frequently exceed T_c. Therefore, it causes many false alarms.To overcome this drawback, we present a so called second order derivative method.

Formally, let x_k denote the feature of the kth frame and $d(x_k, x_{k+1})$ denote the feature variation between the kth and the $(k+1)th$ frame. In traditional method,feature variation $d(x_k, x_{k+1})$ is directly compared with T_c . In second order derivative,instead of $d(x_k, x_{k+1})$, $d(x_{k+1}, x_{k+2}) - d(x_k, x_{k+1})$ is compared with T_c. The new scheme not only works well for Figure 2(a),but also can effectively eliminate the false positives of Figure 2(c).

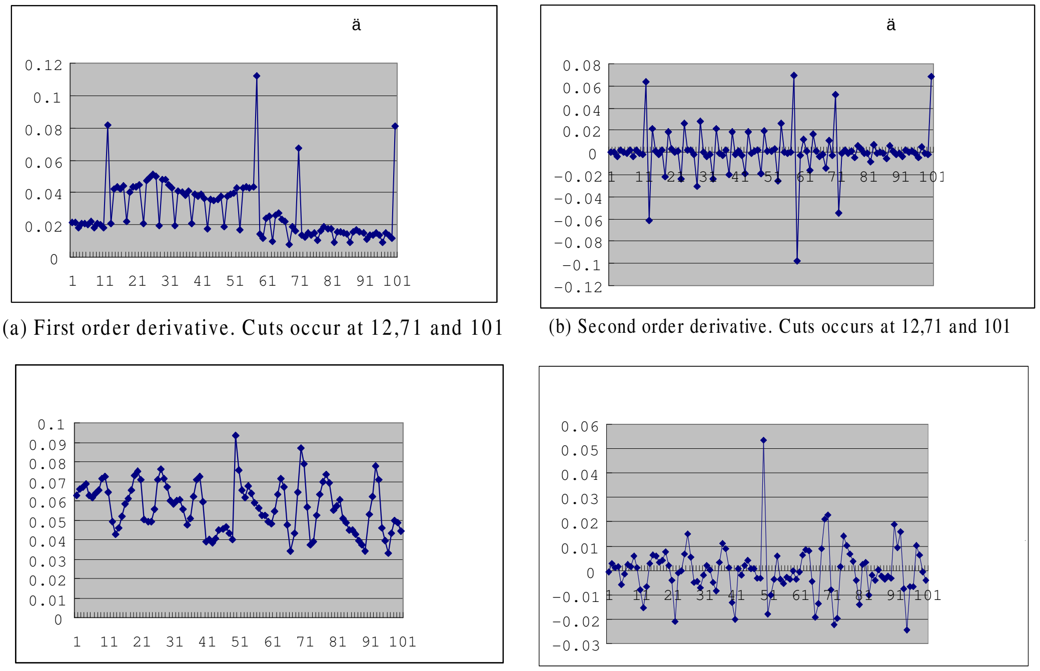
On the one hand, for Figure 2(a), if the kth and $(k+1)th$ frame belong to the same shot and there is a cut transition between the $(k+1)th$ and $(k+2)th$ frame, $d(x_k, x_{k+1}) \ll T_c$ and $d(x_{k+1}, x_{k+2}) > T_c$ hold. Therefore it is very likely that $d(x_{k+1}, x_{k+2}) - d(x_k, x_{k+1})$ still exceeds threshold T_c , that is to say, new scheme can detect the cut transition in video sequences of Figure 2(a). Figure 2(b) illustrates the effectiveness of second order derivative method.

On the other hand, for Figure 2(c), although there is no cut transitions within this sequence, relations $d(x_k, x_{k+1}) > T_c$ and $d(x_{k+1}, x_{k+2}) > T_c$ hold almost everywhere. For relatively large $d(x_k, x_{k+1})$ and $d(x_{k+1}, x_{k+2})$, $d(x_{k+1}, x_{k+2}) - d(x_k, x_{k+1})$ usually becomes smaller than T_c. Thus,many false alarms can be eliminated. Figure 2(d) shows the second order derivative feature variation of (c).

In conclusion, second order derivative scheme can depress the mean value of feature variation and is very effective to boost the precision for “unstable” video sequence. However, second order derivative may decrease the recall measure, because it will neglect the first cut when two cuts occur within three consecutive frames. Fortunately, such situation seldom appears.

Post Processing Module

Besides cut transition, abrupt illumination change like flashlight effect or large movement of object/camera sometimes will also lead to feature variation above T_c. To reduce these false positives, we design specific modules, including flashlight detector and gradual transition filter, to sift the cut candidates.



(a) First order derivative. Cuts occur at 12,71 and 101

(b) Second order derivative. Cuts occurs at 12,71 and 101

(c) First order derivative. No cut occurs.

(d) Second order derivative. No cut occurs.

Figure 2: The global threshold T_c is 0.06.In(a),there is an insertion and a deletion,the same situation happens to (b). In (c) there are a lot of insertions, but there is only one in (d).

