RECENT ADVANCES AND CHALLENGES OF SEMANTIC IMAGE/VIDEO SEARCH

Shih-Fu Chang[1], Wei-Ying Ma[2], and Arnold Smeulders[3]

[1] Columbia University, New York, USA
[2] Microsoft Research Asia, Beijing, China
[3]University of Amsterdam, Netherlands

ABSTRACT

We present an overview of recent advances and major challenges in image and video search, with a specific focus on large-scale semantic concept detection and indexing. Such semantic indexing paradigm has been driven by the increasing availability of the large resources of corpora, novel labeling approaches, innovative image features, and machine learning techniques for visual content recognition. We will discus key approaches, recent results, and novel applications in text-to-concept semantic search and multi-modal retrieval models. Open issues and major opportunities will also be presented.

Index Terms — semantic indexing, image and video search, content labeling, statistical modeling

1. INTRODUCTION

Due to the explosive growth of visual data (both online and offline) and the phenomenal success in Web search, there has been increasing expectation for search technologies for images and videos. Although many technical challenges remain, in the past several years we have seen exciting progress and new potential for advancing the state of the art in this area. Many new ideas and results have been shown, culling knowledge from multi-modal content analysis, machine learning, information retrieval, and user interaction. The goal of this paper is to provide a review of the important directions, key results, and remaining open issues.

The main science challenge is understanding media by bridging the semantic gap between the bit stream on the one hand and the visual content interpretation by humans on the other. Hence, our focus here is on semantic concept detection and its application in image and video search. Specifically, we discuss large-scale concept lexicons, image/video corpora, labeling approaches, popular image features and recognition models, and multimodal video search. We discuss each of the above topics and present a short list of important open issues at the end.

2. VISUAL CORPUS AND LEXICON

One of the major forces driving the advancement of image search techniques is the increasing availability of large visual corpora and the well-defined evaluation methodologies associated with them. Unlike the conditions in the 90's, researchers now can easily gain access to millions of images or hundreds of thousands of video shots for research with some form of annotation either via formal processes or indirect association.

For example, TRECVID [1] in its 6th year of evaluation currently provides hundreds of hours of broadcast news video from multi-lingual channels. Video programs are segmented into individual shots and video shots in the development set have been annotated with a number of semantic description labels. To define a suitable set of semantic concepts, a recent effort has also been completed to define a Large-Scale Concept Ontology for Multimedia (LSCOM) [2]. Through joint discussion and evaluation by information analysts, librarians, and researchers, about 1000 concepts were selected from various categories such as event, object, scene, people, location, and production. 449 of the LSCOM concepts were then manually annotated over a subset of TRECVID videos (more than 80 hours). The resulting annotation set available at [3] is probably the largest video annotation data set available to date for researchers, both in terms of the number of concepts and the number of the samples for each concept.

Many other significant corpora have also emerged in different domains. In ImageClef [4], images and associated text documents have been used for evaluating medical image retrieval and recently images from other domains like Web have been added. In CalTech 101 [5], images from 101 object categories have been collected from Web image search engines in order to evaluate performance of various object recognition methods. To tap into the large image repository provided on the Web, clustering experiments in [7][8] have used a large number of images (thousands to millions) from online photo sharing sites or search engines. In [6], images of 1000 isolated objects are collected to test robustness of recognition methods against varying appearances and recording conditions.

The corpora mentioned above differ in several important aspects, including content diversity, image quality, quality of annotations, and numbers of image samples. One of the most important aspects affecting semantic classification is

the annotation quality. Usually, sets like TRECVID using exhaustive manual annotation enjoy a higher annotation accuracy than others. Images obtained from Web search engines are almost by definition more realistic but may quickly demonstrate inconsistent and diverse content when the size of the returned set becomes large. On the other hand, textual tags from online social sites may not be reliable. Some empirical analysis showed that precision as low as 15% may be possible [9]. Therefore, special cares are needed in utilizing low-quality labels in training content recognition models, analogous to the case of using unreliable transcripts to train automatic speech recognizers [24].

