Asian Journal of Information Technology 7 (4): 133-137, 2008
ISSN: 1682-3915

A Novel Approach Towards Keyframe Selection for Video Summarization

[1]Chitra A. Dhawale and [2]Sanjeev Jain
[1]Department of MCA, H.V.P. Mandal, Amravati (MS), India
[2]Department of Computer Science and Engineering, LNCT, Bhopal (M.P), India

Abstract: For the short summary of large video, first step is to split the longer video into shots, then the representative frames (key-frames) for each shot are selected. These representative frames together form the video summary. In this study we have considered the video key-frame selection. As a part of research work and the minor research project, we have studied and implemented various methods for shot detection and key-frame selection using Matlab and C++. In this study the improved algorithm for histogram based approach is proposed for multiple visual descriptor features of video key-frames selection for compact video representation.

Key words: Video summary, video shot detection, key-frame features, key-frame extraction

INTRODUCTION

Due to rapid advancement in digital video technology and the increasing availability of computing resources, there is an explosion of digital video data in last few years especially on the Internet. However, the increasing availability of digital video has not been accompanied by an increase in its accessibility. This is due to the nature of video data, which is unsuitable for traditional forms of data access, indexing, search, and retrieval, which are either text based or based on the query-by-example paradigm. Therefore, techniques have been sought that organize video data into more compact forms or extract semantically meaningful information (Dimitrova *et al*., 2002).

As video contains much redundant information, for the easy retrieval in short time, the repeated information in video can be reduced by creating video summary which is a small collection of salient images extracted or generated from the underlying video source. A video summary can be built much fast, since generally only visual information is utilized and no handling of audio and textural information is needed. Therefore, once composed, it is displayed more easily since there are no timing or synchronization issues.

To create such video summary, first step is to split the video into smaller unit called shots and then the key-frame are selected for each shot to represent the compact video summary. In literature, various shot detection methods and their comparative study has been published (Hampapur *et al*., 1995; Otsuji and Tonomura, 1993; Zabih *et al*., 1995; Boreczky and Rowe, 1996; Lienhart, 1999; Zhang *et al*., 1993).

RELATED WORK

Assuming that the video has already been segmented into shots and extract the key-frames from within each shot detected. One of the possible approach by Tonomura *et al*. (1993) for keyframe selection is to choose the first frame in shot as key frame. Rui *et al*. (1998) uses first and last frames where video is time sampled regardless of shot boundaries. As per Hanjalic *et al*. (1996), difference between consecutive frames in terms of color histogram for visual features is compared w ith certain threshold value. And the key-frame is selected for this value greater than threshold. Zhuang *et al*. (1998) group the frames in clusters and the key frames are selected from the largest clusters.

In Girgensohn *et al*. (2000) constraints on the position of the key frames in time are also used in the clustering process; a hierarchical clustering reduction is performed, obtaining summaries at different levels of abstraction. In Gong *et al*. (2000) the video are summarized with a clustering algorithm based on Single Value Decomposition (SVD). The video frames are time sampled and visual features computed from them. The refined feature space obtained by the SVD is clustered, and a key frame is extracted from each cluster. In CueVideo toolkit by IBM Ulbaden (Niblack *et al*., 2000), Authors have extracted the I-frame from the video and these I-frames are considered as a keyframes.

The drawback to most of these approaches is that the number of representative frames must be set in some manner a priori depending on the length of the video shots for example. This cannot guarantee that the frames selected will not be highly correlated. It is also difficult to

Corresponding Author: Chitra A. Dhawale, Department of MCA, H.V.P. Mandal, Amravati (MS), India

set a suitable interval of time, or frames: large intervals mean a large number of frames will be chosen, while small intervals may not capture enough representative frames, those chosen may not be in the right places to capture significant content. Some approaches uses clustering of those methods. Still other approaches work only on compressed video, are threshold-dependent, or are computationally intensive.

FRAME FEATURES

As mentioned above some papers use single visual descriptor for frame comparison but this cannot capture all the pictorial details needed to estimate the change in visual contents of frames which can be used for the accurate comparison between frames in shot. The I-frames selected also not capture the visual details of all the frames in a shot. While, few papers are published which considers the multiple features for the frame comparison at the cost of computation time and memory requirement for storing the multiple features. In this study, we have considered two features edge direction and texture of frame. Also the texture feature is extended to capture the color too. Thus by using minimum features, we have tried to capture detail and precise information in minimum time and memory.

