International Journal of Computer Applications (0975 - 8887)
Volume 60 - No. 6, December 2012

Unsupervised Approach for Retrieving Shots from Video

M. Kalaiselvi Geetha
Dept. of Computer Science and Engg.,
Annamalai University

S. Palanivel
Dept. of Computer Science and Engg.,
Annamalai University

ABSTRACT

Acquiring the video information based on user requirement is an important research, that attracts the attention of most of the researchers today. This paper proposes an unsupervised shot transition detection algorithm using Autoassociative Neural Network (AANN) for retrieving video shots. The work further identifies the type of shot transition, whether abrupt or gradual. Keyframes are extracted from the detected shots and an index is created using *k*-means clustering algorithm for effective retrieval of required shots based on user query. The approach shows good performance in retrieving the shots, tested on five popular genres.

General Terms:

Shot Transition, Shot Retrieval

Keywords:

Shot transition detection,Autoassociative neural network, *k*-means clustering algorithm, Shot retrieval

1. INTRODUCTION

The growth of multimedia database is ever increasing due to its content richness. Researchers, students, academecians are approaching this media for their required information. One example of distribution is the Internet which is a fast growing medium through which distribution/sharing can be made. It is absolutely clear that manual management of such a voluminous data is unbelievable and unrealistic, which necessitates automation. Acquiring the required information from the ever increasing voluminous database is a challenging task. This paper proposes an unsupervised approach for retrieving required video shots. The method uses five popular genres for analysis viz., cartoon, sports, news, commercial and serial. The approach begins with identifying shots, followed by indexing the shots for retrieval.

1.1 Related Work

Shot transition detection (STD) is the fundamental step for various video processing applications such as indexing and video retrieval techniques. STD aims at organizing the video information into meaningful segments [1]. A basic method to detect shot transition is to look for discontinuity between two adjacent frames by assigning a discontinuity metric [2] and [3] between successive frames. The techniques applied for shot boundary are surveyed in [4]. A unified model for detecting different types of shot transitions is presented in [5].
Indexing and retrieval are closely related issues. Retrieval of desired video content from the huge video database requires proper organisation, i.e., indexing. Video indexing enables effective organization and retrieval of large amount of video information [6] and archives. Inappropriate organisation of video database hinder the effective usage of the same. While indexing a video database, the three major issues [7] are 'what to index', 'how to index' and 'which to index'. The Video retrieval systems such as Virage [8] are based on similarity computation.
This paper is organized as follows. The major issues related to retrieval are discussed in Section 2. Section 3 explains the proposed shot transition detection algorithm. *k*-means clustering algorithm is discussed in Section 4. Experimental results are discussed in Section 5 and finally Section 6 concludes the paper.

2. ISSUES RELATED TO SHOT RETRIEVAL

Huge volume of digital video is available for navigation today. It is a rich source of information and it is a challenge to manage the video data as an information resource. Managing the video information actually includes analysing, summarising, indexing, retrieving and searching the video database for the required information based on understanding an image/video, what has been generally considered to be an extremely difficult task. The semantic content of the video is widespread and all the above mentioned tasks largely depend on the domain of the video. For example, in sports domain, its semantic content varies from that of movie video or a cartoons video. To make the task more amenable, a unique approach applied is to split the video into shots for further processing.

2.1 Video Shot Transition Detection

Video can be viewed as a hierarchical structure consisting of scene, shot, and frame from top to bottom increasing in granularity, while a shot is a unit of the video that can be used for analysis. A shot is a sequence of frames captured by the single camera action, without any break/interruption. If an interruption occurs between shots, it is called as shot transition. Shot transition detection is the primary issue in video analysis [9], as it is intrinsically and inextricably linked to the way the video is produced. Shot gives a condensed representation of the video which in turn support the extraction of characteristics of either independent frames or short sequences in a video.
Modern video is made up of different kinds of shot transitions. If the transition occurs suddenly between two consecutive frames it corresponds to abrupt transition; whereas if the transition sustains through some frames in the video sequence it is said to be a gradual transition. The basic idea of shot transition detection is to find the discontinuity in visual content. The literature reports many techniques for detecting the shot transitions. But, automatic shot transition detection is difficult, since any kind of shot transition can be easily mistaken with camera and object motion which occurs in a video often. Particularly, detecting

gradual transition is more challenging. However, recently [10] proposed a method to distinguish gradual transitions and camera motions effectively.
A survey of techniques on recent literature can be seen in [11]. Video shot is an affluent source of information that captures the basic story line of an entire video. Suitable identification, illustration, and continuity of this information are needed for further analysis. In that rationale, a feature is a descriptive aspect extracted from the frames. Therefore the main objective is to:

