WHICH THOUSAND WORDS ARE WORTH A PICTURE? EXPERIMENTS ON VIDEO RETRIEVAL USING A THOUSAND CONCEPTS

Wei-Hao Lin and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.

ABSTRACT

In contrast to traditional video retrieval that represents visual content with *low-level* features (e.g. color and texture), emerging *concept*-based video retrieval allows users to search video archives by specifying a limited number of *high-level* concepts (e.g. outdoors and car). Recent studies have demonstrated the feasibility of concept-based retrieval, but a fundamental question remains: what kinds of concepts should we index? We analyze a large video archive annotated with more than a thousand high-level concepts, and develop guidelines for choosing concepts of high utility to video retrieval.

1. INTRODUCTION

Video retrieval aims at finding shots in a video archive that are relevant to a query. Early systems require humans to annotate video with text descriptions [1], which is time-consuming and does not scale to large video collections. On the contrary, content-based video retrieval (e.g. [2]) employs technologies from image processing and computer vision to automatically index visual content with *low-level* image features (e.g. color and texture). Although this removes the burden of manual indexing, users are required to prepare image examples or specify esoteric image parameters. Recent work in video retrieval combine advantages of both indexing schemes to index video content with *high-level* semantic concepts (e.g. outdoors and car) automatically derived from low-level features [3]. For example, A user with information need, "find shots of one or more roads with lots of vehicles", can directly specify concepts such as "road" or "vehicle" in the query.

Which concepts should be automatically indexed, however, is still an open research question. The characteristics of visual concepts have been studies in the context of news story tracking [4], but it remains unclear how these characteristics apply to video retrieval. One might simply disregard the concept selection problem, and propose to index as many concepts as possible, for example, adopting the Thesaurus for Graphic Materials from the Library of Congress. Building automatic detectors is by no means trivial. Some high-level concepts, for example, "face", take decades of research. As a compromise between coverage and research effort, we argue that the near-term research goal of concept-based retrieval should give priority to concepts that are likely to benefit as many queries as possible. The effort of developing "helpful" concept detectors can then be amortized over a large number of queries.

To identify concepts of high utility to video retrieval, we analyze a large video archive annotated with more than a thousand concepts and relevance judgment, as described in Section 2. In Section 3 we describe how concept utility is estimated using mutual information. We present a series of analysis in Section 4, and finally develop specific guidelines for choosing video concepts of high utility to video retrieval.

2. VIDEO ARCHIVE AND ANNOTATIONS

The video archive used in our analysis is from the 2003 TREC Video Retrieval Evaluation (TRECVID) [5]. The TRECVID 2003 development set consists of 62.2 hours of ABC World News Tonight, CNN Headline News, and C-SPAN programs. We assess the relevance of all shots to 20 search topics in TRECVID 2003[1]. The TRECVID topics are designed to represent a variety of search types. The average number of relevant shots of a topic is 21.5 (min 3, max 49).

The video archive was annotated collaboratively [6] by TRECVID 2003 participants. Annotators tagged each video shot with concepts from a 133-concept ontology, and can also type in free text to describe content not covered by the predefined set. Ultimately annotators added additional 935 concepts, resulting in a total of 1068 concepts. There are a total of 201757 annotations for 48098 shots, and the average number of annotation per shot is 4.19.

This material is based on work supported by the National Science Foundation under Grant No. IIS-0535056 and by the Advanced Research and Development Activity under contract number H98230-04-C-0406.

[1] which are 25 official topics minus Topic 106, 114, 116, 118, and 119 that have no relevant shots in the development set.

1-4244-0367-7/06/$20.00 ©2006 IEEE

Category	Examples
Program	advertisement, baseball, weather news
Scene	indoors, outdoors, road, mountain
People	NBA players, officer, Pope, president Clinton
Objects	rabbit, car, airplane, bus, boat
Activities	walking, women dancing, cheering
Events	crash, explosion, gun shot
Graphics	us weather map, NBA scores, program schedule

Table 1. Examples of concepts in each category.

We classify the 1068 concepts into eight categories[2] proposed by the Large Scale Concept Ontology for Multimedia (LSCOM) workshop [7]. The total number and three examples of each category are shown in Figure 1 and Table 2, respectively. The largest category is Objects (33.1% of the 1068 concepts), followed by People (16.6%) and Scene (15.4%).

