3—20

Extracting Story Units from Video using Contextual Information

Hang-Bong Kang *
Dept. of Computer Engineering
The Catholic University of Korea

Abstract

In this paper, we propose a story unit extraction method using contextual information in video shots. We divide the contextual information into two categories: local and global contextual information. We represent the local contextual information in the video shot as foreground region's information, shot activity, and background color semantics. The global contextual information is determined by local contextual similarity and time locality. It also takes into account the effect of surrounding or neighboring shots. Using contextual information, we extract desirable story unit boundaries from video shots.

1 Introduction

An important task in video content analysis is to extract structures from video to facilitate efficient browsing and retrieval. However, it is not an easy task because video usually has a large volume and an unstructured format. To efficiently extract structures, it is desirable to segment a long video into shots first and then select representative frames for each shot. The representative or key frames are meaningful frames in a video shot. Using similarities between key frames, video shots can be clustered into episodes or story units. The story unit is characterized either by a single event such as dialog, action scene, etc. or by several events which take place in parallel. This unit is useful in the movie retrieval because a movie usually consists of a huge number of shots.

Several research works have been done in segmenting video into story units. Yeung et al.[1] propose shot-based organization structures in which the story flow is shown in scene-transition graphs. Rui et al.[2] show a method for constructing Table-of-Contents by extracting story units. Hanjalic et al.[3] segment movie into story units automatically. These approaches use visual content similarity based on low level features, and time locality for clustering criteria. For a visual similarity measure, they use a color histogram. To compute time locality, they use a local window or an attractive function which is a decreasing function along the time axis. Sometimes, however, these approaches have problems in detecting story units. For example, if two video shots have similar contexts but do not have the same color histograms, they cannot belong to the same story unit. In addition, because of the different sizes of foreground objects, similar shots are not clustered into the same story unit. Furthermore two different shots are merged into one story unit because of similar color histograms.

To deal with these problems, we propose a new approach using contextual information which reflects semantics. This paper is organized as follows. Section 2 discusses contextual information in video shots. Section 3 presents our story unit extracting method based on contextual information. Section 4 shows the experimental results using our approach.

2 Contextual Information

We divide the contextual information in video into two categories: local and global contextual information. The local contextual information in a video shot refers to the foreground regions' information, shot activity, and background color semantics. The global contextual information refers to the video shot's environment or its relationship with other video shots. Figure 1 shows the contextual information in video.

2.1 Local Contextual Information

To compute local contextual information from video shots, it is desirable to divide a video frame into a foreground layer and a background layer. The foreground layer describes the temporal occurrences of a simple object or a few regions. The background layer covers a whole area except foreground regions in a video frame. Separating a video frame into foreground regions and background regions is done in two steps. First, we detect whether there is a camera motion in video shots. Then, we use a spatial color segmentation result and a change detection mask based on motion information.

*Address: 43-1 Yokkok 2-dong, Wonmi-gu, Puchon city, Kyungki-do 420-743 Korea. E-mail: hbkang@www.cuk.ac.kr

To detect camera motion, we use Lucas-Kanade gradient decent method for optical flow. We quantize the phase of the motion vector into eight directions and compute the dominant motion intensity and the motion phase. Using the dominant motion intensity and the motion phase in each frame, we detect camera motions such as "zooming", "panning", "tilting", and "no camera motion" in a shot[4]. The dominant motion intensity and the motion phase in a shot is used to represent shot activity.

If there is no camera motion in a video shot, moving objects can be the candidates of foreground regions. Generally, the motion of moving object entails intensity changes in magnitude so that intensity changes are an important cue for locating moving objects. To detect moving objects, we use a method based on statistical hypothesis proposed in [5]. As in [5], we make a change detection mask between two frames within a video shot. The change detection mask refers to a binary image that indicates foreground(moving regions) and background(stationary regions). For spatial color segmentation of a frame, we use HSV color space because the similarity of two HSV colors is determined by their proximity in the HSV color space. We then segment a frame using multi-resolution recursive shortest spanning tree algorithm[6]. Finally, we superpose the change detection mask on the top of the segmentation result. When the majority part of a spatially segmented region is covered with the foreground parts of the change detection mask, the whole region is declared as the foreground. Otherwise, the whole region is declared as background.

If there is a camera motion such as "panning", "tilting", and "zooming", we choose the region whose center is located around the center of the frame as one of the foreground regions. After deciding foreground regions, we compute foreground region's normalized color distribution.

From background regions, we compute the saturated dominant colors of the background using 12 quantized color bins such as red, orange-red, orange, or-yellow, yellow, green-yellow, green, green-blue, purple-blue, purple and purple-red. We use three most largest color bins as dominant colors. Then, the dominant colors are classified into warmth(red-orange environments), cold(yellow-green environments) and contrast(light-dark, warm-cold) classes to represent semantics [7]. Using foreground region's information, shot activity and background color semantics, we represent video shot's local contextual information.