According to the ideal flash model and ideal cut model described in [14], a simple flashlight detector is designed. The mean value of each frame's pixel intensities is employed to reflect its luminance. Once the increment of luminance between two successive frames is above a particular threshold, flashlight detector will examine the luminance of next several frames. If the luminance of one frame falls to the relatively low value again, the cut candidate is taken as a flashlight effect, otherwise, it is declared as a cut.

Gradual transition filter module is adopted to reduce the false positives caused by object/camera movement. Large movement usually causes great feature variation similar to cut transition. However, feature variation caused by movement usually lasts a number of frames while that of cuts only occurs at a spot. Based on this observation, a gradual transition filter module is designed to distinguish the cut transitions and false positives caused by movement. If feature variations surrounding the cut candidate are unstable and most of the magnitudes are comparative to T_c, the candidate will not be declared as a cut.

2.3.3 Gradual Transition Detector

Gradual transition detection remains a hard problem due to its gradualness and versatility[10]. During the transition process, two shots are superimposed together (dissolve), or the former gradually becomes totally black, then the latter shot gradually appears (wipe or FOI). A transition process may last more than 100 frames.

Twin comparison method proposed by [9] works well for short gradual transition detection. However, it has some difficulties in detecting long gradual transition. On the one hand, camera/object motion is likely to cause false alarms. On the other hand, truncation of long gradual transition may frequently occur, because the detection process terminates once the difference of adjacent frames doesn't exceed the lower threshold.

To overcome such shortcomings, we divide gradual transitions into two categories: *short gradual transition* and *long gradual transition*. Short gradual transition refers to transition whose length is less than 6 frames,and long gradual transition for others. In addition, we treat short and long gradual transition respectively, traditional twin comparison method for short ones and an improved twin comparison algorithm for long ones. In the improved method, the lower threshold is self-adaptive based on the motion feature from compressed domain. In addition, the improved approach, described by a finite state automata model, possesses more robust conditions for both entering and exiting of the gradual transition process.

Motion Based Self-adaptive Threshold

Feature variation caused by object/camera movement exhibits similar characteristics to that of gradual transition. To boost the performance of gradual transition detection, the system has to successfully reduce the disturbance of motion. A straightforward idea is to adjust the lower threshold self-adaptively according to the magnitude of motion, more specifically, the system heightens it if large movement occurs or depresses it if small or no movement occurs. An implementation of this idea can be roughly described as:

Step 1.If current frame is P or B frame, extract motion vectors of all macro-blocks from compressed domain of MPEG video.
Step 2.Calculate the mean values of horizontal motion vectors and vertical ones respectively. As to I type frame, just copy the previous one's features.
Step 3.Get the current threshold according to a linear relation between the threshold and the horizontal/vertical motion features. The coefficients of the linear function are heuristically determined.

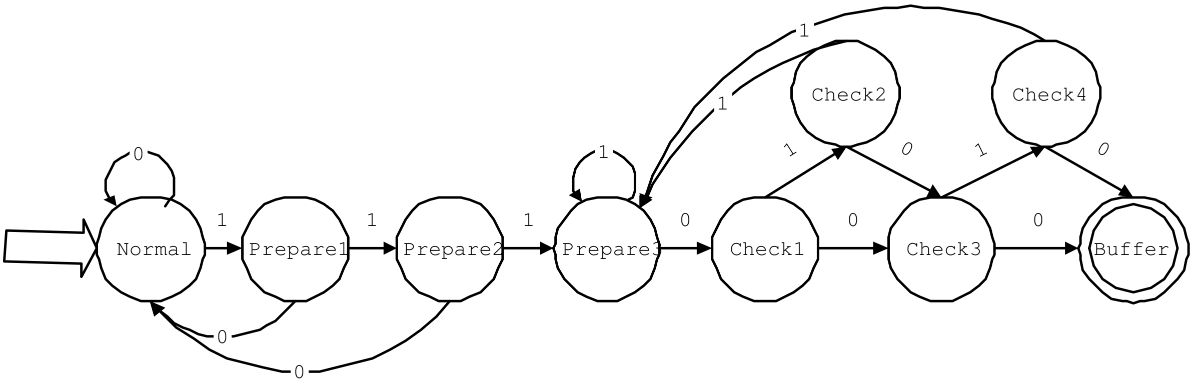


Figure 3: Finite state automata model.

Finite State Automata Model

The principle of the improved twin comparison method can be demonstrated by a finite state automata (FSA) model depicted in Figure 3. As the figure shows, there are 9 states in the FSA model and the model's alphabet set includes only 0 and 1. The model has only one accept state $Buffer$. The states $\{Prepare1, Prepare2, Prepare3\}$ are entering states of gradual transition.Obviously,to reach $Prepare3$, at least three 1 signals are required.The states $\{Check1, Check2, Check3, Check4\}$ are exiting states. To go from $Prepare3$ to $Buffer$, at least three 0 within 5 continuous signals are required.Otherwise, the model will cycle at state $Prepare3$. Thus, its working process is summarized as follows.

