Scene Detection In Hollywood Movies and TV Shows

Zeeshan Rasheed and Mubarak Shah
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, Fl 32816
zrasheed,shah@cs.ucf.edu

February 27, 2003

Abstract

A scene can be defined as one of the subdivisions of a play in which the setting is fixed, or when it presents continuous action in one place. We propose a novel two-pass algorithm for scene boundary detection which utilizes the motion content, shot length and color properties of shots as the features. In our approach, shots are first clustered by computing Backward Shot Coherence (BSC); a shot color similarity measure that detects Potential Scene Boundaries (PSBs) in the videos. In the second pass we compute Scene Dynamics (SD), a function of shot length and the motion content in the potential scenes. In this pass, a scene merging criteria has been developed to remove weak PSBs in order to reduce over segmentation. We also propose a method to describe the content of each scene by selecting one representative image. The segmentation of video data into number of scenes facilitates an improved browsing of videos in electronic form, such as video on demand, digital libraries, Internet. The proposed algorithm has been tested on a variety of videos that include, five Hollywood movies, one sitcom, and one interview program and promising results have been obtained.

1. Introduction

The availability of audio visual data in the digital format is increasing every day. This information includes documents, audio-visual presentations, home made videos and professionally created contents such as sitcoms, TV dramas and feature movies. Movies alone constitute a large portion of the entertainment industry. Every year around 4,500 motion pictures are released around the world spanning over approximately 9,000 hours of video (Wactlar [1]). With the digital technology getting inexpensive and popular, there has been a tremendous increase in the availability of videos through cable and Internet such as *video on demand*. For feasible access to this huge amount of data, there is a great need to annotate and organize this data and provide efficient tools for browsing and retrieving contents of interest.

Digital video is a rich medium as compared to text material. There could be many possible ways to index it. A basic approach of video annotation is to detect the shots and use a set of key frames to represent the shot content. The second level of abstraction could be to combine similar shots together and form *scenes* or *story units*. The organization of videos in this fashion is more meaningful than presenting the shots alone. Recently, DVDs are being made with options to view a particular scene in the movie. To obtain such a representation, a human observer is required to sequentially watch the video and locate the important boundaries or *scene edges*. However, a manual content analysis is not feasible for huge amount of data as it is slow as well as expensive.

A large amount of work has been reported to structure videos resulting in several interactive tools to provide navigation privileges to the viewers, [2, 5, 6, 8]. Yeo et al.[3] were the first ones to propose a graphical representation of video data by constructing a *Scene Transition Graph*. Each node in the graph represents a shot and edges represent the transitions within shots. The scene transition graph is then split into several sub graphs using *complete-link* method of hierarchical clustering such that each sub graph or *scene* satisfies a color similarity constraint. Hanjalic et al. [4] uses a similar approach of shot clustering and finds *logical story units* in MPEG compressed domain.

The graphical approach works well for videos which are shot in restricted environments (for example videos made inside a studio by stationary cameras including News videos, talk shows, game shows, sitcoms etc.). These programs are often shot with multiple cameras which switch back and forth, repeatedly showing the same contents within a scene. Feature movies, however, are often filmed in open and dynamic environments using moving cameras and have continuously changing contents. Furthermore, directors use different camera techniques and effects that make it difficult to create suitable graph for scenes. Therefore, two major problems are encountered:

• A false color match between shots of two different scenes may wrongly combine the scenes (and the intermediate scenes as well) into one segment causing an under segmentation.
• Action scenes may be broken into many scenes for not satisfying the color matching criterion producing an over segmentation.

Ngo et al. [7] proposed a motion based approach to represent shots and to cluster similar shots to form scenes. Spatio-temporal slices of video sequences are constructed and local orientation of pixels are computed using structure tensors. With this information each shot is represented by either one or more key frames or by constructing mosaics. Finally, shots are clustered into scenes by analyzing the histogram intersection. Adams et al. [9] present another way to parse videos by estimating *tempo* in the feature movies. Camera motion parameters and shot length are used to construct a *tempo* plot. The proposed method detects edges in the tempo function and identifies instances where the tempo of the movie changes with time. However, this method can not detect boundaries between the scenes in which the tempo is not changing.

In order to improve the scene segmentation, we believe that one must also incorporate other attributes of movies together with the color similarities. These features may include audio information, motion of the camera and objects, shot transition rate etc. The knowledge of post processing of these videos such as composing and editing techniques (referred as montage in film literature) can also be utilized to improve the segmentation task.

2. Proposed Approach

In Webster dictionary, a scene is defined as follows [10]:
• A subdivision of an act in a dramatic presentation in which the setting is fixed and the time continuous OR
• One of the subdivisions of a play; as a division of an act presenting continuous action in one place.