2.2 Global Contextual Information

The global contextual information refers to the video shot's environment or its relationship with other shots. It can be determined by the coherence between shots and by the effect of surrounding or neighboring shots. Before computing the coherence value between two shots, we first select key frames which represent each shot. The key frames are selected using the camera motion and foreground region information[4]. When camera motion is "panning", or "tilting", we choose the first and the last frame as the key frames because the intention of the camera director who is taking the shot is usually shown in the first and the last frame. Then, we select more frames based on the amount of content changes in the shot. In the zooming shot, we choose the last frame as the key frame because it is the most important frame in the shot. When there is no camera motion, we choose the frame as a key frame which has large foreground regions at its center. Then, we select more frames by computing the amount of content changes from this key frame using bidirectional search on the time axis.

The coherence value between two shots is computed using the key frames of each shot. It is defined by local contextual similarity, and time locality which represents temporal coherence. We use a fuzzy representation model in computing similarities because this is useful to manage the vagueness of similarity properties. The local contextual similarity is defined as the disjunction of two membership values such as foreground similarity and background similarity. Foreground similarity is computed from the normalized color distribution of foreground regions, whereas background similarity is computed with a weighted sum of shot activity and dominant color semantics. Time locality has a membership function which is a decreasing function of time because similar shots should be close to each other temporally. So, we define the coherence value of two shots as the conjunction of local contextual similarity and time locality between two shots. Because each shot is represented with a few key frames, the coherence value of two shots is set to the maximum value of the possible pair of key frames which belong to each shot.

The effect of surrounding or neighboring shots is considered in computing global contextual information. Even if the coherence value between two shots is high, it is desirable not to link the two shots when the number of shots located between two shots is large. This is reflected in our story unit extraction method.

3 Story Unit Extraction

A story unit can be a single event or several events taking place in parallel. In our approach, it is constructed by grouping similar shots using global contextual similarity. The grouping procedure has three

steps: initial linking, refinement, and adjustment. In the initial linking step, we determine whether a link can be established between two shots. When the coherence value of two shots is large, we make a link between two shots. If the current shot k is linked to a shot $k + n$, all intermediate shots automatically belong to the same story unit. If we find a shot k which has a link with one of the previous shots and has no link with subsequent shots, the shot k is at the boundary of the story unit. In the initial linking step, roughly similar shots can be grouped as the same story unit.

The second step is a refinement step which controls appropriate grouping. We use global contextual information in this step. If we find that the number of shots located between the two linked shots is larger than the threshold, and if these shots have no other links with previous shots or subsequent shots, we disconnect the initial link. Figure 2 shows this process.

The final step deals with shots which do not belong to any adjacent story unit boundaries. These shots have no previous and subsequent links. In this case, we compute the coherence values of the shots with adjacent story units. For example, if the shot q shown in Figure 2 has a larger coherence value of shots belonging to the story unit m than that of the shots belonging to the story unit $m + 1$, we decide that the shot q belongs to the story unit m.

4 Experimental Results

Simulations have been performed on the movies and TV dramas. First, we extract DC image sequences and detect shot boundaries using regions' flow and color histograms. To select key frames, we detect the camera motion using optical flow in each video shot. Based on camera motion information, we classify the video shots into panning shots, tilting shots, zooming shots and no camera motion shots. And then we select one or a few key frames for each video shot[4].

To compute local contextual information for each shot, we divide the video frame into a foreground layer and a background layer. Then, we extract foreground regions using the color segmentation result and a change detection mask. For local contextual similarity, we compute normalized color distribution, shot activity and background color semantics.

For a global contextual information, the coherence value of two shots is determined from a local contextual similarity and time locality. Using this information, we detect initial story unit boundaries. Then, we refines the linking based on the effect of surrounding or neighboring shots. Finally, we adjusts story unit boundaries by dealing with shots which do not belong to any adjacent story units.

In our approach, even though the shots may have different sizes of foreground regions, they are clustered into the same story unit because they have similar normalized foreground color histograms. In addition, the shots which have similar context but have different color histograms can be linked to the same story unit because the background color semantics is similar.

5 Conclusions

A new story unit extraction method using contextual information is proposed. The contextual information is divided into local and global contextual information. The local contextual information refers to the foreground region's information, shot activity, and background color semantics. The global contextual information refers to the video shot's environment or its relationship with other video shots. We determine global contextual information using the coherence value between two shots and the effect of surrounding or neighboring shots. Using contextual information, we can extract desirable story units regardless of low-level feature like color histogram.

References

[1] M. Yeung, B. Yeo, and B. Liu, "Segmentation of Video by Clustering and Graph Analsis," *Computer Vision and Image Understanding*, Vol. 71, No. 1, pp.94-109, 1998.