Step 1. When the detection process starts, it is assumed at $Normal$ State by default.
Step 2. The lower threshold T_l is determined according to the global motion feature of current frame.
Step 3. If the discontinuity value exceeds T_l, the input signal will be set as 1, otherwise it is set 0.
Step 4. According to the current state and input signal, the FSA goes to next state.
Step 5. If FSA reaches accept state $Buffer$, go to step 6, else go to step 2.
Step 6. Compare the accumulated variation with the higher threshold T_h. Only if the variation exceeds T_h, the candidate is declared as a real gradual transition.

Overall, there are three advantages of FSA model over traditional twin comparison method.Firstly, by motion based self-adaptive threshold, the system can reduce the disturbance of object/camera motion. Secondly, the system possesses more robust condition of entering the gradual transition detection process, since only more than three consecutive discontinuity values that exceed the lower threshold indicate a suspect gradual transition.Finally, the condition of exiting the gradual transition is also more robust. The system considers gradual transition is over only when there are three discontinuity values, within 5 consecutive frames, below the lower threshold. Therefore, the false positive and truncation are both suppressed to a certain extent. The evaluation result shows that performance on gradual transition of this system outperforms those of all the other systems.

2.4. Collaboration of Separate Modules

Up to now, we have introduced three main modules. They are not working independently. The collaboration or interaction of those modules can be described by Figure 4.

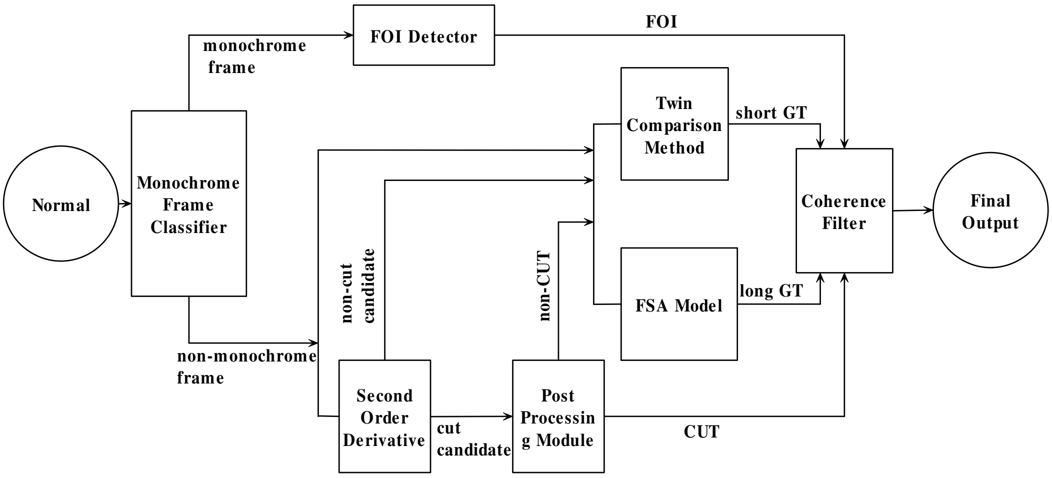


Figure 4: Finite state automata model.

As Figure 4 shows, FOI detector has highest priority. On a frame arriving, it is firstly recognized whether it is a monochrome one or not. If it is, the FOI detector will be activated and runs until FOI transition is over. Otherwise, the frame will be processed by the next three parallel modules, i.e. CUT detector, short GT detector and long GT detector. In cut detection process, the non-cut candidates are not discarded but sent to GT detectors for further recognition. At the last stage, a coherence filter is employed to break the tie of short GT and long GT, consequently, only one appropriate shot boundary is declared at a position.

2.5. Experiment and Evaluation

The system is developed based on the collections of 2003 SBD task. One video named “$19980203\_CNN.mpg$” is selected as training data, the other seven ones as testing data. To investigate the contributions and drawbacks of each module, we develop and test various modules separately. Then assemble these modules in different manner to constitute a system. By evaluating these system, we can select one system with optimal architecture. Before submitting the results, all the eight videos are treated as training data to tune parameter settings of the system, and therefore optimal parameter settings are obtained. This system is treated as a baseline system.

Ten runs, most of which correspond to variations of global parameters of baseline system, have been submitted. Except by tuning parameters, several other submissions are produced by removing a specific module from the baseline system. The evaluation results and their corresponding analysis are listed in Table 1 and Table 2.

Ranked by F-measure, of the total 132 submissions, our best run $thuai15$ outperforms all the other non-Tsinghua ones in overall performance(The performance of $it09$ is almost the same as that of $thuai15$, but $thuai15$ has better tradeoff between recall and precision). Moreover, our 10 submissions are among the top 20 by overall performance. Although 9 of our runs are in the top 30 of the cut measure, the best one $thuai10$ only ranks $16th$. In gradual measure, top 10 runs are of Tsinghua.Moreover, our 10 submissions also perform excellent gradual frame accuracy, since they outperforms almost any other non-Tsinghua runs.

