Content-based Video Indexing and Retrieval Using Corr-LDA

Rahul Radhakrishnan Iyer[1], **Sanjeel Parekh**[2], **Vikas Mohandoss**[3], **Anush Ramsurat**[4], **Bhiksha Raj**[1] **and Rita Singh**[1]

[1] Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Technicolor, Cesson-Sévigné, France
[3] Information Technology, NITK Surathkal, Mangalore, KA, India
[4] Computer Science and Engineering, NIT Tiruchirappalli, Tiruchirappalli, TN, India

Abstract

Existing video indexing and retrieval methods on popular web-based multimedia sharing websites are based on user-provided sparse tagging. This paper proposes a very specific way of searching for video clips, based on the content of the video. We present our work on Content-based Video Indexing and Retrieval using the Correspondence-Latent Dirichlet Allocation (corr-LDA) probabilistic framework. This is a model that provides for auto-annotation of videos in a database with textual descriptors, and brings the added benefit of utilizing the semantic relations between the content of the video and text. We use the concept-level matching provided by corr-LDA to build correspondences between text and multimedia, with the objective of retrieving content with increased accuracy. In our experiments, we employ only the audio components of the individual recordings and compare our results with an SVM-based approach.

1 Introduction

Multimedia data typically entails digital images, audio, video, animation and graphics together with text data. Multimedia content on the internet is growing at an exponential rate. It is for this reason that indexing and retrieval of multimedia has become a pertinent issue and a hot topic of current research. Video indexing and retrieval have a wide spectrum of promising applications, motivating the interest of researchers worldwide. Our work proposes a novel solution to the problem of video indexing and retrieval. It is motivated by the premise that current multimedia indexing and retrieval techniques, which are largely based on sparse-tagging, often lead to erroneous and completely unwarranted results in terms of accuracy of video clippings returned to the user based on the search query. In order to provide better results, we base our video search paradigm on the actual content of the recordings.

Related Work

Video query by semantic keywords is one of the most difficult problems in multimedia data retrieval. The difficulty lies in the mapping between low-level video representation and high-level semantics. In (Naphide and Huang 2001), the multimedia content-access problem is formulated as a multimedia pattern recognition problem. A probabilistic framework is proposed to map the low-level video representation to high level semantics using the concepts of *multijects* and their networks called *multinets*.

A widely used structure for a stream of video data is to store them as contiguous groups of frames called segments. Each segment comprises of collections of frames. Usually, video comparisons compare each frame to find the similarity between two video streams. Research conducted in this domain has mostly focussed on structural analysis of video, which includes methods such as shot boundary detection, key frame extraction and scene segmentation, extraction of features including static key frame features, object features andmotion features, video datamining, video annotation, video retrieval including query interfaces, similarity measure and relevance feedback, and video browsing.

In recent years research has focused on the use of internal features of images and videos computed in an automated or semi-automated way (Fablet, Bouthemy, and Prez 2000). Automated analysis calculates statistics, which can be approximately correlated to the content features. This is useful as it provides information without costly human interaction.

The common strategy for automatic indexing had been based on using syntactic features alone. However, due to its complexity of operation, there has been a paradigm shift in the research concerned with identifying semantic features (Fan et al. 2004). User-friendly Content-Based Retrieval (CBR) systems operating at semantic level would identify motion-features as the key besides other features like color, objects etc., because motion (either of camera motion or shot editing) adds to the meaning of the content. The focus of present motion based systems had been mainly in identifying the principal object and performing retrieval-based on cues derived from such motion. With the objective of deriving semantic-level indices, it becomes important to deal with the learning tools. The learning phases followed by the classification phase are two common envisioned steps in CBR systems. Rather than the user mapping the features with semantic categories, the task could be shifted to the system to perform learning (or training) with pre-classified samples and determine the patterns in an effective manner. A concise review of these techniques is provided in (Geetha and Narayanan 2008;

Hu et al. 2011).