—Extract feature from each frame that gives motion invariant representation, and competent enough to capture the global information in a shot with less computational complexity. Each shot in a video sequence explains a distinct scene in the video. For segmenting the shots, the prominent characteristics that are representative of the shots must be assembled together. Thus, the second objective is to:

—Group the frames in accordance to the individual shot semantics, to identify transition, by capturing the distribution of features over the entire video sequence. While trying to capture the continuity of the frames, if discontinuity is identified, the next step is to analyze whether the discontinuity corresponds to a shot transition or not. The literature reports techniques such as fixed thresholds [12], statistical detection methods [12] and adaptive thresholds [13] for measuring the discontinuities. Thus the third objective is to:

—Formulate a decision technique to examine whether a shot transition actually occur at the identified frame or not. The final complication and the most challenging task to handle is the fact that not all shot transitions are abrupt. Eventually, it is evident that gradual transition occurs over multiple frames. Hence, the final objective is to:

—Determine the type of transition, whether it is abrupt or gradual.

2.2 Video Indexing

Video indexing is the process of attaching concept terms to segments in a video. It enables the user to access video based on the requirement. An index can be created using any one of the modalities. For example, names of actor or players can be used to build the index, while indexing a movie video or a sports video. Indexing the huge volume of video data manually is an extremely slow and tiresome process even with coarse grain indexing. With finer granularity, the cost of manual indexing is still prohibitive. Thus, automatic video indexing is highly essential in the current developing environment for effectively acquiring the required video information.
Present work aims at building an index at the shot level. The basic task of shot indexing is video segmentation which segments the video into shots. The frames in a video shot explain its salient content. Including all the frames in a shot is again a tedious and time consuming task. If the index is to be created at the shot level, key frames which are representative frames of a shot can be utilized for building index. Thus, the objective of the approach is to:

—Extract the key frame from a shot that distinctly explicates the semantic content of the shot sequence. Video information is a sequence of shots. For indexing purpose, shots having similar semantic content are to be grouped together. Thus, the next step is to capture the distribution of feature vectors extracted from the shots in a video sequence. Therefore, the second objective is to:

—Construct a shot specific model to represent the distribution of features over the entire shot and to identify clusters of the feature points level by level in the indexing procedure.

2.3 Video Retrieval

Visual information retrieval creates unique challenges due to the multimodality, visual complexity and its dynamic behavior. For efficient searching and locating of required video data, an index that describes the video content is essential. The index produced, greatly improves the feasibility of searching a video data. Such index tries to match the search-by-keyword paradigm, the method popularly used by the users. The keyword queries are found to be effective, while searching for generic objects or scenes (e.g., car, green valley). Using text queries alone is not sufficient while retrieving the multimodal visual information. Thus, image based similarity measures emerged and retrieval is performed using the feature vectors extracted from the visual media, primarily using the key frames to build the index.
While the research on video analysis extract low level features that represent the content in a video, the retrieval process aims to get a high level query from a human user. Therefore, retrieval approach is subjective to how human beings interpret the visual data. The low level feature cannot capture the high level semantics entrenched in user query. Building the semantic gap between the high level query from a human and the low level feature extracted is the main challenge in the video retrieval process. The objectives are,

—A general feature which is not domain specific and capable of capturing the semantic content in a shot need to be developed. Retrieval methods are heavily dependent on the user query. Current techniques largely use keyword query. Searching the multimedia content using a keyword is unrealistic.

—A query paradigm that correlates the user requirement and the shot content need to be developed. Generally, feature based-indexing supports query based retrieval by image/frame similarity. For example, given a user query, the retrieval system should search the indexed shots which have the closest characteristics with that of the user query.

—Use a similarity measure to perform a similarity search for retrieving the shot having the similar attributes with that of the user query.

3. UNSUPERVISED SHOT TRANSITION DETECTION

For detecting a shot transition (ST), the analysis focuses on the characteristics of adjacent frames. Hence, it is clear that the model applicable here must be competent enough to uncover the variations that exist between the consecutive video frames. Hidden Markov Model(HMM) is good at capturing the temporal behavior but, it needs a huge volume of training data and loses optimization information. Training by optimization over the entire pattern space gives better discriminative power to the models since the models now learn patterns that need to be discriminated. Support vector machine (SVM) is found to be effective at this type of learning since training involves optimization over entire pattern in linear space. But, video information is made up of shots and is viewed as a non-linear model. Gaussian Mixture Model(GMM) can be used to confine the distribution of the data, but the components are considered to be Gaussian and the number of mixture components is also set in advance. Autoassociative Neural Network(AANN) captures the distribution of the data points depending on the constraints imposed by the structure of the network; just as the number of mixtures and Gaussian functions do in the case of GMM and hence are exploited in this work.

