# ONTOLOGY DESIGN FOR VIDEO SEMANTIC THREADS

*John R. Kender*

Department of Computer Science
Columbia University
New York, NY 10027
jrk@cs.columbia.edu

*Milind R. Naphade*

IBM T J Watson Research Center
Business Informatics Department
Hawthorne, NY 10532
naphade@us.ibm.com

## ABSTRACT

We propose that, at the highest level of video understanding, the human needs for meaning and the methodologies to extract it are both universal and generic. One must develop an ontology, then develop analyzers that learn the statistical correlates of that ontology, and finally use the analyzers to tie together common occurrences across individual videos. The first step towards adapting the ontology to the genre is the design of automated tools to assist in the annotation of the ground truth; these tools in turn provide feedback on the appropriateness of the filters and the ontology.

We support this hypothesis by presenting and discussing some experiments conducted on the NIST TRECVID 2003 video corpus. We also validate this hypothesis by showing the connection between story tracking in our multimedia news and topic detection and tracking in the NIST TDT natural language effort. At the highest level, we find that our annotation tool shows that semantic concepts tend to cluster reliably into a few significant semantic dimensions. For news videos specifically, two of these clusters measure ”presidentiality” and ”outdoor-ness”.

## 1. INTRODUCTION

Video analytical requirements differ from domain to domain, and become increasingly dependent on semantic analysis as the level of video editing increases. The structural spectrum, stretching from surveillance through home, educational, sports, news, documentaries, to dramas, requires increasingly sophisticated semantic filters. Although there are semantic classes that are common across these domains, the mapping from features to classes does not often generalize well.

Fortunately, at the highest level of semantics, abstractions which involve the detection of semantic boundaries across multiple heterogeneous feeds over time (sometimes called ”stories” or ”themes”) are common and necessary. These abstractions depend less on low-level features and more on the domain ontology. Their proper devices are not detectors but unsupervised techniques such as machine learning, generic methods that are applicable across many domains.

This paper reports on a case study of tracking stories using a well-defined ontology, and on what the successes and failures of this effort say about ontology design. In Section 2 we describe the VideoAnnEx visual concept ontology. In Section 3 we examine the aggregate properties of the resulting XML database of shot annotations, using an approach similar to that used by statistical natural language processing to analyze linguistic corpora. In Section 4 we purge annotator biases. In Section 5 we explore the time-dependency of these concepts, and develop a similarity metric that allows news episodes to be semantically matched across time and across channels. In Section 6 we apply a particular low-level computer vision segmentation method to aggregate stories, and we visualize and interpret the apparent structure of the cognitive space underlying the news story domain. In Section 7 we recommend how visual concept ontologies and video annotation tools can be improved.

## 2. THE TOOL AND THE ONTOLOGY

From April to July of 2003, 111 researchers from 23 institutions worked together for 422 man-hours to associate 433K of ground-truth semantic labels to 62.2 hours of videos. These videos were taken from the National Institute of Standards and Technology (NIST) TRECVID 2003 video development set, a set that includes 241 30-minute programs of ABC World News Tonight and CCN Headline News, plus a smaller amount of C-SPAN segments. These were recorded from January through June 1998, by the Linguist Data Consortium (LDC) hosted by the University of Pennsylvania, for use in the NIST Topic Detection and Tracking Phase 2 experiment (TDT2)[1]. Fully half of all these news programs

This work was completed while the first author was supported on sabbatical at IBM Research. This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

were segmented into shots, with each shot annotated with an average of 9.2 visual concept labels. Researchers in this Video Annotation Experiment ("VideoAnnEx") used the IBM MPEG-7 Annotation Tool [2], which presents a hierarchy of 133 suggested visual concept labels, but also accepts new user-created ones. It outputs for each video an MPEG-7 XML file of shot-synchronized visual concept annotations. The resulting massive XML database comprised approximately 47K shots (including commercials) annotated with a semantic ontology that grew to 1038 different visual concept labels.