3. Feature Extraction

3.1. System Overview

Our multimodal system is shown in Figure 5, and this system uses visual, text and time information for concept detection.

As Figure 5 shows, the visual features are extracted on shot keyframes provided by C-CLIP and processed in the visual modal detector to generate the baseline list. In this detector, support vector machine (SVM)[3] classifiers are trained on

Table 1: Evaluation result of the ten submissions(Ranked by F-measure)

Sysid	All transitions			Cuts			Gradual transitions			GT frame accuracy		
	Rcl	Prc	F#	Rcl	Prc	F#	Rcl	Prc	F#	Rcl	Prc	F#
thuai15	0.884	0.896	0.890	0.928	0.931	0.929	0.792	0.82	0.806	0.824	0.86	0.842
thuai14	0.888	0.89	0.890	0.928	0.93	0.929	0.803	0.807	0.805	0.829	0.848	0.838
thuai10	0.888	0.89	0.890	0.925	0.939	0.932	0.809	0.79	0.799	0.839	0.835	0.837
thuai16	0.896	0.881	0.888	0.929	0.926	0.927	0.827	0.79	0.808	0.824	0.848	0.836
thuai17	0.903	0.872	0.887	0.929	0.923	0.926	0.846	0.775	0.809	0.82	0.848	0.834
thuai07	0.889	0.885	0.887	0.928	0.93	0.929	0.808	0.791	0.799	0.839	0.836	0.837
thuai02	0.888	0.885	0.886	0.925	0.931	0.928	0.808	0.791	0.799	0.839	0.836	0.837
thuai05	0.884	0.888	0.886	0.919	0.938	0.928	0.811	0.787	0.799	0.839	0.836	0.837
thuai19	0.902	0.864	0.883	0.902	0.864	0.920	0.821	0.789	0.805	0.817	0.859	0.837
thuai08	0.898	0.866	0.882	0.898	0.866	0.919	0.801	0.801	0.801	0.838	0.848	0.843

Table 2: Description and analysis of the submissions

Name	Description and Analysis
$thuai02$	It is produced by the baseline system, with default architecture and default parameter settings.
$thuai05$	Second order derivative method is employed in short GT detection module. In theory, this would lead to higher precision and lower recall of GT, but the actual result is not accordant with inference.
$thuai07$	This run loosens the condition of short GT detector. Thus, recall of short GT should increase, while precision should decline. But the effectiveness is not evident.
$thuai08$	This run removes the gradual transition filter module. Thus, recall of cut should increase, while its precision should decline.Opposite affection is expected on GT detection. The evaluation result confirms the above guess.
$thuai10$	This run prolongs the duration of post-cut processing modules.In $thuai02$, before declaration of a cut, the post-cut module firstly examines the next 4 frames of cut candidate. Only if the next 4 frames satisfy some specified conditions, the candidate is assumed to be a real cut. In the thuai10, the post-cut module examines the next 5 frames of cut candidate before declaring a cut. Thus, cut precision ascends.
$thuai14$	$thuai14$ is a new baseline submission derived from $thuai10$. In this run, the coefficient of motion based self-adaptive threshold is tuned higher than thuai10. Thus, the recall of gradual transition declines, while the precision increase.
$thuai15$	Even higher coefficient of self-adaptive threshold.Therefore, the gradual transition recall declines further, and precision increases further.
$thuai16$	Lower threshold of post-gradual transition module.So the GT recall increases, while precision declines.
$thuai17$	Even lower threshold of post-gradual transition module. So GT recall increases further, while precision declines further.
$thuai19$	This run replace the second order derivative with first order derivative in cut detection module, so the recall of cut abruptly increases, while precision abruptly declines.

visual features. Here we combine some visual features to test the early feature fusion method, which is the opposite of late classifier ensemble fusion. The features for testing are classified by their corresponding classifiers and then ordered shot lists are generated. Each list is comprised of ordered items and each item is a $[S_i, C_i, R_i]$ triplet, where S_i is the shot Id, C_i is the confidence value, and R_i is the rank value for item i. The well-behaved lists are selected and fused with an AP-based Borda count to form the baseline list.

Filtered out the unwanted shots of commercials and headline anchors, the baseline list is split into two sublists and reranked in each sublist to utilize the text and timing cues which may be preserved only on the sub-dataset. A linear weighting scheme (section 3.3.3) is adopted for the reranking process. After that, the two sublists are further fused by the insertion algorithm introduced in section 3.3.1 and that completes the Split-Rerank-Insert (SRI) fusion process.

The final list is obtained after the shot clustering algorithm (section 3.3.4) applied.

The fusion methods adopted are comprised of early feature fusion, AP-based Borda Count, Split-Rerank-Insert process and linear weighting scheme. These different techniques are applied in different parts of the system since the concept detection task relies on heterogeneous information sources and thus requires different operations for different kinds of information.

The TRECVID 2003 corpus is partitioned into four non-overlapping subsets to train the SVM classifiers, to select and generate the baseline list, to tune and to validate the fusion parameters. The submitted runs are run on the TRECVID 2004 corpus using these validated parameters.