3.1 Autoassociative Neural Networks

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and

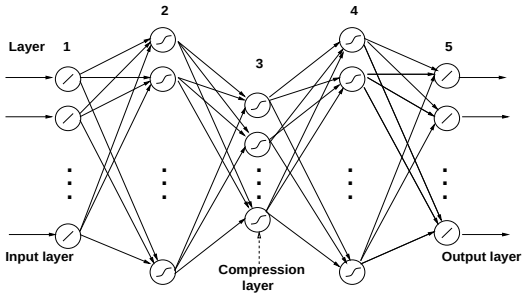

Fig. 1. A five layer AANN model.

are used to capture the distribution of the input data. The distribution capturing ability of the AANN model is described in this section. Consider the five layer AANN model shown in Fig. 1, which has three hidden layers. In this network, the second and fourth layers have more units than the input layer. The third layer has fewer units than the first or fifth. The processing units in the first and third hidden layer are nonlinear, and the units in the second compression/hidden layer can be linear or nonlinear. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hyper surface obtained by the projection onto the lower dimensional space.

Artificial Neural Networks has been used successfully in many research fields for solving problems which require learning and classification capabilities in a non-linear space. Finding the information that discriminates two shots in a set of correlated video segment is a complex non-linear problem to solve. Obviously, the video frames that correspond to a shot have unique characteristics. If the core information can be incorporated in to the network input variables, then the unique and identical characteristics can be captured by the subspace of the AANN embodied by the transformation at the hidden layers. Once the AANN is trained with the video frames, the frames that shares the identical characteristics will result in a minimum error at the output layer, which shows that, the frames are from one single shot. Meanwhile, if the frames from two different shots are trained, the result will have a large error at the output layer, showing the existence of a shot transition. Thus, with an appropriate threshold, AANN can be used to detect the shot transitions.

3.2 Autoassociative Neural Network Misclustering Rate (AMR) Algorithm

The five layer autoassociative neural network model as shown in Fig. 1, is used to capture the distribution of the features. The structure of the AANN model used in this work is 64L 90N 10N 90N 64L for capturing the distribution, where L denotes a linear unit, and N denotes a nonlinear unit. The nonlinear units use $tanh(a)$ as the activation function, where a is the activation value of the unit. The back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

3.2.1 **AMR Algorithm:**. The AMR algorithm for detecting STs is summarized as follows:

(1) Histogram features are extracted from each frame in the video.

(2) In a video sequence, consider a slide window of size w. Select a new frame p in the window w, where p is the frame that is the center of the slide window w. Selection of p splits the window w into two equal sized windows w_1 and w_2 of size $w/2$ as shown in Fig. 2.

(3) Histogram features of the video frames from the window w_1 are used to train the AANN model. The features from the window w_2 are fed to the AANN model, for testing the presence of shot transition.

(4) The output of the model is compared with the input, to compute the normalized squared error. The normalized squared error (e) for the feature vector ($\boldsymbol{y}$) is given by,

$$e = \frac{\|\boldsymbol{y} - \boldsymbol{o}\|^2}{\|\boldsymbol{y}\|^2} \quad (1)$$

where $\boldsymbol{o}$ is the output vector given by the model. The error (e) is transformed into a confidence score (c) using c = exp(-e). The average confidence score is calculated for the trained features.

(5) If the average confidence score from the AANN model is less than a threshold, this indicates that p^{th} frame position is the shot transition point.

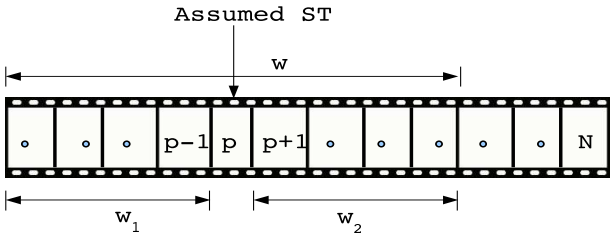

Fig. 2. Proposed slide window approach: Two windows w_1 and w_2 split by assumed shot transition at p.

The novelty of the proposed approach lies in the fact that, it identifies the presence of a ST by comparing the output of the model with the input and does not need any training. Thus, it can be argued that the proposed AMR algorithm does not require training video samples beforehand and can be said to be unsupervised. The work flow of the AMR shot transition detection algorithm is illustrated in Fig. 3.

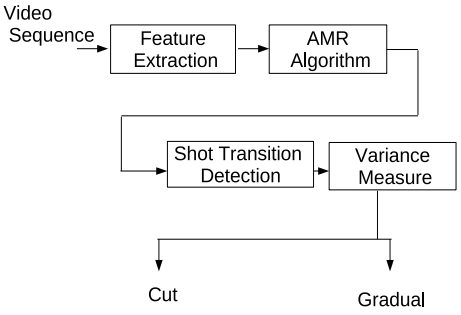

Fig. 3. Workflow of the shot transition detection method.

