

How many high-level concepts will fill the semantic gap in video retrieval?

Alexander Hauptmann
Language Technologies
Institute
School of Computer Science
5000 Forbes Ave
Pittsburgh, PA 15213
U.S.A.
alex@cs.cmu.edu

Rong Yan
Language Technologies
Institute
School of Computer Science
5000 Forbes Ave
Pittsburgh, PA 15213
U.S.A.
yanrong@cs.cmu.edu

Wei-Hao Lin
Language Technologies
Institute
School of Computer Science
5000 Forbes Ave
Pittsburgh, PA 15213
U.S.A.
whlin@cs.cmu.edu

ABSTRACT

A number of researchers have been building high-level semantic concept detectors such as outdoors, face, building, etc., to help with semantic video retrieval. Using the TRECVID video collection and LSCOM truth annotations from 300 concepts, we simulate performance of video retrieval under different assumptions of concept detection accuracy. Even low detection accuracy provides good retrieval results, when sufficiently many concepts are used. Considering this extrapolation under reasonable assumptions, this paper arrives at the conclusion that “concept-based” video retrieval with fewer than 5000 concepts, detected with minimal accuracy of 10% mean average precision is likely to provide high accuracy results, comparable to text retrieval on the web, in a typical broadcast news collection. We also derive evidence that it should be feasible to find sufficiently many new, useful concepts that would be helpful for retrieval.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information System—*video*

General Terms

Experimentation

Keywords

Concept-based Video Retrieval, Semantic Gap, High-level Semantic Concepts, LSCOM

1. INTRODUCTION

Digital images and motion video have proliferated in the past few years, ranging from ever-growing personal photo collections to professional news and documentary archives. In searching through these archives, digital imagery indexing based on low level image features like color and texture, or manually entered text annotations

often fail to meet the user information needs, i.e., there is a persistent semantic gap produced by “the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation” [24].

The image/video analysis community has long struggled to bridge this semantic gap from successful, low-level feature analysis (color histograms, texture, shape) to semantic content description of video. Early video retrieval systems [10, 25, 27] usually modeled video clips with a set of (low-level) detectable features generated from different modalities. These low-level features like histograms in the HSV, RGB, and YUV color space, Gabor texture or wavelets, and structure through edge direction histograms and edge maps can be accurately and automatically extracted from video. However, because the semantic meaning of the video content cannot be expressed this way, these systems had a very limited success with this approach to video retrieval for semantic queries. Several studies have confirmed the difficulty of addressing information needs with such low-level features [14, 22].

To fill this “semantic gap”, one approach is to utilize a set of intermediate textual descriptors that can be reliably applied to visual content concepts (e.g. outdoors, faces, animals). Many researchers have been developing automatic semantic concept classifiers such as those related to people (face, anchor, etc), acoustic (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports) and production (camera motion, blank frames) [4]. The task of automatic semantic concept detection has been investigated by many studies in recent years [2, 16, 11, 31, 12, 8, 28], showing that these classifiers could, with enough training data, reach the level of maturity needed to be helpful filters for video retrieval [5, 18].

The idea of indexing pictures has been approached by many archivists, and we can draw on their vocabularies for finding concepts and meaningful indexing terms for image material. Perhaps the largest index of image material is at the Library of Congress, and indexed by the Thesaurus of Graphic Materials (TGM), using manually selected index terms for every library item [1]. The TGM is sponsored by the Library of Congress, Prints & Photographs Division and in the form of TGM-I (Subject Terms) offers 69,000 terms for indexing subjects shown in pictures and also what pictures are about. Additionally, the companion TGM-II (Genre and Physi-

cal Characteristic Terms) document offers more than 650 terms for types of photographs, prints, design drawings, and other pictorial materials, with new terms added regularly. The Art and Architecture Thesaurus by the Getty Research Institute [7] is another standard source for image and video description with a “structured vocabulary of more than 133,000 terms, descriptions, bibliographic citations, and other information relating to fine art, architecture, decorative arts, archival materials, and material culture.” With slightly fewer indexing terms, the Australian Pictorial Thesaurus [20] has “15,000 subject terms for the indexing of images and other original material collections.” Additionally, there are numerous smaller controlled vocabularies for stock photography collections, museums and news organizations.

One might simply disregard the concept selection problem, and propose to index as many concepts as possible, for example, taking all index terms from the Thesaurus for Graphic Materials. However, we believe concepts should be carefully chosen. Building automatic detectors is by no means trivial, and developing concept detectors of very low retrieval utility will waste effort, or even worse, degrade the performance of large-scale concept-based retrieval systems. Which specific concepts should be automatically indexed is still an open research question.

