OVER-COMPLETE REPRESENTATION AND FUSION FOR SEMANTIC CONCEPT DETECTION

Apostol (Paul) Natsev, Milind R. Naphade, John R. Smith

IBM Thomas J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532
{natsev,naphade,jsmith}@us.ibm.com

ABSTRACT

Automatic semantic concept detection in images is a promising tool for alleviating the user effort in annotating and cataloging digital media collections. It enables automatic identification of people, places, and objects, for enhanced indexing and searching of home photographs, for example. While constructing robust semantic detectors has been shown feasible for global generic concepts with a sufficient number of good training examples (e.g., *indoors*, *outdoors*), many interesting concepts, such as *face*, *people*, occur at subpicture granularity, occupy only a portion of the image and therefore frequently have training examples with a reduced signal-to-noise ratio. Such regional concepts are harder to detect due to imperfections in automatic image segmentation algorithms leading to inaccurate object boundaries and low-level feature ambiguities.

In this paper we focus on the problem of boosting detection performance of existing regional concept detectors by exploiting detection redundancy. Specifically, we propose to use the same detector multiple times to evaluate and combine multiple detection hypotheses for the same content—but at different content granularities—in order to reduce detection sensitivity to segmentation errors. We validate the approach using Support Vector Machine classifiers for 14 regional semantic concepts from the NIST TRECVID 2003 common annotation lexicon, and show performance improvements of multi-granular detection and fusion.

1. INTRODUCTION

Semantic understanding of multimedia content is critical in enabling effective access to it. With the proliferation of digital cameras, camcorders, and content authoring tools for the PC, this is especially true for the average digital photography user and home video enthusiast who is typically not a trained expert in cataloging, searching, and managing such content. Typical consumer applications place the content annotation and indexing burden almost entirely on the end user, therefore limiting their ability to create and manage large collections of digital media assets. To overcome this limitation, there is a need for fully automatic or semi-automatic solutions for semantic analysis of digital media.

While general-purpose semantic understanding of images and video is a daunting task, successful specialized approaches have been developed for several domains, including broadcast news, sports videos, situation comedies etc. The annual NIST TRECVID benchmark has also shown the feasibility of detecting generic concepts, such as indoors, outdoors, people, faces, etc. Currently, statistical semantic modeling approaches, such as Support Vector Machines, Gaussian Mixture Models, and Hidden Markov Models, are among the most successful generic methods for solving the above problem [1, 2, 3, 4]. Their success, however, depends to a large extent on the quality of the low-level visual features used in the modeling and detection process. While training data may be well segmented and labeled manually—thus providing accurate training features—such accurate object segmentation and feature representations are rarely available at detection run-time due to imperfections of existing segmentation algorithms. Therefore, a significant gap exists for regional concepts between the quality of features used for training and for detection.

In this paper, we propose to use content representation redundancy at detection time in order to alleviate the effect of the above object boundary/feature ambiguities. In particular, we aim to improve the performance of existing regional concept detectors by evaluating them on multiple granularity variations of the same content and fusing the resulting detection hypotheses into a single more robust detection hypothesis, which is less sensitive to object segmentation and feature extraction errors. Note that the modeling stage remains unchanged, and is independent of the proposed detection framework, which makes the approach a feasible alternative for boosting detection performance without the need for additional training or more accurate training data.

There is a large body of work on over-complete content representations and the exploitation of content redundancy for general image analysis, compression, content-based retrieval, and to a smaller extent, concept detection. Work in this area is too rich to list here exhaustively but includes

0-7803-8554-3/04/$20.00

Laplacian pyramids, multi-resolution wavelet analysis, and hierarchical hidden Markov modeling, among others.Most of the above approaches have been focused primarily in applications other than concept detection, however, or when used for detection purposes, they have been an integral part of the training phase and the specific concept modeling framework. Our focus in this paper is on a generic and simple framework that applies at detection time only and is independent of the training phase and the modeling approach. We show that even with very simple redundant representations, such as regular grids and spatial layouts, we can achieve a significant performance boost at concept detection time, without relying on accurate image segmentation or the need to train complex multi-resolution models.

