Static and Multiresolution Feature Extraction for Video Summarization

J. Kavitha\textsuperscript{a}, Dr. P. Arockia Jansi Rani\textsuperscript{b}

\textsuperscript{a}Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, 627012, India
\textsuperscript{b}Assistant Professor, Manonmaniam Sundaranar University, Tirunelveli, 627012, India

Abstract

The objective of this paper is to develop a video summarization system to extract significant frames of interest from a given video. To meet the objective, it is proposed to consider both the static features and the wavelet features. Visual attention integrated from the static and wavelet feature set are combined using a prioritized fusion method. Experimental results show that the static features dominate in certain videos and wavelet features dominate in certain videos. Hence, the proposed fusion approach is suitable for slow motion videos and fast moving videos. Further, the performance of proposed work outperforms the state-of-art methods for video summarization.

1. Introduction

The volume of digital video data has been increasing significantly in recent years due to the wide use of multimedia applications in the areas of education, entertainment, business, and medicine [1]. Hence, a system needs to effectively and efficiently manage and store the huge amount of audio visual information, at the same time providing a user friendly access to the stored data [2]. Video summarization is widely used in video cataloging, indexing and retrieving. It is a short representation of an original video content. Generally video summarization techniques are classified into two categories namely video summary and video skimming. Video summary is a small collection of salient images extracted from the underlying video...
source. Video skimming is a collection of image sequences as well as the corresponding audio abstract extracted from the original sequence.

A video summary considers only visual information and it doesn’t handle audio and textual information. It is displayed more easily since there are no timing or synchronization issues. So the users are able to grasp the video content more quickly [3]. It is for these reasons the proposed method works belongs to the video summary categorization.

Many of the algorithms developed for video summarization by researchers aimed at extracting salient key frames from a video. A complete survey on video abstraction techniques can be found in [4] state of the art. Zhuang proposed an unsupervised clustering scheme to extract key-frames from shots [5]. It requires a pre-defined threshold parameter to control the density of keyframes in a shot. Girgensohn proposed a time constrained clustering approach to filter out some clusters and to determine the representative frame for a cluster [6]. Dufaux developed a technique is composed of three steps. Shot boundaries detection, shot selection, and key frame extraction within the selected shot. The shot and key frame are selected based on measures of motion and spatial activity and the likeliness to include people [7].

Static attention model and motion attention model are two important methods for detecting all salient key frames in video. Static attention model was derived from saliency-based visual attention model for static scene analysis [8]. Then, some changes were made in Itti et al.'s model for satisfying the requirements of inherent characteristic of videos by Yu-Fei Ma et al [9]. Also a saliency map was generated from each frame by the three channel saliency maps computation, color contrasts, intensity contrasts, and orientation contrasts. Chia-Chiang Ho et al developed a system that consider intensity and color features for low level features in the YCbCr color space [10]. Thus the computation of low-level features yields two saliency maps, i.e., the intensity saliency map and the color saliency map which is similar to that of the Itti’s method. Yun Zhai and Mubarak Shah proposed a spatial saliency maps using the color statistics of images [11]. It was designed with a linear computational complexity with respect to the number of image pixels. Visual saliency was measured through a spatiotemporal attention model driven by intensity, color and orientation [12].

The remainder of the paper is organized as follows. In section 2, the system architecture of proposed method is described. The proposed fusion technique combines static features and transform domain features are explained in section 3. Experimental results are given in section 4. Conclusion and future work are in section 5.

2. System Architecture

Figure 1. depicts the overall system architecture of the proposed method. It combines both the static attention method and the Discrete Wavelet Transform (DWT) based method for extracting salient key frames from a video.

![Fig.1. Overall System Architecture](image-url)
First, the input video split into frames. Then, it finds the edges in two consecutive video frames using sobel edge detection algorithm. Based on these edges, it compares blocks of the video frames with one another. If throughout of comparison exceeds a specified threshold, then it shows that the scene has been changed [13]. The change of scene is considered as a shot. Each shot contains relevant key frames. The numbers of key frames are changed based on the threshold value. The Static attention method is used to find color opponency of each frame in the each shot. Based on the color opponency values, static attention method gives some set of key frames. In DWT method, the video frames are transformed into wavelet sub-bands and the standard deviation between two consecutive frames is computed. Based on the threshold values, it gives some set of key frames. Finally, the results of two methods are combined by using priority fusion method. It eliminates the unwanted and redundant frames. The resultant key frames are visually more attractive and meaningful.