Novel Labeling Approaches

As discussed above, manual labeling processes are often used in order to obtain reliable and complete annotations. It was also found annotation mechanisms and tools used significantly affected the quality and throughput of the resulting annotations. Some lessons can be drawn from the past efforts when the annotation process is repeated in new domains. In LSCOM [3], about ten thousand hours of human efforts were used to generate about 33 million labels, each indicating presence or absence of a specific concept in a video shot. By forcing annotators to give binary-value labels of a concept at a time, it produced much more accurate and consistent annotations than the alternative method that allowed the annotator to enter multiple tags relevant to the image under review. In [10] an interesting annotation interface was developed to explore the capacity of human visual perception. Annotators maximize the rate of labeling by viewing the fixed locations on the screen while panels of images are rapidly displayed, leading to a multi-fold speedup. In [23], a visual iconic language was developed to support a video annotation process in which users browse and compound over 2200 iconic primitives, each representing certain concept categories in the video stream. Finally, an innovative framework was developed in [11] to transform the tedious annotation process into a Web-based interactive game, in which randomly paired users at different sites try to come up with identical annotations. Such a game playing paradigm is novel and effective by motivating humans to help complete tedious tasks, and in this case, generate voluminous labels with a high quality.

3. SEMANTIC CONCEPT CLASSIFICATION

The massive visual data and associated annotations have facilitated development of novel techniques for indexing images and videos at the semantic level. By training a statistical detector for each of the concept in the visual lexicon, a pool of semantic concept detectors can be constructed to generate multi-dimensional descriptors in the semantic space. A large number of baseline concept detectors have demonstrated using generic features such as color, texture, and edge [14][15].

After the concept detection, each image or video shot can be represented by a semantic concept vector [12], whose elements indicate the confidence scores or likelihood values in detecting different concepts. The key innovation here is to go for a weak representation of many concepts that yields a much better insight in the semantics of the video than insisting on accurate representation of just a few of them. Such a representation is intuitive – analogous to the term frequency vector used for indexing text documents. However, there exist subtle but important difference between the interpretations of the term frequencies in a document and the confidence scores of concepts in an image or video. Such issues will affect the strategies for concept search, which will be discussed in Section 5.

Recent advances in image feature extraction and matching have also brought exciting opportunities for improving the concept detection performance. Novel features using local interest points or parts [16] capture salient information in the image, with invariant local features extracted from each local interest part. Aggregative attributes (like bags of quantized parts) or spatial graphs of the parts can then be used to represent the overall visual content. Novel methods for matching such parts-based representation have also been used to compute image similarity, from which discriminative classification models are developed. Different local image descriptors were compared in [17] in their performance of object recognition.

Semantic concept detection has greatly profited from advances in machine learning. Discriminative classifiers such as Support Vector Machines (SVM) over image (and sound) features, though straightforward, have been shown effective for detecting a number of visual concepts [13]. Large pools of concept detectors are also demonstrated in [14] and [15], covering 491 and 374 concepts respectively. Software for feature extraction and the SVM models used in [15] are available for public research use.

Another notable direction for semantic labeling of visual content is to explore the relations among image content and the textual terms in the associated metadata. Such metadata are abundant but are often incomplete and/or noisy. By exploring the co-occurrence relations among the images and the words, the initial labels may be filtered and propagated from initial labeled images to additional relevant ones in the same collection. A cross-media relevance model was proposed in [18] to model the joint probabilistic distributions of the words and the visual tokens in each image. Such joint distributions are then used to estimate the likelihood of detecting a specific semantic concept in a new image. In [25], a unified graph-based learning framework was developed to integrate features from multiple modalities (keywords and visual features) for web image classification, retrieval, and clustering.

4. IMAGE LABEL BY WEB SEARCH

Manual annotation of image or video data is costly and difficult to scale up to a large set of concepts. On the other hand, images from the Web repositories, e.g., Web search engines or photo sharing sites, come with free but less reliable labels. In [19], a novel framework called AnnoSearch was proposed to explore such Web-based resources. The task is to automatically expand the text labels of an image of interest, using its initial keyword and image content. The seed keyword was first used to find relevant images on the Web, whose textual metadata were then clustered in order to discover new keywords for the image at hand. The newly discovered words were further filtered by checking the content similarity between the target image and the images from the Web. Such keyword expansion mechanism is fully automatic, utilizing the Web resources in a novel way. However, its scalability general images remains to be proved as quality of the expanded labels may depend on the image type and the availability of appropriate seed keywords.