AANN performs an identity mapping of the input space and it is used to capture the distribution of the input data [14]. For example, initially if there are 25 frames in the slide window w, if a frame p is considered as the center frame which is assumed to be the shot transition point, this splits the original window w into two windows w_1 and w_2, each with 12 frames. AANN is good at capturing the distribution of features [14]. If the features in these two windows come from two different shots, the distribution created by this assumption will have significant differences and hence AANN cannot cluster these data properly. Conversely, if the distribution created by this assumption comes from one single shot AANN can effectively cluster these data. The approach is very effective in detecting shot transition in a video data of less than 2 seconds duration. The pseudocode for AMR shot transition detection algorithm is given in Section. *3.2.2*.

3.2.2 Pseudocode of AMR Algorithm

(1) *Input Video sequence*
(2) *Extract histogram features*
(3) *Consider a set of frames that form a window of size w*
(4) *Select a frame p which is the center frame of window w, where p is the assumed shot transition point*
(5) *p splits the window w into two windows w_1 and w_2 of equal size, $w/2$*
(6) *Features from w_1 are used for training the AANN model*
(7) *Features in window w_2 are fed to the AANN for testing whether the assumed p is actually a shot transition or not, using a threshold value*
(8) *Use variance measure to determine the type of ST*
(9) *Window w is moved one frame rightwards and the AMR algorithm is iteratively performed from step 4.*

As seen in Fig. 2 'hypothetical ST' is assumed at the center frame p (the middle point of two adjacent windows) and tested to determine whether it matches with the characteristics of a ST. After one 'hypothetical ST' has been tested, the window w is moved one frame to the right. That is why it is named as, 'slide window'. After moving, $(p+1)^{th}$ frame is assumed as the shot transition point at the next iteration. At this stage, the window w_1 have frames from 2 to p and $(p+1)^{th}$ frame is assumed as the shot transition point. Window w_2 have frames from $(p+2)$, and step 3 of the AMR algorithm shown in Section. 3.2.2 begins again. These steps are iteratively performed until the rightward window reaches the end of the video sequence. This technique affords autonomous training and AMR computation for every two consecutive windows circumvents the error-broadcasting problem.

4. INDEXING AND RETRIEVAL USING *K*-MEANS CLUSTERING

The enormous availability of video information makes it a tedious procedure to annotate and index manually by fast forward or rewind. Clustering is applied to group the shots to form an index. The proposed shot retrieval approach consists of the following modules:

—**Shot Indexing:** *An index has been build by extracting keyframes from each shot.*
—**Shot Retrieval:** *Enables a specific shot to be retrieved from the video database. User can choose any one query frame, and can playback the desired shot.*

4.1 *k*-Means Clustering

k-means clustering algorithm is used to classify or to group the objects based on attributes/features into k clusters, where k is positive integer number. It aims at classifying data items into a fixed number of clusters starting from an initial partition. Grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. k-means algorithm, the first one of the clustering algorithms proposed, is based the following idea:
Assignment step: *Given a set of initial clusters, each point is assigned to one of them.*
Refitting step: *Each cluster center is replaced by the mean point on the respective cluster.*
The indexing procedure is adopted to index the shots detected using the AMR shot detection algorithm discussed in the previous section. The flowchart of the indexing procedure used in the present work is shown in Fig. 4. For indexing with k-means clustering algorithm, video is segmented into shots and each shot is represented by keyframes. In this paper, the first frame of every shot is considered as the keyframe of the respective shot. The step by step procedure is given below.

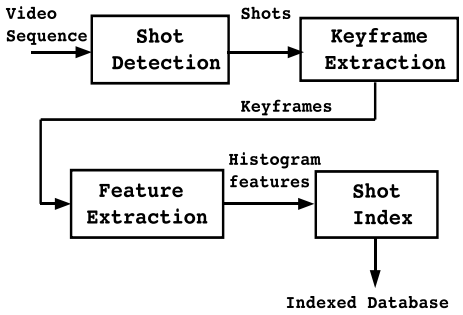

Fig. 4. Indexing procedure.

—*Histogram is extracted from the keyframes.*
—*k-means algorithm is applied to form shot clusters. Mean centers of these clusters are obtained.*
—*Euclidean distance is used to find the distance of each of the feature vectors extracted from the keyframes with the mean centers.*
—*The minimum distance obtained gives the cluster to which does the feature vector belong.*

For retrieving the required video shot, given a query frame, histogram is extracted from the query frame. Euclidean distance is used as a similarity measure to match the query frame with the keyframes. The retrieval procedure applied in this work is shown in Fig. 5.

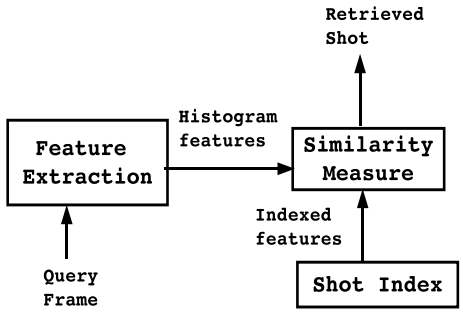

Fig. 5. Retrieval procedure.