[2] Y. Rui, T. Huang and S. Mehrotra, "Constructing Table-of-Content for Videos," *ACM Multimedia Sys. Jour.*, Vol. 7, pp. 359-368, 1999.

[3] A. Hanjalic, R. Lagendijk and J. Biemond, "Automated High-Level Movie Segmentation for Advanced Video Retrieval Systems," *IEEE Trans. Cir. and Sys. for Video Tech.*, Vol. 9, No. 4, pp. 580-588, June 1999.

[4] H. -B. Kang, "Key frame Selection using Region Information and Its Temporal Variations," *Proc. IMSA'99*, pp.33-37, Oct. 1999.

[5] M. Kim, J. Choi, D. Kim, H. Lee, M. Lee, C. Ahn and Y. Ho, "A VOP Generation Tool: Automatic Segmentation of Moving Objects in Image Sequences Based on Spatio-Temporal Information," *IEEE Trans. Cir. Sys. for Video Tech.*, Vol. 9, No. 8, pp. 1216-1226, Dec. 1999.

[6] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Data bases," *Computer Vision and Image Understanding*, Vol. 75, July/August, pp. 3-24, 1999.

[7] J. Corridoni, A. Bimbo, and P. Pala, "Image Retrieval by Color Semantics," *Multimedia Systems*, Vol. 7, pp. 175-183, 1999.

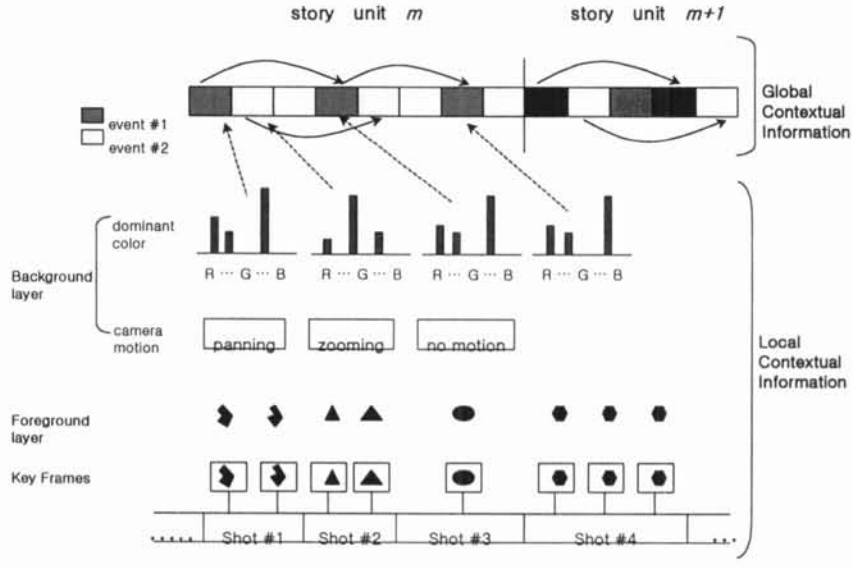

Figure 1. Contextual Information in Video

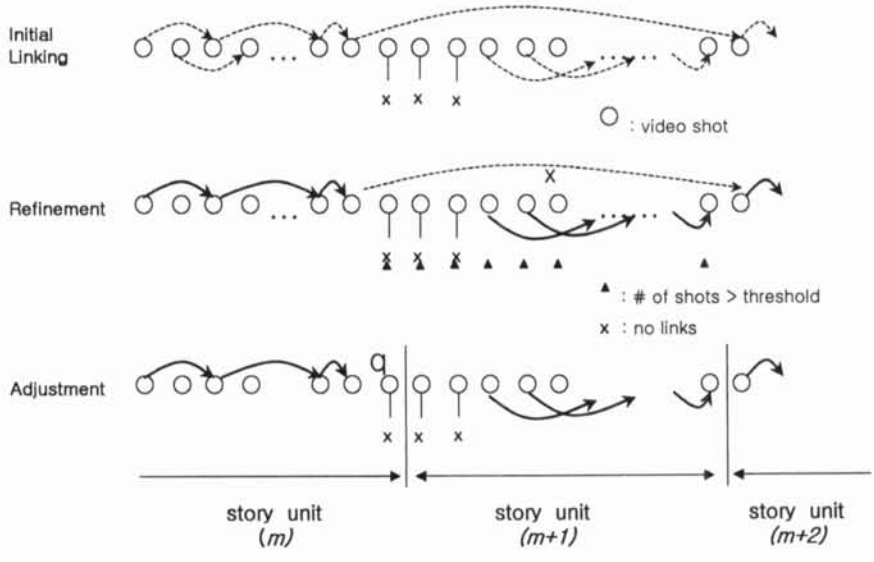

Figure 2. Extraction of Story Units