Translating Related Words to Videos and Back through Latent Topics

Pradipto Das
CSE Dept., SUNY Buffalo
Buffalo, NY 14260
pdas3@buffalo.edu

Rohini K. Srihari
CSE Dept., SUNY Buffalo
Buffalo, NY 14260
rohini@cedar.buffalo.edu

Jason J. Corso
CSE Dept., SUNY Buffalo
Buffalo, NY 14260
jcorso@buffalo.edu

ABSTRACT

Documents containing video and text are becoming more and more widespread and yet content analysis of those documents depends primarily on the text. Although automated discovery of semantically related words from text improves free text query understanding, translating videos into text summaries facilitates better video search particularly in the absence of accompanying text. In this paper, we propose a multimedia topic modeling framework suitable for providing a basis for automatically discovering and translating semantically related words obtained from textual metadata of multimedia documents to semantically related videos or frames from videos. The framework jointly models video and text and is flexible enough to handle different types of document features in their constituent domains such as discrete and real valued features from videos representing actions, objects, colors and scenes as well as discrete features from text. Our proposed models show much better fit to the multimedia data in terms of held-out data log likelihoods. For a given query video, our models translate low level vision features into bag of keyword summaries which can be further translated using simple natural language generation techniques into human readable paragraphs. We quantitatively compare the results of video to bag of words translation against a state-of-the-art baseline object recognition model from computer vision. We show that text translations from multimodal topic models vastly outperform the baseline on a multimedia dataset downloaded from the Internet.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

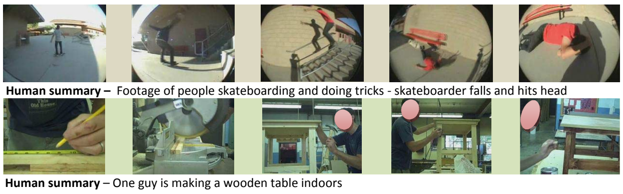
Algorithms, Experimentation

Keywords

multimedia topic models; video to text summarization

1. INTRODUCTION

In recent years there has been an abundance of multimedia data in the form of video contents from television networks, video uploads to websites and so on. However, organizing such data by integrating the semantic content of the videos is a very difficult problem. On the other hand, summarizing videos directly into text summaries can lead to significant improvements in many end-user applications—multimedia search experience, content based advertisements [33], helping the visually impaired and so on. In this paper we concern ourselves with two language agnostic tasks: **i)** Building a topic modeling framework to model multimedia documents consisting of videos and textual metadata and **ii)** use the topic modeling framework to predict bag-of-word summaries for a new video belonging to a previously known category. The first task helps us discover semantically related concepts in the text through latent topics and translating them to topically related videos or frames. The second task takes a video and generates intermediate text keywords ideal for natural language generation.



Human summary – Footage of people skateboarding and doing tricks - skateboarder falls and hits head

Human summary – One guy is making a wooden table indoors

Figure 1: An example of the task of video summarization

As a further addition we also experiment with efficient natural language sentence generation from the predicted bag of words (henceforth BoW) using language models and confidence of syntactic parse tree generation following a simple template. The first two tasks, though, are the focus of this paper. Fig. 1 shows some keyframes of two sample videos from our training dataset and the short summaries written by human annotators. This dataset is discussed in Section 1.1. *In our paper, translation from video to text is synonymous to summarizing a video with a set of textual keywords.*

It seems intuitive that a topic model which incorporates low level vision features representing objects, actions, color and scenes and correspond those to the text summaries should have a better chance of describing the multimedia data. We thus seek representations of visual data that mimic the subject-verb-object-scene quadruplet structure of English sentences in terms of subject, object and scene nouns as well as verbs. Additionally illumination gives rise to color which differentiates one object from another and is often expressed as adjectives. Objects, actions and color can be visualized

WSDM'13, February 4–8, 2012, Rome, Italy.
Copyright 2013 ACM 978-1-4503-1869-3/13/02 ...$15.00.

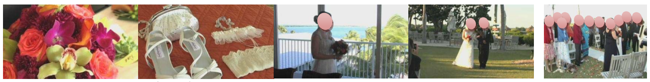
using specific word concepts and can be counted over time thus lending themselves to quantization but scene represents global energy distributions which pervade the arrangement of objects and is thus better represented as real values.

In the context of this paper, our training data consists of videos and associated summaries. The training data is also available with event category labels for e.g. "boarding event" and topic modeling video documents in each event can allow us to discover "sub-topics" e.g. "skateboarding", "snowboarding" and "surfing". Of course, our topic models do not include any event label bias and can be applied on the overall dataset as well. However we will see shortly that doing so can be very unappealing to end users when summaries need to be generated.

Apart from the topical analysis of video documents, the problem of generating summaries directly from videos also has significant end user appeal. We emphasize that the video summarization/translation task in this paper is to describe an entire video firstly as a bag of salient keywords i.e. BoW, and then, as a further addition, use simple Natural Language Generation (henceforth NLG) techniques to summarize the bag of words into a human readable paragraph of text wherever possible. Translating and generating summaries from a video can always be looked upon as finding the right information need which is paramount to any search problem.

Video summarization in our context is different from image annotation mentioned in early literature [27] and in more modern ones [18]. If we perfectly annotate all objects and actions in each frame correctly, we can probably use a standard topic modeling technique [5] to figure out the themes in the multimedia documents. However, detecting objects [15] and scenes [22] in images and actions [11] and keyframes [14] in videos in a reliable manner are open problems in the computer vision community [18]. Most of these detection techniques fall in the domain of supervised learning and require large amounts of annotated data. Video annotation task is a particularly laborious process even though tools like VATIC [29] have been written to ease the effort (for an interactive demo of the tool, see VATIC's website[1]). Also firing thousands of detectors to accurately label even a single keyframe of an unknown video leads to many false positives. This is particularly true of the videos *in-the-wild* i.e. videos downloaded from the Internet where there are plenty of resolution problems, severe motion blurs, camera shakes etc. These are the videos that we experiment with in this paper.