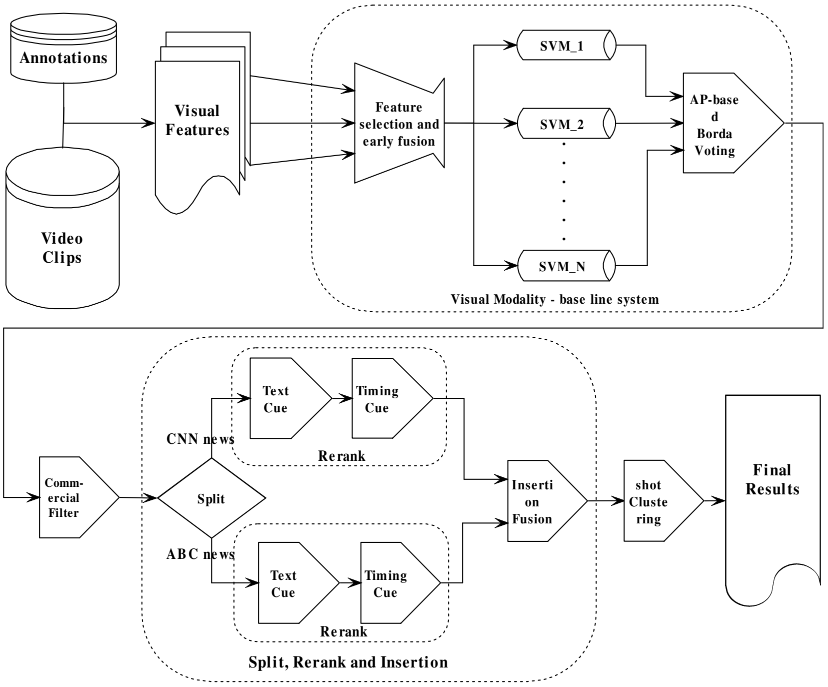


Figure 5: The Tsinghua Concept Detection System.

3.2. Visual Features and Visual Modal Detector

3.2.1 Visual Feature Extraction

In the visual modality, several image features of color and texture are extracted on keyframes. The features are extracted either for the whole image or the sub-images in 2×1 or 3×1 grids. Most of our color features are based on the non-uniformly quantization of HSV color space into 36 bins[15], which is abbreviated as HSV36. The following visual features are extracted:(1)Color histogram of HSV36 [15] (CHHSV36 36 bins),(2)Color coherent vectors of CHHSV36 (CCV 72 bins), (3)Auto-correlogram of HSV36 with distance set $D = \{1, 3, 5, 7\}$ (HSVCorr 144 bins), (4)Tree-structured wavelet transform feature (TWT 104 bins), (5)Pyramid-structured wavelet transform feature (PWT 48 bins), and (6)Edge orientation histogram quantized at 5 degree interval (72 bins).

3.2.2 Selection and Early Fusion of Visual Features

We use support vector machine (SVM)[3]as the classifier. For detection of certain concept, the confidence of each sample is determined by the normalized distance to the decision boundary, based on which a rank list can be produced for different feature with its optimal SVM. Then the effectiveness of this feature for the concept can be evaluated by the average precision (AP) of its rank list.

As there are many different visual features available, we devise a procedure of feature selection to find out the most effective features for every concept. The low dimensionality of our visual features makes it possible for us to further try different combination schemes and encouraging results are given in Figure 6.

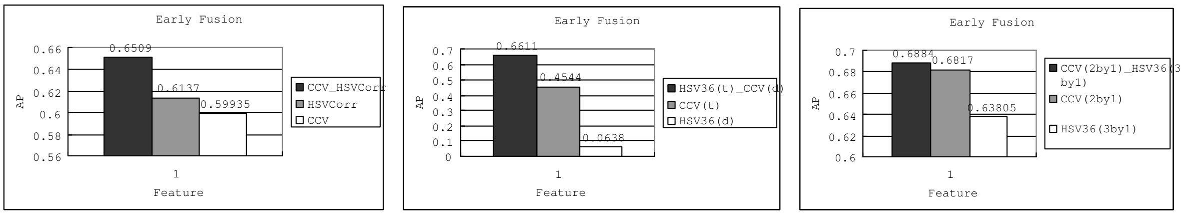


Figure 6: The early fusion experiments. The experiments are conducted on the TRECVID 03 data set. AP value is obtained through cross validation on the data set. The different features are described as follows: CHHSV36:Color histogram of HSV36; CCV:Color coherent vector; HSVCorr: Auto-correlogram of HSV36; HSV36(t): HSV36 extracted from the top half of the keyframes; CCV(d): CCV extracted from the bottom half of the keyframes; CCV(2by1): combination of two CCVs extracted from each block of the keyframes partitioned by 2by1 grids; CHHSV36(3by1): combination of three CHHSV36s extracted from each block of the keyframes partitioned by 3by1 grids; "_" in the feature name means combination of different features.

