

TSUBAME 共同利用 平成 24 年度 学術利用 成果報告書

利用課題名 マルチメディア内容解析に関する研究
英文: A Study on Multimedia Content Analysis

利用課題責任者 佐藤 真一
First name Surname Shin'ichi Satoh

所属 国立情報学研究所
Affiliation National Institute of Informatics
URL <http://www.nii.ac.jp>

邦文抄録(300 字程度)

本研究では、映像等のマルチメディアコンテンツの主として視覚情報を解析し、その意味内容情報を自動抽出する手法について検討する。特に、映像・画像からの特徴量抽出、学習データを用いた意味内容識別器の学習、並びに評価用データの識別・評価という、一連のマルチメディア意味解析処理の TSUBAME における効率の良い実現を目指す。

英文抄録(100 words 程度)

We study multimedia content analysis, namely, analyzing visual information and delineating semantic content information. Especially, we try to investigate efficient implementation of multimedia semantic content analysis pipeline on TSUBAME, including extraction of features from images and videos, training semantic content classifiers using training data, classification and evaluation using test data.

Keywords: 5つ程度

マルチメディア内容解析、特徴抽出、識別器
Multimedia Content Analysis, Feature Extraction, Classifiers

1. Background and Purpose

We have developed NII-KAORI-SECODE [1], a general framework for semantic concept detection, and use it to participate several benchmarks such as IMAGE-CLEF [2], PASCAL-VOC and TRECVID [3]. The purpose is to evaluate performance of various visual representations for concept detection-like task. In this framework, first features are extracted from keyframes, then concept detectors using these features are learned by using SVM with chi-square RBF kernel. The probability output scores of the learned concept detectors are used for ranking.

We evaluate both global features and local features. The global features include color

moments, color histogram, edge orientation histogram, and local binary patterns. The local feature is based on the BOW model in which the SIFT descriptor is extracted at interest points detected by affine invariant and multi-scale dense sampling detectors.

For such benchmarks as IMAGE-CLEF, PASCAL-VOC, TRECVID, to achieve high detection performance, it requires to use as many features as possible. Therefore, a huge computational resource is needed. For example, BBNVISER team (USA) used a shared 309 node, Dual Quad Core, Intel(R) Xeon(R), 32-48GB RAM/node cluster to run for more than 1,000

hours to detect 10 events in 4,000 hours of video.

The purpose of using TSUBAME is to investigate the implementation of the KAORI-SECODE for such large scale experiments. Specically, we are studying how to efficiently handle large I/O requests when processing a large number of images and their associated features; large number of jobs (e.g. training and testing classifiers) needed to run in parallel so that the experiments can be finished in a short time;

2. Technical Details

A general framework for building semantic content detectors is shown in Figure 1 (Courtesy of IBM-Columbia Team in TRECVID 2011).

We use VLFEAT library [4] to extract local features such as DSIFT and PHOW using dense sampling at multi-scales. For each image frame, the number of local features is approx. 10,000. The total number of image frames is approx. 1,000,000. These local features are quantized into visual words using 1,000-word codebook. After quantization, soft assignment is used to form a feature vector for each frame.

We use these features to train and test 100 classifiers for ImageCLEF – Photo Annotation Task, 346 classifiers for TRECVID – Semantic Indexing Task, and 20 classifiers for TRECVID – Multimedia Event Detection Task.

The huge number of features, images, and classifier models require a huge disk space to store and I/O requests to process. When multi-jobs (usually hundreds to thousands of jobs) access the data for I/O at the same time, it causes deadlock at disk storages, network bandwidths.

By using TSUBAME, we would like to study how these problems can be solved.

3. Results and Discussions

We report here experiments for ImageCLEF Photo Annotation Task. The feature configurations are shown in Table 1. The detection performance is shown in Table 2, which is consistent with our submission last year.

It took about 2 hours to extract one type of feature (SIFT-variations) for 1,000 images. Quantization and soft-assignment take about 3.5 hours. It takes 1 to 4 hours to train a classifier (depending on the number of training samples), and 2 hours for testing the trained classifier to test images.