5. EXPERIMENTAL RESULTS

The performance of the proposed AMR shot transition detection algorithm is evaluated on a database as seen in Table 1. The video clips are captured at the rate of 25 frames per second at 320×240 resolution. For analysis, 5 popular video games like cartoon, sports, commercials, news and serials are utilised.

5.1 Exploring the Best Parameters

Extensive experiments have been conducted to decide the best parameters of the proposed algorithm.

—**Window Size:** *Window size plays a major role in STD. Various window sizes of 25, 50 and 75 frames are used in the experiments. The analysis is curtailed at window size of 75, as more than one ST is found in most cases in the database for the window size of 100.*
—**Number of Bins:** *In the histogram feature extraction, the features are extracted using various bin values. b dimensional feature vectors extracted using b bins, obtained by varying the*

Table 1. Video clips used for shot transition detection experiments.

	Duration (min)	No. of Cuts	No. of Graduals
Cartoons	66	229	32
Sports	52	177	75
Commercials	55	131	84
News	59	289	23
Serials	68	148	54
Total	300	974	268

values of b as 8, 27, 64 and 125 are used to evaluate the performance of the proposed algorithm.

—**STD Threshold:** *Experiments are carried out to examine the effect of various threshold values. The value that significantly reduces the number of false alarms and mis-detections is set as the threshold value. In the present work, 0.8 is fixed as the threshold value.*

—**AANN Network Structure:** *The structure of AANN model used is (64L 90N 10N 90N 64L), where, L - Linear unit, N-Nonlinear unit ($tanh(a)$), and integers are varied to evaluate the performance. It is seen that the network structure (64L 90N 10N 90N 64L) gives optimal performance.*

—**Epochs:** *Epochs of the AANN model is fixed to 100 after examining its performance with the network structure.*

5.2 Performance Evaluation

To evaluate the performance of the algorithm, the approach uses precision(P) and recall(R). as defined by (2).

$$Recall(R) = \frac{T_c}{T_c + T_m} \qquad Precision(P) = \frac{T_c}{T_c + T_f} \tag{2}$$

where, T_c indicate the number of shot transitions correctly detected, T_m indicate the number of mis-detections and T_f indicate the number of false detections.
Recall indicates how many shot transitions (cut or gradual) are detected by the algorithm among all the transitions present. Precision indicates how many true transitions (cut or gradual) among all the transitions are detected by the algorithm. A good STD algorithm should show high scores for both precision and recall values. To simultaneously assess the number of false alarms and mis-detections, the present work used F-score measure with geometric mean as defined by (6) as the combined measure of precision(P) and recall(R).

$$F_\alpha = \frac{2PR}{P + R} \tag{3}$$

where α is weighting factor. For harmonic mean of P and R, α=0.5 is used.
For conducting experiments, b dimensional histogram features are extracted using b bins by varying the values of b as 8, 27, 64 and 125. High F-score is obtained with 64 bins with a window size of 50 as shown in Fig. 6. Window size of 50 gives better performance since in most cases, window size of 25 does not report any ST. And, window size of 75 reported more than one ST in a few instances. The window size of 50 with 64 bins appreciably reduced the number of false alarms and mis-detections.
False alarms and mis-detections relative to window size and number of bins are shown in Fig. 7 and Fig. 8. The experiments conducted with a threshold value of 0.8 appreciably reduced the number of false alarms and mis-detections. Consequently, this study used a window size of 50 with 64 bins and a threshold of 0.8 for optimal ST detection.