The first definition of *scene* emphasizes the fact that shots belonging to one scene are often taken with fixed physical settings. Several cameras capture the video with different angles while the background remains the same (for example, a scene filmed inside a studio). A characteristic of this category of scenes is their repetitive structure due to the switching between the same cameras with fixed view. However, this definition may not hold for all kinds of scenes. For example an outdoor scene, where background may change. Other examples are the scenes which are shot with the cameras mounted on trucks or a trolley. In this case, a scene may be defined by the continuity of ongoing actions performed by the actor(s).

Movie directors, while filming the scenes, also control the pace of film in order to sustain the viewers interest.

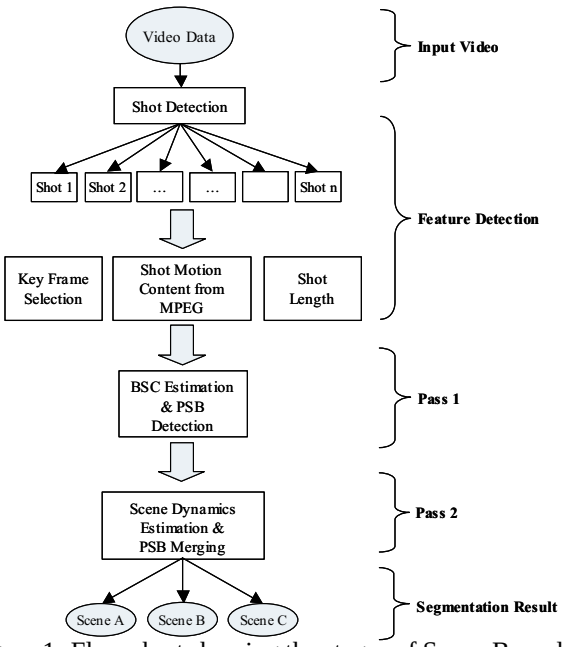

Figure 1: Flow chart showing the stages of Scene Boundary Detection algorithm.

Two important factors which have been known to influence the pace of a movie are the Montage (editing) and the motion, [13]. For a given scene, these factors are kept consistent such that the the viewer's attention is always engaged. We use this knowledge to develop a two-pass algorithm for scene boundary detection suitable for feature movies. The video is initially parsed into shots by camera break detection. Each shot is represented by one or more key frames depending upon the shot activity (Section 2.1). For each shot, its length and motion contents are also estimated as features (Section 2.2). In the pass one of our algorithm which is motivated by the first definition, a color similarity measure of shots is computed called *Backward Shot Coherence*(BSC). It describes how well a shot matches with the previously seen shots. We find valleys in BSC and detect several *Potential Scene Boundaries* (PSB). An *action* scene with changing contents may split into many scenes for not satisfying color similarities. To improve the segmentation, we merge scenes during the pass two by deleting weak PSBs. This is achieved by computing *Shot Dynamics* (SD) of each scene detected in the first pass. SD is a function of shot length and motion contents of the shots in a scene (Section 3). Figure 1 shows the flowchart of complete algorithm. We also propose a method to describe the contents of each scene by selecting one representative image (see Section 4). Section 5 discusses results of our algorithm obtained by processing 5 Hollywood movies, one episode of a sitcom and one hour of a famous talk show. Section 6 concludes this paper.

2.1. Shot Detection and Key Frame Selection

The video track is first divided into shots which is defined as a sequence of frames taken by a single camera. A histogram intersection technique has been used. A 16 bin HSV normalized color histogram is computed for each frame with 8 bins for Hue and 4 bins each for Saturation and Value [11]. Let f^x be the x^{th} frame of the video and $D(f^i, f^j)$ represents the intersection of histograms of frames i and j respectively, then:

$$D\left(f^i, f^j\right) = \sum_{b \in allbins} min\left(H_i\left(b\right), H_j\left(b\right)\right), \quad (1)$$

where H_i and H_j are the histograms of corresponding frames. A shot boundary is flagged when:

$$D(f^i, f^{i-1}) < T_{color}, \quad (2)$$

where T_{color} is a threshold that captures the allowed tolerance between color statistics of two shots. Let a and b are the indices of the first and the last frames of the i^{th} shot S_i is a set of frames, such that:

$$S_i = \left\{f^a, f^{a+1}, ..., f^b\right\}. \quad (3)$$

Each shot is represented by a set of key frames K_i, such that all frames are distinct. Initially, the middle frame of the shot is selected and added to the null set, K_i. Next, each frame within a shot is compared to every frame in the set K_i. If the frame differs from all previously chosen key frames by a fixed threshold, it is added in K_i, otherwise ignored. This algorithm can be summarized as:

STEP 1: Select middle frame as the first key frame
$K_i \leftarrow \{f^{\lfloor (a+b)/2 \rfloor}\}$

STEP 2: for $j = a$ to b
if $max\left(D\left(f^j, f^k\right)\right) < Th \quad \forall f^k \in K_i$
Then $K_i \leftarrow K_i \cup \{f^j\}$

where Th is the minimum frame similarity threshold. This approach selects multiple frames for the shots which have higher dynamics and temporally changing visual contents as compared to relatively static shots.