In this paper, we examine the use of high-level semantic concepts [17] to assist in video retrieval and its promise to fill the semantic gap by providing more accessible visual content descriptors. We argue that a few thousand semantic concepts [15] that have reasonably reliable detection accuracy can be combined to yield high-accuracy video retrieval. If we can define a rich enough set of such intermediate semantic descriptors in the form of a large lexicon and taxonomic classification scheme, then robust and general-purpose content annotation and retrieval will be enabled through these semantic concept descriptors. Although a huge number of successful semantic concept detection approaches have been developed and evaluated, a number of questions remain to be answered, e.g., how many semantic concepts are necessary? How accurate do they need to be? Can we reasonably expect to identify these useful concepts and build automatic detectors for them.

To gain a deeper understanding on these open research issues, we have designed and conducted a set of retrieval experiments on large-scale video collections, which constitute first steps towards these questions and provide guidelines for future work.

The paper is organized as follows: We begin in Section II with a case study showing that increasing the number of semantic concepts improves video retrieval with experiments on concept counts needed to get good retrieval. We also provide an estimate on the minimal number of concepts that will be sufficient to construct a highly accurate video retrieval system. Then, in Section III, we analyze whether it is feasible to find several thousand such concepts to be added to the current lexicon. We end with a discussion, summary and additional open questions that are begging for answers.

2. VIDEO RETRIEVAL USING SEMANTIC CONCEPTS

To illustrate the usefulness of constructing a large number of high-level semantic concepts to enhance video retrieval quality, we provide a case study based on a large-scale TRECVID video collection [23] and two recently developed video concept lexica, namely the Large-Scale Concept Ontology for Multimedia (LSCOM) [15] and the MediaMill challenge concept data [26], both of which also

include an “annotation corpus” for the TRECVID 2005 video collection, where for each concept in the lexicon and every shot in the collection, it was manually determined whether the concept was absent or present in the shot. Our video retrieval experiments are conducted on 83 use cases queries, developed as part of LSCOM (also with concept truth annotations available on TRECVID 2005 video) to uncover the relationship between the number of concepts, detection accuracy and retrieval performance.

2.1 Description of Video Archive

In this section, we briefly describe the video collection, the set of high-level semantic concepts and the use cases (query topics) adopted in our case study.

2.1.1 TRECVID 2005 development set

The National Institute of Standards and Technology (NIST) has sponsored the annual *Text REtrieval Conference* (TREC) as a means to encourage research within the information retrieval community by providing the infrastructure and benchmark necessary for large-scale evaluation of retrieval methodologies. In 2001, NIST started the TREC Video Track (now referred to as TRECVID [21]) to promote progress in content-based video retrieval via an open, metrics-based evaluation, where the video corpora have ranged from documentaries, advertising films, technical/educational material to multilingual broadcast news. The international participation of TRECVID has rapidly grown from 12 companies and academic institutions in 2001 to 62 participants in 2005. The core TRECVID evaluation emphasizes a search task, where video clips relevant to a specific query (e.g. Shots of emergency vehicles in motion) need to be retrieved, either automatically or interactively.

Precision and recall are two central criteria to evaluate the performance of retrieval algorithm. Precision is the fraction of the retrieved documents that is relevant. Recall is the fraction of relevant documents that is retrieved. NIST also defines another measure of retrieval effectiveness called non-interpolated average precision over a set of retrieved documents (shots in our case). Let R be the number of true relevant documents in a set of size S ; L the ranked list of documents returned. At any given index j let R_j be the number of relevant documents in the top j documents. Let $I_j = 1$ if the j^{th} document is relevant and 0 otherwise. Assuming $R < S$, the non-interpolated average precision (AP) is then defined as $\frac{1}{R} \sum_{j=1}^S \frac{R_j}{j} * I_j$. Mean average precision (MAP) is the mean of average precisions over all queries.

As the largest video collections with manual annotations available to the research community, the TRECVID collections have become the standard large-scale testbeds for the task of multimedia retrieval. Each TRECVID video collection is split into a development set and a search set of roughly equal size. The development sets and their annotations are used as the training data to develop automatic multimedia indexing/retrieval algorithms in low-level video feature extraction, high-level semantic concept extraction and search tasks. The search sets mainly serve as the testbeds for evaluating the performances of retrieval systems.

In our case study, we used only the development data set from the TRECVID 2005 corpus [21]. This consists of broadcast news videos captured in the months of October and November 2004 from 11 different broadcasting organizations across two continents and several countries. The video collection includes multilingual news video captured from MSNBC (English), NBC Nightly News (English), CNN (English), LBC(Arabic), CCTV(Chinese) and NTDTV

(Chinese). The development set used here comprises about 70 hours of video. The development corpus was segmented into 61901 representative shots or units of retrieval with each such unit associated with a keyframe from the shot.