An examination of the literature suggests that the video aspect of news stresses human facial reactions and at-the-site visual reportage [3], two aspects not as prominent in audio, textual, or close-captioned sources. The ontology, however, was based on a retrospective analysis of three years' worth of (much smaller) prior NIST VIDEOTREC annotation attempts. This four-level hierarchy consisted in its first two levels of: 35 visual event concepts (Person-Action, People-Event, Sport-Event, Transportation-Event, Cartoon, Weather-News, Physical-Violence), 38 visual scene concepts (Indoors, Outdoors, Outer-Space, Sound), and 49 visual object concepts (Animal, Audio, Human, Man-Made-Object, Food, Transportation, Graphics-and-Text), where 11 sound concepts were distributed amongst the scene and object concepts. This ontology is clearly uneven in both its breadth and its depth. Most significantly, these visual concepts do not indicate any human reactions, have only the most generic of site descriptions. Additionally, they included only three named objects (that is, political figures).

## 3. ANALOGIES TO STAT-NLP AND IR

To mine and test the statistics of visual concept use, we adopted the following analogy to the methods of the fields of Natural Language Processing (NLP) and Information Retrieval (IR). News stories, which are reported in various episodes over various days, are similar to document classes, with the episodes themselves similar to documents. Video shots act like paragraphs, and therefore visual concepts act like words. Thus, a news episode can be modeled as a bag of visual concepts. But in contrast, a typical corpus in NLP consists of 20 to 50 millions of words, or about 100 times as much data.

It has been noted by Zipf and others that the frequency of use of a word is inversely proportional to the word's rank order. Stated in logarithms, Zipf's law says $log(f(i)) = c - log(r(i))$, where $f$ and $r$ are frequency and rank, respectively, and $i$ is the index for the $i^{th}$ word. A plot of the visual concepts in VideoAnnEx (including all additional annotator-supplied concepts) found that there is no such fit. Plotting the 814 concepts that are used only by the ABC and CNN broadcasts, it is only when a generalized form of Zipf's law called Mandelbrot's law is used, does there appear any reasonable inverse power law relationship. The optimal three-parameter least-squares fit is instead: $log(f(i)) = 2.3 - 2.2log(r(i + 3))$. This suggests the visual conceptual ontology is not yet mature, is too generic, and is too sparse. If one assumes that the more frequently used concepts are correctly assigned, and if one flattens the observed curve so that its slope approaches that of the -1.0 of Zipf, one finds that the intercept for rank would approach that of $log(r(i)) = 6 * 2.2 = 13$. This gives a total vocabulary of $e^{13}$, or about 400K concepts, a number close to that reported for the proprietary manually-annotated visual concept hierarchy used by CNN, of which 90% are named entities such as specific people or places.

We also attempted to see if there were semantically meaningful major clusterings of shots. We adopted the Dice metric for comparing shot labelings, with each shot reduced to a bit string of concept presence. [4]. Shots were then clustered according to the average link metric for inter-cluster distance [5]; however, the use of simple link or complete link gave similar results. We found only four major clusters. The one subsuming the most concepts corresponded to shots involving the anchor person and the anchor setting. The second was a "presidential" cluster, corresponding to concepts related to the activities, official and unofficial (as this was the year of Monica Lewinsky) of the concept Bill-Clinton, just as predicted by Gans. A third cluster reflected outdoors and transportation settings, and the fourth sports.

Agains following statistical NLP, we studied the collocations of visual concepts, that is, pairs of concepts that occurred more frequently than the product of their independent probabilities would suggest. This often suggests semantic "idioms", that should be considered concepts in their own right. For relatively small data sets such as ours, the IR community recommends the $G^2$ likelihood ratio as the most accurate measure of collocations[6]. We found many, mostly because annotators appear to fixate on a few personal favorite visual concepts which they tend to use together. However, some collocations pointed out defects in the proposed ontology, for example, the common simultaneous labelings of "people" and "people-event" suggest that these concepts probably be merged into a single one in a revised ontology.. Conversely, the avoidant concept pairs we detected, such as "indoors" and "transportation", can suggest ways in which an ontology can be more accurately factored into disjoint concept categories.