We view the video to text translation/summarization problem in the light of multidocument summarization of plain text documents which has been popularized by the Text Analysis Conference[2] (TAC). In the multidocument summarization track of TAC, participants are given document sets (docsets) of newswire articles typically belonging to 5 major event types like "Health and Safety," "Accidents and Natural Disasters" etc. and are asked to generate a fixed length fluent summary of the documents in each docset. A docset in the TAC setting is unique in that it contains a set of documents that are relevant for a particular information need like "Cyclone Katrina." The system summaries are scored in several ways including the most reliable manual way using PYRAMID [21] evaluation but systems usually are tuned w.r.t. the automatic ROUGE [16] scoring. By analogy, we assume that each docset here corresponds to a video and contains a sequence of frames and a set of keyframes. At test time we are given unknown event specific videos without any text summary. For measuring system performance, we generate summaries of videos and evaluate them using the recall oriented ROUGE-1 score to measure the percent overlap of the words in the short ground truth summaries.



Human Summary: Montage of clips from an outdoor wedding

Predicted bag of words summary: birthday wed indoor outdoors mob dance flash cake parade ceremony fish

Figure 2: An example of vocabulary intrusion in the task of video summarization. Best viewed with magnification

A key concern in generating a BoW summary of a video is the vocabulary intrusion problem. Fig. 2 shows an example of vocabulary intrusion in the task of video summarization that arises out of topic modeling on the entire vocabulary of the corpus. If we consider a vocabulary of V words—the probability of getting the top L words correctly in the summary is $(1/V)(1/(V-1))...(1/(V-L+1))$. If V is large (such as 2000) then the probability is very low. Further, if the entire vocabulary is used, then intrusive words describing other but related event categories like "birthday, flash mob, dance, parade, fish" can appear with high probability (see a possible predicted BoW summary in Fig. 2 from a topic model(Fig. 3b) trained over all events with number of topics set to 200). This problem is mitigated by first classifying the test video into its corresponding event category (Section 4.4) and then using a topic model to predict the BoW summary. In the absence of the event labels, this direction improves readability and is much faster.

The novelty in our new approach to topic modeling video documents with textual metadata is the use of the right features for the videos and augmenting basic topic models for joint modeling with those features along with text. We represent each video in terms of objects, actions, color (represented with discrete distributions) and scenes (represented with Normal distributions with unknown means and variances) and try to find a translation space that translates the pattern of these features to a permutation in language vocabulary. Such a representation of a video is both intuitive and logical. We observe that the interplay of the full spectrum of representations (Section 4) indeed yield the highest likelihoods to held out test data than those using partial representations (Section 4.1).

1.1 Dataset Description

The dataset that we use for the video summarization task is released as part of NIST's 2011 TRECVID Multimedia Event Detection (MED) evaluation set[3]. The dataset consists of a collection of Internet multimedia content posted to the various Internet video hosting sites. The training set is organized into 15 event categories, some of which are: *1) Attempting a board trick 2)Feeding an animal 3)Landing a fish 4)Wedding ceremony 5) Working on a woodworking project* etc.

We use the videos and their textual metadata in all the 15 events as training data. There are 2062 clips with summaries in the training set with almost equal distribution amongst the events. The test set which we use is called the Transparent Development (Dev-T) collection. The Dev-T collection includes positive instances of the first 5 training events and near positive instances for the last 10 events—a total of 630

[1] http://mit.edu/vondrick/vatic/

[2] http://www.nist.gov/tac/2011/Summarization/

[3] http://www.nist.gov/itl/iad/mig/med11.cfm

videos labeled with event category information (and associated human synopses which are to be compared against for summarization performance). Each summary is a short and very high level description of the entire video and ranges from 2 to 40 words but on average **10** words (with stopwords). We remove standard English stopwords and retain only the word morphologies (not required) from the synopses as our training vocabularies. The proportion of videos belonging to events 6 through 15 in the Dev-T set is much low compared to the proportion for the other events since those clips are considered to be "related" instances which cover only part of the event category specifications. The performances of our topic models are evaluated on those kinds of clips as well. The numbers of videos in events 6 through 15 in the Dev-T set are $\{4,9,5,7,8,3,3,3,10,8\}$ while there are around 120 videos per event for the first 5 events. All other videos in the Dev-T set neither have any event category label nor are identified as positive, negative or related videos and we do not consider these videos in our experiments.

1.2 Evaluation Measures

We measure the predictive performance of the topic models using the Evidence Lower BOunds (ELBO) on held-out test set—the Dev-T collection *with* summaries, as well as the predictive ELBO for BoW summary generation on the held-out Dev-T collection *without* summaries (Section 4.1). ELBO is just log likelihood and is directly related to measuring average perplexity of the model per observed textual word [5, 4]. We also evaluate our BoW summaries using the ROUGE scorer. ROUGE measures the n-gram overlap for system generated summaries to the ones written by annotators and the scores are interpreted in terms of recall. Usually 4 gold standard summaries are needed for evaluation but here we use the base case of using only one short summary as a reference summary per video on this dataset. While summarizing, since our primary task is to evaluate only the BoW summaries generated from a video, we use the ROUGE-1 unigram measure. We evaluate 5 and 10 keywords long BoW summaries respecting the average length of the short human summaries. Since we are considering videos in the Dev-T set with event category information, we can use the ROUGE evaluation setup of multidocument summarization as used in TAC. If the categories are not known, we can multiply the ROUGE scores with the event classification accuracies to obtain lower bounds (see Section 4.4 for lower bounds on classification accuracies). Evaluations with higher order n-grams are not needed for unigram translations. We do not use manual evaluations since the data cannot be released for public verifications.

