AUTOMATIC VIDEO ANNOTATION THROUGH SEARCH AND MINING

Emily Moxley †, *Tao Mei* ‡, *Xian-Sheng Hua* ‡, *Wei-Ying Ma* ‡, *B.S. Manjunath* †

† Vision Research Lab, University of California, Santa Barbara
‡ Microsoft Research Asia

ABSTRACT

Conventional approaches to video annotation predominantly focus on supervised identification of a limited set of concepts, while unsupervised annotation with infinite vocabulary remains unexplored. This work aims to exploit the overlap in content of news video to automatically annotate by mining similar videos that reinforce, filter, and improve the original annotations. The algorithm employs a two-step process of search followed by mining. Given a query video consisting of visual content and speech-recognized transcripts, similar videos are first ranked in a multimodal search. Then, the transcripts associated with these similar videos are mined to extract keywords for the query. We conducted extensive experiments over the TRECVID 2005 corpus and showed the superiority of the proposed approach to using only the mining process on the original video for annotation. This work represents the first attempt at unsupervised automatic video annotation leveraging overlapping video content.

Index Terms— video annotation, video search, data mining

1. INTRODUCTION

The large increase of video data demands effective organization for efficient user retrieval and browsing. Tagging, or annotation, enables text-based querying and content summarization. Although some practical video-oriented sites such as [1] have user-generated tagging, the annotations have not been quality-controlled. Therefore, the annotations are typically incomplete and noisy, containing many incorrect keywords as well as missing vital keywords. An automatic method, in the end, is required that provides both coverage and precision of video tags to allow use of these large databases.

Research on video annotation has proceeded along two dimensions: some use machine learning of visual and text features to perform *supervised* annotation, while others work in an *unsupervised* framework. Typical supervised methods developed through the TRECVID collaboration [2] use Support Vector Machines (SVM) to learn a pre-selected set of concepts. Recent efforts on supervised annotation have focused on correlative tagging, which exploits annotation co-occurrences, such as "mountain" and "outdoor," in the labeling process [3]. Lavrenko *et al.* [4] constructed a joint probability of visual region-based words with text annotations over a training set to annotate videos, incorporating co-occurrent visual features and co-occurrent annotations. Other techniques label a set of concepts using a training set for label propagation on graphs based on various visual features [5]. These supervised approaches, however, cannot learn new annotations. On the other hand, Velivelli *et al.* [6] used the Automatic Speech Recognition (ASR) results to mine a corpus for annotation. However, the mining step is limited to vocabulary creation for the entire database, rather than in creation of a more shot-specific vocabulary. It is observed that existing methods usually suffer from two problems: (1) *supervised* approaches do not address tag discovery, since previously unseen annotations lack training data. As a result, they are limited to a pre-defined concept set; (2) *unsupervised* approaches such as [6] fail to use only similar videos for annotating the query, resulting in annotations appropriate for the entire database rather than specific to a target video.

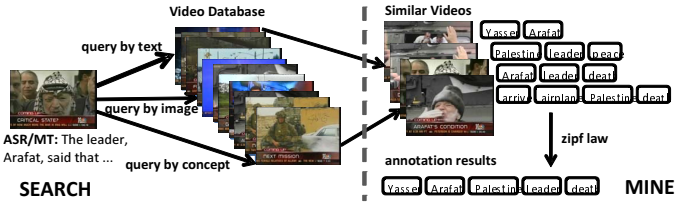

Fig. 1: System diagram. System first collects similar videos using various search algorithms. The ASR/MT transcript of the similar videos are used in a mining step to generate tag annotations for the query video.

Motivated by the work in [7] which discovered annotations for images from the words surrounding similar images on the Web, we propose to leverage search and mining techniques for video annotation with an unlimited vocabulary. The basic assumption is that similar or relevant videos share a common set of tags. However, video annotation by search and mining has significant differences to image annotation. First, video has an audio track which enables us to directly extract ASR or machine-translated (MT) transcripts without using any other information. Second, the transcript is very noisy which indicates direct annotation extracted from the original transcript is also noisy (see Figure 3 for an example). As a result, practical video repositories have exclusively manual annotations, in contrast to many image databases.