Images returned from Web search engines are error-prone. Usually only the images on the first few pages of the returned results are correct. Additionally, even images of the same object or scene may have large variations of appearance, view, scale, and quality. To cope with such issues, [20] extended a probabilistic clustering technique to discover the hidden patterns among the images from the Web and handled unrelated images returned from search engines. Promising performance was demonstrated in using the Web images to train detectors of generic objects such as cars and airplanes.

The diverse content and inconsistent quality associated with Web images tend to have large impact on the robustness of the concept detectors. Estimation of performance difference between detectors using hand prepared annotations and that using free Web data is an important issue studied in [9]. Experiments over 15 named location concepts were conducted and an automatic method was developed to predict the performance degradation caused by the use of noisy images from Web search engines. It was found that cross-domain image similarity and some forms of concept difficulty measures were the most useful features for predicting the performance difference mentioned above. Assessment of such performance gap is important for assessing the tradeoff between annotation cost and detector quality.

5. MULTIMODAL SEARCH

Semantic indexes produced by a large pool of concept detectors greatly improve the feasibility of searching images/videos at the semantic level. Such semantic indexes readily match the search-by-keyword paradigm – the most popular search method used by users today. To do so, user queries are mapped to the predefined semantic concept space, using term matching or some forms of term expansion (e.g., via relations of synonyms or meronyms). Such text-to-concept search methods were shown to be especially effective for the type of queries that search for generic objects or scenes (e.g., building or snowy scenes) [15].

However, using the concept search method alone is not sufficient for satisfying all different types of queries, which often include searches for named persons, sports, etc. Recognizing such deficiency, [21] studied query-class dependent retrieval models, which used adaptive strategies for fusing different search tools (such as concept search, image-based similarity retrieval, and text search). Machine learning methods were used to automatically determine the fusion weights among different tools. In [22], a data mining approach was further developed to automatically discover the distinct query classes and optimal multimodal fusion weights for each class based on past training queries and their search results.

6. OPEN ISSUES

The exciting developments in semantic indexing presented above are accompanied with many challenging open issues. Some of the most important ones are discussed below.

Explore the Full Potential of New Image Features

As mentioned in Section 3, novel local image features and image representations have been incorporated in many emerging object recognition approaches, significantly outperforming conventional methods based on global features like color, texture, and edge. One important question here is whether we have arrived at the right features for image indexing, just like keywords for text document indexing. Can such new features be used to successfully develop a large number of detectors for diverse semantic concepts, such as that covered in LSCOM? Initial promising results have been shown in [20] in recognizing generic image classes using Web images as training data. Parts-based models have also been shown to improve classification accuracy in TRECVID evaluation [15] when combined with baseline models using conventional features. One additional concern of such new features is the high computational complexity involved – requiring at least one order of magnitude more training time than those using conventional features. Hence, improvement of model efficiency while exploring the full power of such new features has become a timely critical issue.

Visual Concept Ontologies

Knowledge resources like WordNet have been used broadly in text-based document retrieval. Such knowledge bases include definitions and ontological relations among words

in a natural language. However, application of such ontologies or thesaurus to the visual domain is not feasible today due to the insufficient coverage and lack of definitions of relations among visual concepts. Some initial efforts have been made in [2] to construct basic structures over a medium set of concepts (about 1000). However, a systematic and scalable solution is still missing today.

Even with an extended thesaurus it may still be the case that the user specifies a search which is not in the detector set. In such case the best thing to resolve is an ontology translating the search into the closest terms the detector set can handle, so separating out the user side in an ontology from the detector side in another ontology. The issue remains that if a notion needs the combination of many detectors, the results are too unreliable per detector yet to arrive at anything but noise in their logic combination. New strategies are needed to combine weak sets of detectors. Additionally, it was found in [15] that a small set of detectors with robust performance and generic concepts is more powerful for video search, compared to a large set of weak detectors. Hence, a very interesting question arises – how to choose the right set of concepts that have good combination of detection performance and relevance to user query topics?

