Concept Detection For Semantic Video Retrieval

Duy-Dinh LE and Shin'ichi SATOH

National Institute of Informatics

Motivation

- **Concept detection** is one of the important tasks in video indexing due to its importance to bridging the semantic gap in multimedia retrieval
- Many methods have been proposed for this task, however finding a method which can **generalize** well for a large number of concepts and is **scalable** for processing **huge video databases** is still challenging.

Proposal

- We introduce a **general framework** for efficient and scalable concept detection by **fusing SVM classifiers** trained by only **simple visual features.**
- We employ the proposed framework for detecting a large number of concepts on **various video datasets** with several thousand hours of video and show the results in our demo.

Framework Overview

Experiments and Evaluations

1. Video Datasets

- TRECVID 2004-2008: 658 hours, 659,322 keyframes.
- NHKNews7 2001-2009: 1,413 hours, 747,529 keyframes.
- JP News and Documentary Video, 2005-2009: 11,335 hours, 680,106 keyframes.
- **Total: 13,406 hours, 2,086,957 keyframes.**

2. Concepts

- LSCOM (http://www.lscom.org/) : **374 concepts.**
- **Six categories** on a top level: objects, activities/events, scenes/locations, people, graphics, and program categories.

3. Annotations for 374 concepts

- Data: 80 hours, 70,000 keyframes of TRECVID 2005 (News programs in US)
- **Human judgements: 28 millions.**
- The number of positive samples for each concept ranges from several hundreds to several thousands.

4. Training Concept Detectors

- SVM with RBF kernel. Optimal params are found by grid search.
- Three types of features: color moments, edge orientation histogram and local binary patterns.
- For each feature, train 4 classifiers for each concept to handle the problem of imbalanced training set.
- Training time: 1-2 hours/classifiers.
- In total, **4,488 classifiers** were trained (374 concepts x 3 features x 4 classifiers).

5. Predicting Concepts in Test Data

- For each shot, one or several keyframes are extracted.
- The three features are extracted from each keyframe and used to form the feature vector.
- Run the 374 concept detectors (12 classifiers/concept detector) on the feature vector of each keyframe. For each concept detector, the scores of the classifiers are fused by taking average.
- Prediction time for 374 concept detectors: 20 keyframes/hours.

6. Results

- Our system ranked second in TRECVID 2007 and the Star Challenge Competition 2008.
- Our approach are used in the Information Grand Voyage Project.
- The results are integrated in NII-KAORI video search engine.

NII

Duy-Dinh LE and Shin'ichi SATOH, National Institute of Informatics
TEL : 03-4212-2583 FAX : 03-3556-1916 Email : {ledduy,satoh}@nii.ac.jp