The task of discovering topically related words is mostly evaluated w.r.t ELBO. We use the topic models from [4] as baselines. We modify the GM-LDA model in [4] following [26] to use discrete visual data and name the model MMLDA—"MM" stands for the multinomials for text as well as the multinomials for the visual words. We implement a deterministic optimization framework for MMLDA instead of the non-deterministic sampling as in [26]. The Corr-LDA model in [4] is also extended by using Normal-Wishart priors and named Corr-MGLDA (M for Multinomials and G for Gaussians). For evaluating video to text summarization based on ROUGE-1 scores, we use a non-topic model based automatic image annotation tool as the baseline for video labeling by using labels aggregated from keyframes. Our topic model based video summarization methods outperform the state-of-the-art image to text translation model [15] applied on video keyframes in terms of ROUGE-1 scores of the predicted keyword summaries.

2. RELATED WORK

Makadia et al. [18] uses nearest neighbor and label transfer techniques to annotate images suitable for the image retrieval task. However, we can not directly apply their methods as the individual frames/keyframes of the videos in our dataset are not annotated. Based on the size and genre of our dataset, such annotations prove very expensive and we do not follow that direction. Further, we are interested in the task of direct natural language summarization of the entire video and not specific annotation of a vast majority of possible objects, actions and scenes in every frame/keyframe of the video. The closest work to our task is by Yang et al. [34] where low level object and scene classifiers are used to obtain object and scene labels in an *image*. These are then combined using background language models and Hidden Markov Models to predict a natural language sentence that automatically includes the best possible verb i.e. action. *We will observe in Section 4.4 that **actions**, which are intrinsic to videos, are important event discriminators.* Further, none of above mentioned methods can discover related concepts as latent topics and translate them into related frames.

In the domain of topic modeling of images with captions, the Corr-LDA model has recently been extended to handle a multinomial feature space in [25] with different number of topics for visual word type and textual word type. The model learns an association from the topic proportions over image domain to those over text domain through a regression formulation. However, during prediction, this dependency needs to be marginalized out anyways. Also, if we quantize every type of real valued vision feature using some clustering algorithm such as K-means into C clusters, then each C represents a parameter of the final model and performance analysis become that much more difficult. Ahmed et al. [1] uses Gaussian feature vectors and mention Normal-Wishart priors but do not use them—they use uniform priors in a non-deterministic sampling framework instead.

On the other hand the Continuous Relevance Model (CRM) [13] and Multiple Bernoulli Relevance Model (MBRM) [8] assume different, nonparametric density representations of the joint word-image space. CRM gets rid of the latent factor representation and achieves non-parameterization. The dataset used in [8] for MBRM has hierarchical word annotations which are handled using multiple Bernoulli models rather than multinomial distributions. In our dataset, multinomial distributions are sufficient since the summaries read like very short documents with repeated word morphologies.

Detecting objects can often be seen as an important step towards identifying the main topic of a video and generating a BoW summary. To that end, Torresani et al. [28] transform an image feature vector into a another lower dimensional feature vector whose values are the outputs of several category classifiers (which are named "classemes" in their paper). We take a similar approach to convert Object Bank [15] (OB) feature vectors to high level sparse histogram of object detectors to be used in our discrete video data representation and as *baseline* for video to text translation. To extract OB features, keyframes are identified to reduce computational time. Keyframe detection is a research topic in its own right, where some recent ones include more involved techniques [14] using Transfer Learning from accompanying text transcripts. However, the keyframes extracted using the change in color histogram [35] satisfy our purposes.

In the domain of topic modeling of videos, the Hidden Topic Markov Model in [32] does not incorporate both text and visual words in a single framework and also does not use a fuller representation of videos as we do. A Markovian assumption is also imposed in [10] for modeling actions and identifying behaviors (with no automatic labeling), however, we can safely ignore frame dependence because our action features are derived using temporal windows and activity tracking is not an objective in this paper. The reformulations of LDA and CTM using class labels and without any temporal dynamics in [30] also target activity classification. To the best of our knowledge this is the first work to use mixed membership topic models for video to text summarization which *can* eliminate frame-wise object annotations.

The proposed models are discussed in the following section in as much depth as possible. The use of asymmetric Dirichlet priors over the topic proportions helps us achieve better sparsity in topics. However, this also leads to singularities in precision matrices conditioned on topics when Normal-Wishart priors for real valued data are not used.

3. THE PROPOSED MODELS

In this section we describe our proposed topic models for multimedia documents. We call the model in Fig. 3d MMGLDA, short for Multinomial-Multinomial-Gaussian LDA, and the model in Fig. 3e to Corr-MMGLDA, short for correspondence MMGLDA. In our context, the correspondence LDA model [4] places a probabilistic constraint on the correspondence of summary words to Gaussian observations—a word is likely to be generated by the topic which is agreed upon by most of the Gaussian instances in the video document. Since the real valued GIST features (see Section 4) "summarize" a scene in an image, we want a stronger influence of the topic of the scene on the summary text. This assumption is relaxed in MMGLDA. Of course the correspondence could have been established between the discrete observations only or both discrete and real valued ones but conditioned on the current dataset, we want more flexibility in topics for sampling discrete observations. We avoid overly complicated topic models and instead go for better data representations and supporting models with just the right amount of complexity.