2. MULTI-GRANULAR DETECTION OF REGIONAL SEMANTIC CONCEPTS

Our approach to detecting visual semantic concepts involves 2 stages. We represent images with a set of low-level visual features, and in the training phase, we learn feature representations corresponding to the binary hypotheses for each concept (presence/absence) using generic supervised machine learning algorithms like Gaussian Mixture Models, Hidden Markov Models and Support Vector Machines [4]. In the detection phase we use the existing models to score target images for the presence/absence of the concept.

Of the various visual semantic concepts we model, those concepts that are present at sub-image levels pose a greater detection challenge. Such concepts can be detected better if the feature extraction is performed at the proper granularity. For example, a semantic concept like *Outdoors* exists at the level of the entire frame. A concept like *Face*, however, exists at a sub-image level. During the training phase we therefore mark concepts in images by manually drawing bounding boxes around regions of interest. The training features are then extracted from the subset of pixels in the bounding boxes. If features for learning a representation for *Face* are extracted at the global image level, the signal to noise ratio will be lower than in the case when the features are extracted from the subset of pixels belonging to the face region. During the detection phase, we cannot expect the bounding boxes to be marked up since the entire detection process is automatic, and the detection of such a region is a difficult computer vision problem.[1] In the past we have relied on image segmentation to extract bounding boxes for the 5 most prominent regions in an image and extracted features from within the bounding boxes to create regional representations. Results based on such an approach can be found in [4]. However this approach has its limitations because automatic image segmentation is error-prone.

In this paper, we investigate an approach to increase robustness of the detection process to segmentation errors. Imperfect segmentation leads to a set of regions, and not a set of objects, which leads to inaccurate feature representations for the true objects in the picture. In order to reduce sensitivity to such errors, we propose to extract features at multiple granularities, and perform detection at all these granularities. Experimental results show that considering even simple regular segmentation schemes, in combination with irregular segmentation leads to improvement in detection performance. We also observe that no single granularity can provide a complete solution.

3. EXPERIMENTAL SETTING

3.1. Dataset and evaluation

For the experiments in this paper we used the development collection from the TRECVID 2003 Concept Detection Benchmark organized by NIST.[2] The collection consists of 60 hours of MPEG video, segmented into shots consisting of about 50,000 keyframes and annotated manually with more than 200,000 labels and their corresponding region bounding boxes, if relevant (see [5] for details of the common annotation effort). Approximately 60% of the collection (28055 keyframes) was used for statistical model training. About 10% of the collection, or 4420 keyframes, was used as a Validation Set I for model optimization and multi-granular fusion parameter tuning. A different set of 5037 keyframes, Validation Set II, was used for final performance evaluation purposes. All experiments reported here are performed on the keyframes as collections of static images and do not use any motion or video information.

For evaluation purposes we use non-interpolated average precision over the ranking of all target images with respect to each semantic concept. For a given concept $\mathcal{C}$, let R be the number of relevant documents in a set of size S, and let L be the list of returned documents, ranked by detection confidence with respect to concept $\mathcal{C}$. At any given index k, let R_k be the number of relevant documents in the top k documents. Let $I_k = 1$ if the k^{th} document is relevant and 0 otherwise. Assuming $R \leq S$, the non-interpolated average precision is then defined as

$$\frac{1}{R}\sum_{k=1}^{S}\frac{R_k}{k} * I_k \qquad (1)$$

When we talk about performance over a set of concepts, the average precision (AP) scores are usually aggregated into a single Mean Average Precision (MAP) score, which is the average of the AP scores over the set of concepts.

[1] In some constrained computer vision applications, it may be feasible to do a brute force search to find the best match for a template-based model. In general, however, this is not feasible to do and the search space of possible region matches is typically reduced through image segmentation.

