Content Based Video Retrieval Framework Using Dual Tree Complex Wavelet Transform

Tahir Jameel, S.A.M. Gilani, Adeel Mumtaz
Faculty of Computer Science and Engineering
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan
saroash_khan@msn.com, {teejay, asif, adeel}@giki.edu.pk

Abstract— **In this paper a novel technique of content based video retrieval is presented. The proposed technique uses Dual Tree Complex Wavelet Transform (DTCWT) based features of video frames for the purpose of shot change detection, key frame selection and video indexing. For shot change detection consecutive frame difference is computed, shot change is reported when the difference exceeds a certain threshold. For keyframe selection a frame is to be selected which is not part of shot transition using k-mean clustering of DTCWT feature vectors. Video shots are indexed using DTCWT features of the selected keyframes. Video query is processed by comparing the features of shot with the features database of the shots. For the purpose of features similarity we have used correlation based distance metric as it produced better results for this kind of feature similarity. The results are compared the results with classical techniques and it is shown how dual tree complex wavelet transform based features performed better. The whole framework uses similar kind of feature which makes it simple and efficient.**

Index Terms— **CBVR, Video Indexing, Shot Boundaries, Key Frames**

I. Introduction

A PICTURE is worth a thousand words is famous proverb. If it is so then video must worth a million words. Video has become a very useful source of multimedia information containing rich, self explanatory and most compelling information presentation source. In this era a revolution has taken place that has transformed the concepts of imaging to the digital media technology paradigm. Enhanced networking capabilities and digital technology has changed physical management of information to an interactive and more effective electronic data management. A large variety of digital video collection is available online about education, entertainment, business, current affairs and medicines. Different video formats are available for compression and video storage. Digital video is an aggregation of frames where each frame is a picture image. These frames are presented sequentially with certain frame rate. This gives an illusion of motion picture. Typical frame rates are 25 or 30 frames per second [1]. With this frame rate we can expect storage required for an hour of digital video. Main concerns about digital video are storage, displaying and searching. Remarkable work has been done in display quality and compression with certain trade off. Most of recent research is being done in efficient search systems for digital videos. Digital video in the perspective of digital images naturally makes possible to apply image processing techniques for compression, storage and searching.

Vide data is hierarchy in structure as shown in Fig. 1. Video sequence is composed of scenes. A scene is a story unit and it is a collection of consecutive shots that have semantic similarity in object, person, space and time [1]. It can be understood as a semantic situation having contents of various location camera shoots. Scene is more towards human understanding because it's close to human feelings and perception. A scene can be a collection of shots. A shot is defined as a continuous camera capture in which a camera can move as well as the objects can move. It is a basic unit of video like a paragraph in a text document. Contents of a shot are quite similar. Ideally there is negligible difference among the frame of a shot. In content based retrieval system, data is usually stored retrieved and demanded in shots. So a video sequence is decomposed into shots and stored in database. Shot boundary detection is usually a pre-processing technique for the video indexing system. In hierarchy a single shot is a collection of temporal correlated frames. Ideally there is unnoticeable difference among the consecutive frames. To reduce redundant information in the shots and to make retrieval system efficient a single frame which can represent the contents of a shot is selected. Such frames are known as keyframes. A shot may contain more than such frames which are known as candidate frames. The technique is to group candidate frames and select a non transient frame. In MPEG [2] video format there are I, P and B frames. I frame is a complete image containing all its data blocks which are referenced in P and B frames. This I frame may be selected as keyframe.

In section II we have shown how DTCWT features are extracted. In section III it is shown that how DTCWT features dissimilarity is utilized for shot boundary detection. Section V narrates video indexing and query shot retrieval. In section IV k-means clustering is applied on DTCWT features for the purpose of key frame selection. Section VI shows the results and finally section VII concludes the whole discussion.

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 449–454.