In MMGLDA, for a multimedia document d, it is possible to have different topics competing for each occurrence of w_m. In Corr-MMGLDA, the number of such modes is constrained to be much fewer. The asymmetric $\boldsymbol{\alpha}$ can yield few additional modes which group co-occurring data dominant in densities or masses in separate latent topics. This phenomenon is observed for a larger number of latent topics.

Table 1 explains the symbols used in the two proposed topic models. Everywhere in this paper, we assume that K is the number of topics. The generative processes for the proposed models are illustrated below:

```
For each video document $d \in 1, ..., D$
    Choose a topic proportion $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$
    For each position $h$ in $d$
        Choose topic indicator $z_h|\theta \sim Mult(\boldsymbol{\theta})$
        Choose a discrete video "word" $w_h|z_h = k, \boldsymbol{\rho} \sim Mult(\boldsymbol{\rho}_{z_h})$
    For each real valued observation $o$ in $d$
        Choose topic indicator $z_o|\theta \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$
        Choose a real valued $\mathbf{w}_o|z_o = k, \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1} \sim \mathcal{N}(\boldsymbol{\mu}_{z_o}, \boldsymbol{\Lambda}_{z_o}^{-1})$
    For each position $m$ in video $d$
        Choose $y_m \sim Uniform(1, ..., O)$ (for Fig. 3e)
        *or* Choose $y_m|\theta \sim Mult(\boldsymbol{\theta})$ (for Fig. 3d)
        Choose a word $w_m \sim p(w_m|z_{y_m}, \boldsymbol{\beta})$ (for Fig. 3e)
        *or* Choose a word $w_m \sim p(w_m|y_m, \boldsymbol{\beta})$ (for Fig. 3d)
```

Symbol	**Meaning** (*r.v. = random variable*)
D	total number of multimedia "documents"
M	total number of discrete text features per multimedia document $d \in \{1, ..., D\}$
H	total number of discrete visual features in a multimedia document $d \in \{1, ..., D\}$
O	total number of real valued visual features per multimedia document $d \in \{1, ..., D\}$
$\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$	r.v. for asymmetric Dirichlet prior for the document level topic proportions
$\boldsymbol{\theta_d}$	r.v. for document level latent topic proportions
$\boldsymbol{\rho}$	corpus level topic multinomials over discrete video features
$\boldsymbol{\beta}$	corpus level topic multinomials for textual words
$\boldsymbol{\mu}$	means of topic Gaussians for the real valued features from videos
$\boldsymbol{\Lambda} = (\boldsymbol{\Sigma}^{-1})$	precision (inverse covariance) matrices of topic Gaussians for the real valued features from videos
y_m in Figs. 3b, and 3d	indicator variable for a sample from $\boldsymbol{\theta_d}$ for discrete text features
y_m in Figs. 3c and 3e	indicator variable for *document level* real valued datum correspondences
z_h	indicator variable for a sample from $\boldsymbol{\theta_d}$ for discrete visual features
z_o	indicator variable for a sample from $\boldsymbol{\theta_d}$ for real valued visual features
w_m	r.v. for textual word at position m in document d; vocabulary size of V
w_h	r.v. for vision oriented discrete feature at position h in document d; vocabulary size $corrV_H$
$\mathbf{w}_o$	r.v. for the o^{th} Gaussian feature vector with a dimensionality of P in document d

Table 1: Meanings of the variables used in the models

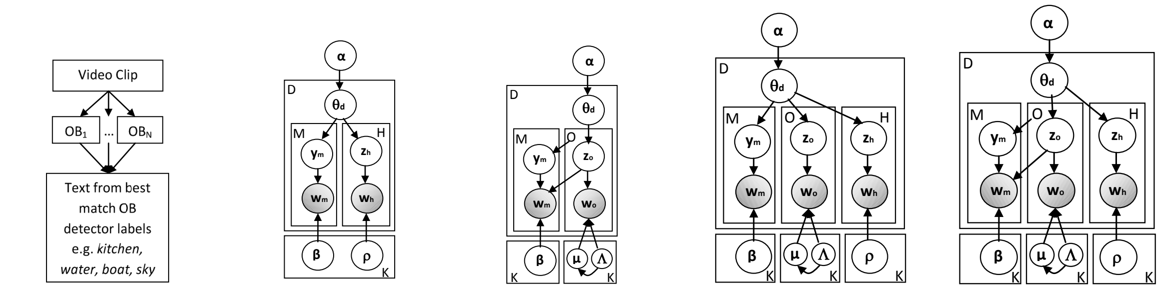
In all further notations, $\mathbf{w}_M$ is the ensemble of observed random variables that represent summary words in the d^{th} multimedia document. Similar notations hold for $\mathbf{w}_O$, $\mathbf{w}_H$ and the indicators $\mathbf{y}$ and $\mathbf{z}$. In this paper, the text vocabularies are event specific and of size 312 words on average.

Fig. 3a shows the Object Bank [15] (OB) baseline that we initially used to translate videos to text. The boxes labeled OB_1, ..., OB_N are the individual object detectors in Object Bank. The positive responses of the detectors lead towards identifying the label of the objects in the keyframes and hence translating the entire video. We choose this baseline to verify the difficult nature of our dataset—there is a 10% overlap between OB's vocabulary and the test set vocabulary, and we should expect to see at least 2-5% recall in ROUGE-1 recall scores for most events based on a 40-50% ROUGE-1 recall achieved by the best 100-word multidocument text summarization systems in TAC competitions.

3.1 Inference on Latent Variables

We use the Variational Bayesian Expectation Maximization (VB-EM) [2, 31] algorithmic framework as the optimization framework. An advantage of VB-EM is that it is deterministic.

