Zero-Shot Video Retrieval Using Content and Concepts

Jeffrey Dalton, James Allan, and Pranav Mirajkar
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
{jdalton, allan, pranav}@cs.umass.edu

ABSTRACT

Recent research in video retrieval has been successful at finding videos when the query consists of tens or hundreds of sample relevant videos for training supervised models. Instead, we investigate unsupervised zero-shot retrieval where no training videos are provided: a query consists only of a text statement. For retrieval, we use text extracted from images in the videos, text recognized in the speech of its audio track, as well as automatically detected semantically meaningful visual video concepts identified with widely varying confidence in the videos. In this work we introduce a new method for automatically identifying relevant concepts given a text query using the Markov Random Field (MRF) retrieval framework. We use source expansion to build rich textual representations of semantic video concepts from large external sources such as the web. We find that concept-based retrieval significantly outperforms text based approaches in recall. Using an evaluation derived from the TRECVID MED'11 track, we present early results that an approach using multi-modal fusion can compensate for inadequacies in each modality, resulting in substantial effectiveness gains. With relevance feedback, our approach provides additional improvements of over 50%.

Categories and Subject Descriptors

H.3.3 [**Selection Process**]: [Information Search and Retrieval]

Keywords

Video Retrieval, Information Retrieval, Information Extraction

1. INTRODUCTION

Retrieving videos in response to a query is a long-standing research challenge. It has been studied frequently and is the underlying problem in the decade-old TREC Video Retrieval Evaluation (TRECVID) program [13]. The problem has taken many forms, with queries being sets of videos [2, 15], images, text, or combinations of those [14].

In TRECVID's Multimedia Event Detection (MED) task the goal is to identify potential events in a collection of multimedia material [4]. In the MED track, a collection of sample relevant videos and a detailed event description are provided. The goal is to identify a high-recall set of videos matching the information need. A MED information need is a complex activity involving people interacting with people and/or objects and is directly observable (e.g. rock climbing). The event "query" includes a free text description of the event and, similar in style to past TREC filtering tasks [12], a large number of videos declared "on topic."

We change this task slightly to be "zero shot" retrieval that does not leverage sample relevant videos. We are given only a textual description of an event which we convert into a query for retrieval. Our goal is to leverage both the textual video representations from automatic speech recognition (ASR) and optical character recognition (OCR) output as well as a representation of the video using high-level object and action visual concepts.

A key challenge is to automatically identify the relevant semantic concepts for a particular text query. For this study, we have a large collection of videos from TRECVID's MED evaluation that are automatically annotated with 531 concepts. As described in detail in Section 3, we introduce a method based upon concept expansion, where we use large external sources of text to construct text language models for each concept. For example, a *face* detector will be deemed useful if the query includes words or concepts that are related to faces (e.g. nose, eyes, mouth). We use the resulting ranked list to create a weighted concept query.

Our approach works broadly as follows. Given an information need expressed in text we generate a query to search for videos directly in the ASR and OCR content. In addition, we take the text query and use it to retrieve video concepts. We use the retrieved concepts to generate a weighted concept query for retrieval against indexed visual concepts, resulting in a ranked list of videos. The result is video results across the different modalities, which can optionally be combined using metasearch fusion.

In this study we present preliminary results for this approach on the MED'11 track. We find that concept-based retrieval is unstable for high precision but provides superior recall effectiveness. We use cross-modal fusion to combine the ranked lists from OCR, ASR, and concept retrieval (Section 5.1). The fused results significantly outperform individual modalities (24% gain in mean average precision (MAP) and 13% in precision at ten, over the best method for each). Finally, in Section 5.2 we also explore relevance feedback approaches, first within each modality – where we show consistent and significant gains – and then across the modalities, with another round of fusion. In the final run, we demonstrate a 25% gain over fusion for pseudo-relevance feedback and a 55% gain in MAP if the feedback is the result of human judgment of the top ten returned results.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
ACM 978-1-4503-2263-8/13/10.
http://dx.doi.org/10.1145/2505515.2507880 .