2.1.2 Semantic concepts

We studied three sets of high-level semantic concepts, where each set is larger than the previous ones and completely includes the smaller sets.

LSCOM-Lite The first set is generally referred to as LSCOM-Lite [15]. It is composed of 39 concepts and was released by NIST in conjunction with the TRECVID 2005 data set. The TRECVID 2005 participants jointly annotated ground truth for these concepts on the TRECVID 2005 development set.

MediaMill The second set of concepts was created by the MediaMill group in the Netherlands [26]. They annotated 101 concepts on the same data set, with a challenge to other researchers to use these concepts, annotations and low-level features to develop better concept detection and general video retrieval systems. Since only 75 of these concepts were present in our largest LSCOM concept lexicon, we limited our evaluation to those 75 concepts to achieve direct comparability.

LSCOM The largest set of concepts contained 300 concepts, developed as part of the LSCOM effort [15]. While the full LSCOM set contains over 2600 concepts, many of them are unannotated or contain no positive instances in the TRECVID 2005 collection. For this study we used 300 LSCOM concepts that were annotated with at least several positive instances. LSCOM is a collaborative effort of multimedia researchers, library scientists, and end users to develop a large, standardized taxonomy for describing broadcast news video. These concepts have been selected to be relevant for describing broadcast video, feasible for automatic detection with some level of accuracy and useful for video retrieval. LSCOM additionally connects all its concepts into a full ontology. However, we only used the LSCOM concepts as a flat lexicon in our experiments. All shots relevant to all concepts had been annotated for the TRECVID 2005 development set¹.

2.1.3 Use cases

To measure the usefulness of large numbers of concepts for video retrieval, we utilized a set of “use cases” that had been defined by the LSCOM activity [15]. Here, in consultation with user communities of broadcast news professionals and intelligence analysts, a number of scenarios were identified, which were well covered in the TRECVID 2004 data set. The scenarios covered unexpected breaking news events such as natural disasters, as well as long-standing news stories, such as “The oil crisis” or “US Elections.” Each of these events had aspects with associated video clips that the users would have liked to access and review in detail.

The use cases were designed to be (a) independent of the concept lexicon used in the annotation corpus and (b) a realistic setting for an information seeker in the sampled user communities. For example, the use cases included the following query topics:

¹This data is available at <http://www.ee.columbia.edu/ln/dvmm/lscom>

- *Afghanistan*: battles, disarmament, demobilization, and reintegration
- *Iraq*: Fallujah, car bombs, improvised explosive devices, assassinations
- *Eritrea*: War by proxy
- *Africa*: Various conflicts
- *Pakistan*: Terrorist attacks

For each use case, a set of specific topics was derived, which would be answerable by video documents in the collection. The topics were intended to emphasize the visual aspect of the video, not the accompanying audio narrative. Thus for the “Afghanistan” use case scenario, some example topics were:

- Battles/violence in the mountains
- Camps with masked gunmen without uniforms
- Armored vehicles driving through barren landscapes
- Mountainous scenes with openings of caves visible
- People wearing turbans and carrying missile launchers

These topics, in turn, were mapped into very specific, visual queries, e.g. looking for shots of:

- A person with head-scarf greeting people
- Military formations engaged in tactical warfare, or part of a parade
- Crowds protesting on streets in urban or rural backgrounds with or without posters/banners
- Military hospital with vehicles in foreground
- Military vehicles or helicopters
- Masked men with guns and/or any other weapons

To carry out some quality assessments of the concept utility, LSCOM also derived relevance judgments for the use-case queries. LSCOM adopted a labeling strategy similar to the “pooling” used in TRECVID [21]. In this strategy, the top results from various systems are labeled, while the bottom results are simply assumed to be negative. LSCOM approximated this approach by feeding the query topics to interactive retrieval systems and having annotators perform “interactive searches” over the collection by essentially issuing text and image queries and finding as many relevant shots as possible during a 30 minute period. This provided an approximate, but high-quality set of labels, useful for evaluating the quality of various concept-based search methods for the queries.