## 4. PURGING THE GROUND "TRUTH"

In a first step to explore how well the ontology allowed stories to be tracked, we merged the VideoAnnEx XML annotations of shots, together with the existing LDC TDT2 database of ground truth time stamps for story episode bound-

aries and ground truth story classifications. Our data merger generated a database of 487 episodes of 42 named stories. However, we quickly discovered, using two separate methods (Dice and Okapi[7]) that the ground "truth" tended to cluster together news episodes that occurred within the same broadcast. We traced the problem to the way in which annotators had been assigned their tasks. Each had been given a number of entire half-hour programs, in a round-robin allocation scheme. All of the episodes within a single broadcast were more easily associated through their concept signatures with their annotator, again due to annotator cognitive fixity.

Requiring that concepts be used on more than two separate days removed much of this bias, but also removed 90% of the ontology. However, even this approach was still insufficient to deal with the erratic use by annotators of those visual concepts that were the most common. These "production values" concepts, such as "text-overlay", "human", "studio-setting", "monologue", etc., suffered from annotator sloppiness. A principal component analysis of the use of these 110 concepts surviving the purge showed that the major variation in their use was due to arbitrary failures of annotators to include these annotations appropriately.

The final purge was achieved by adopting from the machine learning community a form of feature selection based on information gain[8]. We binarized the presence of concepts within each episode, and then by using the TDT2 ground truth, we grouped together all the episodes of a given story. We then defined for each story the information gain for each concept, and selected the highest ones. What finally survived were mid-frequency concepts. For an information gain of .02, 14 reliable visual concepts are selected: Road, Airplane, Rock, Forest, Meeting, Sport-Event, Golf, Hockey, Ice-Skating, Meeting-Room-Setting, Snow, Outer-space, Music, and Bill-Clinton. Surprisingly, all but one are visual concepts that refer to settings, rather than actors, objects, or events.

## 5. TEMPORAL DEPENDENCIES

Although temporal variation in news episode selection has been noted in the journalism literature, there does not appear to be any quantified model of the life cycle of a news story or its component visual concepts. We explored this issue statistically on the episode level, since there was reliable ground truth from the TDT2 database about named story episode occurrences and durations.

We found in general that the two channels available for VideoAnnEx, ABC and CCN, differed substantially in their basic temporal statistics. However, with neither channel did it appear that the duration of episode $N$ in a story predicted the number of days until episode $N+1$ or the duration of episode $N+1$; in both channels these three variables were correlated only with $|\rho| < .07$. Additionally, the episodes presented by either channel correlated only weakly with each other on a daily basis.

There was one interesting pattern, however. Empirically, it appears that the decay of the probability of appearance of further new episodes for a given named story are essentially the same, in simple shape and in parameter, in both channels. In both, the likelihood of finding an additional episode after exactly $d$ days (where $d = 0$ is possible) is approximately proportional to $1/(d+1)$. That is, episode decay follows a power law with a significantly long tail, which may reflect a common editorial judgement of viewer aggregate boredom over time, regardless of a channel's policy for same-day repetitions.

Given this model of temporal dependency, we were able to combine the semantic and temporal measures of story similarity in a computationally straightforward way. We defined the similarity of two episodes, $i$ and $j$, to be: $S(i,j) = Dice(i,j)/(1+Gap(i,j))$, where $Gap(i,j)$ returns the difference in days between the two episodes.

## 6. SUBSPACES FROM NORMALIZED CUTS

Having finally purged our ontology, incorporated time, and populated a full similarity matrix, $S$, we sought a clustering method for the video episodes that would maximize inter-story cluster separation while at the same time maximizing intra-cluster cohesion. One spectral clustering method specifically designed for those criteria is the method of normalized cut [9]. It solves the generalized eigen equation $(D-S)x = \lambda Dx$, where $D$ is a diagonal matrix with $D(i,i) = \sum_j D(i,j)$, with $D-S$ called the Laplacian matrix of the problem. This generalized form effectively achieves the desired two optimizations simultaneously. Its eigenvectors should describe a useful submanifold of the visual concept space, that is, those dimensions along which video episodes are shown to have significant, if latent, ontological similarities (within cluster) and differences (between clusters).

