2009 NEM Summit "*Towards Future Media Internet*"

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.

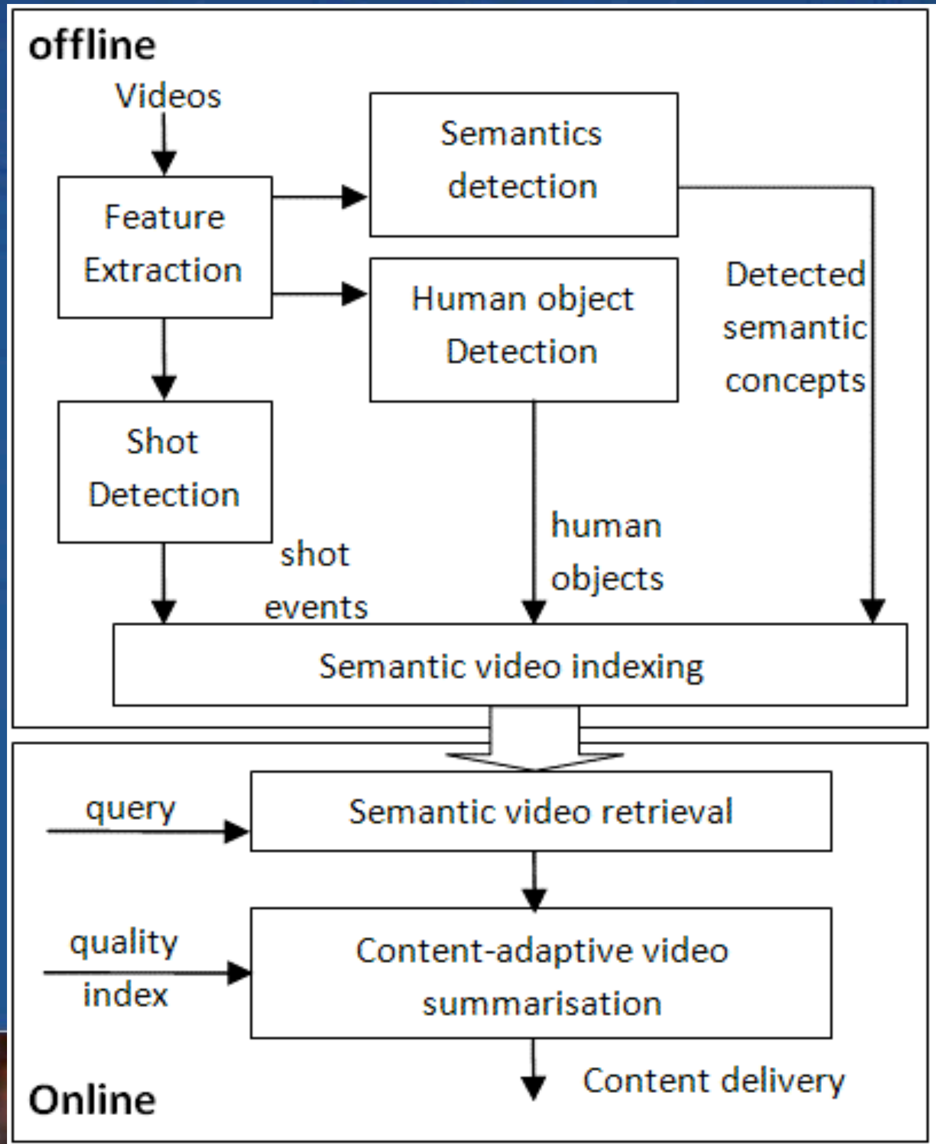
Contents

1. Introduction

- 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;

$$\lambda_1(i) = \mu(i)/4 + 0.5 \qquad \lambda_2(i) = \mu(i)/4$$

  - Motion prediction error and normalised energy:

$$err(i) = C_i^{-1} \sum Y_{dc}^{(i)}(j), \quad 1 \le j \le C_i \qquad E_y(i) = E_{0\_y}^{-1} \sum [Y_{dc}^{(i)}(j)]^2$$

2. Feature Extraction &Video Segmentation

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

$$\ell_i(\mu) = 1 - \mu(i-1)/[3\mu(i)]$$
$$\ell_i(\sigma) = 1 - \sigma(i-1)/[2\sigma(i)]$$
$$\ell_i(p) = sqrt(p(i))$$
$$\ell_i = [\ell_i(\mu) + \ell_i(\sigma) + \ell_i(p)]/3$$

- 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.
- ℓ_i is considered to measure local activity levels.

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%.

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

	Num.	Detect	Missed	False	Pre.	Recall	F1
cut	357	361	5	9	97.5%	98.6%	98.0%
GT	94	105	11	22	79.0%	88.3%	83.4%
All	451	466	16	31	93.3%	96.5%	94.9%

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

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

Table 3: Summarisation ratio vs. average quality index.

Ratio	10%	15%	20%	25%	30%	35%	40%
Quality	31.2%	43.5%	57.3%	66.4%	74.1%	79.5%	84.9%

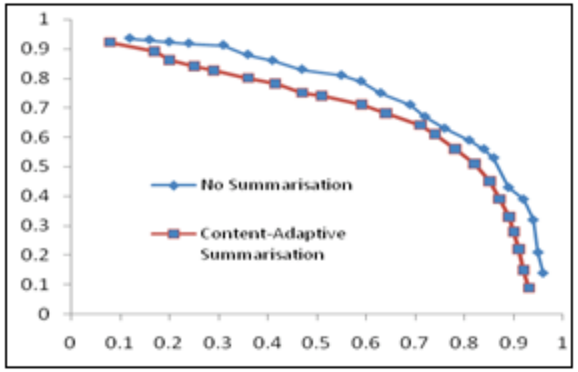


Figure 3: Comparison of retrieval performance using original and summarised video, where x- and y-axis respectively refer to precision and recall rates in terms of semantics retrieved.

- ✓ 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.

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 ?