A Statistical Modeling Approach to Content Based Video Retrieval

Milind R. Naphade, Sankar Basu, John R. Smith, Ching-Yung Lin and Belle Tseng
Pervasive Media Management Group
IBM Thomas J. Watson Research Center
Hawthorne, NY 10532
{naphade,sbasu,jsmith,chingyung,belle}@us.ibm.com

Abstract

Statistical modeling for content based retrieval is examined in the context of recent TREC Video benchmark exercise. The TREC Video exercise can be viewed as a test bed for evaluation and comparison of a variety of different algorithms on a set of high-level queries for multimedia retrieval. We report on the use of techniques adopted from statistical learning theory. Our method depends on training of models based on large data sets. Particularly, we use statistical models such as Gaussian mixture models to build computational representations for a variety of semantic concepts including rocket-launch, outdoor, greenery, sky etc. Training requires a large amount of annotated (labeled) data. Thus, we explore use of active learning for the annotation engine that minimizes the number of training samples to be labeled for satisfactory performance.

1 Introduction

Supporting semantic queries in video retrieval is an important and difficult problem in multimedia analysis [7, 9] and query by keywords is needed for effective utilization of multimedia repositories [10]. Recent work in video retrieval has witnessed an interesting shift from the query by example (QBE) paradigm [5, 10, 2] to the query by keywords paradigm using semantic video objects [7, 9, 3, 12, 6]. The reason behind this shift lies in the desire to build media search engines that are as pervasive and useful from user perspective as their counterparts in text.

To build systems that help users, find what they desire, it is important to address the semantics of the queries. To evaluate effectiveness of the search engines, it is then necessary to have a set of queries and a benchmark database of videos that can be used to conduct experiments. In this light, the emergence of Video TREC [1] as a benchmark is a significant development. The National Institute of Standards and Technology has undertaken an exercise of creating a publicly available database of videos and a set of semantically meaningful queries that need to be answered with respect to this database. In this paper, we discuss a framework for modeling semantic concepts and answering the V-TREC queries on the V-TREC database.

To represent keywords or key-concepts in terms of audiovisual features Naphade et al. [7] presented a framework of multijects. Chang et al. [3] use a library of examples approach, which they call semantic visual templates. In both cases, the implication is that if a lexicon-based approach to retrieval is advocated, there must exist a method to derive these representations from a set of user provided examples. Often the problem as with the QBE paradigm is the lack of a sufficient number of these examples to estimate generic representations that are effective.

2 Learning Semantic Concepts for Video Retrieval

The architecture of our statistical modeling approach to video retrieval is outlined in Figure 1. This system includes three main components: semantic annotation, learning and model construction, and video retrieval. The idea behind using automatically constructed models for video retrieval is to facilitate the paradigm of query by keyword [9, 3].

We now describe modeling a set of semantic concepts by learning them from annotated content. We also describe how these models are used to answer the V-TREC queries.

2.1 The V-TREC Benchmark

The V-TREC benchmarking task could be viewed as extensions of its well known text-only analogue. The system designed, some of which included a human in the loop, were presented with topics-formatted descriptions of an informa-
tion requirement and were asked to return a list of shots from the videos in the test collection which met the requirement.

The topics contained not only text but possibly examples (including video, audio, images) of what is needed. The topics expressed a wide variety of needs for video clips: of a particular object or class of objects, of an activity/event or class of activities/events, of a particular person, of a kind of landscape, on a particular subject, using a particular camera technique, answering a factual question, etc. The units of retrieval was shots.

The test data consisted of about 11 hours of NIST documentary video clips comprising of over 7000 shots. There were 74 semantic queries like "Find all space shuttle liftoffs." Open source videos e.g., from BBC and elsewhere were allowed for training the system. Use of transcripts created by automatic speech recognition (ASR) was allowed but more importantly, participants were asked to report performance obtained not by ASR but by audiovisual content analysis.

2.2 Defining a Lexicon

For annotating video content, we created a lexicon for describing events, scenes, and objects; the following excerpt gives some of the terms:

- Events: water skiing, boat sailing, person speaking, landing, take-off/launch, and explosion;
- Scenes: outer space (moon, Mars), indoors (classroom, meeting room, laboratory, factory), outdoors (nature, sky, clouds, water, snow, greenery, rocks, land, mountain, beach, field, forest, canyon, desert, waterfall), and man-made (road, cityscape);
- Objects: non-rigid objects (animal, deer, bird, duck, human), rigid objects (man-made structure, building, dam, statue, tree, flower), transportation (rocket, space shuttle, vehicle, car, truck, rover, tractor), and astronomy.

2.3 Annotating a Corpus

In order to allow a model-based approach to video retrieval, ground-truth data is needed for training the statistical models. For the purpose of creating training data, one can envision developing a strategy for video annotation that allows the users to annotate each shot in the video sequence. The tool would allow users to identify and label scenes, events, and objects by applying the labels at the shot-level. The tool also allows the user to associate the object-labels with an individual region in a key-frame image.

2.4 Active learning to Enable Faster Annotation

Abundance of video data and diversity of labels make annotation a difficult and expensive task. We formulate the task of annotation in the framework of active learning. We first train a classifier with a small set of labeled data, and propagate persistent annotation by prompting the annotator to provide the ground truth for the most informative, or most uncertain subset of the available data-set. The learner is thus refined at every iteration and the user needs to annotate as few samples as is possible.