The derivations for the MMG class of topic models become sufficiently complicated due to the need for using priors over the parameters governing the real valued observations. Since the nature of the modes for topic proportions is not


Video Clip
OB$_1$
...
OB$_N$
Text from best match OB detector labels e.g. *kitchen, water, boat, sky*
α
θd
D
M
H
O
ym
zh
zo
wm
wh
wo
β
ρ
μ
Λ
K

(a) Object Bank object detection model [15] (b) MMLDA [4, 26] (c) Corr-MGLDA [4] (extended) (d) MMGLDA (proposed) (e) Corr-MMGLDA (proposed)

Figure 3: Graphical model representations of existing topic models and proposed extensions— Figs. 3d and 3e. In this paper, we extend the model in Fig. 3c i.e. the Corr-LDA model in [4] with Normal-Wishart priors over parameters for real valued observations as well.

known in advance, singularities arising out ill-conditioned topic covariance matrices must be handled. This problem is mitigated in a principled way by introducing independent Normal-Wishart priors governing the mean vectors and precision matrices of the Gaussians conditioned on the topics. Since both $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are unknown we cannot factorize $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ directly because the variance of the distribution over μ is a function of $\boldsymbol{\Lambda}$. Instead we use combinations of Normal-Wishart priors on each Gaussian component as:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\kappa_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0) \quad (1)$$

where $\boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Lambda}_k$ is the precision matrix for the k^{th} factor or topic. This is similar to the mixture model used in [19]. To preserve the dependence between the means and covariances, a *partially* factorized tractable q distribution with "free" variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$, $\boldsymbol{\phi}^{(O)}$, $\boldsymbol{\phi}^{(H)}$ (for every multimedia document $d \in D$) is imposed by

$$q(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}_O, \mathbf{z}_H|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\phi}^{(O)}\boldsymbol{\phi}^{(H)}) = \left[\prod_{d=1}^{D} q(\boldsymbol{\theta_d}|\boldsymbol{\gamma_d}) \left[\prod_{m=1}^{M_d} q(y_{d,m} |\boldsymbol{\phi}_{d,m}) \prod_{o=1}^{O_d} q(z_{d,o}|\boldsymbol{\phi}^{(O)}_{d,o}) \prod_{h=1}^{H_d} q(z_{d,h}|\boldsymbol{\phi}^{(H)}_{d,h})\right]\right] \prod_{i=1}^{K} q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) \quad (2)$$

with $\boldsymbol{\theta_d} \sim Dirichlet(\boldsymbol{\gamma_d})$, $z_{d,o} \sim Mult(\boldsymbol{\phi}^{(O)}_{d,o})$ and $z_{d,h} \sim Mult(\boldsymbol{\phi}^{(H)}_{d,h})$. The maximum likelihood (ML) estimates of free parameters are found by optimizing the lower bound on $\log p(\mathbf{w}_M, \mathbf{w}_H, \mathbf{w}_O|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$.

The hyperparameters for α in the asymmetric Dirichlet case (the concentration parameter and the base measure) and κ_0, ν_0, $\mathbf{m}_0$ and $\mathbf{W}_0$ are not shown in Figs. 3c, 3d and 3e and in equ. 2 above. Also $\boldsymbol{\phi}$ are the free parameters of the variational summary_word multinomials over Gaussian_observations in the correspondence multimodal models or summary_word multinomials over topics in the plain multimodal models; $\boldsymbol{\phi}^{(O)}$ are the free parameters of the variational Gaussian_observation multinomials over topics and similarly for $\boldsymbol{\phi}^{(H)}$ for discrete visual features.

The variational posterior distribution $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ does not factorize into the product of the marginals, but we can always write it as $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q(\boldsymbol{\Lambda}_k)$. Then we use the result from mean field theory[24, 31] that says that the log of the optimal solution for factor q_j is obtained by considering the log of the joint distribution over all hidden and observed variables and then taking the expectation with respect to all of the other factors $\{q_i\}$ for $i \neq j$ i.e. for visible and hidden variable ensembles V and H, $\log q_j^*(H_j) = E_{i \neq j}[\log p(V, H)] + const$. This results in a Normal-Wishart distribution and is given by:

$$q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, (\kappa_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k) \quad (3)$$

where $\boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Lambda}_k$ is the precision matrix for the k^{th} factor or topic. The expression in Equ. 3 is obtained by first writing out the expression for $\log q^*(.)$ and selecting those terms that involve $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$. This yields:

$$\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{i=1}^{K} \log p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) + \quad (4)$$
$$\sum_{d=1}^{D}\sum_{o=1}^{O_d}\sum_{i=1}^{K} \phi^{(O)}_{d,o,i} \log \mathcal{N}(\mathbf{w}_{d,o}|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}) + const$$