2. RELATED WORK

For an overview of recent work we refer the reader to surveys of content-based video indexing and retrieval provided by Hu et al. [5] and Snoek and Worring [16]. Previous work mapping text to concepts relies upon exact or approximate string matching or by associating ASR transcripts [10] with the concepts. Snoek et al. [18] use the vector space model to match text queries to concept descriptions which are used for identifying relevant videos. Neo et al. [11] perform expansion of both the text query and the concept description using Wordnet and a sample of external news. Li et al. [8] perform video retrieval using text queries that are mapped to concepts and utilize the video information across modalities. Instead of expansion using a hand-built knowledge source such as Wordnet, we propose a method based upon external source expansion, using the web to build a model of a concept from topically related text documents. Instead of string matching to a small number of precise concepts we use retrieval to rank concepts. This step is novel because new systems use thousands of concepts with sparse descriptions that are unlikely to match a query directly.

Feng et al. [3] use relevance modelling [7] to retrieve and annotate videos with one word labels. In this work we also use relevance feedback, but we use it to perform query expansion, to bridge query-document vocabulary mismatch and improve recall.

The most closely related work to ours is that of Younessian et al. [19] who leverage automatic speech recognition transcripts, acoustic concepts, and visual semantic indexing concepts to rank videos in the MED retrieval task. We use the SIN concept features plus additional action based concepts developed for the MED events. They manually assign concepts to events, while in this work we perform this automatically using retrieval to rank the most likely concepts.

3. CONCEPT MODELING

In this section we describe the heart of our approach to zero-shot retrieval which leverages high-level semantic concepts. Given videos labeled with a collection of hundreds or thousands of concepts, the task is to identify the set of concepts relevant to our original text query and use these concepts to find relevant videos. A fundamental issue is how to construct a representation of the concepts. In this work, we propose a model based on external source expansion. Because the language used in the videos is extremely sparse and noisy, we leverage the web to construct a model of the concepts, hoping to capture the wide variety of vocabulary across a range of sources.

3.1 Concept Source Expansion

For each visual concept, c_i, we create L_i, a model intended to represent the language that is likely to be used to describe that concept. For example, the visual concept *horse* might include words and phrases such as *pony*, *horse galloping*, *stable*, *jockey riding*, and *saddle* with high probability.

Although there are many ways to build the model, for this poster we construct the model by searching the web for content with the same topic as the visual concept. We form a query using 1) the concept name itself (e.g., *hill*) and 2) a list of manually generated related words (e.g., *mountain*, *trek*, *climb*, *landscape*). The query is run on a web search engine. The top k (k=100 in these experiments) results are fetched, and the full-text stored. The model for the visual concept is constructed by concatenating the results (pages are limited to 50,000 characters). From the resulting model, we can generate unigram, bigram, and other statistics. For example high probability terms for the concept *hill* include trek, nepal, camp, mountain, hill, valley, peak; the concept *birds* has the words: nest, wings, species, feathers, song, and flight.

3.2 Text to Concept Queries

To translate an event description (a query) into a set of concepts we first create a text query from the event and retrieve concepts. For these initial experiments we use the sequential dependence retrieval model [9]. The resulting ranked list of concepts is used to generate a concept query.

We experiment using the top 5, 10, 20, or 50 concepts. We hypothesize that the quality of ranked concepts will degrade rapidly. We also explore how the selected concepts should be weighted. We considered uniform weighting, where all of the concepts are treated equally, and a weighting inspired by the Relevance Model [7] where concepts are ranked by how well they match the query. Here we hypothesize that the weighting is not be important with few concepts but is important when more concepts are used.

4. EXPERIMENTAL SETUP

For both the video and concept retrieval experiments in this work we use the sequential dependence model [9] in the Markov Random Field retrieval framework. We use Dirichlet smoothing. Both queries and documents are stemmed using the Porter stemmer. Queries are stopped using a widely used 418 word stoplist. This model is supported by the open source Galago search engine[1] which we use for these experiments. Statistical significance testing is performed using a paired t-test.

4.1 Data

We use the video data from the TRECVID 2012 Multimedia Event Detection (MED) evaluation [4]. The collection contains the combined videos from the MED12 events (Event Kit), training (DEVT), and development data (DEVO). In total there are more than 47,000 videos containing more than 1400 hours of video. On average the videos are short clips which have an average length of approximately 2 minutes.

There are 30 event detection queries provided as part of the evaluation. We perform 3-fold cross-validation by evenly splitting the queries with their number modulus three. We split the queries this way because this method ensures an even distribution of queries across years since they vary in difficulty.

