LARGE SCALE SEMANTIC CONCEPT DETECTION, FUSION, AND SELECTION FOR DOMAIN ADAPTIVE VIDEO SEARCH

JIANG YU-GANG

DOCTOR OF PHILOSOPHY

CITY UNIVERSITY OF HONG KONG

SEPTEMBER 2009

CITY UNIVERSITY OF HONG KONG
香港城市大學

Large Scale Semantic Concept Detection, Fusion, and Selection for Domain Adaptive Video Search
大規模語義概念的檢測、融合及選擇進行數據域自適應視頻檢索

Submitted to
Department of Computer Science
電腦科學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

JIANG Yu-Gang
姜育剛

September 2009
二零零九年九月

ABSTRACT

This thesis investigates the problem of video search based on semantic concepts. We present approaches to handle three correlated issues that are critical to this problem: (1) *how to construct an effective feature representation for semantic concept detection,* (2) *how to exploit semantic context to improve the detection of these concepts,* and (3) *how to select the most suitable concept detectors to answer user queries.* In particular, as the target videos may come from different domains (genres or sources) with distinctive data characteristics, for each of the issues, we will need to cope with the domain changes.

Video frames are represented by bag-of-visual-words (BoW) derived from local keypoint features, which are invariant to rotation, scale and illumination. We first conduct a comprehensive study on the representation choices of BoW, including vocabulary size, weighting scheme, stop word removal, feature selection, spatial information, and visual bi-gram. The aim is to offer practical insights in how these choices will impact the performance of BoW for semantic concept detection. We also show how to further augment the BoW representation by exploring the linguistic and ontological aspects of visual words. A visual-word ontology is constructed to hierarchically specify their hyponym relationship, which is incorporated into BoW for improved video frame representation.

To exploit semantic context, we develop a novel and efficient domain adaptive semantic diffusion algorithm. Inter-concept relationship is modeled using a semantic graph, which treats concepts as nodes and the concept affinities as the weights of edges. It is then applied to refine the initial detection results through a function level graph diffusion process, aiming to recover the consistency and smoothness of the detection results over the graph. To handle the domain change between training and test sets, our algorithm involves a graph adaptation process which iteratively refines the concept affinity based on the target domain data characteristics. This algorithm is efficient and scalable to large scale data sets.

For the selection of concept detectors, we focus on exploring heterogeneous knowledge sources for better measurement of query-detector similarity. Instead of using WordNet as in most existing works, we exploit the context information associated with Flickr images to estimate the similarity between queries and concept detectors. This similarity measure, named FCS, reflects the word correlation in images rather than text corpora. With an initial detector set selected by FCS for each query, we further propose a semantic context transfer algorithm that adapts the query-detector similarity to a target data set. The adaptation process is highly efficient, satisfying the critical requirement of online video search.

We evaluate all the proposed techniques on large scale video search benchmarks provided by TRECVID from years 2005 to 2008. Experimental evaluations demonstrate promising results of our techniques, and their potential to be applied to other applications such as visual object categorization and web scale image retrieval.

CONTENTS

List of Figures **x**

List of Tables **xiv**

1 Introduction **1**

1.1 From Annotating to Searching Videos 1

1.2 Thesis Overview 4

1.2.1 Concept-based Video Search 4

1.2.2 Methods and Contributions 5

1.3 Organization 9

1.4 Publications 10

2 Related Work **11**

2.1 Large Scale Semantic Concept Detection 11

2.1.1 General Framework 12

2.1.2 Feature Representation 14

2.1.3 Context-based Concept Fusion 17

2.2 Concept Selection for Video Search 19

2.3 Coping with Data Domain Changes 21

2.4 Background: TREC Video Retrieval Evaluation (TRECVID) . . . 22

3 Representations of Keypoint-Based Concept Detection: A Comprehensive Study **25**

3.1 Bag-of-Visual-Words (BoW) Feature 26

3.2 Representation Choices 28

3.2.1 Vocabulary Size 28

3.2.2 Weighting Schemes 29

3.2.3 Stop Word Removal 30

3.2.4 Feature Selection 31

3.2.5 Spatial Information 32

3.2.6 Visual Bi-gram 33

3.3 Kernel Choices of BoW Classification 34

3.4 Empirical Study 36

3.4.1 Experimental Setup 37

3.4.2 Weighting Schemes and Vocabulary Sizes 40

3.4.3 Stop Word Removal 42

3.4.4 Feature Selection 42

3.4.5 Spatial Information 46

3.4.6 Visual Bi-gram 48

3.4.7 Kernel Choice 50

3.5 Discussion 52

3.5.1 Generalizability to Other Data Sets 52
3.5.2 Fusion with Color/Texture Features 56
3.6 Cross-Domain Concept Detection 57
3.7 Summary 59