2.2. Motion Contents and Shot Length

We associate two features with each shot; *the shot length* and *the shot motion content*. These attributes of shots provide cues for the nature of the scene. Generally, dialogue shots span on large number of frames with relatively little camera/actor movement. On the other hand, shots of fight and chasing scenes change rapidly and the camera motion is jerky and haphazard with higher movements of actors (Arijon [12]).

2.2.1. Computation of Shot Motion Content

Motion in shots can be divided into two classes; *global motion* and *local motion*. Global motion in a shot occurs due to the movements of the camera. On the other hand, the local motion is caused by the relative movement of objects with respect to the camera. We define *shot motion content* as the amount of local motion in a shot. We exploit the encoded information in MPEG-1 compressed video. First, a global affine motion model is estimated by using a least square method on motion vectors decoded from MPEG file. An affine model with six parameters is represented as follows:

$$u = a_1 \cdot x + a_2 \cdot y + b_1$$
$$v = a_3 \cdot x + a_4 \cdot y + b_2, \quad (4)$$

where u and v are horizontal and vertical velocities respectively, $a_1 - a_4$ capture the camera rotation, shear and scaling, $b_1 - b_2$ represent the global translations in horizontal and vertical directions respectively, and $\{x, y\}$ are the coordinates of block's centroid. Once a global affine transformation is obtained, the velocities of blocks are reprojected and the goodness of the fit is measured by examining the difference between the actual and the reprojected velocities of the blocks. In case of global motion, the difference between the two is zero or very small. However, when the motion is not solely due to the camera and objects move relative to the camera, the motion vectors cannot be approximated by an affine model. Therefore, the magnitude of the error is utilized as a measure of shot motion content. Let u_k and v_k be the encoded velocities and u'_k and v'_k be the reprojected velocities of k^{th} block in j^{th} frame, then the error ϵ_j in the fit is measured as:

$$\epsilon_j = \sum_{k \in motionblocks} \sqrt{\left(u'_k - u_k\right)^2 + \left(v'_k - v_k\right)^2}. \quad (5)$$

The shot motion content of shot i is the aggregation of ϵ of all P frames in the shot:

$$SMC_i = \sum_{j \in S_i} \epsilon_j, \quad (6)$$

where SMC is the *shot motion content*. Figure 2 shows the motion content values for three different shots.

3. Scene Boundary Detection Algorithm

In this section, we discuss our two pass algorithm for video scene segmentation. In the first pass, *Potential Scene boundaries* are detected based on the color properties of shots. The second pass deals with the removal of weak scene boundaries by analyzing the shot length and motion contents of shots of potential scenes.

3.1. Pass One: *Color Similarity Analysis*

The first pass of the algorithm deals with the detection of *Potential Scene Boundaries* (PSBs) in the videos. We define a PSB as a possible instance of the beginning and/or

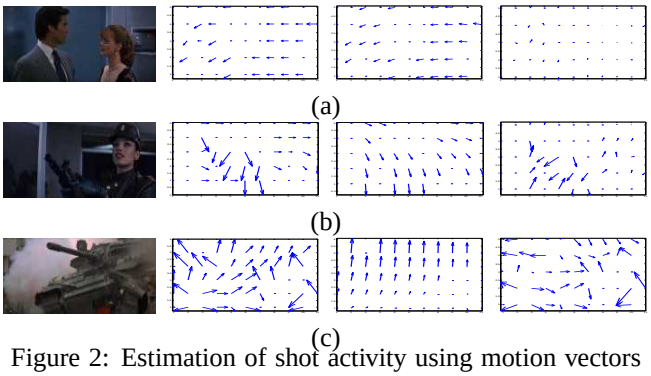

Figure 2: Estimation of shot activity using motion vectors for three P frames. Left to right: *P frames*, Motion vectors extracted from MPEG, Reprojected flow vectors, Difference between the original and reprojected flow vectors. The activity computed (a) 9.8, (b) 46.64 and (c) 107.03.

ending of a scene in a movie. This is achieved by estimating a feature for each shot, called *Backward Shot Coherence* (BSC); a similarity measure of a given shot with respect to the previously seen shots. We first compute the *shot coherence* of the shot i in a window of previous shots (say N) which is defined as the color similarity between two shots. Let SC_i^j expresses the shot coherence of shot i with shot j where shot i and j contain n and m key frames respectively, then:

$$SC_i^j = \max_{f^x \in K_i, f^y \in K_j} \left(D(f^x, f^y) \right). \tag{7}$$

Backward shot coherence for shot i is then computed by taking the maximum shot coherence in a window of length N, that is:

$$BSC_i = \max_{1 \le k \le N} \left(SC_i^{i-k} \right), \tag{8}$$

where BSC_i is the *shot coherence* of shot i. Figure 3 shows

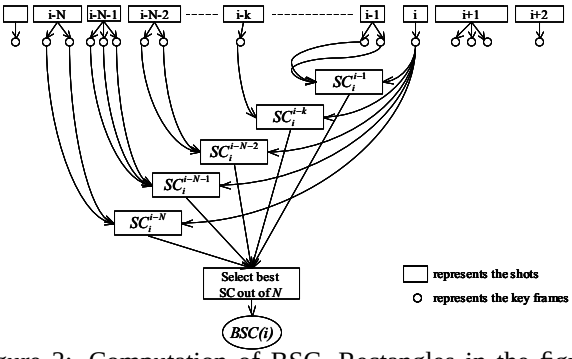

Figure 3: Computation of BSC. Rectangles in the figure represent shots and circles indicate key frames.

a graphical representation of this step. A scene is defined as a collection of contiguous shots in time which are taken at the same location and show similar visual content (according to the first definition, Section 2). In the beginning of a new scene, the initial shots do not resemble with the shots of the previous scene due to the dissimilarities in the physical settings. BSCs of these shots are very small and show a poor match with the shots of the previous scene. As the scene progresses, the shots are repeated. Similar contents are seen in shots and therefore BSCs of those shots attain higher values. This continues until the start of a new scene. The beginning of a new scene can be detected by locating valleys in the plot of BSC. We call these valleys *potential scene boundaries* (PSBs) as they are the candidates for the starting point of a new scene. In some cases, a PSB may be detected within a scene as an outlier. A shot in the middle of a scene which is unique for a given scene can cause a false valley in the BSC. For example, a flash back shot or a shot of a new actor entering into the scene may cause a false alarm. To suppress the outliers and prevent over segmentation in this phase, we compare the color attributes of key frames of neighboring potential scenes. If a pair of key frames of two adjacent potential scenes are found to be similar, then the scene boundary between the two is removed and the scenes are merged into one scene. Let k and $k+1$ be the two potential scenes, the PSB between the scenes will be removed if:

$$D\left(f^i, f^j\right) \ge T_{color}, \tag{9}$$

where $f^i \in Scene_k$ and $f^j \in Scene_{k+1}$. Figure 4(a) shows the plot for BSC of first 300 shots of the movie *Top Gun* for pass one. These shots span over five scenes as segmented by the human observer (See Table 2, scenes 1-5). Solid (green) vertical lines indicate the PSBs. Dotted (blue) vertical lines are the weak PSBs that were removed after adjacent scene merging by comparing key frames similarity, Eq. 9 (See attached media files for color images). Figure 4(b) shows the first key frame of some shots. Note that a transition from shot 9 to shot 10 results in a sharp valley as the latter shot was never shown before and hence a PSB is detected. Similarly a transition from shot 71 to 72 is also identified as a scene boundary. Several outliers were successfully removed by comparing the key frames of adjacent scene boundaries. Note that first three segment (solid lines at shot number 10, 68 and 72 in Figure 4(a)) correspond to the first three scenes in the ground truth. The fourth scene in the movie is about flying training of pilots which is broken into a large number of potential scene boundaries (solid blue lines from shot 76 to 273). This occurs due to the high dynamics and un-repetitive structure of the scene.

The computation of BSC is controlled by the selection of window size N. It can be considered as a memory parameter which mimics a human's ability to recall a shot seen in the past. When we watch a video, we relate the newest shot with the previous ones to determine if it is another shot of

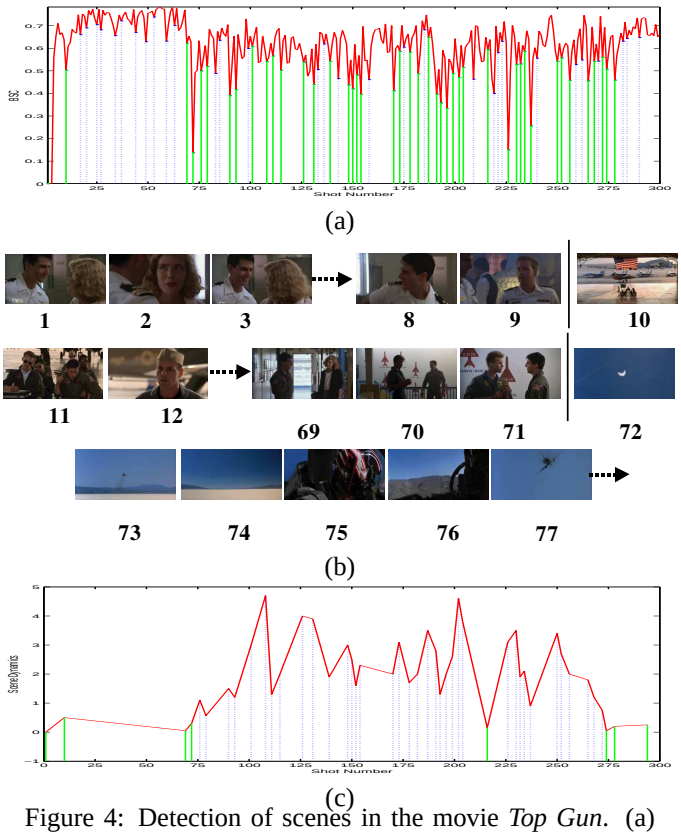

Figure 4: Detection of scenes in the movie *Top Gun*. (a) The plot of BSC is shown for 300 shots. (b) First key frame of each shot is shown with shot index. (c) Plot of *Scene Dynamics* of potential scenes detected in Pass one.