2.2 Retrieval Experiments

To evaluate the retrieval utility of the LSCOM-Lite, the MediaMill and the full LSCOM concept sets, we designed experiments based on the guidelines of the automatic retrieval task in TREC video retrieval evaluation (TRECVID), which requires an automatic video retrieval system to search relevant documents without any human feedback. As the baseline, we generated the standard text retrieval output by searching the speech transcript for automatically extracted text keywords from each use case. Given the annotations of high-level concepts and relevance judgments of use case queries at our disposal, we can linearly combine the semantic concept predictions into the text retrieval outputs so as to determine how much these use cases would be affected. In more detail, we compared text retrieval results to the results after incorporating (a) the LSCOM-Lite set of 39 concepts, (b) 75 concepts from the 101 concept MediaMill challenge which overlapped with the LSCOM concepts, and (c) 300 of the LSCOM concepts that had at least 10 positive instances in the annotations.

Let us begin our discussions with the most ideal case, where we assume that the semantic concept detection is perfect (equivalent to directly using the concept truth annotation) and that the combination method of concepts is optimal. Although impractical, the results reported in this setting can serve as a theoretical upper bound to indicate how useful the concepts can be. In previous work [30], we developed a theoretical framework for studying the performance limits over both monotonic and linear combination functions. In brief, the approach computes the optimal linear combination performance with the semantic concepts f_i fixed. Let us denote the linear combination $F(D, Q) = \sum_{i=1}^N \lambda_i f_i(D, Q)$ and $AP(F(D, Q))$ as the average precision of order list σ where σ is determined by retrieval score $F(D, Q)$ with respect to D . Therefore our task can be rewritten as a bounded constrained global optimization problem,

$$LLB = \max_{\lambda_i} AP \left(\sum_{i=1}^k \lambda_i f_i(D, Q) \right),$$

where we consider LLB as the locally fixed linear bound. Note that this bound allows different concept combination weights for different queries.

To handle the bounded constraint optimization problem, we use a global optimization algorithm called the MCS algorithm, proposed by Huyer et al. [6]. Their method combines both global search and local search into one unified framework via multi-level coordinate search. It is guaranteed to converge if the function is continuous. Also in our implementation, multiple start points are tried for each query to avoid local minima problems.

To get a more realistic estimate (as opposed to the perfect “oracle” detection) of the concept utility with state-of-the-art concept detection approaches, we repeated the experiment after introducing noise into the perfect concept prediction (but still with an oracle combination). The results from TRECVID 2006 semantic concept classification evaluation [19] show that the current best semantic concept classification systems can average slightly less than 0.2 MAP over the LSCOM-Lite concepts. Because mean average precision is a ranked-based measure and difficult to simulate, we approximated this MAP with a breakeven precision-recall point at 0.2². This was

²Breakeven precision-recall is usually a good approximation for mean average precision. They are equivalent to each other if the precision-recall curve is mirror symmetric to the line of precision = recall.

easily achieved by randomly switching the labels of positively annotated shots to be (incorrectly) labeled as negative and conversely switching some negatively labeled shots as incorrect positive examples, until we achieve the desired breakeven point where precision is equal to recall. This made the concept labels appear roughly equivalent to a detector with MAP of 20%.

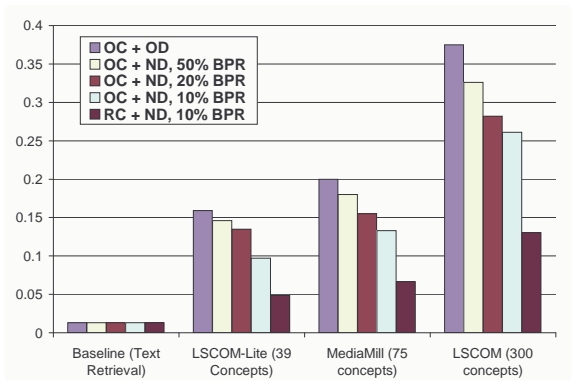


Figure 1: The oracle detection and oracle combination (OC + OD) results on the use cases for text-only retrieval and retrieval with the LSCOM-Lite, MediaMill and the larger LSCOM concept sets. The results using simulated “noisily detected” (OC + ND) concepts are also shown at 50%, 20% and 10% breakeven precision recall (shown as “50% BPR”, “20% BPR” and “10% BPR” respectively). Realistic combination estimates (RC) assume the combination algorithm is only half as effective as perfect combination and are shown with 10% breakeven precision recall as “RC + ND, 10% BPR”.