4 Visual Word Proximity and Linguistics 61
4.1 Motivation 61
4.2 Modeling Visual Linguistics 64
4.2.1 Bag-of-Visual-Words Ontology 64
4.2.2 Linguistic Similarity of Visual Words 66
4.3 Proximity Weighting With Visual Linguistics 68
4.4 Linguistic Matching with CEMD 70
4.4.1 Constraint-based EMD 71
4.5 Experiment I: Semantic Concept Detection 73
4.5.1 Data Set and Experimental Setup 73
4.5.2 Visual Word Weighting 74
4.6 Experiment II: Near-Duplicate Keyframe Retrieval 75
4.6.1 Data Set and Experimental Setup 76
4.6.2 Effect of Linguistic Similarity Measure 77
4.6.3 Comparison of CEMD and EMD 79
4.6.4 Effect of Word Weighting 79
4.6.5 Performance on TRECVID 2006 Test Set 80
4.7 Summary 82

5 Contextual Concept Fusion 84
5.1 Motivation 84
5.2 Preliminaries 87
5.2.1 Notations 87
5.2.2 Graph Diffusion 88
5.3 Efficient Diffusion of Semantic Context 89
5.3.1 Semantic Graph 89
5.3.2 Semantic Diffusion 90
5.4 Domain Adaptive Semantic Diffusion 92
5.5 Experimental Setup 94
5.6 Results and Comparison 97
5.6.1 Effect of Parameters 99
5.6.2 Effect of Negative Correlation 100
5.6.3 Effect of Concept Affinity Estimation Method 100
5.6.4 Effect of Baseline Performance 101
5.6.5 Comparison with the State of the Arts 102
5.6.6 Run Time 103
5.7 Summary 103

6 Domain Adaptive Concept Selection for Online Search 105
6.1 Motivation 105

6.2 Flickr Context Similarity 109
6.3 Semantic Context Transfer 112
6.3.1 Offline Detector Refinement 114
6.3.2 Online Adaptation of Query-Detector Similarity 116
6.3.3 Algorithm Summary and Discussion 119
6.4 Experiments 121
6.4.1 Data Sets and Evaluation 121
6.4.2 Query-Detector Similarity Measure 122
6.4.3 Effect of Semantic Context Transfer 124
6.4.4 Performance based on Query Types 127
6.4.5 Comparison with the State of the Arts 129
6.5 Summary 130

7 Conclusion **131**
7.1 Summary of Contributions 131
7.2 Future Directions 133

Appendix **136**

A VIREO-374: LSCOM Semantic Concept Detectors **136**
A.1 Implementation Details 136
A.2 Fusion with Columbia374 (CU-VIREO374) 138
A.3 Download Statistics 141
A.4 List of Concept Names 142