3. The Proposed Method

The proposed method considers both the static features and the wavelet features. Visual attention integrated from the static attention method and modified discrete wavelet transform based method are combined using a prioritized fusion method.

3.1 Static Attention Method

The static attention method as depicted in fig.2 considers color contrast, and intensity of each frame [14]. The procedure is as follows: First, divide each frame into $8 \times 8$ macro blocks in the shot. Convert the color space of each block from RGB to LMS. LMS is a color space represented by the response of the three types of cones of the human eye, named after their sensitivity at long, medium and short wavelengths [15]. In the LMS conversion, first convert the color space of each block of RGB to XYZ. Then convert it into LMS color space approach. The transformation from LMS signals to the opponent signals serves to de correlate the color information carried in the three channels, thus allowing more efficient signal transmission. The three opponent pathways also have distinct spatial and temporal characteristics that are important for predicting color appearance [6]. The color contrast and intensity of each block is calculated by using LMS signal and color opponent signals such that Red-Green and Blue-Yellow opponency.

---

**Fig.2. Static Attention Method**

Divide Each Frame into $8 \times 8$ macro block

Convert RGB to LMS

Calculate Color Opponency (RG, BY), Contrast and Intensity

Calculate Center Surround Difference and Visual Attention

Key Frames
The static attention value is calculated based on the center surround difference of each block (C_i) and Gaussian falloff weight (W_i) of the saliency block. The center-surround difference is the sum of intensity and color contrast difference of each block with its neighborhood. Center-surround difference of each block is computed using the formula in equation.

Then, calculate visual attention index (A_s) of each block is calculated using the following formula.

\[ A_s = \frac{1}{N} \sum_{i=1}^{N} (W_i \cdot C_i) \]  \hspace{1cm} (1)

Where, \( W_i = \exp \left( \frac{|p_i - p_{\text{center}}|^2}{2\delta_w^2} \right) \) \hspace{1cm} (2)

\[ C_i = \sum_{q \in \Omega} q|p_i(I) - q(I)| + \beta |p_i(H) - q(H)| \] \hspace{1cm} (3)

Where \( p_{\text{center}} \) – Center of frame

\( \delta_w^2 \) - Variance is set to be the one third of the frame width

Using the above formula, static attention value of each frame in the shot is calculated. The saliency map of each frame in the shot is generated. Finally, it gives set of salient key frames for each shot if the frame has highest color contrast and intensity values.

### 3.2 Modified DWT Based Method

This work is the extension of DWT based summarization [16] The DWT is applied to each shot and statistical features are extracted. This result is used to select pixels of interest in each frame in the shot. Fig.3. shows the overall process of the transform based video summarization. Let the video frames in each be transformed into wavelet sub-bands namely \( \beta_H, \beta_V, \beta_D \) of size \( r \times c \). The detail sub-bands \( \beta_H, \beta_V \) and \( \beta_D \) are used in the video abstraction process. Compute the standard deviation \( (\sigma_m) \) between the two consecutive frames \( f_k \) and \( f_{k+1} \) in each detail sub-band using the formula given below.

\[ \sigma_m = \frac{\sum_{n=1}^{N-1} (d_m(n) - M_m)}{N-1} \] \hspace{1cm} (4)

Where, \( m = \{ \beta_H, \beta_V, \beta_D \} \)

\[ d_m = \sum_{i=1}^{r} \sum_{j=1}^{c} (f_{k+1}(i,j) - f_k(i,j)) \] \hspace{1cm} (5)

\[ M_m = \frac{\sum_{n=1}^{N-1} \sum_{i=1}^{r} \sum_{j=1}^{c} (d_m(n))}{N-1} \] \hspace{1cm} (6)

In the next step, compute the threshold \( (T) \) suitable to select the pixels of interest \( (P) \) in each frame,

\[ T_m = M_m + \alpha \sigma_m \] \hspace{1cm} (7)

Where, \( \alpha \) is a constant

\[ f(P_n) = \begin{cases} 
1 & \text{if } (d_1(N) > T_1 \text{ & } d_2(N) > T_2) \\
1 & \text{if } (d_2(N) > T_2 \text{ & } d_3(N) > T_3) \\
1 & \text{if } (d_3(N) > T_3 \text{ & } d_1(N) > T_1) \\
0 & \text{otherwise}
\end{cases} \] \hspace{1cm} (8)
Where, N is number of frames in video

The DWT attention value is selected using high resolution value of the frame. It mostly detects all events of interest in the input video.