the continued scene or the first shot of a new scene. This judgement is limited by the ability to remember old shots. With the time, the probability of matching a newer shot with a very old shot is decreased. That is why we have put a constrain on the number of shots to be used for estimating BSC. However, the choice of N greatly affects the initial segmentation. If this value is too large, it may span over several scenes and a wrong estimate of BSC may be obtained. On the other hand, if N is very small, the shot may not be compared with sufficient number of shots within the scene. As a result, an over-segmentation of video may be obtained in this pass. We set $N = 10$ for our experiments and found this value to be appropriate for our data set.

3.2. Pass Two: *Scene Dynamics Analysis*

Most non-action scenes such as dialogue scenes with repetitive structure are well segmented during the first pass. However, scenes with weak structure are often broken in several scenes. In particular, action scenes are divided into several scenes due to non-repetitiveness of shots. The poor match among the shots causes an over segmentation of video. For a semantically meaningful segmentation, these potential scenes are needed to be merged together. The over segmentation in pass one implies that the use of only color information is not enough for an appropriate segmentation of videos. Therefore, we have incorporated the shot length and shot motion contents as useful features to analyze scene properties. A characteristic of such scenes is their high motion activity and small shot length. Therefore, a weight *Scene Dynamics* (SD), is computed for each potential scene as follows:

$$SD_i = \frac{\sum_{j \in Scene_i} SMC_j}{\sum_{j \in Scene_i} L_j}, \tag{10}$$

where SD is the Scene Dynamics of the scene i, SMC_j is the shot motion content of j^{th} shot in the scene and L_j is the length of corresponding shot. The large values of SMC and smaller values of L_j in dynamic scenes cause SD to be large. On the other hand, relatively calm scenes which span on numerous frames with small SMC returns a very small value. The scene dynamics of every pair of adjacent potential scenes is analyzed. The PSB between two consecutive scenes k and $k+1$ will be removed if SD of both scenes exceed a fixed threshold. See Figure 4(c) which shows the final scene boundaries for first 300 for *Top Gun*. Note that PSBs (blue dotted lines) were removed where the scene dynamics are relatively high for consecutive scenes (See attached media files for color images).

4. Scene Representation

A scene representation using one or multiple images is crucial for building an interactive tool for video browsing. In case of DVDs of Hollywood movies which are available with the chapters selection option, each chapter is represented by one key frame. The creators, who have complete access to the script of the movies, manually pick a frame that adequately reflects the scenario. Since this is a subjective process, the choices of frames may vary from individual to individual. However, the main objective of the key frame is to give a hint of the height of drama, suspense and/or action of the scene. In this section we address the issue of automatic selection of key frames for scenes. In our approach, we first compute a shot goodness measure as a function of shot coherence, shot length and shot activity. A shot is a good representative when:

- the shot is shown several times (higher SC with other shots),
- the shot spans over longer period of time (larger shot length) and,
- the shot has minimal motion content (smaller SMC).

We have further noticed by analyzing the key frames in DVDs that images with multiple faces are preferred over one face or with no face for scene representation. For example, in case of a scene where two actors are talking, a

frame showing both is chosen over frames with single person. Therefore, a fixed number of shots with the highest shot goodness value is selected as candidates for scene key frame. These shots are then tested for the presence of faces in the first key frames. The shot with the maximum number of faces is selected as the representative shot for the corresponding scene.

4.1. Measuring Shot Goodness

The shot goodness is computed by analyzing three properties of every shot which includes *Shot Coherence*, *Shot Length* and *Shot Activity*. For each shot in the scene, its coherence with every other shot is computed. For a scene with N shots, a correlation matrix of dimension $N \times N$ is constructed where element (i, j) is the coherence of shot i with shot j. When a shot is shown several times, the summation of column values is large. On the other hand, shots seen fewer times will have smaller values. Let $C(i)$ be the correlation sum of shot i, then:

$$C(i) = \sum_{j \in Scene} SC_i^j. \tag{11}$$

We associate a weight with each shot as follows:

$$W(i) = \frac{C^2(i) \times L_i}{log(SCM_i + \alpha)}, \tag{12}$$

where W is the shot goodness and α is used to prevent a division by zero. The squared value of $C(i)$ is used to give more emphasis to shot coherence, whereas the log for shot motion content is incorporated to reduce its effect on shot goodness. In our experiments the mean of SMC of all shots in a scene was used for α, i.e.

$$\alpha = \frac{1}{K} \sum_{j=1}^{K} SMC_j, \tag{13}$$

where K is the total number of shots in the scene. In the second step, three shots with the highest $W(i)$ are selected as candidate shots and face detection is performed.