References **145**

LIST OF FIGURES

1.1 Three video clips from different news broadcasting channels. From top to bottom: CCTV, MSNBC, and LBC. 2
1.2 Highly diversified instances of two semantic concepts. 3
1.3 System diagram of concept-based video search. 5
1.4 Multiple semantic concepts co-exist in the same video frame. . . . 7
2.1 A general framework for semantic concept detection. 12
2.2 Local invariant features. Salient patches are detected over multiple image scales (Hessian affine detector in this example) and then described by SIFT (128-dimensional vector). The detector and descriptor are designed to be stable over certain variations, such as orientation, scale and etc. 15
3.1 Image representation using bag-of-visual-words. 27
3.2 Keyframe examples of 20 semantic categories in TRECVID 2006 data set. 37
3.3 Image examples of different categories in PASCAL VOC 2005 data set. 38
3.4 Classification performance with stop word removal. 43
3.5 Classification performance on TRECVID 2006 under vocabularies pruned using various feature selection criteria. 44
3.6 Classification performance on PASCAL VOC 2005 under vocabularies pruned using various feature selection criteria. 45
3.7 Classification performance on TRECVID 2006 using region-based features computed from different spatial partitions. 47
3.8 Classification performance on PASCAL VOC 2005 using region-based features computed from different spatial partitions. 48
3.9 Performance on TRECVID 2006 using fusion of visual bi-grams and visual words. 49
3.10 Instances of *flag-US* with different backgrounds in TRECVID 2006 data set. 52
3.11 Concept detection performance on TRECVID 2008 and PASCAL VOC 2007 data sets using region-based BoW features computed from different spatial partitions. At each single partition choice, the MAP performances are similar. However, the fusion of different spatial partition choices significantly improves the performance (16% and 10% MAP improvements for TRECVID 2008 and PASCAL VOC 2007 respectively). 54
3.12 Performance comparison with the top-50 (out of 161) official Type-A runs submitted to TRECVID 2008. 55
3.13 Performance comparison with the 17 official submissions of PASCAL VOC 2007. 55

3.14 Within-domain and cross-domain concept detection performance on TRECVID 2007 using BoW and global features (CM+WT) respectively. The concepts are in descending order of frequency in TRECVID 2007 test set (from left to right). 59
4.1 Comparing keyframe similarity using bag-of-visual-words. 62
4.2 Visual word ontology (b) as a bridge between expensive keypoint matching (a) and simple bin-to-bin word comparison (c). 64
4.3 Proximity-weighting scheme: (a) keypoint p_j is *softly* assigned to multiple visual word v_{j1} and v_{ji}; (b) information content (cluster sizes of v_1 and v_2) is used to assess the importance of v_2 to keypoint p. 69
4.4 Sensitivity to parameter α. 74
4.5 Examples of near-duplicate keyframes: (a) different acquisition time; (b) lens variations; (c) video editing. 76
4.6 NDK retrieval performance on Columbia data set: (a) Comparison of linguistic measures; (b) Performance of CEMD versus EMD; (c) Effect of word weighting on CEMD. 78
4.7 NDK retrieval performance on TRECVID 2006 test set. 81
4.8 Examples of near-duplicate keyframes retrieved by CEMD-Soft. The left most five examples are query keyframes, followed by the most similar retrieved keyframes. The true positives are marked in red boxes. 82
5.1 Illustration of context-based video annotation. (a) Top 5 video shots of concept *desert* according to the annotation scores from an existing pre-trained detector, in which the semantic context was not considered. (b) Refined shot list by semantic diffusion. The subgraph on the top shows two concepts with higher correlations to *desert*. Line width indicates graph edge weight. (c) Refined subgraph and shot list by the proposed domain adaptive semantic diffusion (DASD). The graph adaptation process in DASD is able to refine concept relationship which in turn can further help improve the annotation accuracy. 86
5.2 A fraction of the semantic graph before and after domain adaptation. Thick edges represent strong correlation between concepts, quantified by the values nearby the edges. (a) The initial concept relationship computed using the manual annotations on TRECVID 2005 development set; (b) The updated concept relationship for TRECVID 2007 test set after performing domain adaptation (more explanations in Section 5.6). 90
5.3 Example keyframes of the concepts evaluated in TRECVID 2005–2007. The upper part shows examples from broadcast news videos in 2005 and 2006, while the lower part shows examples from documentary videos in 2007. Note that the appearance of the same concept from the two data domains may be visually very different. 96