3.3 Priority Fusion Method

Combination of the DWT based method and static attention method must give human perception based on salient key frames. Generally, viewers have more attention in the object moving frames rather than static one. Sometimes the human viewers are interested in still images or background even though it does not contain moving object. For these reasons, priority fusion method is proposed in this paper. In static attention method, maximum static attention value is selected. The key frames are selected in DWT method if DWT attention value is greater than zero. Then, the selected key frames are combined and sequence order of frames is maintained.

4. Experimental Results with Discussions

The implementation is done using the programming language MATLAB 2013a. The performance of the proposed work is analyzed using the metrics such as recall, precision, F1 score and compression ratio. The videos are downloaded from www.open-video.org and http://cvrr.ucsd.edu/aton/testbed/... The videos are edited by video cutter for selecting maximum 600 frames. All videos are in 240×352×3 dimensions. The main aim of this work is to detect more attractive and meaningful key frames in the input video and to eliminate all redundant frames. To achieve this, it is desired to get a maximized precision, recall and f1 score. The results are compared with open video summary and user summary. The true positive rate measures the proportion of actual positives which are correctly identified and is complementary to the false negative rate. The true negative rate measures the proportion of negatives which are correctly identified and is complementary to the false positive rate [17]. Precision measure is the fraction of retrieved instances that are relevant, while recall measure is the fraction of relevant instances that are retrieved. If the precision value is high then it is understood that the algorithm returned substantially more relevant results than irrelevant, while high recall means that the algorithm returned more of the relevant results.

\[
\text{Recall} = \frac{n_{TP}}{n_{TP} + n_{FN}} \quad (9)
\]

\[
\text{Precision} = \frac{n_{TP}}{n_{TP} + n_{FP}} \quad (10)
\]
nTP, nFP and nFN refer to a number of true positive, false positive and false negative.

F1 measure can be used to reflect the averages of Recall and Precision and is defined as

\[ F_1 = 2 \times \frac{\mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \] (11)

The accuracy is the proportion of true results (both true positives and true negatives) in the population.

\[ \text{Accuracy} = \frac{nTP + nTN}{nTP + nTN + nFP + nFN} \] (12)

Compression ratio is computed by dividing the number of key frames in the result by the number of frames in the original video.

\[ C_R = \frac{T_{\text{Keyframes}}}{T} \] (13)

Where \( C_R \) is the compression ratio, and \( T_{\text{Keyframes}} \) is number of key frames in the result video and \( T \) is the number of frames in the original video.

The performance of the proposed work is analyzed using various videos and the results are tabulated in Table 1. It shows the number of frames (N) and target frames (T) in the input video file. The metrics including number of detected frames (dF), detected target frames (dT), precision (P), recall (R), F1 score (F1), Accuracy (A), computation time and compression ratio (CR) are computed. It shall be noted that the proposed work identified desired key frames. The number of detected frames (dF) increases with respect to the shots in the input video.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Input Video</th>
<th>N</th>
<th>T</th>
<th>dF</th>
<th>dT (%)</th>
<th>P (%)</th>
<th>R (%)</th>
<th>F1 (%)</th>
<th>A (%)</th>
<th>CT  (sec.)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>boatframes1.mpg</td>
<td>214</td>
<td>3</td>
<td>3</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>99</td>
<td>91</td>
<td>0.01</td>
</tr>
<tr>
<td>2.</td>
<td>shipv23.mpg</td>
<td>303</td>
<td>4</td>
<td>7</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>80</td>
<td>99</td>
<td>163</td>
<td>0.02</td>
</tr>
<tr>
<td>3.</td>
<td>face5.mpeg</td>
<td>300</td>
<td>4</td>
<td>6</td>
<td>75</td>
<td>43</td>
<td>75</td>
<td>55</td>
<td>98</td>
<td>128</td>
<td>0.02</td>
</tr>
<tr>
<td>4.</td>
<td>facev60.mpg</td>
<td>372</td>
<td>3</td>
<td>7</td>
<td>100</td>
<td>43</td>
<td>100</td>
<td>60</td>
<td>99</td>
<td>178</td>
<td>0.02</td>
</tr>
<tr>
<td>5.</td>
<td>manv631.mpg</td>
<td>601</td>
<td>3</td>
<td>6</td>
<td>100</td>
<td>50</td>
<td>100</td>
<td>67</td>
<td>99</td>
<td>258</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 1. Performance Analysis of the proposed work

