
This version is available at https://strathprints.strath.ac.uk/29416/

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Unless otherwise explicitly stated on the manuscript, Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Please check the manuscript for details of any other licences that may have been applied. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (https://strathprints.strath.ac.uk/) and the content of this paper for research or private study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to the Strathprints administrator: strathprints@strath.ac.uk
Extracting Semantics and Content Adaptive Summarisation for Effective Video Retrieval

Jinchang Ren and Jianmin Jiang

{j.ren, j.jiang1}@bradford.ac.uk
Digital Media and Systems Research Institute
University of Bradford, Bradford, U.K.
Contents

1. Introduction
2. Feature Extraction and Video Segmentation
3. Extracting Human Objects and Semantic Concepts
4. Content-adaptive Video Summarisation for Retrieval
5. Results and Discussions
6. Conclusions and Future Work
1. Introduction

- Content-based Information Retrieval (CBIR) has been widely investigated to overcome in text-based systems;
- Automatic extraction of semantics is one of the fundamental tasks for CBIR applications;
- It is particularly important to extract objects/semantics from content-rich video sources for effective retrieval;
- Content-adaptive summarisation is useful in achieving effective data representation and transmission;
1. Introduction

- Overall diagram is given
  - Two main blocks, i.e. online and offline parts;
  - Offline part includes low-level video processing and extraction of high-level semantics for content-based video indexing;
  - Online part includes video retrieval and content-adaptive summarisation for effective content delivery.
2. Feature Extraction & Video Segmentation

- Feature extraction on the basis of DC-images $Y_{dc}^{(i)}$, $U_{dc}^{(i)}$, $V_{dc}^{(i)}$
  - Inter-frame DC-differencing image $D(i)$
    $$D(i) = \sum_{ch} |ch_{dc}^{(i)} - ch_{dc}^{(i+1)}| / 3, \quad ch = Y, U, V$$
  - Mean and standard derivation of $D(i)$, $\mu(i)$ and $\sigma(i)$
  - $p_1(i)$ and $p_2(i)$ to denote two proportions of macroblocks whose changes in $D(i)$ above two adaptive thresholds;
    $$\lambda_1(i) = \frac{\mu(i)}{4} + 0.5 \quad \lambda_2(i) = \frac{\mu(i)}{4}$$
  - Motion prediction error and normalised energy:
    $$err(i) = C_i^{-1} \sum Y_{dc}^{(i)}(j), \quad 1 \leq j \leq C_i$$
    $$E_y(i) = E_0^{-1} \sum [Y_{dc}^{(i)}(j)]^2$$
2. Feature Extraction & Video Segmentation

- Detect cuts using extracted likelihoods, thresholding it followed by phase correlation on DC-images for validation;

- Detect gradual transitions by appearance-based modelling, such as fade out/in may lead to a V-shape in measuring the frame energy; dissolve has large prediction errors and its boundary frames are as different as a cut.

- $\ell_i$ is considered to measure local activity levels.

\[
\ell_i(\mu) = 1 - \frac{\mu(i-1)}{3\mu(i)}
\]

\[
\ell_i(\sigma) = 1 - \frac{\sigma(i-1)}{2\sigma(i)}
\]

\[
\ell_i(p) = \sqrt{p(i)}
\]

\[
\ell_i = \frac{\ell_i(\mu) + \ell_i(\sigma) + \ell_i(p)}{3}
\]
3. Extracting Human Objects & Semantics

- Human objects are detected via statistical modelling. Each colour entry is attached with probability as skin \( p_s(c) \) or non-skin \( p_n(c) \). Maximum likelihood strategy is then used for classification.

- SVM based supervised learning is employed to extract several semantics concepts like building, indoor/outdoor, and the sky. Colour and edge are the main features used for this purpose.
4. Content-adaptive Video Summarisation

- Content-adaptive criterion is employed to re-sample the original video for summarisation, where high activity levels are assigned with finer sample rates.

- Overall workflow for retrieval include
  I. For each video, detect shot boundaries as structuring events;
  II. Within each cut, detect human objects & semantic concepts;
  III. Using these events and semantics for shot-level content-based video indexing;
  IV. Video retrieval via specifying certain semantics;
  V. The retrieved videos are summarised for efficient network transmission and content delivery.
5. Results and Discussions

- Evaluation criteria: recall, precision and F1.
- Shot detection results shown in Table 1, overall performance 95%.
- Results on extracting semantics are shown in Table 2 and Fig. 2, average accuracy is about 85%.
- Results on video summarisation and retrieval are shown in Table 3 and Fig. 3 where a summarisation ratio of 20-30% is suggested to keep about 60-75% of the contents.

### Table 1: Average performance in terms of precision and recall rates for shot detection.

<table>
<thead>
<tr>
<th></th>
<th>Num.</th>
<th>Detect</th>
<th>Missed</th>
<th>False</th>
<th>Pre.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>cut</td>
<td>357</td>
<td>361</td>
<td>3</td>
<td>9</td>
<td>97.5%</td>
<td>98.6%</td>
<td>98.0%</td>
</tr>
<tr>
<td>GT</td>
<td>94</td>
<td>105</td>
<td>11</td>
<td>22</td>
<td>79.0%</td>
<td>88.3%</td>
<td>83.4%</td>
</tr>
<tr>
<td>All</td>
<td>451</td>
<td>466</td>
<td>16</td>
<td>31</td>
<td>93.3%</td>
<td>96.5%</td>
<td>94.9%</td>
</tr>
</tbody>
</table>

### Table 2: Average performance in terms of precision and recall rates for semantics extraction.

<table>
<thead>
<tr>
<th></th>
<th>Num.</th>
<th>Detect</th>
<th>Missed</th>
<th>False</th>
<th>Pre.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>outdoor</td>
<td>832</td>
<td>821</td>
<td>147</td>
<td>130</td>
<td>83.4%</td>
<td>82.3%</td>
<td>82.9%</td>
</tr>
<tr>
<td>indoor</td>
<td>1096</td>
<td>1055</td>
<td>171</td>
<td>136</td>
<td>87.7%</td>
<td>84.4%</td>
<td>86.0%</td>
</tr>
<tr>
<td>building</td>
<td>1214</td>
<td>1188</td>
<td>122</td>
<td>96</td>
<td>91.9%</td>
<td>90.0%</td>
<td>90.9%</td>
</tr>
<tr>
<td>sky</td>
<td>904</td>
<td>923</td>
<td>152</td>
<td>171</td>
<td>81.5%</td>
<td>83.2%</td>
<td>82.3%</td>
</tr>
<tr>
<td>Average</td>
<td>4046</td>
<td>3987</td>
<td>592</td>
<td>533</td>
<td>86.6%</td>
<td>85.4%</td>
<td>86.0%</td>
</tr>
</tbody>
</table>

### Table 3: Summarisation ratio vs. average quality index.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>31.2%</td>
<td>43.5%</td>
<td>57.3%</td>
<td>66.4%</td>
<td>74.1%</td>
<td>79.5%</td>
<td>84.9%</td>
</tr>
</tbody>
</table>
6. Conclusion and future work

Main contributions

- An effective system is presented for semantic video retrieval, which enables automatic extraction of human objects and several semantic concepts from low-level features.
- Using rule-based reasoning and machine learning, over 85% of semantics can be detected.
- Content-adaptive summarisation provides effective delivery of retrieval results while maintaining a high relevance score ranked by users.

Further investigation includes detection of more semantics and improvements in detecting gradual transitions;

Acknowledgement: The work is supported by the EU IST Framework Research Programme under projects HERMES (IST-216709) and LIVE (IST-4-027312).
Thank you!

Any Questions?