Note that the variance of the distribution over $\boldsymbol{\mu}_k$ is a function of $\boldsymbol{\Lambda}_k$. The random variables $\mathbf{m}_k$ and $\mathbf{W}_k$ can be thought of as surrogates to $\mathbf{m}_0$ and $\mathbf{W}_0$ and that κ_k and ν_k surrogates to κ_0 and ν_0 but conditioned on latent topic k. The expressions for these variables, which are also used in the M-Step updates, can be found in Equs. 23, 24, 22 and 25. These expressions are obtained by matching the moments of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ to the Normal and Wishart distribution expressions. The optimal solution for $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ depends on the moments evaluated with respect to the distributions of other variables, and so the variational update equations are coupled and must be solved iteratively. Following [5, 4], let us now write down the objective functional, $\mathcal{L}_{(.)}$, to be maximized which acts as the lower bound to the true data log likelihood. We have **for the MMGLDA model**,

$$\begin{aligned}\mathcal{L}_{MMG} &= E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\mathbf{y}_M|\boldsymbol{\theta})] \\ &+ E_q[\log p(\mathbf{w}_M|\mathbf{y}_M, \beta)] + E_q[\log p(\mathbf{z}_O|\boldsymbol{\theta})] \\ &+ E_q[\log p(\mathbf{w}_O|\mathbf{z}_O, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + E_q[\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}_0, \mathbf{W}_0, \kappa_0, \nu_0)] \\ &+ E_q[\log p(\mathbf{z}_H|\boldsymbol{\theta})] + E_q[\log p(\mathbf{w}_H|\mathbf{z}_H, \boldsymbol{\rho})] - E_q[\log q(\boldsymbol{\theta}, \mathbf{y}_M, \\ &\mathbf{z}_O, \mathbf{z}_H, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\phi}^{(O)}, \boldsymbol{\phi}^{(H)}, \mathbf{m}, \mathbf{W}, \boldsymbol{\kappa}, \boldsymbol{\nu})] \quad (5)\end{aligned}$$

We only highlight the derivations for the expressions: $E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}\left[\frac{\ln|\boldsymbol{\Lambda}_i|}{2} - \frac{(\mathbf{w}_o-\boldsymbol{\mu}_i)'\boldsymbol{\Lambda}_i(\mathbf{w}_o-\boldsymbol{\mu}_i)}{2}\right]$, $E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}[\log p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)]$ and $E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}[\log q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)]$. In the variational Bayesian setting, the expression:

$$E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}\left[(\ln|\boldsymbol{\Lambda}_i|)/2 - ((\mathbf{w}_o - \boldsymbol{\mu}_i)^\top\boldsymbol{\Lambda}_i(\mathbf{w}_o - \boldsymbol{\mu}_i))/2\right]$$

needs to be evaluated in the log likelihood calculation for every video document d to update the free distributions given the current parameter values. The term $\left[\frac{\ln|\boldsymbol{\Lambda}_i|}{2}\right]$ is the normalization factor of the Gaussians and its expectations can cause the log likelihood to be positive. We therefore only evaluate $E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}\left[-((\mathbf{w}_o - \boldsymbol{\mu}_i)^\top\boldsymbol{\Lambda}_i(\mathbf{w}_o - \boldsymbol{\mu}_i))/2\right]$

for the per document updates and subtract the log of the exponentials of the aggregations as an approximation. We independently derive and mention only the final expressions for the following variables due to space constraints and self containment of this paper:

$$E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\left[\ln|\boldsymbol{\Lambda}_i|\right] = \sum_{p=1}^{P}\psi\left(\frac{\nu_i+1-p}{2}\right)+P\ln 2+\ln|\mathbf{W}_i| \quad (6)$$

$$E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\left[(\mathbf{w}_o-\boldsymbol{\mu}_i)'\boldsymbol{\Lambda}_i(\mathbf{w}_o-\boldsymbol{\mu}_i)\right]$$
$$= P\kappa_i^{-1}+\nu_i\left((\mathbf{w}_o-\mathbf{m}_i)'\mathbf{W}_i(\mathbf{w}_o-\mathbf{m}_i)\right) \quad (7)$$

$$E_{q_{[\boldsymbol{\mu},\boldsymbol{\Lambda}]}}\left[\log q(\boldsymbol{\mu},\boldsymbol{\Lambda})\right] = \sum_{i=1}^{K}\left\{\frac{1}{2}\ln\hat{\boldsymbol{\Lambda}}_i+\frac{P}{2}\ln\frac{\kappa_i}{2\pi}-\frac{P}{2}-H[q(\boldsymbol{\Lambda}_i)]\right\} \quad (8)$$

$$H[q(\boldsymbol{\Lambda}_i)] = -\ln Z(\mathbf{W}_i,\nu_i)-\frac{(\nu_i-P-1)}{2}\ln\hat{\boldsymbol{\Lambda}}_i+\frac{\nu_i P}{2}, \text{ where} \quad (9)$$

- $Z(\mathbf{W}_i,\nu_i) = |\mathbf{W}_i|^{-\nu_i/2}\left(2^{\nu_i P/2}\pi^{P(P-1)/4}\prod_{p=1}^{P}\Gamma\left(\frac{\nu_i+1-p}{2}\right)\right)^{-1}$
- $\ln\hat{\boldsymbol{\Lambda}}_i = E_q[\ln|\boldsymbol{\Lambda}_i|] = \sum_{p=1}^{P}\psi\left(\frac{\nu_i+1-p}{2}\right)+P\ln 2+\ln|\mathbf{W}_i|$

Note that Ψ is the digamma function. For the expression $\sum_{i=1}^{K}E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\left[\log p(\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)\right]$, we have:

$$\sum_{i=1}^{K}E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\left[\log p(\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)\right] = \frac{1}{2}\sum_{i=1}^{K}\Big\{P\ln(\frac{\kappa_0}{2\pi})+\ln\hat{\boldsymbol{\Lambda}}_i-\frac{\kappa_0 P}{\kappa_i}$$
$$-\kappa_0\nu_i(\mathbf{m}_i-\mathbf{m}_0)'\mathbf{W}_i(\mathbf{m}_i-\mathbf{m}_0)\Big\}+K\ln Z(\mathbf{W}_0,\nu_0)$$
$$+\frac{\nu_0-P-1}{2}\sum_{i=1}^{K}\ln\hat{\boldsymbol{\Lambda}}_i-\frac{1}{2}\sum_{i=1}^{K}\nu_i Tr(\mathbf{W}_0^{-1}\mathbf{W}_i) \quad (10)$$