A diagram of the proposed video annotation system is shown in Figure 1. First, a database is searched using different modalities for a query video, and then the ASR of similar videos is mined to identify keywords for this query video. This approach addresses the shortcoming of limited vocabulary in previous methods since the process is not reduced to machine learning of certain annotations. Instead, a general unsupervised process is used where visual, text, and concept features are used to find similar videos, and then textual analysis used to mine ASR for annotations. Furthermore, the information used to annotate is not limited to information in the query video alone nor to general information from the entire collection as in [6]. This approach designates a certain group of similar videos for improving completeness and accuracy in the annotation process. It diverges from the work in [7], which uses search and mining techniques for image annotation, in medium (video), search techniques, mining techniques, and dataset qualities such as text cleanliness and

This work was performed when the first author was visiting Microsoft Research Asia.

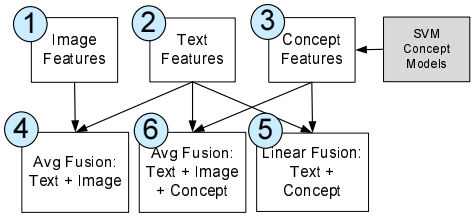

Fig. 2: Diagram of search modalities.

the size of the dataset.

In the remainder of this paper, an explanation of the search and mining algorithm can be found in Sections 2 and 3, followed by experiments and conclusions in Sections 4 and 5, respectively.

2. SEARCH FOR SIMILAR VIDEOS

The goal of the search step is to find videos with content similar to that in the query, such that the words associated with the search results inform the original video. Video search is a well-studied topic, and the performance of the annotation will advance as video search improves. Most existing video search systems rely on some combination of transcript, keyframe, and concept detection similarities.

This paper proves that using current search algorithms, mining of the search results can yield information useful for the original. Intuitively, it would seem the algorithm is robust to significant noise as irrelevant search results will all be different while relevant results will share commonalities that are extracted in the mining step. However, experiments presented in this paper will reveal that this expectation is true only to a limited extent, as the irrelevant search results are not random noise but are correlated in some way.

The search performed in this paper has several different modalities, based on image, text, concepts, and combinations of those three individual modalities, as shown in Figure 2.

1. Image alone, where global image features rank shots. This modality is also called query by example, or QBE.
2. Text alone, where ASR/MT transcripts rank shots.
3. Concepts alone, where scores from SVM models for 39 TRECVID concepts rank shots.
4. Average fusion of text and image modalities.
5. Linear fusion of text and concept modalities.
6. Average fusion of text, image, and concept modalities.

Further description of them can be found in [8]. The 39 general concepts consisted of the TRECVID concept set and were learned from the development set of the 2005 TRECVID data by SVM.

3. MINING FOR VIDEO ANNOTATIONS

The ASR/MT transcripts of the similar videos can be mined as documents for annotations. The noise and errors in current ASR/MT technology makes key*phrase* extraction impossible, since nearly any relevant phrase has an error in at least one of the words. After stemming and stop-list application [9], a term frequency vector is created for each video clip representing the number of times each term is spoken in the clip.

Given a ranked list from a specific search modality, a similar set $\mathcal{T}$ is first extracted to supplement the query video's transcript. The cutoff for this set was determined heuristically but was applied uniformly for all search rankings. Namely, videos were only considered sufficiently similar for inclusion if they were in the top 50% of the range of the top 100 results. Consider a video i with similarity score S_i to the query. The indicator function for inclusion in set $\mathcal{T}$ is:

$$I_i = \begin{cases} 1 & \text{if } S_i \geq m, \\ 0 & \text{if } S_i < m \end{cases} \tag{1}$$

where

$$m = S_{rank=100} + \alpha \times (S_{rank=1} - S_{rank=100}) \tag{2}$$

and $\alpha = 0.5$ for initial experiments. The resulting term frequency vector f_q, a vector populated by the frequency of each word in shot q's transcript, is formulated as a weighted combination of the videos in similar set $\mathcal{T}$:

$$f_q = \sum_{i \in \mathcal{T}} w_i f_i \tag{3}$$

Two methods of formulating w_i, for weighting transcripts from $\mathcal{T}$, are tested.