In the past several researchers have considered the problem of building semantic relations or correspondences for modeling annotated data. We mention here several important contributions to this area. In (Barnard et al. 2003) authors investigate the problem of auto-annotation and region naming for images using Mixture of Multi-Modal Latent Dirichlet Allocation (MoM-LDA) and Multi-Modal Hierarchical Aspect Models. Use of Canonical correlation analysis (CCA) for joint modeling is proposed in (Rasiwasia et al. 2010) and (Hodosh, Young, and Hockenmaier 2013). The idea is to map both images and text into a common space and then define various similarity metrics. This can be used for both indexing and retrieval. Kernel CCA is used by (Hodosh, Young, and Hockenmaier 2013) where the problem is formulated as that of maximizing an affinity function for image-sentence pairs. Rank SVM based approach is proposed in (Hodosh and Hockenmaier 2013) where linear classifiers are trained for an image with relevant and irrelavant captions. Many other novel methods have been proposed for establishing relations between images and text (Carneiro et al. 2007; Farhadi et al. 2010; Gupta and Davis 2008; Datta, Li, and Wang 2005). The Labeled LDA model is introduced by (Ramage et al. 2009) to solve the problem of credit attribution by learning word-tag correspondences.

We propose to use an extension of LDA (Blei, Ng, and Jordan 2003) called the Correspondence-LDA, familiarized by Blei and Jordan (Blei and Jordan 2003) to bridge the gap between low level video representation and high level semantics comprehensively. Our approach is significantly different from the discussed approaches because we model the problem in the probabilistic framework where both captions and videos are said to be generated from a set of topics. Moreover, we use a bag-of-words representation for both video content and text. Particularly, we differ from Blei's usage of Corr-LDA for image annotation and retrieval in the following two aspects:

1. Blei et al. segment an image into different regions and the feature vectors for each of those regions are computed. We do not perform any such segmentation for videos

2. We use the bag-of-words representation for feature vectors. As a consequence, we use multinomial distribution where as Blei et al. assume that visual words are sampled from a multivariate gaussian distribution

Paper Organization

The paper is organized as follows. We formulate the problem and explain the technical approach in sections (2) and (3) respectively. The experimental results obtained for a set of videos and their implications are discussed in section (4). We present a few limitations of our work in section (5) and finally draw conclusions and explore possibilities of future work in the last section (6) of the paper.

2 Problem Formulation

We assume a collection of videos, Ω, a fixed-size textual dictionary $D^{Dict} = [D_1, D_2, \ldots, D_n]$ and a sensory word dictionary $S^{Dict} = [s_1, s_2, \ldots, s_m]$. The textual dictionary is built from a list of all possible words for search queries, indexed alphabetically. An annotated video $V_i \in \Omega$, then, has a dual representation in terms of both the sensory words $F^i = [f_1, f_2, \ldots, f_m]$, which are individualized components of the multimedia file under consideration, and textual descriptors $D^i = [D_{i_1}, D_{i_2}, \ldots, D_{i_k}]$, where F^i represents the frequency of occurrence of the sensory words in the video V_i i.e. f_j denotes the frequency of occurrence of s_j in the video V_i considered, and D^i represents the annotation of the video V_i, where $D_{i_j} \in D^{Dict}$, for $j = 1, 2, \ldots, k$. The relevant parameters required for the tasks mentioned below are estimated from the model. The tasks can be clearly defined as :

- *Video Retrieval*: This refers to the task of retrieving videos relevant to a text search query. In formal terms, for a query q, we determine $P(q|V) \ \forall \ V \in \Omega$; this represents the probability of the query being associated with a given video. We rank these set of videos based on the obtained set of probabilities and retrieve the ones above a set threshold.

- *Video Indexing*: This task refers to annotating a video that is without captions. Formally, given an unannotated video $V^* \notin \Omega$, we estimate $P(D_i|V^*) \ \forall \ D_i \in D^{Dict}$, after which we rank the textual descriptors in decreasing order of probabilities and return the ones above a threshold, which best describe the video in consideration.

3 Technical Approach

In this paper, we approach the problem in the *Correspondence Latent Dirichlet Allocation* (Corr-LDA) framework. The main idea is to learn correspondences between multimedia (video in this case) and text. We later use the learned statistical correspondence to retrieve videos in response to text queries, by computing :

- $P(V, D)$: Required for clustering and organizing a large database of videos.

- $P(D|V)$: Required for automatic video annotation and text-based video retrieval.

The tool which facilitates in creating these correspondences is the Corr-LDA. In order to explain our approach first we introduce certain necessary terminology in the next section.

