A Novel Rough Sets Based Video Shot Clustering Algorithm

1Wei Zhe, 1Li Zhan-Ming, 2Qi Yan-Fang and 2Zhao Li-Dong
1College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
2Department of Computer Engineering, Lanzhou Polytechnic College, Lanzhou 730050, China

Abstract: A novel image clustering algorithm was proposed in compressed domain. Firstly, DCT coefficients and DC coefficients are extracted from image consequences and an Information System is achieved by using DC coefficients. Secondly, Information System is reduced by introducing reduction theory of Rough Sets (RS), so the concise representation of the image is obtained by reduced DC coefficients. Finally, by introducing subdividing theory of RS, the image is clustered objectively. Based on the experimental results obtained in this study, the proposed algorithm enjoys following advantages. (1) the image is processed in compressed domain, the computing complexity is decrease. (2) RS is introduced to prepare the image data, the effectiveness is enhanced.

Key words: Video clustering, compressed domain, rough sets, DCT coefficients, DC coefficients

INTRODUCTION

Content-based image retrieval has become a very active research area since the introduction of automatic video retrieval system (Ferman and Tekalp, 1998). Advances in communications, multimedia and computer technologies have made video information producing too easily in our daily life, whereas, the sheer volume of video makes it extremely difficult to use and analysis. So it is necessary to develop a efficient technology to analyse, process and index video data automatically. Video mining is typically and effectively method to solve it. Video Shot clustering is the key technology in the video mining which breaks the massive volume of video consequence into smaller chunks. Each shot represents an event of actions.

Many theories and technologies are presented for shot clustering, especially in uncompressed domain, i.e., all video must be decompressed before methods are processed. Since operations on uncompressed video do not permit rapid processing because of the amount of data, so it is very time consuming. At the same time, more and more videos are in compressed forms according to MPEG national standard, either in storage or in communication. So it is urgent need to develop algorithms to clustering directly in compressed domain.

The study proposes a novel video shot clustering algorithm that operate directly in compressed video for shot clustering. The algorithm overcomes the limitations of previous approaches and increases the efficiency obviously.

CONVENTIONAL ALGORITHM OF VIDEO CLUSTERING

There have been many considerable work reported on shot clustering. Different methods and technologies are used to examine the changes between successive frames and determine whether changes have taken place (Antani et al., 2002). At present, approaches may be categorized into following classes: (1) visual feature difference based technology, include difference of gray-level sums, sum of gray-level differences, difference of color histograms, colored template matching, difference of color histograms and $\chi^2$ comparison of color histograms (Lo and Wang, 2001); (2) motion analysis and fuzzy clustering based technology, it can use motion information to detect the discontinuities of frame and draw a conclusion by using threshold (Yi et al., 2006; Dhillon et al., 2009); (3) model classification or neural network based technology, by constructing model or neural network train and capture the different type of shot transitions(Kim et al., 2005; Lee et al., 2006); (4) key frame and video summarization based technology, by optimal key frame representation scheme or video summarization simplify video shot boundary detection (Sze et al., 2005; Money and Agius, 2007); (5) object based technology.

Corresponding Author: Wei Zhe, College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

1056
employing the coding scheme of MPEG-4, extract the object of video to examine the boundary of shots such as localized and global lighting changes, variations in object size, occlusions and complex object motion and so on (Lei and Xu, 2006); (6) background based technology, exploits the fact that shots belonging to one particular scene often have similar backgrounds, although part of the video frame is covered by foreground objects (Chen et al., 2008); (7) event based technology, use of dimensionality reduction for video event detection without explicitly using motion estimation or object tracking (Tziakos et al., 2008; Choi et al., 2002) and (8) compressed domain based technology, instead of working on the original image sequences, technique to detect changes directly on intra-frame JPEG coded compressed data is develop. A fixed number of connected regions are selected and some predetermined collections of AC coefficients from the 8×8 DCT blocks in the regions are used to form a vector (De Bruyne et al., 2008; Bhandarkar and Chandrasekaran, 2004).

The above methods have their merits respectively; however, the previously mentioned methods all suffer from following limitations: (1) more above methods processed in uncompressed domain, so it needs expensive computation to decoding. This is very time consuming and (2) The shot detection algorithm is subjective or need some additional condition or threshold. These conditions or thresholds are difficulty to determine generally.

Rough Sets is a powerful and intelligent tool for data analysis, it has successfully been used in many applications such as Machine Learning, Expert Systems and Pattern Classification it can classify object without any prior knowledge. So it can detect shot boundary easily and efficiently.