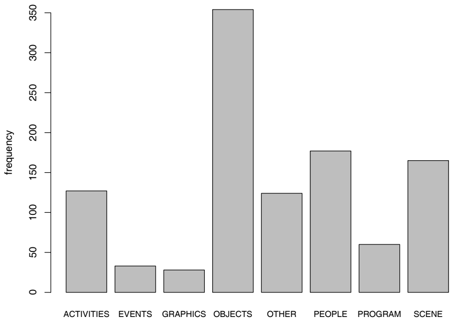

Fig. 1. The 1068 concepts are grouped into eight categories.

3. DETERMINING CONCEPT UTILITY

We employ an information-theoretic notion, mutual information (MI) [8], to determine if a concept is “helpful” in retrieving the shots relevant to a query. MI has been an effective in feature selection in other tasks such as text categorization [9]. Denote the relevance of a shot as R, and the presence or absence of a concept in a shot as C, R and C are binary random variables. The mutual information between R and C, $I(R; C)$ is defined as follows,

$$I(R; C) = \sum_{r,c} P(r, c) \log \frac{P(r, c)}{P(r)P(c)}$$

where $r \in \{\texttt{presence}, \texttt{absence}\}$, $c \in \{\texttt{relevance}, \texttt{irrelevance}\}$. MI can be interpreted as how much randomness of R, i.e. entropy, is reduced from the knowledge of C. If one become more certain about the relevance of a shot after knowing the presence or absence of a concept, i.e. MI is greater than zero, the concept is defined as a helpful concept for the topic. In practice it is very difficult to achieve zero mutual information when the data set is not extremely large, and thus we define a concept C as helpful only when the entropy of R using Maximum Likelihood Estimates is reduced more than 1%, which is the minimal threshold that can filter out most spuriously helpfulness from rare concepts that never occur with relevant shots.

We further divide helpful concepts into two types: *positively* helpful concepts (P-concept) and *negatively* helpful concepts (N-concept). The presence of P-concept in a shot increases the degree of relevance. On the contrary the presence of N-concepts decreases the degree of relevance. N-concepts often are employed as filters to narrow search space. P-concepts and N-concepts are determined by *pointwise mutual information*, defined as follows,

$$I_P(r; c) = \log \frac{P(r, c)}{P(r)P(c)}$$

If $I_P(\texttt{presence}; \texttt{relevance})$ of a concept is greater than $I_P(\texttt{absence}; \texttt{relevance})$, it is a P-concept for a topic, and an N-concept otherwise. For example, for the topic “find shots of an airplane taking off”, “sky” is a P-concept and “animal” is a N-concept.

4. ANALYZING VIDEO ARCHIVE

Given an unannotated, large video archive, how many concepts are there? As a practical question, how much video do annotators have to watch before a reasonable set of concepts are identified? To answer these questions we first plot concept frequency, i.e. the number of shots where a concept appears, against the rank of a concept by concept frequency, as shown in Figure 2. The most frequent concept in TRECVID 2003 is “male speech” (22148), followed by “text overlay” (20540) and “music” (15847). The linear relationship between concept frequency and rank approximately follow Zipf’s Law [10], as first observed by [4]. The good news is that top-ranked concepts are extremely frequent and a set of common concepts may be quickly collected without browsing through

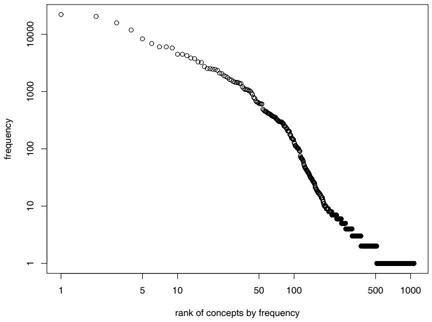

Fig. 2. Concept frequency approximately follows Zipf’s Law. Note that x and y axes are in log scale.

[2] Ambiguous concepts are grouped in the “Other” category.

much of a archive. To give a quantitative answer, we simulate the following scenario: an annotator browses a video archive from the first shot of the first video, and write down new concepts right after they appear. The results are plotted in Figure 3, where x axis is the number of shots that an annotator has watched, and y axis is the accumulated concept frequency, i.e. the number of unique concepts identified so far so far times the frequency of each concept. The result is very encouraging: By watching merely 1.2% (2723 seconds) of the archive an annotator can gathers a set of concepts that account for 90% of occureneces of concepts in the 62.2-hour video collection.