Reference

- [1] Duy-Dinh Le, Shin'ichi Satoh: A Comprehensive Study of Feature Representations for Semantic Concept Detection. ICSC 2011: 235-238.
- [2] Duy-Dinh Le, Shin'ichi Satoh: NII, Japan at ImageCLEF 2011 Photo Annotation Task. CLEF (Notebook Papers/Labs/Workshop) 2011.
- [3] Duy-Dinh Le, Sébastien Poullot, Xiaomeng Wu, Bertrand Nouvel, Shin'ichi Satoh: National Institute of Informatics, Japan at TRECVID 2010. TRECVID 2010.
- [4] VLFEAT: <http://www.vlfeat.org/>

Figure 1: A general framework for building semantic content detectors (Courtesy of IBM-CU team at TRECVID-MED 2011).

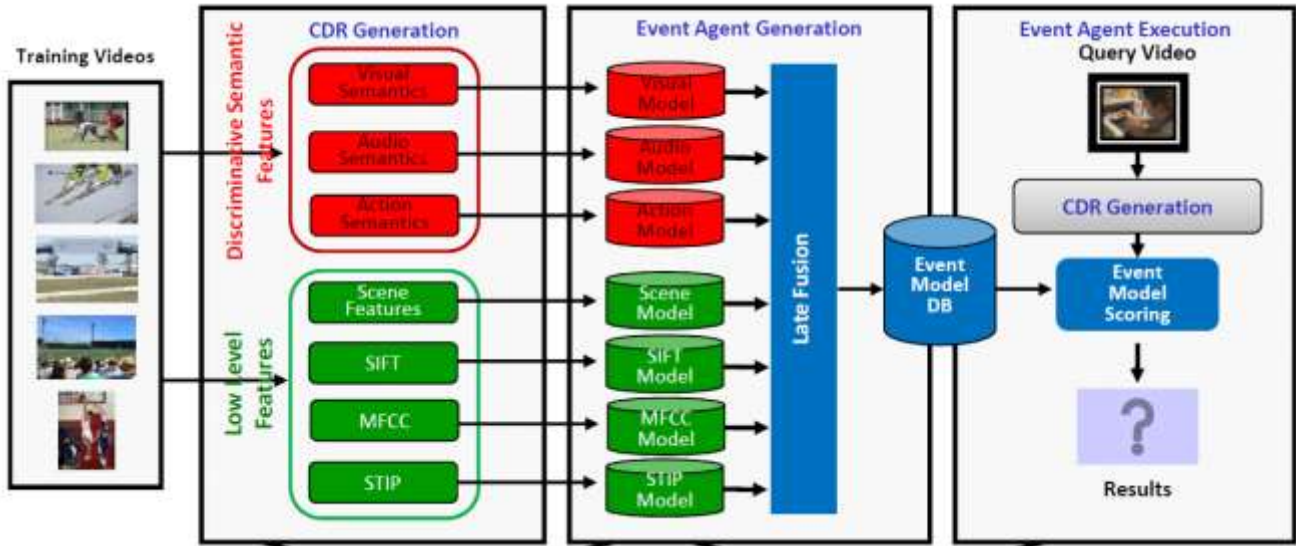


Table 1. Feature configurations.

Feature Type	Granularity	Color Space	Quantization (#Bins)	Total Confgs
Color moments (CM)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	3x3	20
Color histogram (CH)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	8x3	20
Local binary patterns (LBP)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	10, 30, 59	15
Edge orientation histogram (EOH)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	12, 18, 36, 72	20
Local features (harhes, harlap, heslap, haraff, hesaff, dense, phow)	1x1, 2x2, 1x3, 3x1	GRAY	500 visual words	44

Table 2. Performance of feature confgs.

Feature Config	MAP (%)	Note
Local Feature Full + Global Feature	35.8854	Bug free
Local Feature Full + Global Feature - Bug	33.7069	Submitted run NII.R1 but having bugs
Local Feature Full	34.4895	Bug free
Local Feature Full - Bug	33.0984	Submitted run NII.R2 but having bugs
Local Feature Light	33.4979	Submitted run NII.R4 but having bugs
Local Feature Light + Global Feature	32.3061	Submitted run NII.R3 but having bugs
Global Feature	27.8498	Submitted run NII.R5