TRECVID Automated and Interactive Search by NUS/ICT

Shi-Yong Neo, Yan-Tao Zheng, Hai-Kiat Goh, Tat-Seng Chua
School of Computing, National University of Singapore

Huanbo Luan, Juan Cao, Qiaoyan He, Sheng Tang, Yongdong Zhang
Institute of Computing Technology, Chinese Academy of Sci.
Overview

- Performed two tasks: Automated search & Interactive search

- Automated search:
  - process text and multimedia query
  - perform retrieval

- Interactive search:
  - Perform flexible relevance feedback, active learning, locality inference
  - Use motion icons (m-icon)
Automated Search
Auto Search Overview

- **Challenge:** *ASR and MT are not good,*
- **Solution:** incorporate multi-modal features to complement text
  - Effective query analysis and retrieval using HLF, motion and visual features.

**Framework**

- **Step 1:** induce and extract query-information
  - *query-class, query-HLF from the text query;*
  - *Query motion & visual features from available example keyframes/shots*
- **Step 2:** perform retrieval and ranking
Query Analysis

- Analyze queries to learn:
  - Query-class, Query-HLF, Query-image-feature and Query-shot-motion

- **Query-class**
  - Showed to be important functions by many prior works
  - Identified by heuristic rules using combination of noun, noun phrases, verbs, NE, etc
  - Function as a guide to fuse multi-modal features effectively.
  - Determined by a set of firing rules for each class:
    - We exploit {Scene, People, Object, Action, Unknown}.
      - \{Unknown\} class is to accommodate the queries that do not belong to any of the first four classes.
      - Other classes cover 19 out of 24 queries
Query Analysis: Query-HLF

- Query-HLF suggests possible HLFs that are important to the query in terms of visual requirements.
- Employ morphological analysis and selective expansion using WordNet on HLFs descriptions and query.
  - Stronger the match between HLF descriptions and query => the more important the HLF is to the query.
- Infer query-HLF from sample keyframes and shots
  - A sample image containing one of the HLFs could explicitly mean that the particular HLF can be important.
- Combine inference from text query and video shots to obtain a better and more representative query-HLF for query.
Query Analysis: Query-image-feature

- **Query-image-feature** ($Q_{IMG}$) corresponds to video features extracted from sample keyframes and video shots.

- **Step 1**: extract three visual features from all the sample keyframes
  - A 320-dimensional vector of edge histograms ($EH$) on 5 regions;
  - A 166-dimensional color histogram ($CH$) vector in HSV space;
  - A set of visual words ($VW$) constructed based on 128-dimensional SIFT vector

- **Step 2**: learn three nonparametric LDA models based on above three visual features ($CH$, $EH$, $VW$)
  - Obtain the latent topic distribution of every shot.
Query Analysis: Query-motion features

- A number of query topics are highly associated with motions.
  - For example,
    - Query “finding shots of train in motion” and “find shots in which a boat moves past” tend to present large horizontal translational global motions in the shot,
    - Query “find shots of a road taken from a moving vehicle through the front windshield” tends to present zoom-like diffusing global motions,
  - We use 2 descriptors for global motion patterns
    - 8-dimensional vector of motion directions: up, down, left, right, up-left, up-right, down-left and down-right
    - 1D global motion intensity: still, median, etc
  - The motion cues are extracted from motion vectors stored in p-frames in compressed domain
    - High efficiency: processing around 50-hour testing videos in approximately 40 hours.
Shot Level Retrieval

➢ Fuse the ASR & MT text, Query-HLF, Query-image-feature and Query-shot-motion

\[
\text{Score} (Q, \text{Shot}_j) = \beta_c \cdot \text{Text} (Q, \text{words} \mid \text{words} \in \text{Shot}_j) + \\
\gamma_c \cdot \sum_{\text{HLF}_m \in \text{shot}_j} [\text{Conf} (\text{HLF}_m) \times \text{Sim}_\text{Lex} (Q_{\text{HLF}}, \text{HLF}_m)] + \\
\delta_c \cdot \max_{\text{image}_n \in \text{IMG}} (\text{image}_\text{sim} (\text{image}_n, \text{shot}_j)) + \\
\chi_c \cdot \max_{\text{image}_n \in \text{IMG}} (\text{motion}_\text{sim} (\text{image}_n, \text{shot}_j))
\]
Experimental Results

- Performed 5 runs to progressively evaluate effect of HLF, visual and motion features