Edge direction feature: For detecting the edge direction feature of frames, two sobel filters are are applied to obtain the gradient of the horizontal and vertical edges of the luminance frame image. As per Ciocca, these values are used to compute the gradient of each pixel and those pixels that exhibit a gradient over a predefined threshold are taken to compute the gradient of angle and then the histogram.

Wavelet coefficient feature for texture: Many researchers have devoted attention to studying texture using multi-resolution analysis, especially the wavelet transform (Campisi *et al*., 1990). The main advantages of the wavelet transform, as a tool for analyzing signals, are orthogonality, good spatial and frequency localization, and ability to perform multi-resolution decomposition.

The weighted standard deviation descriptor: The weighted variance texture feature vector from a grayscale image, are the following:

- Subject the grayscale image to an L-level discrete wavelet decomposition. We use the Debauchee wavelet for this purpose.
- At each ith level (i = 1, 2, …, L), there are three detail images (LH, HL, and HH). There is an additional approximation image in the Lth level. Calculate the standard deviations of all these images. Also, calculate the mean of the approximation image.
- The weighted standard deviation feature vector is defined as follows:

$$f = \left\{ \begin{matrix} \sigma_1^{LH}, \sigma_1^{HL}, \sigma_1^{HH}, \frac{1}{2}\sigma_2^{LH}, \frac{1}{2}\sigma_2^{HL}, \frac{1}{2}\sigma_2^{HH}, ..., \\ \frac{1}{2^{L-1}}\sigma_L^{LH}, \frac{1}{2^{L-1}}\sigma_L^{HL}, \frac{1}{2^{L-1}}\sigma_L^{HH}, \frac{1}{2^{L-1}}\sigma^{A}, \mu^{A} \end{matrix} \right\} \quad (1)$$

where, F_i^{MM} is the standard deviation of the MM (stands for HL, LH, or HH) detail image, in the ith level of decomposition; F^A is the standard deviation of the approximation image and μ^A is the mean of the approximation image. Note that the standard deviation of each sub-band image at level i is weighted by the factor $(1/2^{i-1})$. The motivation for this approach is the fact that the standard deviations of the sub-band images give a measure of the amount of detail in that sub-band. Furthermore, since texture mainly consists of quasi-periodic spatial variations, we expect the higher frequency sub-bands (lower levels of decomposition) to contain more texture information. Naturally, we benefit by giving a higher weight to these sub-bands. The mean of the approximation image gives intensity information about the image. For an L-level decomposition, the length of the feature vector is 3L+2. we first show how we use the WSD texture descriptor to compactly describe both color and texture in images.

The content of an frame image is described using the WSD content descriptor, as follows:

- Isolate texture and color information by mapping the image from the RGB space to the YCrCb space. The Y-matrix contains the grayscale component, and consequently the texture information. The Cr and Cb matrices contain the color information.
- Extract the WSD feature vector from the Y, Cr and Cb matrices using three levels of Haar wavelet decomposition. The length of each feature vector is 11. The content descriptor for the image is a 33-dimensional vector formed by concatenating the feature vectors of the Y, Cr and Cb matrices.

The main idea in using a texture descriptor to describe color, is that the texture of the Cr and Cb components provide detailed information about the distribution of color in images. The 33-dimensional

content descriptor compactly describes both texture and color in images. An advantage of this descriptor is that one can give weights to its texture and color components, tailoring it to the type of images to be retrieved. The first 11 elements of the descriptor provide texture information, and the rest provide color information.

Frame difference measure: To compare the two frame descriptors a difference measure is used to evaluate the texture and color feature histogram and edge histograms. The difference between two color histogram (d_H) is calculated by using histogram intersection measure. The distance between 2 edge direction histogram (d_E) is computed using city block distance and wavelet statistics (d_W) using euclidean distance. These three values are then combined to form the final frame difference.

$$d_{HWE} = (d_H . d_W) + (d_W . d_E) + (d_E . d_H) \quad (1.2)$$

significant change in color, texture and edge feature values result high values in d_{HWE}.

KEY-FRAME SELECTION

The cumulative graph is constructed for the frame difference values. This graph shows how the frames visual content changes over entire shot, the sharp slope indicate significant changes in the visual content due to a moving object, camera motion. These cases are considered as the interesting event points that must be considered in selecting the key-frame to include in the final shot summary. The representative frames are those corresponding to the mid points between each pair of consecutive curvature point (Chetverikov *et al*., 1999).