THE PROPOSED ALGORITHM

Rough sets theory used in this algorithm: Let $U \neq \emptyset$ be a universe of discourse and $X$ be a subset of $U$. an equivalence relation, $R$, classifies $U$ into a set of subsets $U/R = \{X_1, X_2, ..., X_n\}$ in which the following conditions are satisfied:

- $X_i \subseteq U, X_i \neq \emptyset$ for any $i$
- $X_i \cap X_j \neq \emptyset$ for any $i$, $j$
- $U_1, U_2, ..., U_n, X_i = U$

Any subset $X_i$, which called a category, class or granule, represents an equivalence class of $R$. A category in $R$ containing an object $x \in U$ is denoted by $[x]_R$. For a family of equivalence relations $P \subseteq R$, an indiscernibility relation over $P$ is denoted by $\text{IND}(P)$:

$$\text{IND}(P) = \bigwedge_{i \in P} \text{IND}(R)$$

(1)

The set $X$ can be divided according to the basic sets of $R$, namely a lower approximation set and upper approximation set. Approximation is used to represent the roughness of the knowledge. Suppose a set $X \subseteq U$ represents a vague concept, then the $R$-lower and $R$-upper approximations of $X$ are defined by Eq. 4 and 5:

$$RX = \{x \in U : [x]_R \subseteq X\}$$

(2)

Equation 4 is the subset of $X$, such that $X$ belongs to $X$ in $R$, is the lower approximation of $X$:

$$\bar{RX} = \{x \in U : [x]_R \cap X \neq \emptyset\}$$

(3)

Equation 5 is the subsets of all $X$ that possibly belong to $X$ in $R$, thereby meaning that $X$ may or may not belong to $X$ in $R$ and the upper approximate $\bar{R}$ on contains sets that are possibly included in $X$. $R$-positive, $R$-negative and $R$-boundary regions of $X$ are defined respectively by Eq.6-8:

$$\text{POS}_R(x) = RX$$

(4)

$$\text{NEG}_R(x) = U - \bar{RX}$$

(5)

$$\text{BNR}(x) = \bar{RX} - RX$$

(6)

Attributes reduction and core: In RS theory, an Information Table is used for describing the object of universe, it consists of two dimensions, each row is an object and each column is an attribute. RS classifies the attributes into two types according to their roles for Information Table: Core attributes and redundant attributes. Here, the minimum condition attribute set can be received, which is called reduction. One Information Table might have several different reductions simultaneously. The intersection of the reductions is the Core of the Information Table and the Core attribute are the important attribute that influences attribute classification.

A subset $B$ of a set of attributes $C$ is a reduction of $C$ with respect to $R$ if and only if:
Table 1: Information system constructed using DC coefficients

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC1</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.41572</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC2</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.41573</td>
<td>0.35354</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC3</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35356</td>
<td>0.19134</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC4</td>
<td>0.35356</td>
<td>0.35356</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35351</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC5</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35336</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC6</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.35356</td>
<td>0.19134</td>
<td>0.35345</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC7</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC8</td>
<td>0.23761</td>
<td>0.23761</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
</tbody>
</table>

- $POS_b(R) = POS_{c}(R)$ and $POS_{b}(R) = POS_{c}(R)$ for any $a \in B$

And the Core can be defined by Eq. 9:

$$CORE_c(R) = \{\varepsilon \in C | \forall \varepsilon \in C, POS_{b-c}(R) \neq POS_{c}(R)\}$$ (7)

Main steps of algorithm

**Extraction of DCT coefficients from DCT domain:** In term of MPEG national standard, the video sequences in compressed domain consist of I, P and B frame. The I frame is the base of video sequences, which use DCT to compress in spatial, the processing of DCT and IDCT are showed as expressions (8) and expression (9), so the DCT coefficients can be easily extracted from video sequences directly. We can represent this process as representation Eq. 10:

$$F(u,v) = \frac{2}{N} e^{-\frac{\pi i}{2N}} \sum_{c=0}^{N-1} \sum_{j=0}^{N-1} f_c(j) \cos \left[ \frac{(2j+1) \pi v}{2N} \right] \cos \left[ \frac{(2j+1) \pi u}{2N} \right]$$ (8)

$$f_c(j) = \frac{2}{N} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} e^{-\frac{\pi i}{2N}} \cos \left[ \frac{(2j+1) \pi v}{2N} \right] \cos \left[ \frac{(2j+1) \pi u}{2N} \right] \Psi(F(u,v),t) \rightarrow \text{DCT coefficients}$$ (9)

where, $\Psi(F(u,v),t)$ denotes the video sequences. Table 1 shows part of $8 \times 8$ block DCT coefficients extracted from video sequences.