We use the automatic speech recognition (ASR) transcripts from the Janus [17] ASR engine. We use word-level output. It contains non-empty output for 5,212 videos. The average ASR transcription length (when there is one) is 873 words.

In addition to speech, optical character resolution (OCR) is performed at a sample rate of 10 frames per second [1]. We use a word-level representation. There are 4,311 videos with non-empty OCR output with an average of 340 words per video.

When combined, there are 8,460 videos with some textual representation, approximately 18% of the total videos. That means that more than 80% of the videos have no chance of being retrieved using text content alone.

4.2 Visual Concept Annotations

We use two types of visual concept detectors. The first consists of 346 Semantic INdexing (SIN) features for each video developed for the TRECVID Semantic Indexing track [6]. Broadly, these are concepts that are likely to occur in any set of videos. Example of these concepts include *Airplane*, *Bicycles*, *Canoe*, *Church*, *Computers*, *Dolphin*, *George Bush*, *Gun*, *Motorcycle*, *Road*, *School*, *Truck*, and *Whale*. These were detected on key frame images with no motion using SIFT based features using a boosted SVM classifier. There is no overlap between the videos in the SIN dataset and

[1] `http://www.lemurproject.org/galago.php`

the MED video collection. The detections are aggregated using the mean value to produce one score per video.

We also include a set of 185 visual concept developed specifically for the event detection queries [1]. These are action orientated concepts such as *group walking*, *landing with the board*, *reeling in*, and *vehicle moving*. The detectors use dynamic features, in particular Dense Trajectory Features (DTF) combined with Histogram of Orientated Gradient (HOG). From these features binary SVM classifiers with a Histogram Intersection Kernel are used for concept classification. Classification is performed on a 130 frame sliding window within each video, providing one detection score per concept for every window. To be comparable to the SIN concepts, we aggregate frame-level information to video-level annotations. Based upon effectiveness on the training data, we found that using the mean detection value plus one standard deviation was an effective indexing technique. We also apply a threshold, $\theta = 0.01$ to discard concepts with low probability.

4.3 Query representation

Given an event, our methods require that the system generate a query. Like commonly used TREC topics, MED events are provided by a complex text description, consisting of a number of fields that provide different levels of detail about the event. We apply the sequential dependence model to the full text of the event, including the name, the definition, and additional fields. The average length of this form of the query is 243 words. We calculate a score for each field and then combine them linearly with weights of 0.45, 0.2, 0.05, and 0.3 for the *name*, *definition*, *explication*, and *evidence* sections, respectively. The weights were chosen by inspection of training data results.

4.4 Evaluation Metrics

For evaluation, we have binary relevance judgments. We include precision oriented measures, such as P@10. We also report Mean Average Precision (MAP) because it incorporates both precision and recall. In the MED surveillance scenario, recall is critical. The evaluation used in the MED'11 evaluation uses Missed Detection (MD) at a given False Alarm (FA) rate, MDFA. (MD is one minus recall.) Specifically, we report MD values at a false alarm rate of 4%. Note that for MDFA04 a smaller value is better because a smaller fraction of the documents are missed.

5. RESULTS

We start with approaches that use just extracted text (OCR) and recognized speech (ASR), though we consider them independently. Less than 20% of the videos have any OCR or ASR text at all, so we expect recall-oriented measures to perform poorly. We found the smoothing parameter μ to be stable across all training folds with ASR, $\mu = 3500$ and for OCR $\mu = 500$ being optimal.

The results for ASR/OCR retrieval are included in Table 3. The OCR runs are all statistically significant better than the ASR runs in precision and MAP, but have similar or slightly worse recall measured with MDFA04. OCR retrieval gets approximately five relevant documents in the top 10 on average, which is promising. However, the recall and overall MAP scores are quite low. In some cases no relevant documents are retrieved: 3 queries for ASR and 1 query for OCR. For ASR, 10 queries returned zero relevant documents in the top 10.

Overall, OCR outperforms ASR in precision and is comparable in recall. However, on their own neither achieves satisfactory MAP or MDFA04 results.

We now evaluate the effectiveness of using automatically selected concepts to retrieve videos. The smoothing parameter $\mu =$ 19000 is optimal across all folds. Table 1 summarizes the results. For uniform weighting, using additional concepts beyond five improves recall, but hurts precision. There is no significant difference between uniform and relevance model concept weighting for five concepts, but for ten the weighting outperforms uniform.