Preliminaries

A *corpus* is a collection of documents. A *document* is defined as a collection of words. The generative probabilistic framework used to model such a corpus is the Latent Dirichlet Allocation (LDA). It models documents as mixtures over latent topics, where each topic is represented by a distribution over words (a histogram).

Video files and textual descriptors are represented as *bags of words* (BoW) in our work. Just as a textual document consists of words, a multimedia document consists of
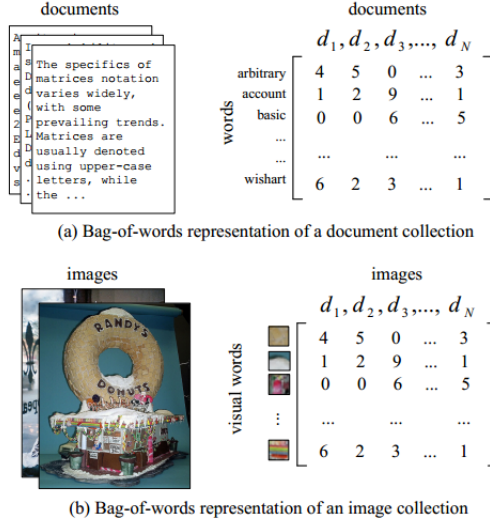
Figure 1: Understanding *Bag-of-words* representation for multimedia (Xiao 2011)

Figure 2: Graphical model representations: (**Left**) Original Corr-LDA (Blei and Jordan 2003). Nodes represent random variables; shaded nodes are observed random variables, unshaded nodes are latent random variables. Edges denote possible dependence between random variables; plates denote replicated structure. Note that the variables y_n are conditioned on M. (**Right**) Modified Corr-LDA. (**Bottom**) Variational distribution used to approximate the posterior in Corr-LDA. (Xiao 2011)

sensory words. This model is a simplifying assumption used in natural language processing and information retrieval wherein, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order. From a data modeling point of view, the bag-of-words model can be represented by a co-occurrence matrix of documents and words as illustrated.

Thus, both the multimedia in consideration and its textual description have BoW representation. We also represent the textual search query as a BoW.

The Corr-LDA is a generative probabilistic model that is an extension of LDA for modeling documents that consist of pairs of data streams. In particular, one data type can be viewed as an annotation of the other data type. Examples of such data sets include images and their associated captions, papers and their bibliographies etc. Like LDA, Corr-LDA can be viewed in terms of a probabilistic generative process that first generates the region descriptions and subsequently generates the annotation words. Here, we modify the original Corr-LDA model (Blei and Jordan 2003) by replacing the Gaussian distribution by multinomial distribution, which leads to a new generative process and parameter estimation algorithm. In particular, for images, it first generates region descriptions from an LDA model (with multinomial distribution in this case). Then, for each of the annotation words, one of the regions is selected from the video/image and a corresponding annotation word is drawn, conditioned on the factor that generated the selected region.

Under the Corr-LDA model, the regions of the image/video can be conditional on any ensemble of factors but the words of the caption must be conditional on factors, which are present in the video/image. Also, the correspondence implemented by Corr-LDA is not a one-to-one correspondence, but is more flexible: all caption words could
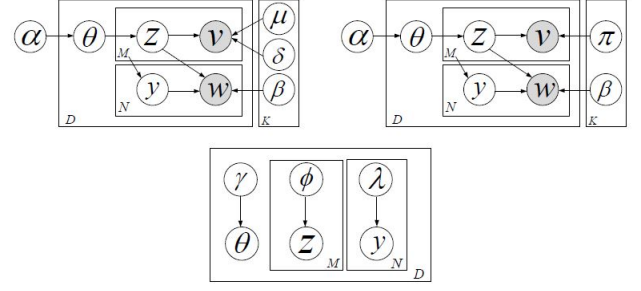
come from a subset of the video/image regions, and multiple caption words can come from the same region.

Corr-LDA can provide an excellent fit of the joint data as well as an effective conditional model of the caption given a video/image. Thus, it is widely used in automatic video/image annotation, automatic region annotation, and text-based video/image retrieval.

