International Conference on Intelligent Information Hiding and Multimedia Signal Processing

A Video Summarization Approach based on Machine Learning

Wei REN Yuesheng ZHU
The Key Laboratory of Integrated Microsystems,
Shenzhen Graduate School, Peking University

Abstract

Video summarization is not only the key to effective cataloging and browsing video, but also as an embedded cue to trace video object activities. In this paper, a video summarization approach based on machine learning is developed for automatic video transition prediction. Several novel features are extracted to characterize video boundary, including cut, fade in, fade out and dissolve for facilitating the understanding content structure and domain rules of a video. These features not only can be used to filter negative false alarms caused by illumination changes but also to improve recognition rate of the key-frames. Our approach provides a good view on temporal continuity of video event. Our results have shown that our approach can accurately predict the transitions in a video sequence and would be a practical solution for automatic video segmentation and video summarization.

1. Introduction

Advance in video compression technology, broadband communication networks, storage devices, and consumer electronics have resulted in the sizes of multimedia data collection drastically expanded. Accessing, manipulating and transmitting the information with such a tremendously large amount of data not only have become a heavy burden to storage and bandwidth, but also makes an I/O access bottleneck. Effective and efficient techniques need to reduce the amount of data transmission and access.

Raw video is an unstructured data stream, physically consisting of a sequence of video shots. A video shot is composed of a number of frames and its visual content can be represented by key-frames. Video summarization defines as a collection of key-frames extracted from a video. In general, content-based video summarization is therefore a two-step process. The first step is partitioning a video into physical shots, called video segmentation or video shot boundary detection. The second step is to find these representative frames. Thus, video can be organized as video, shot, key-frames hierarchy. Video summarization can provide a simple and effective way to abstract a long video sequence. They can be a generated as storyboards and video abstractions. Key-frames can act as the most representatives of video shot for video indexing, browsing, and retrieval. Video summarization is indispensable processing for video management. After video is structural organized hierarchically, thus, video can be stored and transmitted by shots as minimum components and indexed by sequential key-frames, and be reassembled in receive end. When transmission errors happen, only relative shots therefore need to be resent. By using key-frames, in addition, complex video retrieval task is transformed into simple image comparison exercises among the corresponding key-frames. For a user query, the server only needs to compare key-frames and to issue a file I/O operation to retrieve the relative video segment for transmission to the client. Consequently, video summarization can prompt broadband used effectively, the amount of manipulation data-stream reduced, and the time of computation and I/O access saved.

Numerous studies have attempted to solve the problem of finding key-frames for video summarization. The examples include hierarchical clustering algorithm [3], rule based importance score [4], unsupervised clustering [5], the temporal distortion minimization [7], K-th Hausdorff distance for pixel set comparison [1], and histogram analysis [2], finding a turning point from motion acceleration to deceleration [6], etc. The following important observations can be drawn from these studies: Some approaches only pay attention to single characteristics of video content like color histogram or motion. However, video is very complex media; single feature is difficult to have discriminating to cover all variations for video indexing. While some approaches based on clustering techniques take all the frames of a shot together and classify them according to their content similarity. Then, the key-frames are selected randomly, or pre-determined as the representative frame of a cluster without consideration of the temporal continuity. Therefore it is not good enough for classifying long panning or tilting shot, because visual content may change drastically during camera panning or tilting. Visual information appears similar only for temporal neighbor frames, not for whole shot.

In our study, video summarization is defined by removing temporal redundancy and grasping new video objects appearance to illustrate the story or

978-0-7695-3278-3/08 $25.00 © 2008 IEEE
DOI 10.1109/IIH-MSP.2008.296

theme of a video. After video object of key-frame captured, they can be labeled in a temporal interval and tracked through video sequence. An integrated method is developed to adopt a coarse-to-fine strategy to select key-frames by analyzing temporal data stream, structure variance and accumulation color changes of the sequence.

2. Video Summarization Processing

In this paper, a two-step video summarization approach is proposed. First, video is automatically segmented into shot by machine learning. Then, key-frames are selected. Fig. 1 shows the flowchart of our video summarization methodology.