[2] http://www-nlpir.nist.gov/projects/tv2003/

3.2. Feature extraction at multiple granularities

Each keyframe in the collection was processed as a standalone image to extract the following features [4]:

- *HSV color correlogram* (166-D)
- *Edge histogram* (64-D)
- *Gray-level co-occurrence texture* (96-D)
- *Moment invariants for shape* (6-D)

For training, features were extracted for each ground truth label from the corresponding object bounding boxes, if given, or the entire image, otherwise. For detection, each target image was processed at four multiple granularities and the above features were extracted at each granularity, resulting in four sets of the above features for each image:

- One set of features extracted at the global image level.
- One set of features extracted from a 3x3 grid dividing the image into 9 equal-sized rectangles.
- One set of features extracted from each region of the layout that divides the image into 4 equal non-overlapping rectangles covering the image and a fifth box of the same size located at the center.
- One set of features extracted from the bounding boxes of the 5 largest objects derived from automatic image segmentation using color and texture.

In all of the reported experiments, we use early feature fusion by concatenating the four types of features [4].

3.3. Semantic Concept Lexicon

We used a subset of the TRECVID lexicon of 133 concepts, of which more than 60 were regional concepts. For experiments in this paper we used 14 of the most frequently occurring regional concepts in the training set:

- Objects: Airplane, Building, Car, Clouds, Face, Female Face, Graphics, Male Face, People, Person, Text Overlay, Transportation
- Sites: Sky, Water body

3.4. Experimental Methodology

For each concept, we train a set of configurations of binary Support Vector Machine classifiers using the features extracted from the manually labeled bounding boxes of the training set. We use SVMs with Radial Basis Function (RBF) kernels and consider multiple RBF parameter setting configurations (for more details, see [4]). For each granularity, we select the parameter configuration leading to the best average precision performance on Validation Set I as the final statistical model of the given concept at the given granularity. We then evaluate the four selected feature granularity models in order to derive four different detection hypotheses for each concept, on each validation set.

Given the four basic detection hypotheses, we consider multiple strategies for combining their soft decisions into a single more robust detection hypothesis. These include greedy selection of the best granularity hypothesis for each concept, as measured on validation set I, as well as taking the minimum, maximum, product, or weighted average of single granularity detection scores. Min, max, and product aggregation are simply ways to simulate boolean AND/OR logic. For weighted averaging, we use binary weights only, effectively considering simple averaging of all possible subsets of granularity hypotheses. Finally, for each of the above fusion methods, we consider several *a priori* score normalization techniques, as follows:

- No confidence score normalization
- Linear normalization to $[0, 1]$ range
- Statistical normalization (0 mean, std. deviation of 1)
- Gaussian likelihood normalization: statistical normalization followed by exponentiation and inversion

This results in a total of 72 fusion configurations of the four granularity hypotheses, which we evaluate on Validation Set I in order to select the *best predicted* fusion combination and apply it on Validation Set II. For comparison purposes, and in order to judge the generalization capabilities of the above fusion methods, we also evaluated all combinations on Validation Set II and recorded the *best observed* fusion combination for each concept. The performance of the best observed combination is thus an upper bound for what we can hope to achieve using the above fusion methods.

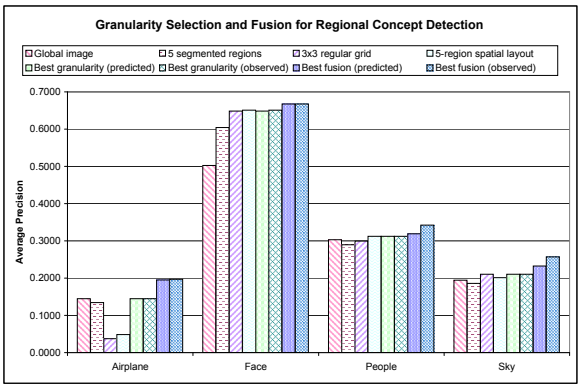

Figure 1: Effect of granularity selection and fusion for regional concept detection for four concepts.