**Extract DCT and DC coefficients:** The DCT coefficients are made of DC coefficients and AC coefficients, DC coefficients denote the average and most important information in video frame. So we can utilize the DC coefficients to represent the video frame. This process can be described as representation (Eq. 11):

$$\text{DCT coefficients} \rightarrow \text{DC coefficients}$$ (11)

**Construct information system:** We have got the DC coefficients of each frame, so we can construct an Information System with it. Each row is a DC coefficient and each column is the frame. This process can be described as representation (Eq. 12):

$$\text{DC} \rightarrow \text{information table} S = \{U, A, V, f\}$$ (12)

where, $U$ is sets, denotes all the object of Information System, $A$ is also a sets, denotes all attributes in Information System, $V$ is the sets of attributes value, $f$ is a function denotes the relations between objects and attributes.

By using above process, we can get Information System as Table 1.

**Reduce information system:** The attributes in the information table can be divided into two types according to their roles: Core attributes and redundant attributes. From table 1, we can see that the frame 2 and frame 8 can be reduced, the reduced Information System is showed as Table 2. If we introduce a threshold, more attributes can be reduced.

The reduced frame called redundant attributes; they have no effect as to subdividing of DCT coefficients, such as frame 2 and frame 8. The attributes that can not reduce called CORESET in information system, which represents the salient content in video sequences.

After the process the volume of video data is reduced dramatically. While the main information of video is remained, so the efficiency of following processing is increased.

**Construct new information system:** After the frame is reduced, the Information System are consists of most important frames, we invert the row and column of Information Table, that is, each row is a frame and each column is the DC coefficients, we can get a new Information System as Table 3.

**Construct model and subdivide new information system:** Use the theory 2.3, we can also classify all frame into some parts, the model used during classification is defined by Eq. 13:
Table 2: The reduced information system

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC1</td>
<td>0.35355</td>
<td>0.27771</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.41572</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC2</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.41573</td>
<td>0.35354</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC3</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.35356</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC4</td>
<td>0.35356</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35351</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC5</td>
<td>0.35356</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35336</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC6</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.35356</td>
<td>0.35356</td>
<td>0.19134</td>
<td>0.35345</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC7</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.19134</td>
<td>0.35353</td>
<td>0.35355</td>
</tr>
<tr>
<td>DC8</td>
<td>0.27761</td>
<td>0.27779</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
</tr>
</tbody>
</table>

Table 3: The inverted information system

<table>
<thead>
<tr>
<th>Frame</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.35355</td>
<td>0.27779</td>
<td>0.27779</td>
<td>0.27761</td>
</tr>
</tbody>
</table>

Table 4: Evaluating result of shot cluster by various video sequences

<table>
<thead>
<tr>
<th>Video type</th>
<th>Correct cluster</th>
<th>Manually cluster</th>
<th>False cluster</th>
<th>Loose cluster</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>65</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>66</td>
</tr>
<tr>
<td>Animation</td>
<td>83</td>
<td>14</td>
<td>2</td>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>Sensory</td>
<td>142</td>
<td>26</td>
<td>5</td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>Story</td>
<td>102</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>78</td>
</tr>
<tr>
<td>News</td>
<td>126</td>
<td>23</td>
<td>3</td>
<td>3</td>
<td>85</td>
</tr>
</tbody>
</table>

Table 5: Total evaluation

<table>
<thead>
<tr>
<th>Video type</th>
<th>Correct detected rate</th>
<th>False detected rate</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>28</td>
<td>14</td>
<td>80</td>
</tr>
<tr>
<td>Animation</td>
<td>15</td>
<td>10</td>
<td>87</td>
</tr>
<tr>
<td>Sensory</td>
<td>19</td>
<td>19</td>
<td>78</td>
</tr>
<tr>
<td>Story</td>
<td>18</td>
<td>18</td>
<td>80</td>
</tr>
<tr>
<td>News</td>
<td>13</td>
<td>15</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
D(l_i, l_{i+1}) = \frac{1}{1024} \sum_{k=1}^{1024} \frac{|c(l_i, k) - c(l_{i+1}, k)|}{\max|c(l_i, k), c(l_{i+1}, k)|}
\]  

(13)

\[
\text{Recall} = \frac{\text{Correct}}{\text{Correct + Missed}}
\]  

(14)

\[
\text{Precision} = \frac{\text{Correct}}{\text{Correct + False Alarms}}
\]  

(15)

Various MPEG video sequences are selected to examine the performance of the proposed algorithm. Table 5 shows the evaluating results of shot boundary detection by various video sequences. By comparison, we can see the proposed algorithm can achieve satisfied results. Table 4 and 5 show the total evaluation of proposed algorithm.

**CONCLUSIONS**

Video shot clustering is a prerequisite for semantic video analysis, the study proposed a novel algorithm for shot clustering in compressed domain. To increase the clustering efficiency and make the results more scientific, the algorithm introduces theory of RS and reduces the redundant information of shot, then classifies the remain frame into cluster without any prior-knowledge. The experimental results proved the validity and feasibility of our proposed algorithm.

**REFERENCES**