Figure 1 shows the retrieval results of combining semantic concepts for the three sets under different noise levels, and we can observe a substantial improvement brought by additional semantic concepts. For example, the video retrieval performance can reach an impressive 37% in terms of MAP after the full LSCOM set of concepts are incorporated, given that the baseline text-only retrieval MAP is only 1%. This surprisingly low baseline of the text-only retrieval, achieved using a configuration typical for video retrieval, can be attributed to the strong visual specificity of the queries, which cannot be answered by merely searching the audio transcript. This series of experiments confirms the huge potential of high-level concepts in helping to build an effective video retrieval system. As a more realistic setting, we also investigate the performance after introducing detection noise into the semantic concepts. In this case, the retrieval performance keeps decreasing when more and more positive shots are switched to negative ones. This shows that detection accuracy of semantic concepts has a noticeable effect on the quality of video retrieval accuracy. However, even if the breakeven precision-recall of these sets of concepts is only 10%, the retrieval MAP can still be boosted to 26% with the full set of LSCOM concepts. This suggests that although the prediction provided by the state-of-the-art automatic concept detection algorithms is far from perfect, they can still retain their potential in augmenting the standard text retrieval output.

It is worth mentioning that all of the above discussions assume the true relevant documents for each query are available and therefore we can use learning approaches to estimate the optimal combination model. However in practice, we will not be able to collect ground truth for every possible query that users may submit.

Hence, it is more reasonable for us to derive the combination models based on some obtainable query properties, such as query description, query context, user profile, or interactive relevance feedback. Based on recent results reported on the official TRECVID 2003 – 2005 retrieval tasks [29], realistic combination models (which are determined by the query description alone) may result in a 30%-50% loss of accuracy compared with the optimal combination models. Thus we modeled the “realistic combination” assumption in Figure 1 (and later in Figure 2) using a 50% degradation over oracle combination. However, even taking this additional discounting factor into consideration, the high-level concepts are still shown to be a potentially useful component for video retrieval given that it can boost MAP from 1% to above 10%.

2.3 How many concepts are sufficient for a good retrieval system?

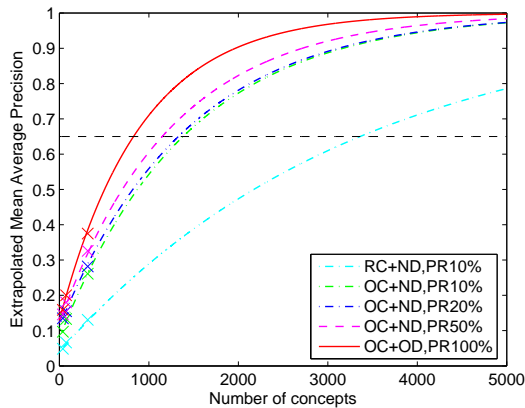


Figure 2: Extrapolated mean average precision vs. the number of high-level concepts. The true MAPs on the three original concept sets are indicated by “x”. The corresponding fitted curves are also plotted, where the solid line represents the oracle detectors (OD) with oracle combination (OC), and the dashed line signifies extrapolation using the noisy detectors (ND) with 50%, 20% and 10% breakeven precision-recall (PR). Realistic combination estimates (RC) assume the combination algorithm is only half as effective as perfect combination and are shown with 10% breakeven precision recall as “RC+ND,PR10%”.

One of the most interesting aspects of this work is that these results can now provide hints to answer the question: how many concepts are sufficient to construct a “good” video retrieval system? To be more rigorous in the rest of the discussions, we define a “good” system as one that can achieve more than 65% MAP. This corresponds to the current MAP accuracy that has been reported for the best web (text) search engines [3]. In order to investigate this problem more thoroughly, we need to extrapolate our MAP numbers on three extant concept sets to some “imaginary” concept sets with larger sizes under the assumption that the additional concepts have a similar quality to the existing concepts. The first step for extrapolation is to determine a reasonable extrapolation function for the given points in order to determine the relationship between MAP and the size of concept set. Since we only have three points to fit, theoretically there are infinite amount of functions that can be used to model the given numbers. However, we can impose a reasonable assumption upon the function space so as to determine a unique extrapolation function, i.e., the maximum MAP increment brought

by a new concept is proportional to the difference between the current MAP and the upper limit 1. In other word, it means that the higher the current MAP is, the less benefit a new concept can offer. According to this assumption, we can come up with the following partial differential equation,

$$\frac{dm}{dx} \propto (1 - m),$$

where m is the value of MAP, x is the number of concepts, and the boundary condition is $m(\infty) = 1$. By solving this equation, it yields,

$$m(x) = 1 - \exp(ax + b),$$

where a, b are two arbitrary numbers which can be determined by curve fitting approaches.