4. RESULTS AND DISCUSSION

Table 1 lists results for the four single granularity hypotheses, the best single granularity greedy selection approach, as well as the best predicted and best observed multi-granular hypothesis from the set of fusion combinations described in the previous section. A representative subset of the results is further illustrated in Figure 1. From the results we can verify that most concepts are detected better at sub-image

Granularity Hypothesis	Global	5 Largest Regions	3×3 Grid	5-Region Layout	Best predicted granularity	Best fusion (prediction)	Best fusion (observation)
MAP score	0.2495	0.2513	0.2575	0.2712	0.2835	0.2962	0.3100
Gain over Global	0.0%	0.7%	3.2%	8.7%	13.6%	18.7%	24.3%

Table 1: Mean Average Precision (MAP) over 14 regional concept detectors evaluated at 4 granularities. Methods compared include concept detection at individual granularities, greedy selection of best predicted granularity (from validation set I), as well as best multi-granular fusion hypothesis based on prediction from validation set I or observation from validation set II.

granularities, although no single granularity provides superior performance across the board. For example, we see that even though we are detecting regional concepts, in some cases the most appropriate granularity for detection is the global image granularity. This can be explained by the fact that for some concepts, such as *airplane*, the background information (e.g., the sky) may be highly correlated to the foreground object (e.g., the airplane) so that the entire image can contribute to better detection of the foreground object. For other concepts, however, such as *face*, the background may not be very correlated to the object of interest and the finer granularity representations, such as the grid-based or layout-based representations, may yield the best results. The same concept can also appear at different granularities so that several granularity hypotheses can be equally important and valid (e.g., *people, sky*).

The above examples validate our hypothesis that performing detection at a single granularity is sub-optimal. However, we also see that the relative importance of the various granularities for any given concept usually carries over from one validation set to the next. The bets predicted granularity from validation set I matches the best observed granularity on validation set II for 8 out of the 14 concepts, and is second best for 5 of the remaining 6 concepts. Greedy selection of the best predicted granularity thus leads to better results than each of the individual granularities, including a performance gain of 13-14% over global detection and irregular segmentation-based detection that we have reported before [4] and 5% improvement over the single best observed granularity (i.e., layout).

Further performance gains can be realized when detection scores at the multiple granularities are combined together, rather than simply ranked and selected greedily. Using the best predicted fusion and normalization for each concept leads to an additional 5% improvement over greedy selection of the best predicted granularity, for a total of 18-19% gain with respect to global detection and segmentation-based detection. The experimental results therefore show that performing detection at multiple granularities can improve mean average precision by 10-19% over single granularity detection, and the proposed granularity fusion approaches generalize well from validation set I to II.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge the IBM TRECVID 2003 team for shot segmentation and annotation. Ching-Yung Lin for keyframe and region extraction.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper we proposed an approach for boosting performance of automatic regional concept detection through the use of multi-granular detection. Automatic visual concept detection enables more effective indexing and searching of multimedia content, such as home photographs and videos. Detecting regional concepts like faces, people, and objects, is a particularly relevant problem to home photograph management but is also more challenging than detecting global concepts due to object segmentation errors. We proposed to evaluate existing detectors at multiple granularities of the same target content in order to reduce sensitivity to segmentation. We considered several approaches for combining detection hypotheses derived from multiple granularities and showed that they lead to significant detection performance improvement on a large corpus of images from the TRECVID 2003 Concept Detection benchmark.

7. REFERENCES

[1] A. Vailaya, A. Jain, and H. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition*, vol. 31, pp. 1921–1936, Dec. 1998.

[2] A. Gupta, T. E. Weymouth, and R. Jain, "Semantic queries with pictures: the VIMSYS model," in *Intl. Conf. on Very Large Databases (VLDB)*, Sep. 1991, pp. 69–70.

[3] I. Buciu, C. Kotropoulos, and I. Pita, "On the stability of support vector machines for face detection," in *IEEE Intl. Conf. on Image Processing (ICIP)*, 2002.

[4] M. Naphade and J. Smith, "Learning visual models of semantic concepts," in *Proc. IEEE International Conference on Image Processing*, Sep 2003.

[5] C. Lin, B. Tseng, and J. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," in *Proc. Text Retrieval Conference (TREC)*, Gaithersburg, MD, Nov 2003.