Visualization and Interactive Access

A video archive includes a huge amount of data and a huge amount of video shots. One of the most effective contributions to digital video archive access is to provide efficient visualization of the data so that users can solve the problem visually as well as textually. Such visualization mechanisms have to be dynamic and compressed as the screen is the bottleneck. Intuitive visualization interfaces have been found critical for successful interactive search of video shots in recent TRECVID evaluation [14]. However important questions remain in how to visualize visual content at higher levels such as stories or topics and how to effectively summarize information across multiple videos.

7. REFERENCES

[1] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/

[2] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, "Large-Scale Concept Ontology for Multimedia," IEEE MultiMedia , vol. 13, no. 3, pp. 86-91, July-September, 2006.

[3] LSCOM Lexicon Definitions and Annotations Version 1.0, Columbia University ADVENT Technical Report #217-2006-3, March 2006. (http://www.ee.columbia.edu/dvmm/lscom)

[4] The CLEF Cross Language Image Retrieval Track (ImageCLEF), http://ir.shef.ac.uk/imageclef/ .

[5] Caltech 101 data sets, http://www.vision.caltech.edu/Image_Datasets/Caltech101

[6] Amsterdam Library of Object Images, http://www.science.uva.nl/~aloi

[7] R. Lienhart and M. Slaney, "pLSA on Large Scale Image Databases," IEEE ICASSP, Hawaii, April 2007.

[8] M. Choubassi, A. Nefian, I. Kozintsev, J. Bouguet, Y. Wu, "Web Image Clustering," IEEE ICASSP, Hawaii, April 2007.

[9] L. Kennedy, S.-F. Chang, I. Kozintsev, "To Search or To Label?: Predicting the Performance of Search-Based Automatic Image Classifiers," In Multimedia Information Retrieval Workshop (MIR), Santa Barbara, CA, USA, 2006.

[10] A. Hauptmann, W.-H. Lin, R. Yan, J. Yang, M.-Y. Chen, Extreme Video Retrieval: Joint Maximization of Human and Computer Performance," ACM Multimedia, Oct. 2006, Santa Barbara, CA.

[11] L. von Ahn, L. and L. Dabbish, "Labeling Images with a. Computer Game," In ACM Conference on Human Factors in Computing Systems (CHI), 2004.

[12] M. R. Naphade, S. Basu, J. R. Smith, C.-Y Lin and B. Tseng, "Modeling Semantic Concepts to Support Query by Keywords in Video," IEEE Intern. Conf. on Image Processing (ICIP), 2002.

[13] A. Amir, et al, "IBM Research TRECVID-2005 Video Retrieval System," NIST TRECVID Workshop, 2005.

[14] M. Worring, C. Snoek, O. de Rooij, G.P. Nguyen, A. Smeulders, "The MediaMill Semantic Video Search Engine, " IEEE ICASSP, Hawaii, April 2007.

[15] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction," NIST TRECVID workshop, Nov. 2006.

[16] D. Lowe, "Distinctive Image Features from Scale Invariant Keypoints," In *IJCV* 60(2):91–110, 2004.

[17] K. Mikolajczyk, C. Schmid, "A Performance Evaluation of Local Descriptors," In *PAMI* 27(10):1615–30, 2003.

[18] J. Jeon, V. Lavrenko, R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models", ACM SIGIR'03.

[19] X.-J. Wang, L. Zhang, F. Jing, W.-Y. Ma, "AnnoSearch: Image Auto-Annotation by Search," IEEE CVPR, New York, June 2006.

[20] R. Fergus, L. Fei-Fei, P. Perona and A. Zisserman, "Learning object categories from Google's image search," ICCV Beijing, China, Oct 2005.

[21] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In ACM Multimedia Conference, New York, Oct. 2004.

[22] L. Kennedy, P. Natsev, S.-F. Chang, "Automatic Discovery of Query Class Dependent Models for Multimodal Search," ACM Multimedia Conference, Singapore, Nov. 2005.

[23] M. Davis, "Media Streams: an Iconic Visual Language for Video Annotation," IEEE Symp. on Visual Languages, 1993.

[24] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115-129, 2002.

[25] H. Tong, J. He, M. Li, C. Zhang, and W. Y. Ma, "Graph based multi-modality learning," *ACM conference on Multimedia*, 2005.