2.1. Automatic Video Transition Predication

A machine learning system is developed that learns to predict video transitions based on statistical information derived from two successive frames. This system is described below:

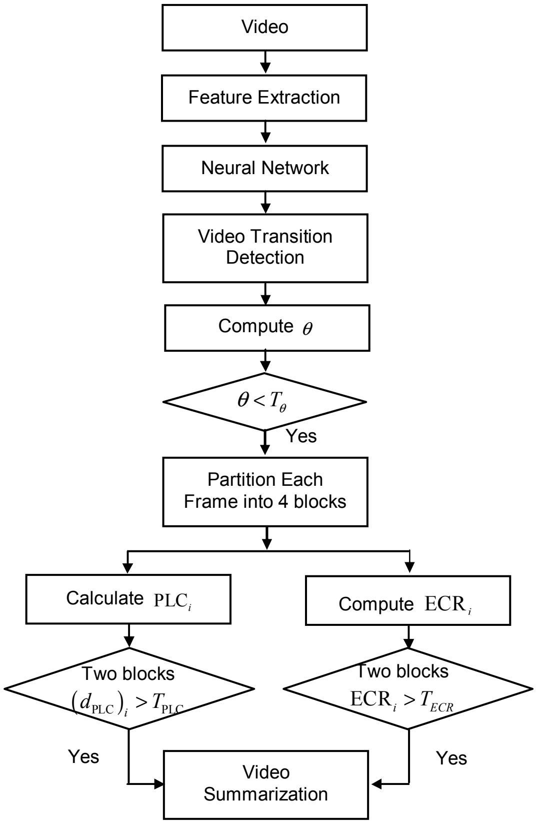

Video
Feature Extraction
Neural Network
Video Transition Detection
Compute θ
$\theta < T_\theta$
Yes
Partition Each Frame into 4 blocks
Calculate PLC_i
Compute ECR_i
Two blocks $(d_{\mathrm{PLC}})_i > T_{\mathrm{PLC}}$
Two blocks $\mathrm{ECR}_i > T_{ECR}$
Yes
Yes
Video Summarization

Fig.1 Flowchart for video summarisation

Several novel features are proposed and used for video transition prediction as follows:

Pixel Likelihood Ratio (PLR)

PLR describes from pixel level how two images are likely similar in energy and is defined by:

$$PLR(i,j)=\frac{\sum_{x,y}\left[P_i(x,y)-P_j(x,y)\right]^2}{\left[\sum_{x,y}P_i^2(x,y)\right]\left[\sum_{x,y}P_j^2(x,y)\right]}$$

Where *p (x,y)* is a pixel value at *(x,y)* position of a image; *i,j* represent frame *i* and frame *j*, respectively. When two frames are similar, PLR value goes to zero.

Edge Change Ratio (ECR)

ECR represents quantity change ratio of these strong energy edges. The edge with strong energy is defined by the edges which hold the energies over the average edge strength $|\vec{\nabla}E|$. $|\vec{\nabla}E|$ is calculated from the magnitude of the gradient field of the 3-channel color image. $|\vec{\nabla}E|$ is given by

$$|\vec{\nabla}E|=\left\{\frac{1}{2}\left[g_{xx}+g_{yy}+\sqrt{\left(g_{xx}-g_{yy}\right)^2+4g_{xy}^2}\right]\right\}^{\frac{1}{2}}$$

$$g_{xx}=\left(\frac{\partial R}{\partial x}\right)^2+\left(\frac{\partial G}{\partial x}\right)^2+\left(\frac{\partial B}{\partial x}\right)^2$$

$$g_{xy}=\left(\frac{\partial R}{\partial x}\right)\left(\frac{\partial R}{\partial y}\right)+\left(\frac{\partial G}{\partial x}\right)\left(\frac{\partial G}{\partial y}\right)+\left(\frac{\partial B}{\partial x}\right)\left(\frac{\partial B}{\partial y}\right)$$

$$g_{yy}=\left(\frac{\partial R}{\partial y}\right)^2+\left(\frac{\partial G}{\partial y}\right)^2+\left(\frac{\partial B}{\partial y}\right)^2$$

Hence, *ECR* is computed by $ECR(i,j)=\left|\frac{Q_i-Q_j}{Q_i+Q_j}\right|$

where Q_i and Q_j is number of edge pixels more than average edge strength, respectively. $|\vec{\nabla}E|$ in consecutive frames. If $Q_i=0$ and $Q_j=0$, then the value of *ECR(i,j)* is set to 1. ECR is insensitive to fade transition.

Correlation Coefficient in Histogram (HCC)

HCC is used to measure how correlate the histograms of two frames. HCC is defined by cross-correlation of histograms:

$$HCC(i,j)=\frac{\mathrm{cov}(H_i,H_j)}{\sqrt{\mathrm{var}(H_i)\,\mathrm{var}(H_j)}}$$

where H_i and H_j is the histogram of frames *i* and *j*, var *(H_i)* and *var (H_j)* is variance respectively, and *cov(H_i,H_j)* is their covariance. In order to eliminate shift in histogram resulted from luminance altered; contrast stretch is performed in advance. Thereby, HCC is insensitive to fade transition.