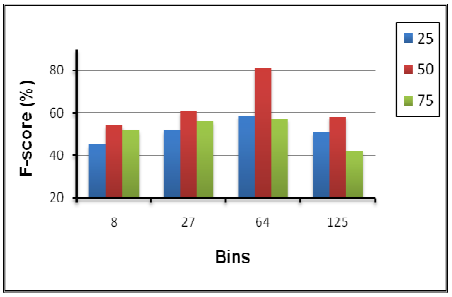

Fig. 6. F-Score

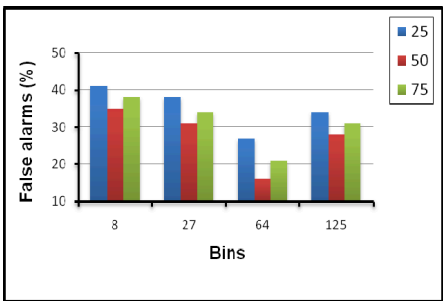

Fig. 7. False Alarms

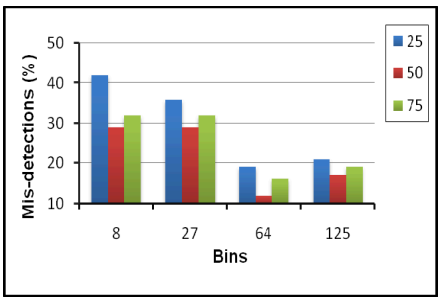

Fig. 8. Mis-detections

5.3 Exploring the Shot Transition Types

Once, the shot transitions are detected using the STD, the subsequent task is to classify the ST whether the detected ST is abrupt or gradual transition. Comparatively, detecting gradual change is difficult than an abrupt cut. The reason is that, a cut shows sharp changes between the consecutive frames, whereas a gradual transition takes place over a number of frames in a video. Though many different types of gradual transitions are found, the proposed approach only aims at identifying whether the detected shot transition is abrupt or gradual transition.
An example of confidence scores are obtained from the AANN model is shown in Fig. 9. As described, if the frames used for testing the presence of a ST comes from the same shot, AANN

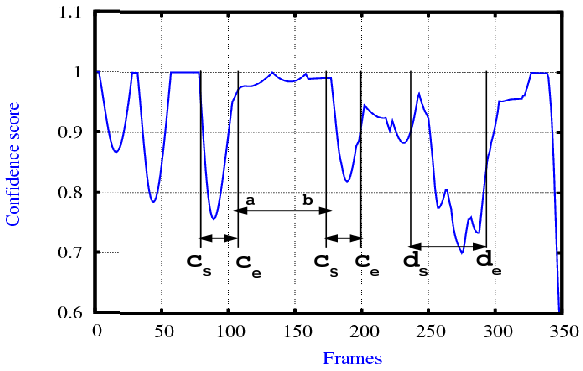

Fig. 9. Variance curve over cut and gradual frames. a to b - no transition. c_s to c_e - cut transition. d_s to d_e - gradual transition.

model can effectively cluster them under one cluster. There will be no significant difference in the obtained confidence scores as seen from a to b in Fig. 9. Rather, if the frames are from two different shots, AANN cannot cluster these frames properly and the obtained confidence scores show significant variations as seen from c_s to c_e and d_s to d_e in Fig. 9, where, c_s and c_e represents the start and end time of a cut transition and d_s to d_e represents the start and end time of a gradual transition. Further, confidence score is inversely proportional to error e, given by c = exp(-e). Hence, under dissimilar values, towards the ST point, confidence score(c), drops to a minimum value thus giving maximum error which indicates that AANN is not capable of capturing the distribution of features under this condition. Thus, STs are identified at minimum confidence score locations as seen in Fig. 9. The assumed ST point is compared with the identified ST along with the threshold value. For declaring ST as the actual transition, a deviation of 0.2 second (approximately 5 frames on either side) is admitted.

After detecting the locations of the STs, the next step is to classify the detected STs, whether abrupt or gradual. The difficulty increases with issues like lighting effects, which results in false positives. Unfortunately, imposing more harder constraints for removing these false positives may as well eliminate the actual gradual transitions. In the proposed approach, for estimating the type of transition, a range of values for which a transition has been already detected are examined using variance measure, along with their ground truth information. Variance is calculated using (4) for the range of scores obtained for the window under consideration, to decide the type of transition. Transitiions reporting high variance are considered as gradual transition.

$$Variance = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{N} \quad (4)$$

where, N is the total no. of frames from c_s to c_e or d_s to d_e, X_i is the i^{th} confidence score and $\bar{X}$ is the mean.

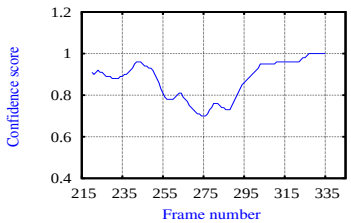

Fig. 10. Gradual transition taken from Fig. 9

The abrupt changes are detected as narrow fall downs, and gradual changes form a parabolic shape as shown in Fig. 9. As observed, the proposed technique is able to detect a cut using the shot length of less than 2 second duration. Further, the variance differences of the start frame at d_s and end frame at d_e are higher than the other frames for a gradual transition. Moreover, the gradual curve extracted from Fig. 9 as shown in Fig. 10 is analysed further for the semantic interpretation of the shot transition. The interpretations obtained is shown in Fig. 11 which demonstrates the efficiency of the approach in detecting the gradual transition. An example of the wipe and dissolve transitions detected as gradual transitions by the AMR algorithm is shown in Fig. 12 and 13.