When applied to the VideoAnnEx episode vectors, now represented by only the 14 visual concepts with the highest information gain, we find that this method does produce as from its strongest eigenvector a visual concept subspace measures something significant, and which might be called "presidentiality". This is, in fact, reminiscent of one of the shot clusters in Section 3). This vector positively weights the visual concepts of Meeting, Meeting-Room-Setting, and Bill-Clinton. It separates, in a measurable way defined below, the episodes of Stories #2 (Lewinsky), #15 (Iraq), and #48 (Jonesboro shooting)—many of which, but not all of which, directly involve Bill-Clinton—from episodes of Story #13 (Olympics) and other stories. The next strongest eigenvector generates a subspace that could be called "outdoors",

as it positively weights Snow, Rock, and Road, and is again somewhat reminiscent of another shot cluster in Section 3). See Figure 1.

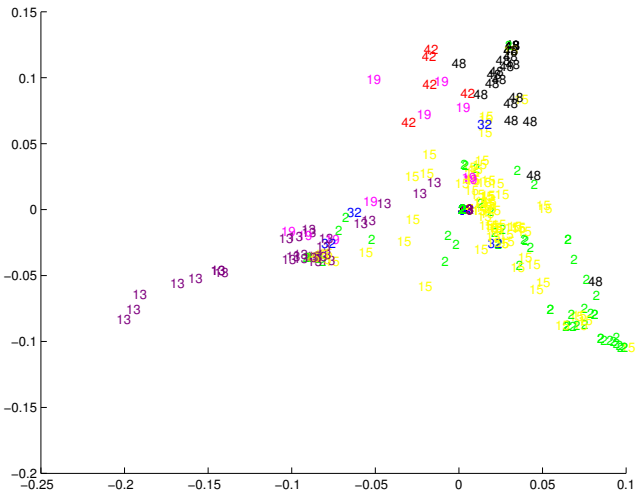


**Fig. 1**. The first two significant eigenvectors returned from the normalized cut algorithm with six stories superimposed by their TDT2 Story number. The X dimension appears to measure "presidentiality", with positive episodes relating to Bill-Clinton; the Y dimension appears to measure "outdoors".

## 7. CONCLUSIONS

We derive from these experiments some recommendations on how to design a more effective visual concept ontology for news stories, and how to design a more effective annotation tool. We suspect that similar recommendations would apply to other general-audience domains as well.

The most effective concepts, those with high information gain, appear to be ones dealing with settings, or with named people. Filters should therefore concentrate on extracting background characteristics, or facial ones. The topmost level of the ontology should probably more accurately reflect more generic categories, particularly since the "presidential" category appears well-established both in our experiments and in the literature. In terms of methodology, the techniques of NLP and machine learning are powerful tools for understanding the strengths and structure of an evolving ontology.

Much better ground truth for news stories is still needed, but it may only be possible if it is gathered with more intelligent annotation tools. The second generation of tools should monitor the statistical record of each visual concept. It should also detect annotator biases, perform spelling and other sanity checks, and require annotators to make forced choices out of each of several disjoint concept categories. New visual concepts should be accepted from annotators only with reluctance, and only if they are not used as a part of detected collocations. Perhaps most critically, the tool should actively encourage annotators to focus on settings and named entities.

One missing component of these investigations is a subtle one. Yet to be explored are effective techniques for determining the proper hierarchical structure of the ontology, particularly with regard to how factored or flat it should be. We suspect that most ontologies for video analysis will be factored, although experiments would probably need two or three orders of magnitude more data than VideoAnnEx to prove this convincingly.

## 8. REFERENCES

[1] Chris Cieri, David Graff, Mark Liberman, Nii Martey, and Stephanie Strassel, "The TDT2 text and speech corpus," in *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[2] Ching-Yung Lin, Belle L. Tseng, and John R. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," in *Proceedings of the NIST TRECVID Workshop*, November 2003.

[3] Herbert J. Gans, *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*, Vintage Books, New York, NY, 1980.

[4] Christopher Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[5] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988.

[6] Ted Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[7] Robertson and Sparck-Jones, "Simple, proven approaches to text retrieval," Tech. Rep. UCAM-CL-TR-356, University of Cambridge Computer Laboratory, 1997.

[8] Eric Xing, Michael Jordan, and Richard Karp, "Feature selection for high dimensional genomic microarray data," in *Proceedings of the International Conference on Machine Learning*, 2001.

[9] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.