	MAP	MDFA04	P05	P10
uniform-05c	0.149	0.607	0.520	0.513
uniform-10c	0.140	0.578	0.540	0.507
rm-05c	0.148	0.603	0.507	0.497
rm-10c	0.157	0.566	0.553	0.533

Table 1: Concept-based (only) video retrieval results. The query term weighting and the number of concepts used is varied.

Feature	Description
Score	minmax normalized score for OCR, ASR, and concepts
Rec. Rank	inverse of the rank for OCR, ASR, and concepts
CombSum	sum of normalized scores of docs
RRSum	sum of inverse ranks of docs
Matches @ K	number of occurrences in the results for OCR, ASR, and concept @ rank cutoff K for 5, 10, 20, 100, 4000

Table 2: Fusion features

	MAP	MDFA04	P05	P10
ASR	0.035	0.865	0.327	0.273
OCR	0.061	0.884	0.580	0.537
Concepts	0.157	0.566	0.553	0.533
Fusion	0.194	0.463	0.673	0.607

Table 3: OCR, ASR, and Concept retrieval and Fusion.

When compared with the OCR and ASR methods in Table 3, we observe that concept based retrieval has significantly higher recall as evidence by the higher MAP and lower MDFA04 scores. These are statistically significant over the ASR/OCR with p<0.05. Concept based retrieval is superior to ASR in P@5 and P@10 and is comparable to OCR.

5.1 Fusing text and concepts

The approach to concept representation described in this poster is successful, but we hypothesize that combining the different modalities may provide the precision benefits of OCR/ASR with the recall of concept-based retrieval. We now explore this direction using rank fusion. We generate features for each run and leverage the RankLib [2] package to learn a LambdaMART-based fusion model. We use a total of 13 features from the three modalities, described in Table 2. The features are pre-processed by shifting and scaling the exponentiated retrieval scores using min-max [0..1] transformation.

The result of fusion is shown in the last row of Table 3. As expected, fusion provides gains at high precision, 24% improvement in MAP over the single best run (Concepts).

5.2 Fusion & Relevance Feedback

As a final step we experiment with combining the output of fusion with relevance feedback. We test both true relevance feedback (rf) and psuedo-relevance feedback (prf) where the top results are

[2] `http://cs.umass.edu/vdang/ranklib.html`

	MAP	MDFA04	P05	P10
OCR output				
text-sdm	0.061	0.884	0.580	0.537
prf	0.071	0.855	0.867	0.640
rf10	0.081	0.852	0.900	0.717
Concepts				
text-sdm-rm-10c	0.157	0.566	0.553	0.533
prf	0.206	0.407	0.587	0.580
rf10	0.261	0.303	0.693	0.693
Feedback-Fusion				
fusion	0.194	0.463	0.670	0.607
fusion-prf	0.242	0.395	0.720	0.645
fusion-rf10	0.300	0.287	0.867	0.823

Table 4: Relevance feedback results for text, concepts and fusion modalities. (ASR is not included because no results are significantly different from baseline in Table 3.)

assumed to be relevant. Because we have three modalities, we create three expansion queries; one for each modality. We use the RM3 relevance model [7] for feedback.

Table 4 summarizes the results of these experiments. We do not include the ASR output in the table because feedback models do not have any significant differences with the baseline retrieval. For OCR, 50 expansion terms are used with $\lambda = 0.4$ (the interpolation parameter in RM3). Both feedback models provide significant improvements over the baseline on all metrics.

We now discuss the concept-based retrieval expansion results. For pseudo-relevance feedback $\lambda = 0.25$ with 10 feedback documents and 20 feedback terms is optimal on all training folds. The results show that PRF significantly improves recall over the baseline retrieval with significant improvements in MAP and MDA04. All results except PRF P5 and P10 are statistically significant over the text-sdm-rm-10c baseline. For true relevance feedback runs we find that $\lambda = 0.1$ with 60 feedback terms is optimal on all training folds. All of the RF runs significantly improve over the concept retrieval baseline. The success of relevance feedback for concepts indicates that there remains a significant vocabulary mismatch between concepts in the query and those in relevant videos.

For example, here are the ten highest weighted expansion concepts for event 13, Parkour: *outdoor*, *vehicle*, *building*, *walking running*, *suburban*, *vegetation*, *windows*, *event*, *animal*, *quadraped*. These concepts are all important for the setting of parkour. The expansion terms indicate some semantic mismatch, with people being recognized as *animals*, and as *quadrapeds* when they are climbing.