- **Run1**: *Required text baseline;
- **Run2**: *Required visual baseline;
- **Run3**: *Fusion without motion using only text query;
- **Run4**: *Fusion with motion using only text query;
- **Run5**: *Fusion with motion using multimedia query;
Experimental Results

- Firstly, the worst performing run (Run1: MAP 0.004) comes from the text baseline.
  - ASR and MT text are not erroneous and thus less predictive than HLF and visual counterparts.
- The visual baseline (Run2: MAP 0.017) in contrast yields much better results.
- Improvements in Run3 and Run4 show that the use of HLF and motion features is effective.
- Run5 (0.061) delivers the highest MAP by multimedia queries.
- Observations:
  - HLFs are one of most important features
  - Motion is effective in certain queries
  - Visual and motion features tend to complement text and HLF features
  - Query content from multimedia counterpart is more discriminating than text alone
Interactive Search
Introduction

- Poor performance of fully auto search
- More intelligent system is demanded
- Solution: interactive search
  - Incorporate user’s feedback to refine the results

- Our emphases for interactive search:
  - Effective UI (User Interface)
    - To maximize user’s annotation speed
  - Multiple feedback strategies
    - To provide multiple refinement options to users
  - Motion icons
    - Design Moving Icons (M-icons) to give info on motion of the shots
Overall Framework

Auto Search Stage

Start

Query

Auto Search

New list of relevant shots

Relevant shots

Interactive Search Stage

Recall-driven

Precision-driven

Re-rank shots in subset

Locality-driven

Neighbor-shots rank list

Strategy Selection

Labeling

Results
Intuitive User Interface

- **UI Design Basis**
  - Fast perception
    - display 3 shots in each row
    - optimum for keystroke action
  - Quick previews of previous & subsequent rank shots
  - Flexible annotation modes
    - manual, semi-auto, auto
    - control flow of shot browsing
  - Query by HLF
  - Retrieval Statistics
  - Self-contained, separated from backend server and Web-enabled
    - *UI developed by Macromedia flash*
Intuitive User Interface

- **UI Design Basis**
  - Fast annotation
  - *keystroke actions, labeling by clicking on keyboard buttons*

- **Efficiency**
  - Approximate 3,500 shots based on motion icons in 15 mins
  - Approximate 5,000 shots based on static icons in 15 mins
Multiple Feedback Strategies I

- **Strategy 1: Recall-directed feedback**
  - Aim: maximize recall performance
  - Extract useful text token and HLF from labeled relevant shots for query expansion
  - Features: text and HLF

- **Strategy 2: Precision-directed feedback**
  - Aim: improve precision of retrieved shots by refining classifier
  - Adaptive sampling strategy for active learning based on SVM
  - Multimodal features: visual, HLF, motion
  - Real time training and classification
Multiple Feedback Strategies II

- **Strategy 3: Semantic coherence (neighborhood inference)**
  - Temporal locality-driven: return neighboring shots of the positive
  - Documentary videos possess high temporal coherency of same topic
  - Neighboring shots tend to be relevant
  - Select neighbors by sliding window
  - **Example:** find shots of street market

![Example images of shots from a street market]
Why Multiple Feedback Strategies?

- More options for users
- More robustness in feedback
- More flexibility for cross-domain annotation
  - For news corpus (TRECV06), recall-driven feedback is effective
    - ASR text is richly available
  - For documentary corpus (TRECV07), neighborhood inference works well
    - Documentary video tends to be of high temporal coherence.
Motion Icons

- **Motivation**
  - Many queries are associated with objects in motion in the video.
  - Static keyframes contain deficient information about video content

- **Our Approach**
  - Construct a summarized clip comprising a sequence of keyframes which can show moving picture information.
  - Motion icon possesses more comprehensive info. than static keyframe
  - Users can have a clearer idea of shot content and identify relevant shots with better confidence
Motion Icons

Example 1: find shots of train in motion

Example 2: find shots of a canal, river, or stream with some of both banks visible
Experiments

- We submitted one run of interactive search
- MAP of 0.251 and 5\(^{th}\) best performing run
- 2 topics achieves highest MAP and 18 out of 20 topics are above median
- 1 query (“Find shots of people and dogs walking”) has no relevant shots found, which lowers overall MAP badly.
Conclusion and Future Work

Focus of Interactive Search

- Efficient UI
- Multiple Feedback Strategies
- M-icon

Future Work

- Can we extend our system to non-expert users?
- Challenges: When to do feedback, which strategy to choose?
- Solution: Recommendation mechanism
  - Analyze experts behavior pattern based on activity log
  - Annotation statistics of non-expert users
Thank You

Q & A