Using the lower bound $\mathcal{L}_{MMG}$, the ML estimations of the hidden variables in video document d can be obtained using Lagrange Multipliers on $\boldsymbol{\phi}^{(H)}$, $\boldsymbol{\phi}^{(O)}$ and $\boldsymbol{\phi}$ as follows:

$$\phi^{(H)}_{d,h,i} \propto \exp\left\{\psi(\gamma_{d,i})-\psi(\sum_{j=1}^{K}\gamma_{d,j})+\log\rho_{i,w_{d,h}}\right\} \quad (11)$$

$$\phi^{(O)}_{d,o,i} \propto \exp\Big\{\psi(\gamma_{d,i})-\psi(\sum_{j=1}^{K}\gamma_{d,j})+E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\big[(\ln|\boldsymbol{\Lambda}_i|)/2$$
$$-((\mathbf{w}_o-\boldsymbol{\mu}_i)'\boldsymbol{\Lambda}_i(\mathbf{w}_o-\boldsymbol{\mu}_i))/2\big]\Big\} \quad (12)$$

$$\phi_{d,m,i} \propto \exp\left\{\psi(\gamma_{d,i})-\psi(\sum_{j=1}^{K}\gamma_{d,j})+\log\beta_{i,w_{d,m}}\right\} \quad (13)$$

$$\gamma_{d,i} = \alpha_i+\sum_{m=1}^{M_d}\phi_{d,m,i}+\sum_{o=1}^{O_d}\phi^{(O)}_{d,o,i}+\sum_{h=1}^{H_d}\phi^{(H)}_{d,h,i} \quad (14)$$

For the Corr-MMGLDA model, $E_q[\log p(\mathbf{w}_M|\mathbf{z}_{\mathbf{y}_M},\beta)]$ expands out to be:

$$\sum_{m=1}^{M}\sum_{i=1}^{K}\left(\sum_{o=1}^{O}\phi_{m,o}\phi^{(O)}_{o,i}\right)\log\beta_{i,w_m} \quad (15)$$

Also,

$$E_q[\log q(\mathbf{y}_M|\boldsymbol{\phi}_{\mathbf{y}_M})] = \sum_{m=1}^{M}\sum_{o=1}^{O}\phi_{m,o}\log\phi_{m,o} \quad (16)$$

and $E_q[\log p(\mathbf{y}_M|O)]$ is constant for all m in d. Equation (15) is a computational bottleneck because finding the confidence of the word w_m on topic i necessitates the elimination of uncertainties of w_m's dependence on $\mathbf{w}_o$ and $\mathbf{w}_o$'s dependence on topic i. This is also a strong point since the marginalization suggests a stronger influence of a topic on a summary word if that influence is justified by most $\mathbf{w}_o$s.

Using a similar lower bound $\mathcal{L}_{Corr-MMG}$ for Corr-MMGLDA, the ML estimations of the hidden variables in video d can be obtained as follows:

$$\phi^{(H)}_{d,h,i} \propto \exp\left\{\psi(\gamma_{d,i})-\psi(\sum_{j=1}^{K}\gamma_{d,j})+\log\rho_{i,w_{d,h}}\right\} \quad (17)$$

$$\phi^{(O)}_{d,o,i} \propto \exp\Big\{\psi(\gamma_{d,i})-\psi(\sum_{j=1}^{K}\gamma_{d,j})+E_{q_{[\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i]}}\Big[\frac{\log|\boldsymbol{\Lambda}_i|}{2}-$$
$$\frac{(\mathbf{w}_o-\boldsymbol{\mu}_i)'\boldsymbol{\Lambda}_i(\mathbf{w}_o-\boldsymbol{\mu}_i)}{2}\Big]+\sum_{m=1}^{M_d}\phi_{d,m,o}\log\beta_{z_{y_m},w_{d,m}}\Big\} \quad (18)$$

$$\phi_{d,m,o} \propto \exp\left\{\sum_{i=1}^{K}\phi^{(O)}_{d,o,i}\log\beta_{i,w_{d,m}}\right\} \quad (19)$$

$$\gamma_{d,i} = \alpha_i+\sum_{o=1}^{O_d}\phi^{(O)}_{d,o,i}+\sum_{h=1}^{H_d}\phi^{(H)}_{d,h,i} \quad (20)$$

3.2 Model Parameter Estimations

Before deriving the expressions for the maximum a posteriori and maximum likelihood estimates of the parameters of the proposed models using moment matching (Section 3.1) and derivatives w.r.t the parameters of the functional $\mathcal{L}_{(.)}$ let us define the following quantities for each topic i:

$$N_i = \sum_{d=1}^{D}\sum_{o=1}^{O_d}\phi^{(O)}_{d,o,i}; \qquad \bar{\mathbf{x}}_i = \frac{1}{N_i}\sum_{d=1}^{D}\sum_{o=1}^{O_d}\phi^{(O)}_{d,o,i}\mathbf{w}_{d,o}$$

$$\mathbf{S}_i = \frac{\sum_{d=1}^{D}\sum_{o=1}^{O_d}\phi^{(O)}_{d,o,i}(\mathbf{w}_{d,o}-\bar{\mathbf{x}}_i)(\mathbf{w}_{d,o}-\bar{\mathbf{x}}_i)'}{N_i} \quad (21)$$