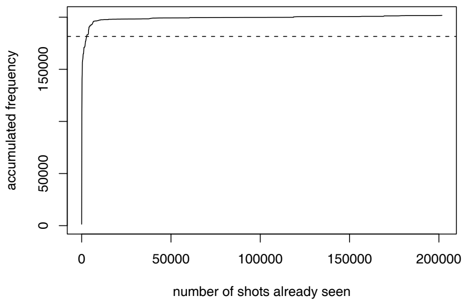

Fig. 3. Common video concepts can be quickly accumulated. The dashed line mark the 90% of total concept occurrences in the archive.

However, the bad news is that most concepts occur very infrequently. 90% of the concepts occur fewer than 100 times in the 48098 shots, which makes it very difficult to develop automatic detectors because statistical learning algorithms require large number of training examples [11]. Are these rare concepts really important for answering video retrieval queries? We investigate how concept frequency is related to video retrieval utility, and plot the number of the search queries that are helped by a concept (both positively or negatively) against concept frequency in Figure 4.

The results clearly show that rare concepts are unlikely to benefit more than one query (the total curve). For example, a rare concept,"mug", occurs only three times and help only one specific query, "find shots of a mug or cup of coffee". Only after concept frequency exceeds 100 can concepts help retrieval for queries of various types. For example, a frequent concept, "outdoors", occurs 3853 times and benefits 18 of 20 topics. We further break down the total number of helped topics by types of help. As concept frequency increases N-concepts are more likely to benefit more queries (the N-concept curve), which is not completely surprising as frequent concepts remove large number of irrelevant shots more effectively than rare concepts. P-concepts demonstrate the similar trend but to a lesser degree (the P-concept curve). Overall frequent concepts are more important because they can benefit more retrieval of search topics, either positively or negatively, than rare concepts.

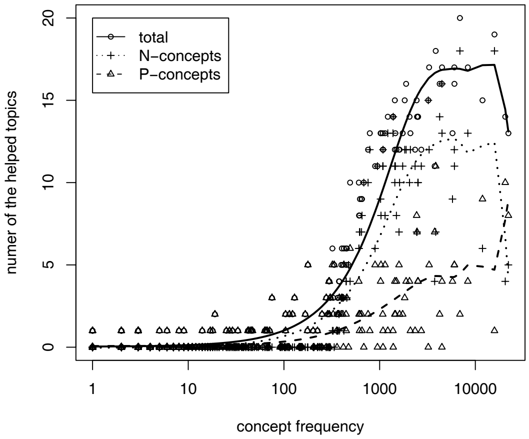

Fig. 4. The number of the topics that a concept can benefit is strongly related to its frequency. Note that X axis is in log scale. We fit three sets of data points with cubic splines.

We further investigate which category contributes more helpful concepts by plotting the accumulated number of the helpful concepts against the eight categories in Figure 5. The results are striking: the large category does not produce the most helpful concepts. While Objects is the biggest category, a smaller category, namely Scenes, contributes the (disproportionally) largest number of helpful concepts. Also unexpected is that the proportion of P-concepts and N-concepts vary from category to category. People concepts appear to be very effective N-concepts, possibly due to their specificity. When a search topic does not mention any people, the relevant shots of the topic are unlikely to contain People concepts (like "Clinton"), and thus concepts in the People category become effective filters.

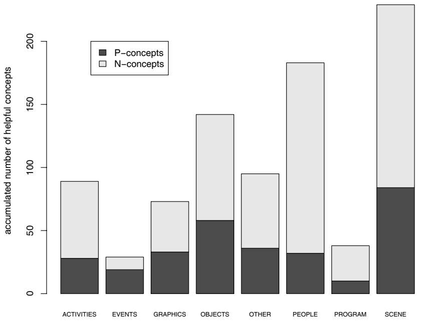

Fig. 5. The accumulated number of the concepts that benefit video retrieval in each category is disproportional to the number of members.

Users can include concepts in the query to concept-based retrieval systems, but how many concepts should be specified? What are the typical number of helpful concepts for a topic? To answer these question we plot the number of helpful concepts against the number of shots relevant to a topic in Figure 6. The total number of the helpful concepts, unfortunately, increases with the number of relevant shots to a search topic. This seems to pose a great challenge to both automatic retrieval system developers and interface designers: how to choose dozens of concepts from more than 1000 concepts that are potentially helpful? However, the scene unfolds very differently after we break down of the total curve into P and N-concepts. While the number of the N-concepts still increases with the number of the relevant shots of a topic, the number of P-concepts levels at around 20 after the number of relevant shots is greater than 10. The increasing N-concepts, similar to the finding in Figure 4, can be partly attributed to the filtering functionality of N-concepts. The more relevant shots a search topic matches, the more concept can become effective filters. The surprisingly steady number of the P-concepts suggests that no more concepts are needed once around 20 P-concepts are specified in the query.