5.4 Per-concept performance before and after semantic diffusion on TRECVID 2006 test set. Consistent improvements are observed for all of the 20 semantic concepts. 98
5.5 The MAP performance on TRECVID 2006 test set under various parameter settings. 100
5.6 Performance of DASD using various baselines on TRECVID 2006 test set. The text under blue and orange bars indicates the feature used in the corresponding SVM detector. 102
6.1 System architecture for domain adaptive video search, illustrated using a query from TRECVID 2007. Flickr context similarity is firstly applied to select a relevant detector set, which is then adaptively refined through transferring semantic context learnt from target data domain. The search performance in terms of average precision over the top-10 retrieved video shots is significantly improved by 164% after domain adaptation. The video shot rank lists are ordered from left to right and top to bottom (false positives are marked in red boxes). 108
6.2 (a) Rich context information associated with a Flickr image. (b) The total number of images returned using keyword-based search in Flickr image context. 109
6.3 The frequency of 374 LSCOM semantic concepts in various sources. Note that the Y-axis is plotted in log scale. 111
6.4 Performance comparison of different query-detector similarity measurements on TV05-08 test sets. 123
6.5 MAP performance on TV05 by varying the number of iterations for online semantic context transfer. 126
6.6 Performance of different query types. The number of queries for each type is shown in the parenthesis. PT and NE stand for personthings and name-entity respectively. 128
6.7 MAP comparison with the top-50 (out of 82) official submissions of the automatic video search task in TRECVID 2008. 130
7.1 Top search results of query *Tiger* from video sharing site YouTube. The three videos are respectively about golf star *Tiger Woods*, animal *tiger* and an entertainment program from *Tiger Beat Magazine*.134
A.1 VIREO-374 detection results: top ranked shots of concept *US_flags* in TRECVID 2006 test set. 137
A.2 Data partitions in TRECVID 2007 and 2008. Both development and test data of TRECVID 2007 are used as development data of TRECVID 2008. 139
A.3 Per-concept performance on TRECVID 2006 test data. The models tested here are trained using TRECVID 2005 development data.140
A.4 Per-concept performance on TRECVID 2007 test data. The models tested here are re-trained using TRECVID 2007 development data. 140

A.5 Performance comparison of VIREO-374, Columbia374, and their fusion with all official TRECVID 2006 concept detection systems. 141

A.6 Performance comparison of VIREO-374, Columbia374, and their fusion with all official TRECVID 2007 concept detection systems. 141

LIST OF TABLES

2.1 Descriptions of TV05–08 data sets. The total number of video shots in each data set is shown in the parenthesis. The 160h (100h) videos from TV05 (07) are used as development data for TV06 (08). 23

3.1 The mean EER on PASCAL VOC 2005 data set using different weighting schemes and vocabulary sizes. The best result at each vocabulary size is shown in bold. 40

3.2 The MAP on TRECVID 2006 data set using different weighting schemes and vocabulary sizes. The best result at each vocabulary size is shown in bold. 40

3.3 The detection performance on TRECVID 2006 and PASCAL VOC 2005 data sets using SVM with different *kernels*. The best results are given in bold. (Note that the evaluation metrics are different on the two data sets) 50

3.4 MAP performance of fusing BoW with color moment (CM) and/or wavelet (WT) on TRECVID 2008 data set. The percentage in the parenthesis shows the degree of improvement over BoW only feature (2rd column). 57

3.5 MAP performance on TRECVID 2007 data set, using models trained on TRECVID 2005 and 2007 respectively. The percentages show the performance drop of cross-domain detection compared to within-domain detection using the same feature representation. 58

4.1 Performance (mean InfAP) of the proposed proximity-weighting scheme on TRECVID 2006 test set. 75

4.2 Per query NDK retrieval efficiency on TRECVID 2006 test set. . . 82

5.1 Overall performance gain (relative improvement) on TRECVID 2005–2007 data sets. SD: semantic diffusion. DASD: domain adaptive semantic diffusion. 98

5.2 Performance comparison of DASD with several existing works. . . 102

5.3 Run time of SD and DASD on TRECVID 2005–2007 data sets. The experiments are conducted on a Intel Core 2 Duo 2.2GHz PC with 2G RAM. 103

6.1 Detector selection using various query-detector similarity measurements. The detectors are selected based on the query words shown in bold. 123

6.2 Search performance on TV05–08. MAP-k means MAP over top k ranked shots[3]. 125

6.3 Significance test based on query types. $x \gg y$ means x is significantly better than y. 128

6.4 Performance comparison on TV05–08 test sets. 129

A.1 Features used in VIREO-374 and Columbia374. 138

A.2 Total number of download requests received for VIREO-374 models and their detection scores on each year's TRECVID data set. . . 141
A.3 List of concept names and per-concept development data (DD) of CU-VIREO374 models. 142