Figure 2 plots the true MAPs on three concepts sets as well as the fitted curves using the proposed exponential function over both perfect concepts and noisy concepts under oracle combination, as well as the noisy detection at 10% breakeven precision/recall and realistic combination, which assumes a 50% degradation in the combination step. It shows that the fitted curves provide a fairly accurate estimation on the target points. These curves also allow us to study the behavior of video retrieval system in depth if there are many more concepts available than the current status quo. To reach the level of a “good” retrieval system, we will only need around 800 concepts if we can obtain perfect detection accuracy for each concept. But more realistically, if overall accuracy of concept detection is only 10%, we will need more concepts to achieve the same level, i.e., around 1,200 to 1,300 concepts. From above results, we can conclude that a few thousand concepts should be sufficient to cover the most crucial contents in video corpora and provide a foundation to construct good video retrieval systems.

However, we also realize that there is a trade-off between the minimal number of concepts and the concept detection accuracy. The lower the detection accuracy, the more concepts we are likely to need to accurately retrieve the video content. Somewhat surprisingly, Figure 2 shows that when using 3000 or more concepts, even a fairly low detection accuracy of 10% for any single concept will be sufficient, even in the realistic combination scenario, which tries to approximate the combination behavior of real systems when compared to oracle combination [29]. While both the number of high-level concepts and concept detection accuracy needs to be taken into account when designing the video retrieval systems, this data suggests that it is better to add more concepts into the mix rather than building very accurate concept detectors, as long as we can find a reasonable method to combine them. We also want to point out that the discounting factor between realistic combination and oracle combination is also dependent on the choice of concept combination methods. If the combination method has to rely on the semantic meaning of the concepts, its retrieval performance may suffer more than a method that ignores explicit concept semantics, especially when the detection accuracy is as low as 10%. This suggests that one should consider semantic-insensitive and learning-based combination methods when the concept detection accuracy is not high enough.

Note also that our above analyses are based on the assumption that additional “imaginary” semantic concepts should have a similar quality (such as detection accuracy, proportion of the positive data and so on) to the existing concepts. The following section offers more analysis to see if it is reasonable to find this many concepts for for such a larger concept vocabulary.

3. FINDING A SUFFICIENT NUMBER OF CONCEPTS

We have shown in previous sections that video retrieval systems equipped with thousands of concept detectors are very likely to perform well, even if the accuracy of the individual detectors is low. The conclusion, however, contains the implicit assumption that thousands of reasonable high-level concepts are detectable in the video archive. The large-scale concept-based retrieval paradigm is clearly not viable unless we have a sufficiently large number of concepts to choose from. So, we try to answer the question: are there really that many concepts in a video archive?

Rather than manually creating and counting concepts, we will estimate the number of concepts in a video collection based on the distribution of the counts of high-level concepts. If one sorts the 300 LSCOM concepts in the TRECVID 2005 video collections in decreasing order of frequency and plots the count vs. rank (i.e., the position in the sorted list) in log-log scale, a surprising pattern emerges. We surely expect a decreasing curve because the concepts are sorted in the decreasing order of frequency, but the curve stays linear, at least for the high-frequency concepts, as shown in Figure 3. This linear frequency-rank relationship in log-log scale is widely known as Zipf’s Law [32], and has been extensively observed in many natural phenomena (e.g., words in a large text collection [13]). The high-level video concepts in TRECVID 2003 video collection also exhibit very similar patterns[9].

Zipf’s Law states that the product of the count and rank is constant,

$$c(i) * r(i) = k \quad (1)$$

where c and r are two functions that return the count and rank of the i -th concept. Equivalently we can rewrite (1) as

$$\log c(i) = -1 \times \log r(i) + \log k. \quad (2)$$

(2) is a straight line of slope -1 in the log-log plot. To estimate the total number of concepts in a video collection, we can set count c to be one, solve (2), and return rank r as an estimate.

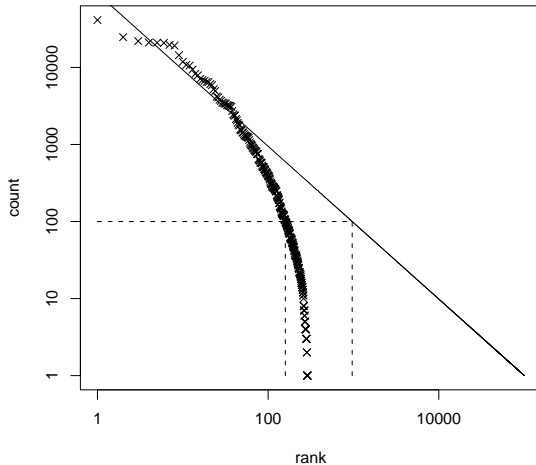


Figure 3: Each cross represents the count (y axis) and rank in the decreasing order of frequency (x axis) of a LSCOM concept in the TRECVID 2005 video collection. Note that both x and y axes are in log scale. The straight line represents an ideal linear relationship based on Zipf’s Law, and is fitted using weighted linear regression with exponential decay weights.