3.2.3 AP-Based Borda Count for Visual Modality

We use a variant of Borda count (voting) method[1], the AP-based Borda voting method[2], to fuse the rank lists based on different visual features. This is a heuristical strategy to assign the weights of lists by their APs, not by parameters derived using the computation-demanding logistic regression. Since APs are between $[0, 1]$, some non-linear mapping can be designed to sharpen the difference between classifiers. More specifically, suppose we have n lists with estimated[2] AP value equal to $w_j, j = 1, ..., n$. We use r_i^j to denote the rank of shot i in list j. Then the AP-based Borda voting can be stated formally as[2]:

$$r_i = \sum_{j=1,...,n} \exp(Cw_j) \cdot (I - r_i^j) \quad , \tag{1}$$

where C and I are some suitable constants. This method offers a 7.2% increase over the best of single SVM list, 18.1% over the mean and 18.3% over the median in this year. See Figure 7.

3.3. Text, Timing information and Split-Rerank-Insert (SRI) Fusion

3.3.1 SRI Fusion

The whole dataset may not be consistent and some special property preserves only in specific subsets, e.g. the timing cue is different in the CNN/ABC subset for "Clinton", see Figure 8. We split the list to different property-preserving sub-lists and apply the special information to rerank them (3.3.2). In the current system, the text and shot position information are sequentially used to rerank the sublists. Then reranked sublists are merged using a rank based list insertion algorithm, which is similar to [16]. The algorithm takes the list with higher AP value as a major list. Given the initial position P_i and insertion distance D, the other list is inserted into this major list starting from the initial position every D shots. D and P_i are determined by experiments. And this algorithm can be easily generalized to more than two lists. We call this procedure as Split-Rerank-Insert.

[2] The AP is evaluated on the training dataset.

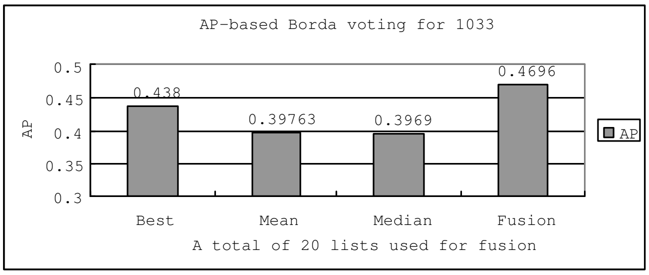


Figure 7: Performance improvements by AP-based Borda Count for "Basket scored" concept this year. The fusion result with AP 0.4696 is our baseline system.

3.3.2 Linear Weighting Fusion Methods

We adopt linear weighting fusion to rerank the visual modal baseline with text and timing information. The linear weighting fusion of two lists is based on rank/confidence value. As the rank values can be converted to "confidence like" scores by certain functions, we do not discriminate the rank and confidence value and use "score" to refer to either of them. The scores are first normalized to $[0, 1]$. Then each normalized list is assigned a weighting factor in $[0, 1]$ and combined using arithmetic operations such as sum, subtraction, multiply, minimum and maximum. For the rank based fusion, only sum operation is used. Since the difference of the estimated performance of two lists may be large, or there may be much more false shots in one list, we can only output the shots in the list which is expected to have higher performance, instead of output all shots in the both two lists. The optimal choice of the weights and the operation are determined by experiments on the training dataset for fusion. Our experiments show that sum and rank are nearly always the best choices.

3.3.3 Detectors and Commercial Filter Based on Text

Two Text-based Detectors

Two different detectors are generated by the text modality using text keyword match and text retrieval techniques. We used two kinds of transcription sources, accurate Closed caption (CC) text and inaccurate Automatic Speech Recognition (ASR) text which has been aligned with the shot boundaries[6]. The text documents are built by aligning CC text with ASR text for each shot. The resulting short documents contain 7 words in average for each shot and many documents are null.

The first text-based detector is the text retrieval detector based on the system developed in [16]. Some keywords are selected manually as query words, based on which a list ranked by the similarity scores between the query and the documents is generated. The common OKAPI formula[12] is adopted for similarity measurement.

The second text-based detector is the rule based keyword match detector, in which the CC text of each shot is selected if it can match a specific rule and then sorted according to its matched degree. For example, the shot may talk about "basket scored" when CC text contains two numbers with range between 60 and 120, e.g. "77 to 89", or when it mentions basketball teams or stars. And these rules can not be effectively represented using simple text retrieval techniques with a bag of words as the basic model. We design each rule manually and assign a score for the matched shots to reflect the matched degree. Thus another rank list is generated.

Rank lists based on the two text detectors are first fused and then fused with the split sublist which is generated from the visual modality using the same linear weighting fusion scheme in Section 3.3.2.

Text Based Simple Commercial filter

An exhaustive text match is carried out to find the commercials and headline anchors because most of them are repeated many times across videos and are nearly identical in ASR-aligned CC text. These shots are gathered to a commercials list L_{CM}. If shot S_i are both in the baseline list L_{BS} and L_{CM}, it is removed from L_{BS} when the concept for detection does not appear in commercials.