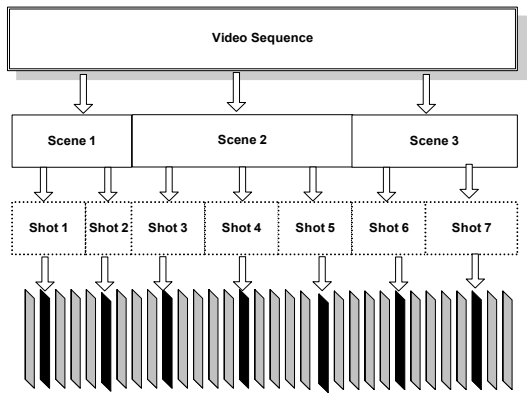

Fig. 1 Video Structure Hierarchy

II. CONTENT BASED VIDEO SIGNATURES

In most of the earlier techniques, video keyword based searching and text based searching have been prominent. But the obvious problem is that sometime keywords can not describe video contents, even the text based methods require tedious job to enter annotations manually. With advent of internet the problem of language have been raised in keyword based searches. It is possible that the most relevant video is available on the network but the language of text/ keyword is different. It was required to attach labels with the video which describe their visual contents. Trivial image description techniques have been applied to describe video contents. Most common are the color histograms, shapes, texture or motion based techniques. For video shot description histogram based techniques produced remarkable results because of temporal correlation, they are robust to object's geometric movements. Also frequency domain techniques have been proposed, amongst them wavelet based techniques have been significant [3] [4].

Kingsbury's [5] dual-tree complex wavelet transform (CWT) is an enhancement to the discrete wavelet transform (DWT), with important additional properties. The main advantages as compared to the DWT are that the complex wavelets are approximately shift invariant (meaning that our texture features are likely to be more robust to translations in the image) and that the complex wavelets have separate sub-bands for positive and negative orientations. Conventional separable real wavelets only have sub-bands for three different orientations at each level, and cannot distinguish between lines at 45º and -45º**.** The complex wavelet transform attains these properties by replacing the tree structure of the conventional wavelet transform with a dual tree. At each scale one tree produces the real part of the complex wavelet coefficients, while the other produces the imaginary parts. A complex-valued wavelet $\psi(t)$ can be obtained as:

$$\psi(t) = \psi_h(t) + j\psi_g(t) \quad (1)$$

where $\psi_h(t)$ and $\psi_g(t)$ are both real valued wavelets.

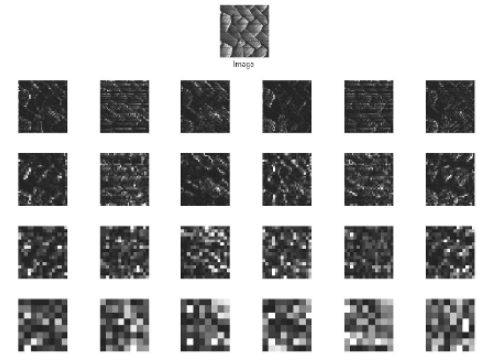

Fig.2. Four-scale CWT of a texture image

CWT like Gabor transform have six orientations at each of four scales (any number of scales can be used, but the number of orientations is built into the method). The main advantage as compared to the Gabor transform is speed of computation. It has a redundancy of only 4 in 2-dimensions and so the post-processing stages (of calculating mean and standard deviation) are also faster as it has less redundancy than the Gabor wavelets. Fig. 2 shows magnitudes of CWT coefficients for a texture image, one can see more details about orientation and scales. Each row represents one scale and columns represent angles within that scale. We performed a four scale (six angles) CWT on each frame. We get 24 real and 24 imaginary detailed sub-bands, and 2 real and 2 imaginary approximation sub-bands. By taking the magnitudes of corresponding real and imaginary coefficients of both approximation and detailed sub-bands we get 26 sub-bands. To calculate the features we measure the mean and standard deviation of the magnitude of the transform coefficients in each of 26 sub-bands, in the same way as [6].

III. SHOT BOUNDARY DETECTION SCHEME

Shot boundary detection is a pre-requisite of various content based video retrieval systems. A video shot is a video sequence that consists of continuous video frames for one camera action [7]. A video sequence can have more than one shot. These shots are merged together such that there boundaries are ill defined. Sometimes it is not obvious to detect the ending or start of a shot.

A number of video editing and mixing software are available to create special effects to merge shots which are known as shot transitions. Shot transitions are of various types [8]:

- A *cut* is an abrupt shot change that occurs in a single frame;
- A *fade-in* starts with a black frame; gradually the image of next shot appears, brightening to full strength.;
- A *fade-out* is opposite of a fade-in.;
- A *dissolve* consists of super imposition of a fade out over a fade in.