The linear relationship between count and rank, however, does not hold after the top 100 concepts (one third of the LSCOM concepts), and one may suspect the validity of estimating the total number of concepts in a video collection based on Zipf’s Law. The reason why the distribution of the mid and low-frequency LSCOM concepts does not closely follow Zipf’s Law is likely due to the idiosyncratic selection process of LSCOM concepts. LSCOM participants included only concepts that met certain criteria. Many low-frequency concepts (e.g., people or objects that would appear only very rarely in broadcast news) are excluded, which accounts for the deviation from Zipf’s Law.

To overcome the selection bias that LSCOM excludes many mid and low-frequency concepts, we fit a straight line using *weighted* linear regression. The weighted linear regression minimizes the following objective function J :

$$J(\beta, \beta_0) = \sum_i w(i) (\log c(i) - (\beta \log r(i) + \beta_0))^2, \quad (3)$$

where $w(i)$ is the weight for the i -th concept, $\beta_0 = \log k$ in (2). We would like to have a weighting function that favors high-frequency concepts (i.e., more weights) but penalizes severely low-frequency concepts (i.e., extremely low weights) because LSCOM may potentially miss very large number of low-frequency concepts. What should such a weighting function look like? One possible choice is an exponential decay function,

$$w(i) = \exp(-\lambda r(i)), \quad (4)$$

where λ is a decay constant. High-frequency concepts (i.e., the value of rank $r(i)$ is small) have large weights; low-frequency concepts (i.e., the value of rank $r(i)$ is large) have very small weights. We choose the λ that makes the estimated β in (3) as close to Zipf’s Law as possible, i.e., close to -1. The fitted linear curve is shown in Figure 3, with $\lambda = 0.0405$, $\beta = -1.0$, and $\beta_0 = 11.42$. By setting count to one in (2) we estimate the total number of concepts in the TRECVID 2005 video collection to be around 100K, which is in the same order of magnitude as the previous study on the TRECVID 2003 video collection [9]. Note also, that this is also similar in magnitude to the numbers of words found in written and spoken documents.

If the TRECVID 2005 video collection were annotated in an unconstrained manner such that the distribution of concepts closely followed Zipf’s Law (we denote this imaginary video concept ontology as LSCOM-Zipf), a total of 100K concepts are clearly more than enough to offer 3K concepts that are needed to achieve good retrieval performance. 3K concepts are shown to be sufficient – even if the detection accuracy is very low – to perform well using the state of the art combination strategy in Section 2.3, but the argument is actually based on the premise that the 3K concepts exhibit similar characteristics (e.g., the number of positive examples) as those in LSCOM. We would not expect retrieval performance to be improved if the 3K concepts from LSCOM-Zipf have significantly lower frequency in the collection than those from LSCOM.

We examine if LSCOM-Zipf could offer concepts of similar quality as LSCOM by comparing the number of training examples in the retrieval experiment in Section 2.3 (the first row of Table 1) and an imaginary retrieval experiment using 3K concepts from LSCOM-Zipf (the second row of Table 1). The LSCOM experiment uses all 300 concepts of LSCOM, while the imaginary experiment uses 3K concepts from a total of 100K concepts. The number of concepts with frequencies above 1000, 500, and 100 examples all show that

Concepts	Size	Used	> 1000	> 500	> 100
LSCOM	0.3K	0.3K	62	89	159
LSCOM-Zipf	100K	3K	97	192	945

Table 1: Comparison of concept frequencies for retrieval experiments based on LSCOM (see Section 2.2) and those based on estimated concepts that follow Zipf’s Law (the second row, denoted as LSCOM-Zipf). The second column is the size of the concept set, and the third column is the number of concepts incorporated into the experiments. The remaining columns list the number of concepts of frequency greater than X in each set (how the numbers of concepts of frequency at least 100 are obtained is illustrated with dashed lines in Figure 3).

LSCOM-Zipf is not only similar to LSCOM but, in fact, could offer more concepts of better quality in terms of frequency. The LSCOM experiment used 62 concepts of frequency greater than 1000, while the imaginary 3K experiment could have as many as 97 concepts at frequencies greater than 1000. Therefore, collecting 3K concepts from a video collection should not only be feasible, but the quality would also be better than for the LSCOM concepts used in our retrieval experiments, which may help retrieval systems achieve even better performance with fewer than 3K of concepts!