The problem at hand is learning correspondence information between videos and text. In other words, we need to estimate the probability of $p(w|V), p(V|w)$, where V^* represents the concerned video and w represents text. Using the modified version of the Corr-LDA model as described below, we apply it to model audio/text and videos/text. We fix the number of topics T in the procedure. Formally, the generative procedure can be expressed through the following sequence pictorially depicted in Fig. 2[1]:

1. Sample topic distribution $\theta \sim Dirichlet(\alpha)$

2. For each sensory word $v_m : m \in (1, \ldots, M)$
   - Draw topic assignment $z_m|\theta \sim Multinomial(\theta)$
   - Draw sensory word $v_m|z_m \sim Multinomial(\Pi_{z_m})$

3. For each textual word $w_n; n \in (1 \ldots N)$
   - Draw discrete indexing variable $y_n \sim Uni(1, \ldots, M)$
   - Draw textual word $w_n \sim Multinomial(\beta_{z_{y_n}})$

Note that in Fig. 2 (**Bottom**), the variational graphical model creates independency between z, θ and y, which enables us to factorize the joint probability.

Approach

Here, our corpus is a collection of videos, a document is a video and the annotations refer to the descriptors. Words

---

[1]We have acquired permission from Han Xiao for reproducing Fig. 1 and Fig. 2 from his master's thesis (Xiao 2011).

can refer to both *auditory words* as well as textual words (BoW). Auditory words are the sensory words obtained from the audio components of the recordings.

The model is trained on a training set of videos. Using the parameters estimated from the trained model, we perform the required actions on test inputs to output required results. Exact inference of the parameters of this model is intractable. So, we use an approximate inference algorithm called *Variation Expectation Maximization (EM)* to estimate the parameters. Due to space constraints, explanation of the inference algorithm has not been included. The reader is referred to (Xiao 2011) for details [2].

The parameters estimated from the learning phase to accomplish the tasks of video indexing, retrieval and calculation of perplexity are (refer to Fig. 2) :

1. θ : Represents the topic-document distribution
2. β : Represents the topic-word distribution

- *Video Retrieval*:

  Given a query consisting of n words $q = [q_1, q_2, \ldots, q_n]$, we compute a score for each video $V_i \in \Omega$ and rank them accordingly. First we remove all the *stop-words* (taken from NTLK) from the search query. Then the scores can be computed as follows:

  $$Score_i = \prod_n P(q_n|V_i) = \prod_n \sum_z P(z|V_i)P(q_n|z) \quad (1)$$

  Here z is the latent variable in the model, which corresponds to the notion of a *topic*. We pick a video V_i from our database Ω and compute the probability of each word of the search query being associated with that video. The total probability of the search query itself being associated with V_i is the product of individual probabilities calculated in (1). The probability of each individual word being associated with that video V_i can be calculated as follows. If we assume that topic z has been chosen, then the probability that the word q_j is associated with that topic is $P(q_j|z)$ (given by β). This multiplied with the probability of picking topic z when considering video V_i (given by θ) gives the probability that q_j comes from topic z when video V_i is considered. But the word q_j could have come from any topic. So, considering all possibilities, we take the sum of above probabilities $\forall z$. This is done for all words q_j in the search query q and their products are taken to compute the score for the ith video V_i. This gives the score for V_i. We calculate such scores $\forall V_i \in \Omega$ and rank the videos based on that.

- *Video Indexing*:

  Using the notation from the previous section, $P(D_j|V^*)$ provides us with relevant annotations for the video V^* (need not $\in \Omega$), $D_j \in D^{Dict}$, where D^{Dict} refers to the textual dictionary. We compute scores for each word D_j in the textual dictionary as follows :

  $$Score_j = P(D_j|V^*) = \sum_z P(z|V^*)P(D_j|z) \quad (2)$$

We calculate the score assuming that the word D_j is taken from the topic z and do it $\forall z$. Thus, $Score_j$ gives the score for the jth word in the textual dictionary being associated with V^*. This is done $\forall$ word D^j in the dictionary. Next, we rank the words in decreasing order of score and return the ones, which have scores above a set threshold. These words are the obtained annotation for the untagged video.