So, 24 low-level features of color, texture, shape, motion and statistical features extracted by comparing difference between frame pairs are employed to train a system that can predict transitions on unseen test data. A neural network is used to automatically detect video transition between video scenes such as cut, fade-in, fade-out, dissolve; at same time, camera motion: pan-left, pan-right, tilt-up, tilt-down, camera-statc are captured and analysed. By estimating camera movement and acquiring local object motion, the trajectories of multiple objects are able to track and automatically label the corresponding regions.

In our system, a random reject filter is also exploited to solve open boundary problem of neural networks to improve the classification accuracy.

2.2. Video Summarization

To select key-frames, the changes in their contour, texture and color between two frames need to be identified. Color and texture change represents an important of visual cue based on psychological studies. Two completely different texture images may have similar color distribution. Texture can reflect the physical surface properties of real objects. It can make images exhibiting a tactile quality. Texture can be captured by variations of intensities. It is defined as a function of the spatial variation in pixel intensities. Except for color and texture, sometimes two image frames with very similar visual content may have significantly different layouts. Such examples can be found easily in sport video in Fig. 6. These frames represent different events or different camera viewpoints and so should be labeled as key-frames.

Sequential comparison is used to determine key-frames. The discontinuity or dissimilarity value is computed between current frame and the last extracted key-frame. The change in contour and texture can be measured in ECR_i, the change in color can be detected by PLC measure, while the variation in structure can be determined based on θ coefficient computed from cross-correlation. θ coefficient is used to measure what degree two images are correlated to each other. The idea behind of ECR_i metric is to track whether new objects appearing in the current frame. Integrating these three measures are able to filter negative false alarms caused by using color information alone, to improve the recognition rate of key-frames, and to generate better results. Adaptive thresholds for key-frames detection are exploited. Thresholds $T_{PLC}, T_{ECR}, T_{\theta}$ in Fig.1 are optimized setting based on the validation videos.

3. Experiment

A neural network is trained using back propagation mode of training for each of the trials. It is optimized for the number of hidden nodes and learning parameters. The training quality is improved using random rejects so that closed decision boundaries are generated.

Fig.2 ~ Fig. 5 shows the normalized feature measurements on an example test video, including PLR, ECR, HCC. In these plots, the top red curve responds to the ground truth of transition with value ‘1’ as “Cut”, with value ‘1.1’ as “Fade-in”, ‘1.2’ as “Fade-out”, ‘1.3’ as “Dissolve” and with value ‘1.4’ as “Camera Operations”. These plots witness that our proposed features can capture precisely video transitions.

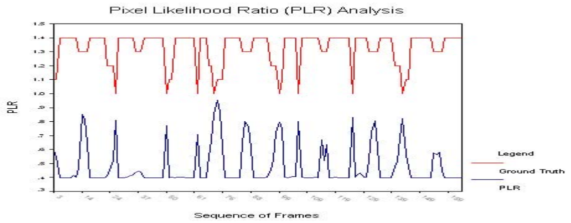


Fig. 2. Pixel Likelihood Ratio (PLR) analysis between successive frames of a test video

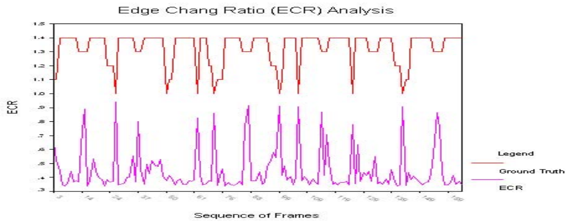


Fig. 3. Edge Change Ratio (ECR) analysis between successive frames of a test video

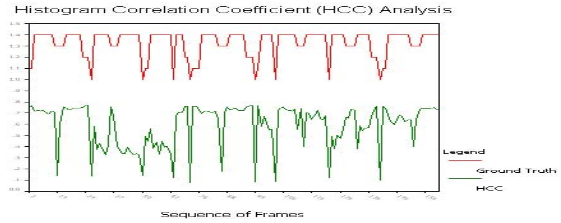


Fig. 4. Histogram Correlation Coefficient (HCC) analysis between successive frames of a test video

Here, the purpose of video summarization is to prompt video object activities tracking and to deliver the outline of video narrative. According with our task, the frame with the half of its size found visual content change over thresholds is selected as key-frames. To evaluate the system performance, in total 10 video clips from the international video benchmark the MINERVA with various video genres are used. Table 1 gives the experimental result for quantitative measure of key-frame selection. The result shows our approach (CF) with higher hit rate, lower missing hit rate and negative false alarm rate, compared to approaches of histogram [2] and motion [6].