4.2. Detection of Faces

Several face detection algorithms have been proposed in the literature,[14]. In case of shots obtained from video tracks of movies, we encountered faces with different levels of scale together with different orientations which makes difficult for face detector to perform with high accuracy. Therefore, we apply a simple but robust method of skin detection approach on the frames to detect faces.

We detect skin by using a method proposed by Kjedlsen et al. [15] that requires training on color space. The middle frame of candidate shots are tested for skin pixels. Each

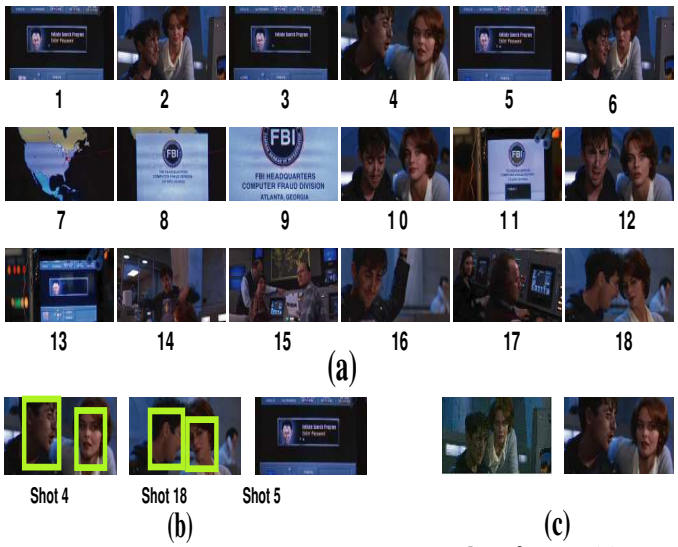

Figure 5: Scene representation using one key frame. (a) 18 key frames, one from each shot. (b) Three shots with highest shot goodness with bounding boxes showing the detection of faces. (c) Final result: left image is extracted from DVD, right image is selected automatically.

isolated segment of skin is considered as face and the frame with the highest votes is taken as the scene key frame. In the case of a tie or when no face is detected in any candidate key frame, the key frame of the shot with the highest goodness value, W, is selected. Figure 5 shows an example. This scene is taken from *Golden Eye* and consists of 18 shots in which Boris and Natalya are having a conversation about breaking into the FBI security system. The key frame of each shot is shown in Figure 5(a). Figure 5(b) shows three key frames with highest values of shot goodness (shot 4, 18 and 5 respectively). Note that shot 4, which is repeated in most of the shots, gets the highest weight. Using the skin detection method, two faces were detected in shots 4 and 18 and none in shot 5. Since shot goodness of shot 4 is higher than shot 18, it is chosen as the representative key frame for this scene. Figure 5(e)(left) is the image extracted from the DVD for this chapter. Compare the similarity of this image with the one on right as selected by our algorithm. Figure 6 shows some results of key frame detection for scene representation for movies (a)*Golden Eye* and (b)*Terminator 2*. Images on left columns are extracted from the DVD whereas frames automatically selected by our algorithm are shown in right columns. Multiple frames are shown when a scene is broken into more than one scene.

5. Experimental Results

We have experimented with video sequences from five Hollywood movies including *Terminator II, Golden Eye, Gone in 60 seconds, Top Gun,* and *A Beautiful Mind*. Each sample

of movie was 35-60 minutes long and taken from the middle of the movie. We have also experimented with one episode of a sitcom, *Seinfeld* (26 minutes of running time), and one complete show of *Larry King Live*. The videos were digitized at 29.97 fps. For each video, a human observer identified the scene boundaries as ground truth. Chapters information from the DVDs have also been incorporated for result evaluation. Table 1 summarizes the data set, the ground truth, and results obtained by our proposed method. This table also lists the number of false+ve and false-ve. To evaluate the performance, we have also listed the Recall and Precision figures for each video.

Table 2 provides detailed scene detection results for *Top Gun*. This video consisted of 89,999 frames (about 50 minutes of running time). Total number of shots found in the video was 1,103. First column in the table shows the Chapters from the DVD. Second column lists the titles for each scene segment identified by a human observer. It should be noted that DVD chapters are a sub set of scenes identified by human observer. Column 3 and 4 provide the number of shots and the number of frames in the segmented scenes. The last column indicates the number of scenes as detected by the algorithm. It should be noticed that the final scene boundaries obtained are more than the ground truth (Table 1). We believe that a slightly over segmentation is preferable over under-segmentation, since split scenes can be combined by further analysis. While browsing a video, it is better to have two segments of one scene rather than one segment consisting of two scenes. There are few missed scenes which are indicated with 'x' in the table. These are the scenes which were wrongly merged in the previous scene boundary and not identified. Please see attached media files for more results.