4. CONCLUSIONS AND FUTURE WORK

Using the TRECVID video collection and the LSCOM truth annotations of 300 concepts, we simulated performance of video retrieval under different assumptions of concept detection accuracy. The experimental results confirmed that a few thousand semantic concepts could be sufficient to support high accuracy video retrieval systems. Surprisingly, when sufficiently many concepts are used, even low detection accuracy can potentially provide good retrieval results as long as we find a reasonable way to combine them. We also provided speculative evidence based on Zipf’s law, that it is feasible to find the required numbers of concepts occurring with sufficient frequency in the video collection. However, we leave unanswered the question which specific concepts should be used

Above all, we hope that this divide-and-conquer approach using large numbers of semantic concepts as an intermediate layer will allow us to develop thousands of concepts that can be somewhat reliably identified in many contexts, and with sufficient numbers of these concepts available, covering a broad spectrum of visible things, users will finally be able to bridge the semantic gap. Ultimately, this paper arrives at the conclusion that “concept-based” video retrieval with fewer than 5000 concepts, detected with minimal accuracy of 10% mean average precision is likely to provide high accuracy results, comparable to text retrieval on the web, in a typical broadcast news collection. These observations may serve as a starting point for the future investigations instantiating such a large collection of concepts for retrieval.

5. REFERENCES

- [1] The thesaurus for graphic materials: Its history, use, and future. *Cataloging & Classification Quarterly*, 31(3/4):189–212, 2001.
- [2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 2002.
- [3] S. M. Beitzel, E. C. Jensen, O. Frieder, A. Chowdhury, and G. Pass. Surrogate scoring for improved metasearch

precision. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 583–584, New York, NY, USA, 2005. ACM Press.

- [4] S. F. Chang, R. Manmatha, and T. S. Chua. Combining text and audio-visual features in video indexing. In *IEEE ICASSP 2005*, 2005.
- [5] A. G. Hauptmann, R. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C. Snoek, G. Tzanetakis, J. Yang, R. Yan, , and H. Wactlar. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proc. of TRECVID*, 2003.
- [6] W. Huyer and A. Neumaier. Global optimization by multilevel coordinate search. *Journal Global Optimization*, 14, 1999.
- [7] G. R. Institute. Art and architecture thesaurus on line, 2006.
- [8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.
- [9] J. R. Kender and M. 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*, pages 1174–1181, 2005.
- [10] M. Lew, editor. *Intl. Conf. on Image and Video Retrieval*. The Brunei Gallery, SOAS, Russell Square, London, UK, 2002.
- [11] C. Lin, B. Tseng, and J. Smith. VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning. In *IEEE International Conference on Multimedia and Expo*, 2003.
- [12] W.-H. Lin and A. G. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 323–326, New York, NY, USA, 2002. ACM Press.
- [13] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.
- [14] M. Markkula and E. Sormunen. End-user searching challenges indexing practices inthe digital newspaper photo archive. *Information Retrieval*, 1(4):259–285, 2000.
- [15] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.
- [16] M. R. Naphade, T. Kristjansson, B. Frey, and T. Huang. Probabilistic multimedia objects (multijets): A novel approach to video indexing and retrieval in multimedia systems. In *Proc. of ICIP*, 1998.
- [17] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

- [18] A. P. Natsev, M. R. Naphade, and J. 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.
- [19] NIST. Overview of trecvid 2006, 2006.
- [20] C. of Australian State Libraries. Australian pictorial thesaurus, 2005.
- [21] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton. Trecvid 2005 - an overview. In *Proceedings of TRECVID 2005*. NIST, USA, 2005.
- [22] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does organisation by similarity assist image browsing? In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 190–197, New York, NY, USA, 2001. ACM Press.
- [23] A. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.
- [24] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE transactions Pattern Analysis Machine Intelligence*, 22 - 12:1349 – 1380, 2000.
- [25] J. R. Smith, C. Y. Lin, M. R. Naphade, P. Natsev, and B. Tseng. Advanced methods for multimedia signal processing. In *Intl. Workshop for Digital Communications IWDC*, Capri, Italy, 2002.
- [26] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Multimedia*, 2006.
- [27] H. Wactlar, M. Christel, Y. Gong, and A. G. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.
- [28] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579, 2004.
- [29] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Carnegie Mellon University, 2006.
- [30] R. Yan and A. G. Hauptmann. The combination limit in multimedia retrieval. In *Proc. of the eleventh ACM international conference on Multimedia*, pages 339–342, 2003.
- [31] J. Yang, M. Y. Chen, and A. G. Hauptmann. Finding person x: Correlating names with visual appearances. In *Intl. Conf. on Image and Video Retrieval (CIVR'04)*, Ireland, 2004.
- [32] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Pub. Co, 1972.