Furthermore, the result also finds that single feature employed such as color or motion is obviously insufficient as a basis for video summarization, because of the large variability in video content.

Table 1 Comparison of video summarization on ten videos

Approach	Hit (%)	Missing Hit (%)	Negative False Alarm (%)
CF	95.3	4.7	3.1
Histogram	87.1	12.9	21.2
Motion	63.3	36.7	34.6

4. Conclusions

In this paper we described a two-step approach for video summarization. A neural network is trained to predicate automatically video transitions. Then, coarse-to-fine strategy is exploited to first filter negative false alarms caused by changes in illumination; following perform fine step analysis to select key-frame. Our experiments have shown that better results are obtained with integrated similarity measures compared to color histogram and motion approach alone. The results have also witnessed that our metrics can robustly characterize visual content for video summarization. Consequently, our approach provides a good view on temporal continuity of video event.

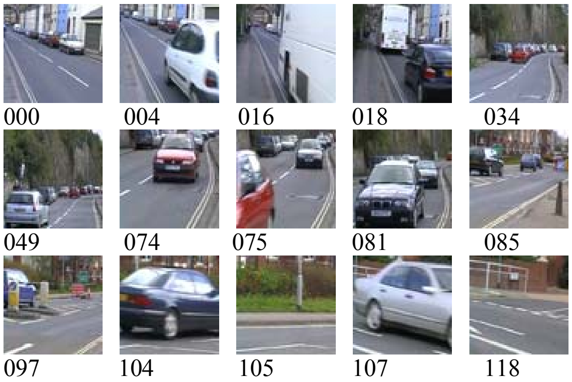
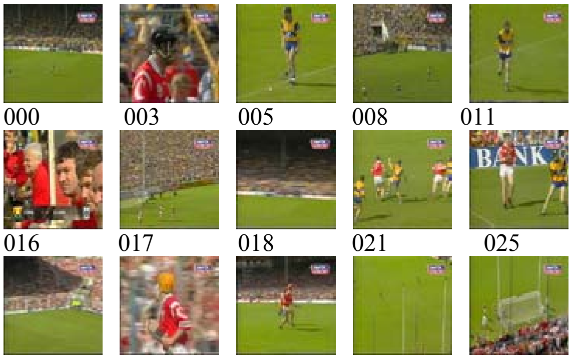


Fig. 5 Video Summarization selected for Traffic video (V248)



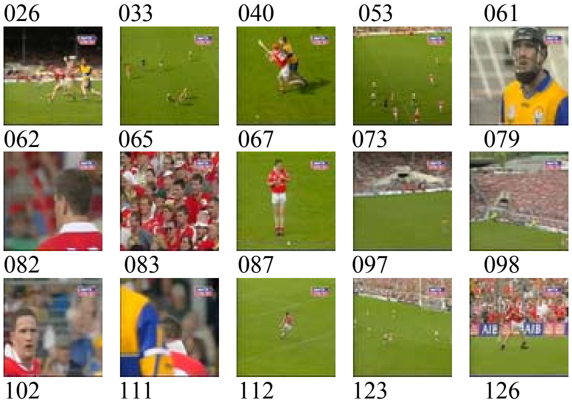


Fig. 6 Video Summarization for Sport Video (V116)

5. Acknowledgements

This work was supported by National Key Technology R&D Program of China under Grant No.2006BAH02A10 and by Science and Technology Program of Shenzhen.

References

[1]. C. Choudary and Tiecheng Liu, "Summarization of Visual Content in Instructional Videos", IEEE Trans. on Multimedia, vol. 9, no. 7, pp.1443–1455, Nov. 2007.

[2]. A. M. Ferman and A. M. Tekalp, "Efficient filtering and clustering and clustering methods for temporal video segmentation and visual summarization", J. of Visual Comm. and Image Representation, vol. 9, no. 4, pp. 336-351, 1998.

[3]. A. Girgensohn and J. Boreczky, "Time-Constrained Key frame Selection Technique", IEEE Multimedia Systems '99, vol. 1, pp. 756-761, 1999.

[4]. A. Girgensohn, J. Boreczky, and L. Wilcox, "Keyframe-Based User Interfaces for Digital Video", IEEE Computer, vol.34, No.9, pp. 61-67, 2001.

[5]. A. Hanjalic , H. J. Zhang, "An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-Validity Analysis", IEEE Trans. on Circuits and Systems for Video Technology, vol. 9, no.8, pp. 1280 –1289, 1999.

[6]. T. Liu, Hong-Jiang Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model", IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, No. 10, Oct. 2000.

[7]. Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," IEEE Trans. Circuits Syst. Video Tech., vol. 15, no. 10, pp. 1245 – 1256, 2005.