To demonstrate that the algorithm presented here works equally well on other video genres, we have also conducted an experiment on one sitcom show *Seinfeld* which belongs to the genre of video with mostly dialogues and very little action content in the shots. Table 1 also lists the scene detection of this show which demonstrates that it performs adequately for a very different genre of videos.

5.1. Scene Detection in Interview Shows

We have also run the same algorithm on one hour show of *Larry King Live* program which was digitized at 10 frames per seconds. The video consisted of 8 segments in which the guest was interviewed by the host. There were 7 segments of commercials between the interview segments. Similar to the movies, the segment of program showing the interview can be considered as one scene, whereas, commercials together can be considered as another scene. The algorithm proposed here worked very well in detecting the scene boundaries between the interview and commercials segments. Due to non repetitive structure of commercials, several clusters were found during pass one. However, in pass two, all commercials are combined together by computing their scene dynamics thus separating program segments from them. The overall result for this video can be found in Table 3. It is worth noting that small clips of news reels shown during the talk show became the part of the commercials.

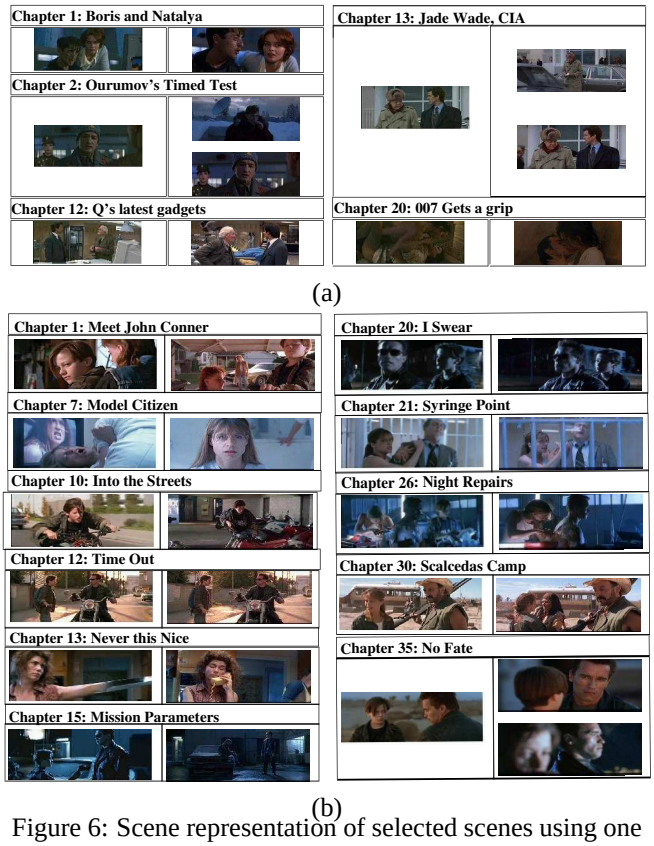

Figure 6: Scene representation of selected scenes using one key frame for (a)*Golden Eye* (b) *Terminator 2*. Images on the left column are the ones obtained from the DVD. Images on the right are the key frames selected by our algorithm. Multiple images have been shown for the chapters for which the scene is broken into multiple scenes by our algorithm.

Video	Duration	#Shots	G.Truth Scenes	Detected Scenes	False -ve	False +ve	Recall	Prec.
Terminator 2	55 min	1,632	36	38	5	7	86.1%	81.6%
Golden Eye	60 min	1,519	25	35	3	13	88.0%	62.9%
Gone in 60 sec.	58 min	1,869	39	43	6	10	84.6%	76.7%
Top Gun	50min	1,103	26	30	3	7	88.5%	76.7%
A Beautiful Mind	36 min	446	17	21	2	6	88.2%	71.4%
Seinfeld	21 min	318	22	27	3	8	86.4%	70.0%

Table 1: Summary of data set and experimental results for five Hollywood movies and one sitcom.