4 RESULTS

Experimental Results

Our experiments were conducted on the MED11 example-based multimedia-event retrieval task (Trecvid 2011). The task here is as follows: we are given a collection of multimedia recordings. We are also provided a small number of examples of recordings representing a category of events such as *repairing an appliance*, *parade*, *parkour*, or *grooming an animal*. The objective is to retrieve all other instances of the same category from the data set, based on what can be learned from the example recordings. In our experiment we only employed the audio components of the individual recordings to perform the retrieval; the visual features were not used. One of the reasons for this is to explore how well an audio can describe a recording. Furthermore, considering only the audio features instead of visual ones in our experiments speeds up execution, saving space and time, without compromising much on the quality. The method used for generating audio bag-of-words is as described in (Chaudhuri, Harvilla, and Raj 2011). Herein, an audio is represented as a set of smaller units called "Acoustic Unit descriptor" (AUD) and the BoW representation is computed over sequences of AUDs. We have used Han Xiao's Corr-LDA implementation for the experiments.

To compare our results for retrieval, we use SVM-based classifiers. We train two sets of SVM classifiers, each with five classifiers: one for each of the video categories, using 500 training videos from our dataset. In both the cases, we train the SVMs on visual features, considering 10 equally spaced frames per video. In one case, referred to as SVM-CTS, we extract 26 low level visual features pertaining to color, texture and structure for each frame (Sergyan 2008). In the second case, we extract dense SIFT features, 1536 per frame, with a step size of 250. This model is referred to as SVM-DSIFT.

The models, both Corr-LDA and SVMs, are trained on a set of 500 videos and tested on a set of 125 mixed category videos.

The following results were observed:

- *Video Retrieval* :

  – We compute precision @ 10, recall @ 10 and mean average precision @ 10 (MAP) for both the models: Corr-LDA trained with 5, 50 and 100 topics (referred to as Corr-LDA-5, Corr-LDA-50, Corr-LDA-100), and SVM-based classifiers. MAP is computed over the queries mentioned in Table 2.

  – As can be seen from Tables 2 and 3, considering each SVM category as an input query, Corr-LDA clearly

Category / Query	SVM-DSIFT		SVM-CTS		Corr-LDA-5		Corr-LDA-50		Corr-LDA-100	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Grooming Animal	0.600	0.220	0.500	0.185	0.300	0.111	0.600	0.222	0.700	0.259
Making Sandwich	0.300	0.125	0.300	0.125	0.600	0.250	0.600	0.250	0.300	0.125
Parade	0.200	0.080	0.600	0.240	0.300	0.120	0.800	0.320	1.000	0.400
Parkour	0.100	0.042	0.800	0.333	0.500	0.208	0.600	0.250	0.700	0.292
Repair Appliance	0.600	0.240	0.600	0.240	1.000	0.400	1.000	0.400	1.000	0.400

Table 2: Comparison of precision and recall at 10 for SVM Based Classifiers and Corr-LDA (for different topics) for Retrieval

	Topics		
Parameter	5	50	100
Number of Iterations	67	639	621
Convergence Threshold	10^{-7}	10^{-7}	10^{-7}
Dirichlet Prior α of Corr-LDA	0.1	0.2	0.2
Number of Topics	5	50	100
Number of Training Videos	500	500	500
Number of Test Videos	125	125	125
Number of Auditory Words	8193	8193	8193
Number of Textual Words	9654	9654	9654

Table 1: Experimental Parameters

Model	Mean Average Precision
SVM-DSIFT	0.121
SVM-CTS	0.160
Corr-LDA-5	0.151
Corr-LDA-50	0.231
Corr-LDA-100	0.243

Table 3: Comparison of MAP at 10 for different models

outperforms the SVM based frame-by-frame classification. We observe from Table 3 that the performance of Corr-LDA depends on the number of topics. SVM-CTS performs better when Corr-LDA is trained with just five topics.

– It is worth noting that although the SVMs used the features from the video itself, Corr-LDA having used just the audio components of the recordings, still outperformed.

– It is also observed that as the number of topics are increased, the accuracy of retrieval of videos based on the search query increases. This is particularly evident with multi-word queries.

– From Figure 3 we see that for two categories namely *parade* and *repair appliance* all of the first few results retrieved using Corr-LDA (100 topics) are relevant. However, the precision is poor for categories *grooming animal* and *making sandwich*. In this respect, retrieval is poorer for SVM-based systems.