3.3.4 Utilization of Timing information

Shot position

Table 3: Final results

Run	Basket Scored		Bill Clinton	
	AP	Gain over baseline	AP	Gain over baseline
AP-based Borda voting –Base Line	0.4696	0.0%	0.0483	0.0%
Commercial Filter	0.4829	2.8%	0.0537	11.2%
Commercial Filter+SRI	0.5478	16.7%	0.0688	42.4%
Commercial Filter+SRI+Cluster	0.5614	19.5%	—	—

In news video, specific type of news such as sports news frequently occurs at specific period of time. This information, denoted as shot position (in video), can be described by probability density function ($p.d.f.$) estimation and further used to generate still another rank list. The density function is estimated using the Parzen-window approach. We choose Gaussian kernel as [8], and the bandwidth is obtained by simply averaging the length of all the relevant shots. The estimated $p.d.f.$ is then smoothed to eliminate zero values. Figure 8 gives two estimated $p.d.f.$s of "Bill Clinton" (32) for each news station (CNN, ABC).

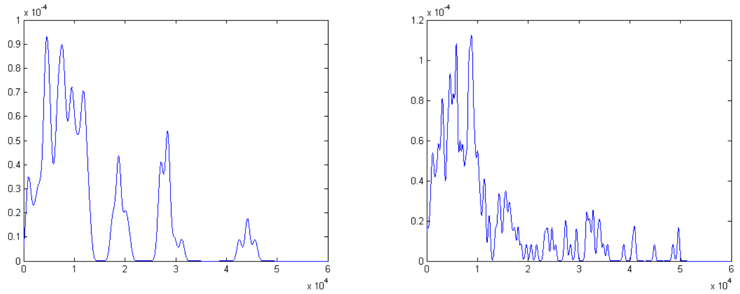


Figure 8: Estimated $p.d.f.$ for concept "Bill Clinton" in CNN and ABC respectively. The X-axis is the shot position in seconds and the Y-axis is the estimated density.

Rank list based on shot position is generated by sorting the shots by the estimated probability in descending order. And this list is combined with the text-reranked visual list using the linear weighting scheme described in Section 3.3.2, which forms the text- and timing- reranked visual list.

Shot Clustering

Tasty mushrooms always grow in cluster. So do the shots containing some concept.

Some concept may not have the fixed position in video, but the corresponding shots often come in cluster. This is what we called shot clustering phenomenon as Figure 9 shows. In Figure 9, the shots are clearly "growing in cluster". In an effort to use such timing cue, we have devised a novel selective shot clustering method described as follows[4].

We use *cluster* to refer a series of shots within which any shot is near to its immediate neighbors. There are 2 parameters to be chosen before doing clustering analysis: *Maximum Intra-Cluster Gap (MICG)* and *Highly Confident Shot Range (HCSR)*.

As a threshold to judge whether two shots are near, MICG is defined as the maximum time interval between temporal adjacent shots in a cluster.

HCSR is a range within which shots from result list are considered to be representatives of positive shots.

When MICG and HCSR are determined, time clustering analysis and result adjustment can be performed in the following steps:

Step 1. Cluster the highly confident shots ,given MICG and HCSR;
Step 2. Absorb non- highly confident shots into clusters. If a result shots is not absorbed by any cluster, it is called an *Isolated Shot*. When a shot is to be absorbed into two different clusters, then merge these two clusters into one.
Step 3. Delete isolated shots from the result list.

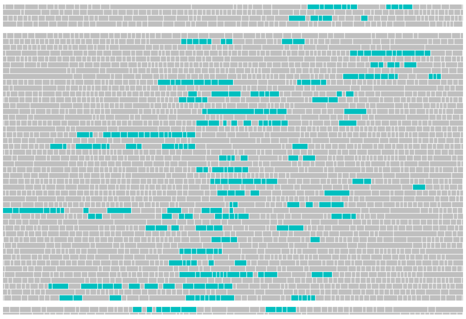


Figure 9: Shot clustering for ”Basket scored”. The shots are shown in time sequence and in proportion to their actual length. Only a small part of the videos is shown and shots containing the concept are in dark.

3.4. Feature Extraction Task in TRECVID 2004

In this year, our feature extraction system is applied to three concepts: “Beach”, “Bill Clinton” and “Basket Scored”. The system is originally developed to detect “basket scored” concept, which achieves the highest AP in all runs submitted this year, see Figure 10. Contributions to the final results from different components are shown in Table 3. Text and time information bring about 20% gain to the visual-information-only result.

Our results of “Bill Clinton” are below the average as shown in Figure 10. The main reason may be that only visual features are used. and some more sophisticated detectors such as face detectors may improve the result significantly. Also we should make more efforts to exploit text information, which is expected to be very effective in detection of “Bill Clinton” [8]. In our system, the text and time information boosts the result more than 40% as shows in Table 3.

As for detection of “Beach” concept, only the component of visual feature selection and early fusion is used. The result is above the average in all runs as shown in Figure 10.