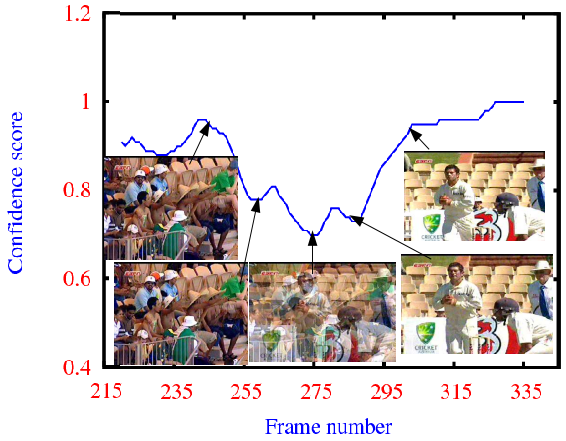

Fig. 11. Semantic interpretation of dissolve transition for the gradual curve
shown in Fig. 10.

Fig. 12. Wipe transition detected by AMR algorithm.

Fig. 13. Dissolve transition detected by AMR algorithm.

In order to demonstrate the supremacy of the proposed approach, the STD algorithm is applied on a test set shown in Table. 1. The database has a total of 974 cuts and 268 gradual transitions. The experimental results are tabulated in Table. 2, which shows that the proposed approach has a higher recall and precision for the database used, that is very much essential for a shot transition detection algorithm.

Table 2. Performance of AMR shot transition detection algorithm using video clips shown in Table 1.

	T_c		T_m		T_f		R	P	Fα
	cut	grad	cut	grad	cut	grad	(%)	(%)	(%)
Cartoons	212	21	9	6	8	5	95.92	96.36	96.14
Sports	162	64	5	7	10	4	97.00	94.19	95.57
Comm	122	76	3	2	6	6	97.60	95.31	96.44
News	276	14	7	5	6	4	97.52	97.87	97.69
Serials	134	42	6	9	8	3	95.71	94.37	95.03
Total	906	217	30	29	38	22	483.75	478.10	480.87
Average							96.75	95.62	96.17

5.4 Retrieval of video shots using k-means clustering algorithm

For the retrieval problem, Euclidean distance is used as the similarity measure to match the query frame with the representative keyframes of the shots. The approach used recall, precision and F-measure metrics to evaluate the system. Manually, how many of the shots in the database are relevant to the query is decided at first. The keyframes used and the performance of the k-means algorithm for each of the genres obtained are tabulated in Table. 3.

The performance of the approaches in retrieving the required shot is further estimated by gradually increasing the duration of the video. This helps to analyse the effect of the system with rise in the size of the database. The analysis shows that, k-means algorithm gives a success rate of 85.79 %, while 4.22 % of the shots are falsely detected and 9.99% of the shots are mis-detected as seen in Fig. 14. The retrieval performance of the approach for the all 5 genres considered is shown in Fig. 15 with increase in the duration of video. This shows that index created using k-means algorithm is showing good performance with increase in the size of the database.

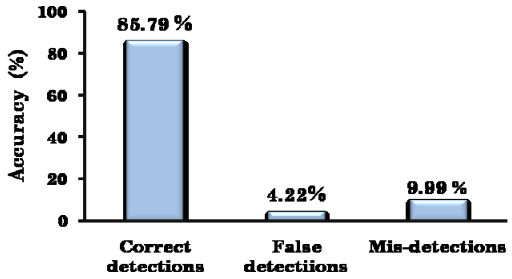

Fig. 14. Percentage of correct, false and mis-detections obtained with k-means Algorithm.

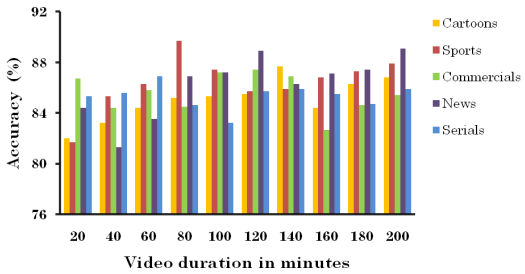

Fig. 15. Retrieval performance of k-means algorithm of 5 genres for different video duration .

5.5 Comparative Study

In this work, indexing is formed using k-means clustering algorithm and Gaussian mixture model (GMM). The performance of the approach is compared with a total of 8631 keyframes extracted from the shots. A comparison of retrieval performance obtained is tabulated in Table. 3. Further, the retrieval performance of GMM for the all 5 genres considered is shown in Fig. 16 with increase in the duration of video. This shows that index created using k-means algorithm and GMM are showing comparative performance with increase in the size of the database.