– This shows that having a higher number of topics helps the Corr-LDA system to learn, differentiate, segregate videos better and hence create concept-level matching with increased accuracy.

• *Video Indexing* :
In our experiments, the quality of the obtained annotations are quantified in two different ways:

1. *Perplexity*, which is given by:

$$Perplexity = \exp \left(\frac{- \sum_{i=1}^{N} \sum_{m=1}^{M_i} ln\, P(d_m | V_i)}{\sum_{i=1}^{N} M_i} \right) \quad (3)$$
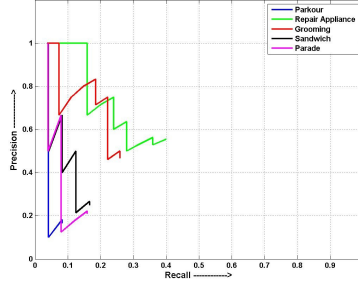
where N represents the number of test videos, M_i the number of words in the caption associated with V_i and d_m represents the mth word of the caption obtained. Note that higher $P(d_m|V_i)$ implies better annotation as there is a higher probability that the word d_m is associated with the video V_i. From the above formula, we can see that as $P(d_m|V_i)$ increases, the value of *perplexity* decreases which implies that annotation quality and perplexity are *inversely proportional*.
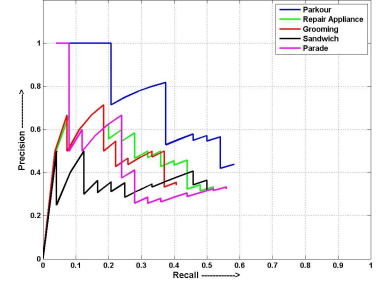
2. *Mean Per-word Precision*, *Mean Per-word Recall* and *F-Score* (Hoffman, Blei, and Cook 2009): Per-word recall is defined as the average fraction of videos actually labeled with w that our model annotates with label w. Per-word precision is defined as the average fraction of videos that our model annotates with label w that are actually labeled with w. F-score is the harmonic mean of precision and recall, and is one metric of overall annotation performance. For our experiments, we considered the top 10 annotations generated for each video. We had a total of 9654 textual words in our dictionary, and computed the per-word metrics for each of these, averaging the results to get the required mean metrics. In some cases, when our model did not annotate any video with the label w, we determined the precision for that word by Monte-carlo methods: Generate sets of random annotations for each of the videos and average over the computed precision. If no video in the test set is labeled with w, then per-word precision and
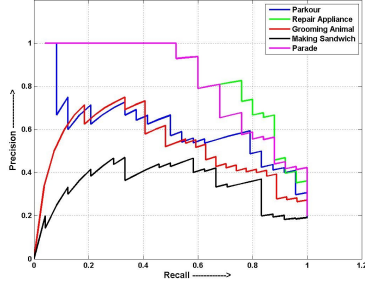
| (a) Corr-LDA with 100 topics | (b) SVM-DSIFT | (c) SVM-CTS |

Figure 3: Precision-Recall curves

Annotation Length	5 topics	50 topics	100 topics
5	33.1512	31.2510	30.5603
10	43.3720	41.6222	40.5960
15	53.0023	51.0888	49.9647
20	64.2810	61.5055	60.2332
25	75.9930	72.2725	71.1561
30	87.7312	83.1981	82.1523
35	99.4043	94.2035	93.3511
40	111.1342	105.4465	104.6538
45	122.8281	116.7421	115.8785
50	134.4591	128.0057	127.0357

Table 4: Video Indexing: Perplexity for different annotation lengths

Model	MPW Precision	MPW Recall	F-Score
Corr-LDA-5	0.02989	0.03066	0.03027
Corr-LDA-50	0.04153	0.03871	0.04005
Corr-LDA-100	0.05537	0.04844	0.05126

Table 5: Comparison of Mean Per-word (MPW) Precision, Mean Per-word Recall and F-Score of the Corr-LDA model with different number of topics

recall for w are undefined, so we ignored these words in our evaluation.