5.2. Scene Representation Results

Figure 6 shows some results for scene representation using a single key frame for *Golden Eye* and *Terminator 2*. For every DVD chapter, images from the DVD are shown in the left columns and images detected by our algorithm are shown in the right columns. Multiple key frames are shown

Scene Boundary detection				
DVD Chapter	**Human Observation**	**#shots**	**#frames**	**# Scenes Found**
Crash and Burn	First encounter with Charlie	9	1,198	1
Charlie	Charlie's presentation	59	5,452	1
N/A	After the presentation	3	1,911	1
Turn and Burn	Flying training	202	6,645	2
No Flexibility	The locker room	20	2,384	1
N/A	In the boss's office	25	4,869	1
Flying Against a Ghost	Maverick and Goose	23	2,920	2
Tempted	Charlie offers a dinner to Maverick	23	3,270	1
Playing with the Boys	Playing volleyball	53	2,695	3
No Apologies	Dinner at Charlie's home	65	9,808	1
N/A	In the elevator	30	3,325	1
N/A	Goose's family arrives	8	1,356	1
Textbook Maneuvers	Training session	60	6,101	2
N/A	In the bedroom	4	2,013	1
N/A	Charlie wakes up alone	2	670	1
The need for Speed	Flight competition	169	8,283	2
Your Attitude	The locker room	10	1,843	1
N/A	Maverick in the bed room	33	5,350	1
Great Balls of Fire	Dinner	x	x	x
Every Point Counts	Goose F-14 crashes	267	11,670	2
Let Him Go	Goose is dead	x	x	x
N/A	I will be here	11	2,071	1
N/A	Goose's room	8	1,401	1
N/A	Maverick and Goose's wife	17	4,399	1
N/A	In the court	x	x	x
Get Him Up Flying	Maverick in the F-14	2	365	1

Table 2: Scene boundary detection on 50 minutes of movie *Top Gun*. *x* shows the missed scene boundaries.

Scene Unit detection			
Human Observation	**# of Shots**	**# of Frames**	**# of Scenes**
Interview Segment 1	38	4,565	1
Commercials + News Reel	86	1,434	1
Interview Segment 2	34	4,377	1
News Reel + Commercials	72	1,593	1
Interview Segment 3	27	3,259	1
News Reel + Commercials + News Reel	65	1,098	1
Interview Segment 4	16	1,863	1
News Reel + Commercials + News Reel	19	779	1
Interview Segment 5	41	4198	1
Commercials 7	30	936	1
Interview Segment 6 + News Reel + Interview Segment	47	3,269	2
News Reel + Commercials	64	1,542	1
Interview Segment 7	20	2,132	1
Commercials	32	318	1
Interview Segment 8	59	3,215	1

Table 3: Scene boundary detection 60 minutes of *Larry King Live* show.

when a scene is split into more than one. In majority of scenes key frames found by the algorithm are very similar to those of DVDs. The selection of one key frame from a scene largely depends on the discretion and choice of the person. Therefore, different people can chose different frames for representing a scene. For example, in *Terminator 2*-Chapter 1 a close shot of John and his friend is shown in DVD. On the other, hand our algorithm selected a frame with three actors; which may be preferred over the original one, as it provides more details about the scene.

6. Conclusion

We presented a two pass algorithm to detect scene boundaries in videos. Our method utilizes several features of video which includes color similarity, shot activity and scene length to perform a higher level segmentation. We have used motion information from MPEG-1 compressed file and developed a method to compute shot activity using this information. We have also proposed a method to represent a scene content using only one key frame. This approach can be applied directly to organize tons of videos without any human intervention and can be utilized to provide browsing facilities to the viewers.

References

[1] Howhard D. Wactlar, *"The Challanges of Continuous Capture, Contemporaneous Analysis and Customized Summarization of Video Content,"*, CMU, USA.

[2] A. Hampapur et al., *"Virage video engine"*, Proc. SPIE, Storage and Retrieval for Image and Video Databases, 1997

[3] Yeung, M.M. et al., *"Segmetation of Videos by Clustering and Graph Analysis"*, CVIU, Vol.71, No:1, 1998

[4] Hanjalic, A. et al., *"Automated high-level movie segmentation for advanced video-retrieval systems"*, IEEE Tran. on CSVT., Vol:9 Issue:4, 1999

[5] R. Smith, *"VideoZoom spatio-temporal video browser"*, IEEE Tran. on Multimedia, Vol:1 No:2, 1999

[6] Smith, M. A. et al., *"Video Skimming for Quick Browsing Based on Audio and Image Characterization"*, IEEE CVPR, 1997.

[7] W. Ngo et al.*"Motion-based Video Representation for Scene Change Detection"*, IJCV, 2001

[8] D. DeMenthon, et al.,*"Relevance Ranking of Video Data using HMM Distances and Polygon Simplification"*, Int. Conf. on Adv. in Visual Info. Systems, 2000

[9] Adams, B. et al.,*"Towards automatic extraction of expressive elements from motion pictures: tempo"*, ICME, 2000

[10] http://www.m-w.com

[11] Removed for reviewers

[12] Daniel Arijon, *"Grammar of the Film Language"*, Hasting House, Publishers, NY, 1976

[13] T. Sobchack and V. Sobchack, *An introduction to Film*. Scot, Foresman and Company, 1987

[14] Ming-Hsuan Yang et al. *"Detecting faces in images: a survey*, PAMI, Vol:24 Issue:1, 2002

[15] Kjedlsen, R. et al.,*"Finding Skin in Color Images"*, Face and Gesture Recognition, 1996