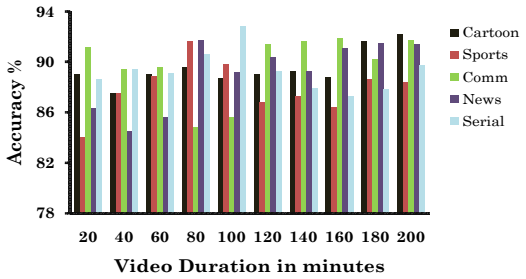

Fig. 16. Retrieval Performance of GMM Model for 5 genres

The time consumption is a major issue in any indexing and retrieval system. Hence, the time consumption of the proposed approach for indexing and retrieval using k-means algorithm is investigated on a PC with Intel PIV 2.53 GHz. processor by varying the size of the database. The time consumption includes feature extraction and indexing steps in the retrieval procedure. The evaluation shows an average computation time of 36 msec with approximately 1000 shots in the database, which is equivalent to about 2 hours of video data. Moreover, increasing the size of the database to 2000 shots increased the computation time by only 5 msec and if the size is 3000, the time linearly increased by 10 msec. Thus the proposed system is scalable.

6. CONCLUSION

The present work proposed an unsupervised algorithm for detecting ST based on a novel AMR algorithm for retrieving video shots. Histograms were extracted as features. The existing methods required a huge amount of training and testing data separately for reliable shot transition detection, whereas, the proposed AMR algorithm entailed only 2 sec. of video for effective performance. Apart from detecting the shot transitions, the approach was able to distinguish abrupt and gradual transitions using variance measure. The approach further uses k-means algorithm for retrieving the required shots. The method shows a better performance when compared to the existing methods. The future work focus on including other modalities of video like audio and text information with the inclusion of more number of genres. Further, furture work also aims at detecting various types

Table 3. Precision-recall obtained with k-means clustering algorithm and F_α for k-means and GMM.

Class	No. of Keyframes	Precision (%)	Recall (%)	F_α for k-means (%)	F_α for GMM (%)
Cartoons	1887	88.25	91.31	89.75	88.47
Sports	1643	84.32	88.17	86.20	85.23
Commercials	1724	87.47	90.72	89.06	88.92
News	1851	86.06	91.53	88.71	87.36
Serials	1526	85.73	87.62	86.66	85.93
Average	1726	86.36	89.87	88.08	87.18

of dissolve transitions that are commonly seen nowadays due to advancement in the modern digital technology.

7. REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval", IEEE Multimedia, Vol. 9, No.3, pp. 42-55, July 2002.

[2] Z.-N. Li, X. Zhong, and M.S. Drew, "Spatial temporal joint probability images for video segmentation", Pattern Recognition, vol. 35, no. 9, pp. 1847-1867, Sep. 2002.

[3] W.K. Li and S.H. Lai, "Storage and retrieval for media databases", Proceedings of SPIE, vol. 5021, pp. 264-271, Jan. 2003.

[4] R. Lienhart, "Reliable transition detection in videos: A survey and practitioners guide", Image Graphics, Vol. 1, No. 3, pp. 469-486, Sept. 2001.

[5] Mohanta, P.P, Saha. S.K, Chanda, B, "A Model-Based Shot Boundary Detection Technique Using Frame Transition Parameters" IEEE Transactions on Multimedia, Volume: 14 , Issue: 1, pp.223-233, 2012.

[6] A. Hampapur, R. Jain, and T. Weymouth, "Feature based digital video indexing", Proceedings of Third Working Conference on Visual Database Systems, Lausanne, Switzerland, pp. 115 - 141, 1997.

[7] C.G.M. Snoek, M. Worring, "Multimodal Video Indexing: A Review of the State of- the-art", Multimedia Tools and Applications, Vol. 25, No. 1, pp. 5 - 35, 2005.

[8] A. Hampapur, A. Gupta, B. Horowitz, C.-F. Shu, C. Fuller, J. R. Bach, M. Gorkani, R. Jain, "Virage video engine", Proceedings of SPIE Storage and Retrieval for Image and Video Databases V, San Jose, CA, USA, Vol. 3022, pp. 188-198, 1997.

[9] C.W.Ngo, H.J.Zhang, T.C.Pong, "Recent Advances in Content-based Video Analysis", International Journal of Image and Graphics, Vol. 1, No.3, pp. 445-468, Dec.2001.

[10] Xiang Fu, Jie-xian Zeng, "An Effective Video Shot Boundary Detection Method Based on the Local Color Features of Interest Points", Proceedings of Second International Symposium on Electronic Commerce and Security, Vol. 2, pp. 25 - 28, May 2009.

[11] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey", Signal Processing: Image Communication, Vol. 16, pp. 477-500, Jan. 2001.

[12] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas, "Video Shot Detection and Condensed Representation-A review", IEEE Signal Processing Magazine, pp. 28- 37, March 2006.

[13] Z. Cernekova, C. Kotropoulos, and I. Pitas, "Video shot segmentation using singular value decomposition", Proceedings of IEEE Int. Conf. Multimedia and Expo, Baltimore, Maryland, Vol. 2, pp. 301-302, 2003.

[14] B. Yegnanarayana and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition", Neural Networks, vol. 15, pp. 459-469, Jan. 2002.