The performance of the proposed work is compared with the existing methods namely DWT-VS (Discrete Wavelet Transform based Video Summarization), HIST-VS (Histogram based Summarization), and DCT-VS (Discrete Cosine Transform based Video Summarization). Table 2 illustrates performance comparison of the proposed work for a video file shipv23.mpg. The video file contains more static images. From table 2 it shall be noted that the percentage of detected target frames (dT) is 100% for the proposed work and static attention method. DWT-VS detects 75% of the target frames, HIST-VS and DCT-VS detect 50% of the target frames. Hence, proposed method and static attention method behave outperform compares to others.

Table 2. Performance Comparison of the proposed work with existing works for a video file shipv23.mpg

<table>
<thead>
<tr>
<th>S. No</th>
<th>Method</th>
<th>dF</th>
<th>dT (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Proposed Work</td>
<td>6</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>80</td>
<td>99</td>
</tr>
<tr>
<td>2.</td>
<td>DWT-VS</td>
<td>6</td>
<td>75</td>
<td>50</td>
<td>75</td>
<td>60</td>
<td>98.7</td>
</tr>
<tr>
<td>3.</td>
<td>Static-VS</td>
<td>6</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>80</td>
<td>99</td>
</tr>
<tr>
<td>4.</td>
<td>H-VS</td>
<td>6</td>
<td>50</td>
<td>33</td>
<td>50</td>
<td>40</td>
<td>98</td>
</tr>
<tr>
<td>5.</td>
<td>DCT-VS</td>
<td>6</td>
<td>50</td>
<td>33</td>
<td>50</td>
<td>40</td>
<td>98</td>
</tr>
</tbody>
</table>
Figure 4. shows the key frames of proposed method, DWT-VS, static-VS, HIST-VS and DCT-VS techniques comparing with target frames (T) for the video file shipv23.mpg. From fig.4 it shall be noted that the proposed and static attention method provides all desirable target frames for the facev60.mpg video file. DWT-VS provides three key frames. HIST-VS and DCT-VS provide two key frames. So it is clearly demonstrated that, the proposed work gives better result in comparison to others.

Table.3 depicts the performance comparison of the proposed work for a video file facev60.mpg. From table.2 it shall be noted that the percentage of detected target frames ($d_T$) is 100% for the proposed work DWT-VS method. Static-VS, HIST-VS and DCT-VS detect 33 % of the target frames.

Table 3. Performance Comparison of the proposed work with existing works for a video file facev60.mpg

<table>
<thead>
<tr>
<th>S. No</th>
<th>Method</th>
<th>$d_f$</th>
<th>$d_T$ (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Proposed Work</td>
<td>7</td>
<td>100</td>
<td>43</td>
<td>100</td>
<td>60</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>DWT-VS</td>
<td>7</td>
<td>33</td>
<td>43</td>
<td>100</td>
<td>60</td>
<td>99</td>
</tr>
<tr>
<td>3</td>
<td>Static-VS</td>
<td>7</td>
<td>100</td>
<td>14</td>
<td>33</td>
<td>20</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>H-VS</td>
<td>7</td>
<td>33</td>
<td>14</td>
<td>33</td>
<td>20</td>
<td>98</td>
</tr>
<tr>
<td>4</td>
<td>DCT-VS</td>
<td>7</td>
<td>33</td>
<td>14</td>
<td>33</td>
<td>20</td>
<td>98</td>
</tr>
</tbody>
</table>
Figure 5. shows the key frames of proposed method, DWT-VS, static-VS, HIST-VS and DCT-VS techniques comparing with target frames(T) for the video file shipv23.mpg. Static-VS, HIST-VS and DCT-VS provide only one key frame. So it is clearly demonstrated that, the proposed work gives better result in comparison to others.

The graph in fig.6 depicts that the average F1 Score for the proposed work is high in comparison with that of the other existing works.
5. Conclusion

The proposed work analyzed the role of static features and the wavelet based statistical features from video frames. The static features are extracted in the LMS color space. The statistical features are extracted in the wavelet domain. Visual attention values in both cases are compared in order to retrieve the frames of interest. The experimental results show that the proposed work is retrieved relevant frames of interest in an efficient manner.

In future, motion vector feature shall be also extracted in order to improve the performance of the proposed work.

References

17. http://en.wikipedia.org/wiki/Precision_and_recall