It is clear from Table 4 that the perplexity values reduce with an increase in the number of topics. It is worth recalling that lower values of perplexities imply better annotation quality. Thus, we see that as more and more of the annotations generated are considered, we get higher/poorer values of perplexity implying poorer annotation quality. It is also clear from Table 5 that the annotation quality improves with increase in the number of topics. Having obtained perplexity values comparable to (Blei and Jordan 2003) and from the precision-recall values and F-scores in Table 5, we can see that the Corr-LDA model success-

fully learns the conditional text-video distributions.

5 Limitations

We have identified the following limitations which we would like to work upon in our future investigations. Firstly, it is important to note that audio alone cannot distinguish a recording. Several dissimilar videos may contain similar audio words. This problem can be remedied by using visual features of the recordings i.e. video BoW. We expect an increase in the performance/accuracy of the model when visual features are taken into account. Secondly, the textual dictionary is built from the training data and hence, a search-query containing terms outside of it cannot be handled. Various smoothing techniques can be used for this purpose.

6 Conclusions and Future work

In this paper we have successfully applied the Corr-LDA model to handle the task of content-based video indexing and retrieval. Our experimental results show that the framework learns the correspondences well with the use of just the audio components of the recordings i.e. using the audio bag-of-words alone. Our method is not just restricted to video and text. Rather, it can be applied in a wide variety of scenarios which includes any form of multimedia, for e.g. audio retrieval using images, video retrieval using audio etc., along with corresponding annotations.

Our work provides an arguably promising new direction for research by building and making use of the inherent semantic linkages between text and video thereby obviating the need for frame-by-frame search. In the future, we would like to further our method by using the video bag-of-words for the stated task. Also, a combination of tag-based query-search and content-based semantic linkages can be used to gain a better performance, by incorporating the merits of both the methods.

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3:1107–1135.

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *26th International Conference on Research and Development in Information Retrieval (SIGIR)*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Carneiro, G.; Chan, A. B.; Moreno, P. J.; and Vasconcelos, N. 2007. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(3):394–410.

Chaudhuri, S.; Harvilla, M.; and Raj, B. 2011. Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In *Interspeech*, 2265–2268.

Datta, R.; Li, J.; and Wang, J. Z. 2005. Content-based image retrieval: Approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '05, 253–262. New York, NY, USA: ACM.

Fablet, R.; Bouthemy, P.; and Prez, P. 2000. Statistical motion-based video indexing and retrieval. In *in Int. Conf. on Content-Based Multimedia Info. Access*, 602–619.

Fan, J.; Elmagarmid, A. K.; Zhu, X.; Aref, W. G.; and Wu, L. 2004. Classview: hierarchical video shot classification, indexing, and accessing. *Multimedia, IEEE Transactions on* 6(1):70–86.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, 15–29. Berlin, Heidelberg: Springer-Verlag.

Geetha, P., and Narayanan, V. 2008. A survey of content-based video retrieval. *Journal of Computer Science* 4(6):474.

Gupta, A., and Davis, L. S. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, 16–29. Berlin, Heidelberg: Springer-Verlag.

Hodosh, M., and Hockenmaier, J. 2013. Sentence-based image description with scalable, explicit models. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 294–300.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.* 47(1):853–899.

Hoffman, M. D.; Blei, D. M.; and Cook, P. R. 2009. Easy as cba: A simple probabilistic model for tagging music. In *ISMIR*, volume 9, 369–374.

Hu, W.; Xie, N.; Li, L.; Zeng, X.; and Maybank, S. 2011. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 41(6):797–819.

Naphide, H., and Huang, T. S. 2001. A probabilistic framework for semantic video indexing, filtering, and retrieval. *Multimedia, IEEE Transactions on* 3(1):141–151.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, 248–256. Stroudsburg, PA, USA: Association for Computational Linguistics.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*, MM '10, 251–260. New York, NY, USA: ACM.

Sergyan, S. 2008. Color histogram features based image classification in content-based image retrieval systems. In *Applied Machine Intelligence and Informatics, 2008. SAMI 2008. 6th International Symposium on*, 221–224. IEEE.

Trecvid. 2011. Dataset, trecvid 2011, http://www.nist.gov/itl/iad/mig/med11. cfm, 2011.

Xiao, H. 2011. Toward artificial synesthesia: Linking images and sounds via words. Master's thesis, Technische Universität München.