Our proposed algorithm does not require processing the whole video, also we have limited the analysis of fixed number of frame difference within a predefined window.

We have tested our algorithm on educational video sequence of our MCA department and various other types of videos. Also compare our results with the key-frame selection using DCT coefficients (I-frames). For the preliminary task of video shot detection, we have applied frame difference and histogram techniques. The results are shown in Table 1 and Fig. 1 and 2.

Table 1: The preliminary task of video shot detection

Video clip	Total frames	Shots	No. of I-Frames	No. of Key-Frames
MCA Office	539	03	35	5
Son Video	173	03	12	4
Football	5402	25	361	76
News	4226	21	234	68

Fig. 1a: MCA Office Tour I-Frames obtained using shot detection

Fig. 1b: MCA Office Tour-Key-frame obtained using our proposed Algorithm

Fig. 2a: Son video clip- I-Frames obtained using shot detection

Fig. 2b: Son video clip-Key Frames obtained using our proposed Algorithm

CONCLUSION

The method has been tested on various video sequences like news programs, sports, academic etc. Each sequence was segmented into shots and then the key-frame are selected .We have compared our results with the I-frames obtained by CueVideo (2000) and found that our method gives better result in much less time and memory.

REFERENCES

Boreczky, J.S. and L.A. Rowe, 1996. Comparison of video shot boundary detection techniques. In Storage and Retrieval for Still Image and Video Databases IV, Proc. SPIE 2664, pp: 170-179.

Campisi, Patrizo, Longari, Andrea, Neri and Alessandro, 1999. Automatic Keyframe Selection Using a Wavelet-Based Approach. Wavelet Applications in Signal and Image Processing VII, Michael A. Unser, Akram Aldroubi, Andrew F. Laine (Eds.). Proc. SPIE 3813: 861-872.

Chetverikov, D. and Z.S. Szabo, 1999. A simple and efficient algorithm for detection of high curvature points in planar curves. Proc. 23rd Workshop of the Austrian Pattern Recognition Group, pp: 175-184.

Dimitrova, N., H.J. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor, 2002. Applications of video-content analysis and retrieval. IEEE. Multimedia, 9 (3): 42-55.

Girgensohn, A. and J. Boreczky, 2001. Time-constrained key-frame selection, technique. Multimedia Tools and Application, 11: 347-358.

Gong, Y. and X. Liu, 2000. Generating optimal video summaries. Proc. IEEE. Int. Conf. Multimedia and Expo, 3: 1559-1562.

Hampapur, A., R.C. Jain and T. Weymouth, 1995. Production model based digital video segmentation. Multimedia Tools and Applications, 1 (1): 9-46.

Hanjalic, A. and R.L. Langendijk, 1996. A new key-frame allocation method for representing stored video streams. Proceeding 1st International Workshop on Image Databases and Multimedia Search.

Lienhart, R., 1999. Comparison of automatic shot boundary detection algorithms. In: SPIE Conf. Storage and Retrieval for Image and Video Databases VII, 3656: 290-301.

Niblack, W., S. Yue, R. Kraft, A. Amir and N. Sundaresan, 2000. Web-based searching and browsing of multimedia data. IEEE International Conference of the Multimedia and Expo, New York, USA, www.almaden.ibm.com/projects/cuevideo.

Otsuji, O. and Y. Tonomura, 1993. Projection detecting filter for video cut detection. Proc. First ACM Int. Conf. Multimedia, pp: 251-257.

Rui, Y., T.S. Huang and T.S. Mehrotra, 1998. Exploring video structure beyond the Shots. Appeared in Proceeding of the IEEE International Conference Multimedia Computing and Systems (ICMCS), Texas USA.

Tonomura, Y., A. Akutsu, K. Otsugi, and T. Sadakata, 1993. VideoMAP and VideoSpaceIcon: Tools for automatizing video content. Proc. ACM. INTERCHI. Conf., pp: 131-141.

Zabih, R., J. Miller and K. Mai, 1995. A feature-based algorithm for detecting and classifying scene breaks. Proc. ACM Multimedia, San Francisco, CA, pp: 189-200.

Zhang, H., A. Kankanhalli and S.W. Smoliar, 1993. Automatic partioning of full-motion video. Multimedia Syst., 1 (1): 10-28.

Zhuang, Y., Y. Rui, T.S. Huang and S. Mehrotra, 1998. Key Frame Extraction Using unsupervised Clustering. Proc. ICIP, Chicago, USA, 1: 866-870.