1. Weight similar video i equally with the original clip q, $w_i = 1 \; \forall \; i \in \mathcal{T}$.
2. Weight original query q with $w_q = 1$, and the similar clips proportional to it's similarity to q. Specifically,

$$w_i = \begin{cases} 1 & i = q, \\ \frac{S_i}{\sum_{i \in \mathcal{T}} S_i} & i \neq q \end{cases} \tag{4}$$

Having found a term frequency vector that incorporates the similar set, f_q, a zipf curve is fit to f_q, sorted in order of decreasing frequency, by finding the best-fit shape parameter as in [9]. The zipf curve models a typical distribution of word frequency in language. By finding the best-fit zipf curve, we are able to determine an appropriate cutoff for the most important words without assuming that a set of keywords have some minimum frequency across all videos. The most frequent terms are kept as keywords, initially those more frequent than the theoretical fifth-ranked word in the best-fit zipf curve.

The use of similar videos "corrects" the errors made in ASR of the video, allowing discovery of new keywords not in the transcript and suppressing errors in the speech recognition for the query video. Combining the term-frequency vectors, either in a weighted or unweighted fashion, of similar videos with the original creates a new f vector leading in more accurate, more complete annotations.

4. EXPERIMENTS

4.1. Data

We conducted experiments over TRECVID 2005 corpus [2], which consists of 169 hours of news video (137 development videos and 140 evaluation videos) in three languages (English, Arabic, and Chinese). Shot is adopted as the basic unit for video annotation in order to provide enough videos for adequate search and mining. In total, there are 89,673 clips in the database. Throughout this section, reference to a "video" indicates one of these 89,673 clips. 112 shots were chosen for querying based on the belief that clip content overlaps with that in other videos. These test shots were selected to be representative of the general content of the database, including commercials, international and domestic news stories. 64 test shots were in English, 25 in Chinese, and 23 in Arabic. The overlapping video in the database was not necessarily expected to be in the same language. Although the video database used in this experiment is far

	shot TRECVID2005_16_17	shot TRECVID2005_16_79	shot TRECVID2005_129_91
Keyframe			
ASR/MT	"He also pointed out that new system is other nuclear powers no problems and make efforts to make up for world women today no revealed that he is referring to what system 6 in the future of Russia is facing what reply it was learned that Russia is developing a new generation of missiles and heavy and those in as many as 10 A nuclear warhead"	"Audience sanctions against those who are ill-health and the Iraqi people to help him again in the sand barrier and monuments have been seriously damaged the Tiger Team Hassan this motion to the husband who kidnapped prosperity if it is true that Hassan prawn expressed the hope that those who kidnapped return Hassan buried the significance if it is indeed true information he will be the first kidnapped in Iraq of those killed foreign women"	"teacher suspect, most of those are telling"
Mining without Search	**nuclear**, **russia**, **system**, develop, effort, face, futur, gener, heavi, learn, **missil**, point, power, problem, refer, repli reveal, todai, **warhead**, women	**hassan**, **kidnap**, true	**suspect**, *teacher*, tell
with Search and Mining	**russia**, **system**, face, **nuclear**, **putin**	**hassan**, **kidnap**, **iraq**, **husband**	host, **leader**, *ozark*, **suspect**, *teacher*, tell, **terrorist**

Fig. 3: Example shots, keyframes, ASR, and tags that are extracted. Bold tags were judged "relevant," italics "incorrect," and all others "irrelevant." In shot 16_17, the de-noising of the approach is evident, while in shot 16_79, the expansion of the approach is evident.

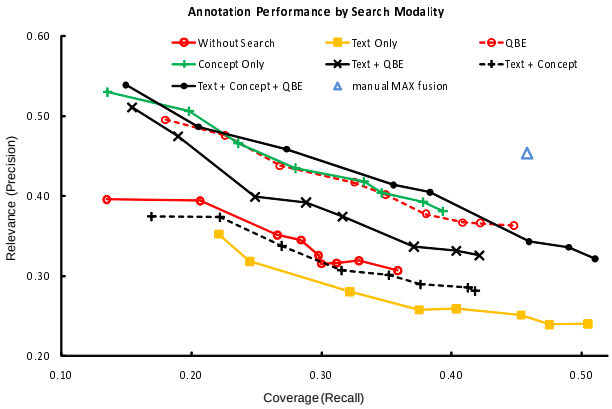

Fig. 4: Search modality comparison. Querying by image and by concept greatly outperforms text search, as text search re-emphasizes the noisy transcript. MAX fusion selects the best performing individual metric for a shot, and outperforms the automatic fusion modalities.

smaller than the image database used for search-based annotation of images in [7], the significant overlap in video content [10] allows search-based annotation to be effective on such a small data set. It is notable that exact duplicates in the dataset are not particularly useful, as the associated ASR/MT text will be identical to the original.