We use a support vector machine as the intrinsic classifier for the active annotator and the active learner. This approach was proposed by Naphade et al [8]. The initial support vector classifier is built on the basis of a very small annotated data-set. Each unseen example is classified by the SVM classifier and if the uncertainty if classification is high, the annotator is prompted to provide the class label. The classifier is retrained after every iteration. If the uncertainty associated with classification is low the label provided by the classifier is propagated. This reduces the number of examples to be annotated by a factor of 10.

2.5 Modeling Concepts

The content modeling system then uses the labeled training video content to classify other video content (in our case, the test V-TREC corpus). Early work on different types of statistical models includes [9]. For the V-TREC task, we have developed models for the following concepts:

- Events: fire, smoke, launch;
- Scenes: greenery, land, outdoors, rock, sand, sky, and water;
- Objects: airplane, boat, rocket, vehicle.
For modeling the V-TREC content, statistical models were used for two-class classification using Gaussian Mixture Model (GMM) classifiers. For this purpose, labeled training data obtained from a video annotation tool was used. The feature vectors associated with training data corresponding to each label were modeled by a mixture of five Gaussians, the parameters (mean, covariance and the mixture weights) of which were estimated by using the standard EM algorithm. The rest of the training data was used to build a negative model corresponding to that label in a similar way. The difference of log likelihoods of the feature vectors associated with a test image for each of these two models was then taken as a measure of the confidence with which the test image can be classified to the labeled class under consideration. A (ranked) list of confidence measures for each of the labels can, thus, be produced by repeating this procedure for all the labels under consideration.

3 Experiments with V-TREC

3.1 Feature Extraction

We analyze the videos at the temporal resolution of shots. Shot boundaries are detected using IBM CueVideo [11]. Key-frames are selected from each shot. From each key-frame we extract features representing color, texture, structure, and shape [9]. Color is represented by 24 dimensional linearized HSV histograms and color moments. Structure is captured by computing edge direction histograms. Texture is captured by gray-level co-occurrence matrix properties. Shape is captured using Dudani’s moment invariants [4]. In addition for query by example, where feature dimensionality is not a bottleneck, as it is in learning representations, we also compute a joint HSV histogram of 166 bins, a wavelet coefficient texture measure and a color composition histogram.

While answering the queries, we compare retrieval using models as opposed to retrieval using one iteration of query by example. During retrieval, we also combine results across features. This is done by allowing linear combinations of model based confidence estimates to answer queries beyond the lexicon.

3.2 Results: Fusing Across Features

It is well known that the performance of statistical models such as the GMM depend to a large extent on the volume of training data. Often, there is relatively small amount of labeled training data available and we therefore adopt a leave one video out strategy for the purpose of experiments. This means that for the purpose of indexing a video clip we exclude this clip from training but include all other available video data.

For each semantic concept, we build a model for the positive and negative hypothesis for each feature type (e.g., color histogram, edge direction histogram, texture etc.). We then need to merge results across features across these multiple classifiers. This, in turn, results in separate confidences with which a test image could be classified into a class. While various ways of combining these can be thought of to produce a single confidence measure, for V-TREC video exercise we considered straightforward methods such as the average and the maximum of the individual confidences.

While this strategy of late feature fusion is the probably the simplest, one can envision other early feature fusion methods such as concatenating different feature vectors into a single feature vector and then building a single GMM. We did not pursue this strategy due to the potential explosion of feature dimension, especially in view of the paucity of available training data. In future however, we plan to use dimensionality reduction techniques. Multimodal fusion resulting from signals belonging to different modalities e.g., speech, text and image/video also falls into this category of issue but we will not report on it in the present paper.

3.3 Supporting Search by examples and keywords

Content-based retrieval is the most amenable to automatic retrieval in the case that the query provides example content. For V-TREC video retrieval, each of the queries provided example content which mostly included a single image or video clip. For automatic content-based retrieval, the following approach was adopted: the query content was analyzed using shot detection, key-frame selection, and feature extraction to...
produce a set of descriptors of the query content. Then, the query descriptors were matched against the target descriptors. We considered two approaches for automatic content-based matching: (1) matching of descriptors of the query and target key-frames, and (2) matching of descriptors for frames from the query and target video, as shown in Figure 2. Figure 2 shows the top 12 retrieved results for the V-TREC query which uses the example video clip of the lift-off of a space shuttle. Figure 3 on the other hand supports query by keywords that correspond to the same V-TREC query. Figure 3 uses the models rocket-launch and the model sky to find all videos in which these two semantic concepts are found with the highest confidence. While the query by example leads to a precision of 16%, the query based on models leads to a precision of 58%. This is just one of the several examples of the V-TREC queries and we are constrained only by the space available for this paper, from providing detailed results.

4 Concluding Remarks and Future Directions

In this paper we examine statistical modeling for content based retrieval in the context of recent TREC Video benchmark exercise. The V-TREC exercise can be viewed as a test bed for evaluation and comparison of a variety of different algorithms on a set of high-level queries for multimedia retrieval. We build models for semantic concepts including Outdoors, Rocket-launch and several other concepts. Using multiple such models we answer the set of V-TREC queries. Some of the V-TREC queries involve complex semantics and we can only use the existing models to narrow down the search in such cases.

Future research aims at improving the performance at the level of active annotation. More importantly we also hope to improve the performance of the individual concept detectors for events and objects through increasingly sophisticated density models and learning algorithms.

References