The Feedback-Fusion section of Table 4 shows the culmination of this process: rank fusion of the relevance feedback runs across the three modalities. Measured by MAP, PRF shows a 25% improvement in MAP and true relevance feedback shows a remarkable 55% gain. In P@10, the respective gains are 6% and 36%.

6. CONCLUSION

In this work we explored zero-shot retrieval using only text queries for retrieving videos leveraging both text and high-level semantic video concepts. We introduced a technique for modeling the textual representation of visual concepts using web search. Retrieval is used to identify and weight relevant concepts for a text query. We demonstrate that automatically selected concept queries can achieve high recall and reasonable precision. When combined with errorful ASR and OCR modalities using fusion the result is both high precision and recall that is greater than any individual modality. For future work, we plan to explore other alternatives for textual representations of visual concepts. Additionally, we plan to incorporate concept detection accuracy and discriminativeness into the concept selection model.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and In part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

8. REFERENCES

[1] H. Cheng, S. A. Jingen, L., O. Javed, Q. Yu, A. Tamrakar, A. Dvakaran, H. Sawhney, R. Manmatha, J. Allan, A. Hauptmann, M. Shah, S. Bhattacharya, A. Dehghan, G. Friedland, B. Elizalde, T. Darrell, M. Witbrock, and J. Curtis. SRI-Sarnoff AURORA System at TRECVID 2012. In *Proc. TRECVID 2012*, 2012.

[2] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu. ClassView: hierarchical video shot classification, indexing, and accessing. *IEEE Trans. Multimedia*, 6(1):70–86, Feb. 2004.

[3] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–1002–II–1009 Vol.2. IEEE, June 2004.

[4] J. Fiscus and M. Michel. 2012 TRECVID Workshop: Multimedia Event Detection Task. Workshop slides at http://www-nlpir.nist.gov/projects/tvpubs/tv12.slides/tv12.med.slides.pdf.

[5] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A Survey on Visual Content-Based Video Indexing and Retrieval. *IEEE Trans. Systems, Man, and Cybernetics*, 41(6):797–819, Nov. 2011.

[6] L. Jiang and A. Hauptmann. Informedia Aurora @TREVID 2012 Semantic Indexing (SIN). In *Proc. TRECVID 2012*, 2012.

[7] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. SIGIR*, pages 120–127, New York, NY, USA, 2001. ACM.

[8] X. Li, D. Wang, J. Li, and B. Zhang. Video search in concept subspace: A text-like paradigm. In *In Proc. of CIVR*, 2007.

[9] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR*, pages 472–479, New York, NY, USA, 2005. ACM.

[10] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proc. Multimedia*, pages 991–1000, New York, NY, USA, 2007. ACM.

[11] S. Y. Neo, J. Zhao, M. Y. Kan, and T. S. Chua. Video retrieval using high level features: exploiting query matching and confidence-based weighting. In *Proc. Image and Video Retrieval*, pages 143–152, Berlin, Heidelberg, 2006. Springer-Verlag.

[12] S. Robertson and I. Soboroff. The TREC 2002 Filtering Track Report. In *Proceedings of TREC 2002*, pages 27–39, 2003.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. Workshop on MIR*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[14] A. F. Smeaton, P. Over, and R. Taban. The TREC 2001 Video Track Report. In *Proceedings of TREC 2001*, pages 52–60, Apr. 2002.

[15] J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proc. ICME*, volume 2, pages II–445–8 vol.2. IEEE, July 2003.

[16] C. G. M. Snoek and M. Worring. Concept-Based Video Retrieval. *Found. Trends Inf. Retr.*, 2(4):215–322, Apr. 2009.

[17] H. Soltau, F. Metze, C. Fugen, and A. Waibel. A one-pass decoder based on polymorphic linguistic context assignment. In *Workshop on Automatic Speech Recognition and Understanding*, pages 214–217. IEEE, 2001.

[18] M. Worring, C. G. M. Snoek, O. De Rooij, G. P. Nguyen, and A. W. M. Smeulders. The mediamill semantic video search engine.

[19] E. Younessian, T. Mitamura, and A. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *Proc. Multimedia Retrieval*, New York, NY, USA, 2012. ACM.