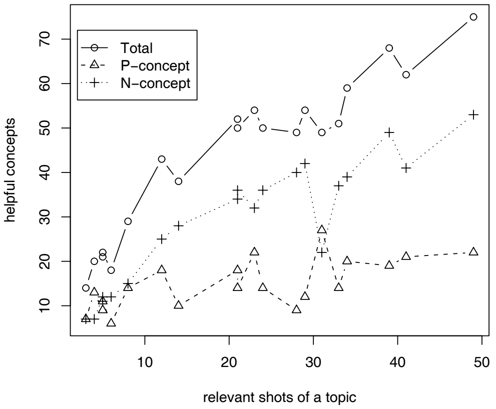

Fig. 6. The numbers of P-concepts and N-concepts vary differently with the number of the relevant shots of a topic.

5. CONCLUSIONS

In this paper we investigate the problem of using a large, fixed set of semantic concepts for video retrieval. We develop several principles for selecting concepts of high utility based on our analysis on a large collection of broadcast news video annotated with more than a thousand concepts and relevance judgment. Firstly, frequent concepts play a more vital role in video retrieval than rare concepts. Unlike rare concepts that benefit none or one specific topics, frequent concepts can help multiple search topics, either by filtering out irrelevant results (N-concepts), or by promoting relevant shots (P-concepts.)

Secondly, we should carefully allocate our resources to developing automatic detectors for different categories. Specifically, concepts Scenes category are shown to be very helpful and should be developed first. Although there are many concepts in the Objects category appearing in the archive, they usually benefit at most single query, making them virtually irrelevant for general search queries.

Finally, our finding that the numbers of P-concepts and N-concepts increase differently with the number of the relevant shots of a topic gives mixed blessing for concept-based retrieval. The good news is that the number of P-concepts appears to be in a manageable size of twenty. Once around twenty P-concepts are specified, users of concept-based retrieval system can stop contemplating more P-concepts. However, the bad news is how P-concepts and N-concepts will be selected from a set of 1000 concepts, either automatically by retrieval systems equipped with machine learning algorithms, or interactively with the help of user interface. Recent user studies [12] show that users have difficulty selecting which concepts would be helpful. Automatic video retrieval systems have yet shown statistically significant improvement over concept combination. Possible solutions to the concept selection problem, also our future work, include designing user interface to facilitate concept selection from a large set, and scaling existing machine learning algorithms to a much larger set of concepts.

6. REFERENCES

[1] Edie M. Rasmussen, "Indexing images," *Annual Review of Information Science and Technology (ARIST)*, vol. 32, pp. 169–196, 1997.

[2] Christos Faloutsos, Ron Barber, Myron Flickner, James Hafner, Wayne Niblack, Dragutin Petkovic, and Will Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems*, vol. 3, pp. 231–262, 1994.

[3] Apostol (Paul) Natsev, Milind R. Naphade, and Jelena Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proceedings of the 13th ACM International Conference on Multimedia*, 2005.

[4] John R. Kender and Milind R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment," in *Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition*, 2005, pp. 1174–1181.

[5] *Proceedings of the TREC Video Retrieval Evaluation 2003*, 2003, `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html\#2003`.

[6] Ching-Yung Lin, Belle L. Tseng, and John R. Smith, "Videoannex: Ibm mpeg-7 annotation tool for multimedia indexing and concept learning," in *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME)*, 2003.

[7] Alexander G. Hauptmann, "Towards a large scale concept ontology for broadcast video," in *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR)*, 2004.

[8] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[9] Yiming Yang and Jan O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML)*, 1997.

[10] George Kingsley Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Hafner Pub. Co, 1972.

[11] Milind R. Naphade and John R. Smith, "On the detection of semantic concepts at TRECVID," in *Proceedings of the Twelfth ACM International Conference on Multimedia*, 2004.

[12] Michael G. Christel and Alexander G. Hauptmann, "The use and utility of high-level semantic features in video retrieval," in *Proceedings of the Fourth International Conference on Image and Video Retrieval (CIVR)*, 2005.