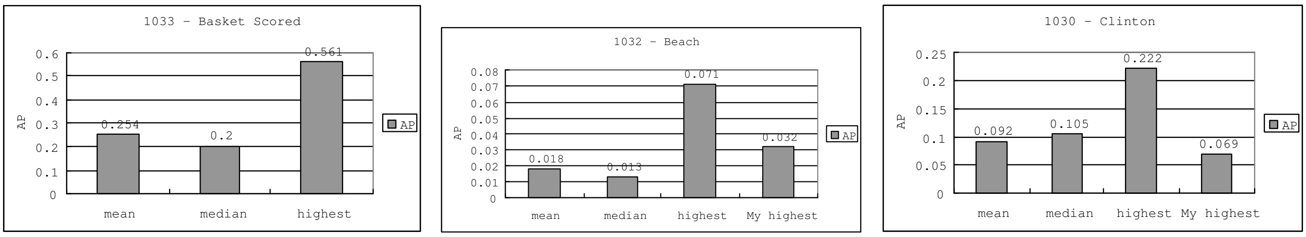


Figure 10: Results of three concepts.

4. Conclusions and Future Directions

In this paper,we have presented our shot boundary detection approach for TRECVID 2004.The evaluation shows that our system is among the best performing groups. By investigating the contribution of each component, we can summarize the advantages of our system as follows.

- Utilizing of multiple complementary features.
- Exploiting the differences of real shot transition and two main disturbances,i.e.illumination change and motion.
- Improvements on traditional algorithms,including second order derivative and finite state automata model.
- Novel manner to utilize motion feature,that is,motion based self-adaptive threshold.

However,there is still much room to improve. For example, most of the thresholds are global and heuristically determined. And flashlight detection method is somewhat straightforward. Moreover, although motion feature is utilized, it is not delicate enough to distinguish global and local movement.

In Feature Extraction task, the visual modal detector forms our baseline, and the text and timing information is further incorporated to boost the performance. Due to the time limit, we have no time for detecting other concepts. And More experiments are underway for applying our system to other concepts and for devising new features and methods to explore the multimodal correlations.

Acknowledgments

Special thank goes to Shaoping Ma, Min Zhang, Yiqun Liu and Le Zhao of Pattern Recognition and Intelligent Retrieval reseach group, State Key Lab of Intelligent Tech. & Sys. for providing the TMiner IR system, v2.1 in text retrieval list generation. We also thank Duo Zhang, Shen Zhang, Jia Li and Huiyi Wang, who are undergraduates of Department of Computer Science, Tsinghua Univ., for their assistance to our research.

This work is supported by National Natural Science Foundation of China (60135010) and National Natural Science Foundation of China (60321002).

References

[1] D. Black. *The Theory of Committees and Elections*. Cambridge University Press, 2nd edition, 1958, 1963.

[2] Le Chen, Dayong Ding, Dong Wang, Fuzong Lin, and Bo Zhang. Ap-based borda voting method for feature extraction in trecvid-2004. In *accepted by 27th European Conference on Information Retrieval (ECIR05)*.

[3] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. In *IDIAP Research Report 02-46, Martigny, Switzerland. (see also www.torch.ch)*, 2002.

[4] Dayong Ding, Le Chen, and Bo Zhang. Temporal shot clustering analysis for video concept detection. In *accepted by 27th European Conference on Information Retrieval (ECIR05)*.

[5] U. Gargi, R. Kasturi, and S. Antani. Performance characterization and comparison of video indexing algorithms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA*, pages 559–565, june 1998.

[6] J.L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. In *Speech Communication*, volume 37, pages 89–108, 2002.

[7] A. Hanjalic. Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, February 2002.

[8] A. Hauptmann and et al. Informedia at trecvid 2003: Analyzing and searching broadcast news video. In *NIST TREC-2003 Video Retrieval Evaluation Conference*, November 2003.

[9] H.Zhang, A.Kankanhalli, and S.W.Smoliar. Automatic partitioning of full-motion video. *Multimedia Syst.*, pages 10–28, 1993.

[10] Chong-Wah Ngo, Ting-Chuen Pong, and HongJiang Zhang. Recent advances in content-based video analysis. *Int. J. Image Graphics*, 1(3):445–468, 2001.

[11] R.Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII,Proc.SPIE,*, volume 3656, pages 290–301, December 1998.

[12] S. E. Robertson and et al. Okapi at trec-4. In *In the Fourth Text Retrieval Conference (TREC-4)*, 1993.

[13] S.Lefevre, J. Holler, and N. Vincent. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, 9:73–98, 2003.

[14] Dong Zhang, Wei Qi, and HongJiang Zhang. A new shot boundary detection algorithm. In *IEEE Pacific Rim Conference on Multimedia 2001*, pages 63–70.

[15] Lei Zhang, Fuzong Lin, and Bo Zhang. A cbir method based on color-spatial feature. In *IEEE Region 10 Annual International Conference*, pages 166–169, 1999.

[16] M. Zhang. *Study on Web Text Information Retrieval(In Chinese)*. PhD thesis, Tsinghua Univ. Beijing China., June, 2003.