4.2. Performance Metric: Relevance-Coverage

A standard precision-recall metric does not accurately reflect the performance of annotation, as annotations do not fit neatly into a true/false categorization. Rather, they fit into a range between "relevant" to "irrelevant," as well as "incorrect." Therefore, we adopted an evaluation metric which provides a tag i with a score, c_i, of $+1$ for a relevant tag, 0 for an irrelevant tag, and -1 for an incorrect tag. "Relevance" and "irrelevance" were judged based on whether a typical user would use that word in a query seeking that video. Thus, a modified *precision* metric, called *relevance* is used and formulated as the average score of the N extracted tags, $P = \frac{1}{N}\sum_{i=1}^{N} c_i$. A similar three-class scoring method has been adopted for image annotation [7], though it assigns a score of 0.5 to uninformative tags.

Additionally, the set of appropriate tags is not limited, and therefore a standard recall metric cannot be used. Instead, a running list is kept of all "relevant" annotations for a video encountered using any search modality and any mining modality incorporated in this paper. Then, we adopted a recall-like metric, called *coverage*, that indicates the percentage of all seen positive annotations $\mathcal{A}$ covered by the method: $R = \frac{|\mathcal{S}\cap\mathcal{A}|}{|\mathcal{A}|}$, where $\mathcal{S}$ is the set of tags extracted using the particular method. The best metric has the greatest area under the relevance/coverage curve, exhibiting high precision without expending coverage. An example set of tags from mining without supplementing through search (using only the initial transcript for annotation, and adopted as a baseline for comparison), mining after supplementing with similar shots from the search list, and the associated ASR can be examined in Figure 3.

4.3. Evaluation on Search Modality

In this analysis we seek to see the performance of different search modalities for the annotation task. Each search modality returns a ranked list of shots for a query based on similarity in that particular mode (image/QBE, text, concepts, or some fusion of them). Intuitively, it seems that image-based querying should perform the best, since text querying returns shots with similar transcripts, and therefore just re-emphasizes the original text. Concept querying is expected to work reasonably as well, since concept-querying uses a 36-dimensional vector that is derived from image features only. Figure 4 shows the six different modalities as compared to annotation without search supplementation, using only the original transcript.

Most notable in Figure 4 is the great improvement using simple QBE search, which resulted in precision improvement of 20% to 25%. Concept and fusion of QBE, text, and concept performed similarly well. As expected, text alone performs quite poorly and is actually worse than the annotation without search since a search using a noisy, incorrect transcript will introduce further errors. Additionally, text fusion with either QBE or concepts greatly decreases annotation performance as compared to either QBE or concept modality search alone. Results are on the order of the improvement found for an image annotation algorithm that has a similar conception [7] despite differences in the annotation target and algorithm.

4.4. Evaluation on Weighting and Similar Set Size

An attempt was made to analyze the sensitivity to the inclusion restraint α, described in equation (2). Rather than keeping those images in the closest 50% of the range of the top 100 results, performance using the top 60%, 70%, and 75% ($\alpha = .4, .3,$ and $.25$, respectively) were also measured. Results are shown in Figure 5. The performance is best when using the top 60% or 50% of the focus range. This result is likely due to the weighting scheme that does not decrease quickly enough from the first return to the last included return. This would result in irrelevant returns at the end of the inclusion list being given a weight that is not significantly different from the (relevant) top return. With irrelevant shots having similar weight to relevant ones, the annotations become muddy, especially because typically irrelevant returns are not random but are from another common story and have common language elements. However, no matter the inclusion rate, all runs outperform annotations without search. In addition, we can observe that in Figure 5, weighting the image shots by their similarity to the query always outperforms equal weighting. Unweighted inclusion of similar shots results in a decrease in performance as compared to without-search annotation.

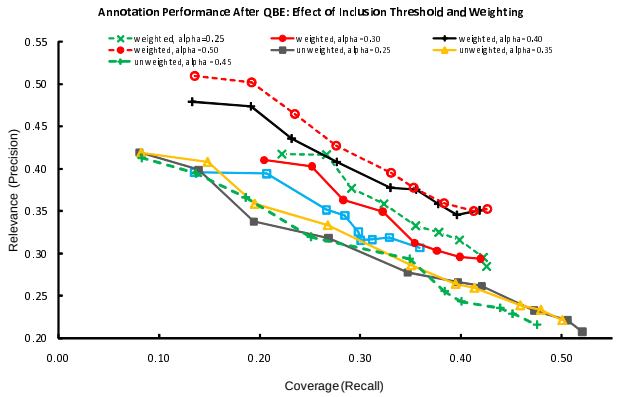

Fig. 5: Performance by inclusion constraint and weighting scheme. Individual runs vary the α parameter in equation (2). Best performances at high α, with fewer included shots. Better performance using weighted similar video inclusion compared to unweighted.

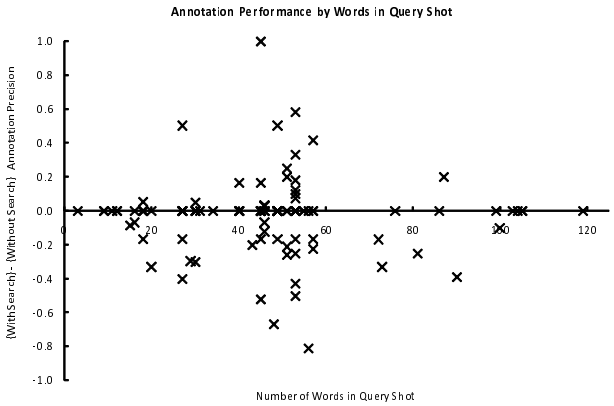

Fig. 6: Difference between annotation extended with text search and without search, using only original transcript, by number of words in query clip. Random scattering shows that successfulness of text-based search does not rely solely on number of words.

4.5. Evaluation on Fusion

If an effective fusion model can be found for the different fusion modalities, it is clear that performance would improve. In fact, a max fusion of the three individual modalities, where after scoring the annotations from each method, the best performing individual method for each query is manually chosen, gives performance well beyond the range of any of the automatic fusion search modalities used. Figure 4 shows a data point using this upper-bound manual fusion of the three different individual modalities.

An attempt was made to exploit the effect of the number of words in query video ASR. Intuitively, it is expected that text search would work best only when there were a large number of words that give a rich textual description of the story. However, Figure 6 shows that the annotation performance change using only the original video ("without" search) compared to that after text search does not seem to be correlated with the number of words in the shot.

It is worth noticing that the performance of annotating commercial shots is rather poor, as the text of commercials did not focus on the particular product. In general, the product was mentioned only briefly and therefore the annotations from transcript are poor. Therefore, commercials perhaps are better annotated using optical character recognition (OCR) or other computer vision techniques that detect logos or product names. In general, only a few shots were needed for improved performance using the described annotation technique.

4.6. Summary

In summary, the best method for extending tags is to weight text of similar videos by its similarity with the original. The search techniques that rely on visual similarity and concept similarity result in the best annotations using this method. Incorporating text similarity in search tends to hinder results. It is best to limit the inclusion set for annotation mining to a smaller group of very similar shots.

5. CONCLUSION AND REMARKS

In this paper, we have presented a novel approach to automatic video annotation by leveraging search and mining techniques. The proposed approach is fully unsupervised and not restricted to any predefined vocabulary. The experiments have proved that search and mining is a robust approach to improving video annotation. To the best of our knowledge, this work represents the first attempt at unsupervised video annotation leveraging overlapping video content. In the future, a more theoretic model will be built for combining the search ranked lists. Additionally, coverage analysis reveals that a robust fusion of the different modalities will produce a single model that effectively annotates videos without relying on analysis of the individual search modalities. A fused theoretic model of search and mining for video annotation shows promise in improving video tagging that allows effective use of video repositories.

6. REFERENCES

[1] YouTube, "http://www.youtube.com/."

[2] TRECVID, "http://www-nlpir.nist.gov/projects/trecvid/."

[3] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *ACM Multimedia*, Augsburg, Germany, September 2007.

[4] V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *Proceedings of ICASSP*, 2004.

[5] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: towards a unified video annotation scheme," in *Proceedings of ACM Multimedia*, 2007.

[6] A. Velivelli and T. S. Huang, "Automatic video annotation by mining speech transcripts," in *Proceedings of CVPRW*, 2006.

[7] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "Annosearch: Image auto-annotation by search," in *Proceedings of CVPR*, 2006.

[8] T. Mei, X.-S. Hua, W. Lai, L. Yang, and et al, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[9] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[10] X. Wu, A. Hauptmann, and C.-W. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," in *Proceedings of ACM Multimedia*, 2007.