By using moment matching techniques on Equ. 3, we obtain the following MAP expressions in general for each topic i:

$$\kappa_i = \kappa_0+N_i \quad (22)$$

$$\mathbf{m}_i = \frac{1}{\kappa_i}(\kappa_0\mathbf{m}_0+N_i\bar{\mathbf{x}}_i) \quad (23)$$

$$\mathbf{W}_i^{-1} = \mathbf{W}_0^{-1}+N_i\mathbf{S}_i+\frac{\kappa_0 N_i}{\kappa_0+N_i}(\bar{\mathbf{x}}_i-\mathbf{m}_0)(\bar{\mathbf{x}}_i-\mathbf{m}_0)' \quad (24)$$

$$\nu_i = \nu_0+N_i \quad (25)$$

Further, using some algebraic manipulations and utilizing Lagrange Multipliers for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ for each topic i, we obtain:

For the MMGLDA model:

$$\rho_{i,j} \propto \sum_{d=1}^{D}\sum_{h=1}^{H_d}\sum_{j=1}^{corrV_H}\phi^{(H)}_{d,h,i}\delta(w_{d,h},j) \quad (26)$$

$$\beta_{i,j} \propto \sum_{d=1}^{D}\sum_{m=1}^{M_d}\sum_{j=1}^{V}\phi_{d,m,i}\delta(w_{d,m},j) \quad (27)$$

For the Corr-MMGLDA model:

$$\beta_{i,j} \propto \sum_{d=1}^{D}\sum_{m=1}^{M_d}\sum_{j=1}^{V}\left(\sum_{o=1}^{O}\phi_{d,m,o}\phi^{(O)}_{d,o,i}\right)\delta(w_{d,m},j) \quad (28)$$

The update for parameters $\rho_{i,j}$ remain the same. To optimize the $\boldsymbol{\alpha}$ parameters, we follow the corresponding expressions in [5] and optimize using Newton's iterative gradient based method using backtracking line search.

For predicting a bag of words summary from an ensemble of low level features of video document d and the learnt $p(w_v|z_v = k, \boldsymbol{\beta})$, we permute the vocabulary V for the new test video as:

$$p(w_v|\mathbf{w}_O, \mathbf{w}_H) \approx \sum_{o=1}^{O}\sum_{i=1}^{K} \phi_{d,o,i}^{(O)} p(w_v|\boldsymbol{\beta}_i) + \sum_{h=1}^{H}\sum_{i=1}^{K} \phi_{d,h,i}^{(H)} p(w_v|\boldsymbol{\beta}_i) \tag{29}$$

4. EXPERIMENTAL SETUP AND RESULTS

Here we briefly mention the descriptors that we use to represent the videos. To represent actions, we use features known as Histogram of Oriented Gradients in 3D (HOG3D) [11]. The gradient directions are binned by mapping them to 10 polar meridian and 6 polar parallel planes and then treating half spaces to be equivalent. We resized the video frames such that the largest dimension (height or width) was 160 pixels, and extracted HOG3D features from a dense sampling of frames. Our HOG3D parameters resulted in a 300-dimensional feature vector using support volumes of dimension 2 × 2 × 5 and 5 × 3 polar co-ordinate bins. We then use K-means clustering to create a 1000-word codebook following [3] from a random sampling of the training data.

Color histogram features are also used as part of the discrete visual data. We use 512 RGB color bins and histograms are computed on densely sampled frames. Due to large deviations in the extremities of the color spectrum, we use the histogram between the 15^{th} and 85^{th} percentiles averaged across a video and counts normalized to lie in [1,100].

Finally we use Object Banks [15] for a histogram pattern of positive object detections. OB transforms an image into a 44604 dimensional concatenated feature vector for each of the 177 *off-the-shelf* object detectors that are currently used. Each entry within a 252 dimensional detection feature vector represents the distance from the decision hyperplane midway within the margins for different scale-space transformations of the image. The object labels in OB cover only about 10% of the summary words (246 out of 2687 for the training set and 166 out of 1219 for the Dev-T set). Keyframes used for these features are extracted using the change in color histogram method [35] and the positive OB responses are quantized following classemes in [28]. Thus $\mathbf{w}_H$ in Figs. 3b, 3d and 3e consists of codebook histograms from HOG3D, color and OB. Needless to say, the contributions of these *off-the-shelf* object detectors are not significant at all.

The real valued features we use in our video representation are those representing scenes as mentioned in [22]. The scene property by itself induces image summarization in a way that is consistent with human perception of vision [23]. A set of perceptual dimensions is proposed along the boundary viewpoint (e.g. depth, openness, expansion, perspective) and along the content viewpoint (e.g. naturalness, roughness, ruggedness, etc.) which represent the dominant *spatial structure* of a scene. These features are named GIST features as is common parlance in computer vision literature. To compute these features, we have used the setup in [7] leading to a 960-dimensional descriptor for each frame. We calculate GIST features for every 10^{th} frame.

To save computational time, the GIST features are projected into lower dimensions using Principle Component Analysis (PCA). PCA is done on the training data across all event categories to remove the dependence of the visual descriptors on specific events. We first visualize the lower (15, 30 and 60) dimensional GIST features in two dimensions using t-statistic based stochastic neighbor embedding (t-SNE) [17], however, the separations look more or less the same and do not yield conclusive evidence of choosing the right number of dimensions. By inspecting the plots from t-SNE, we choose 15 dimensions and validate the choice by both manually inspecting the eigenvalues and experimentally cross-validating with the baseline Corr-MGLDA topic model (Fig. 3c). 30 or 60 dimensional features decreased the ELBO of the model. We do not select further lower dimensions based on significance of the eigenvalues. Each $\mathbf{w}_o$ in Figs. 3c, 3d and 3e represents a frame in 15 dimensions corresponding to a GIST feature vector.

4.1 Model Log Likelihoods and Topics

In this section, we evaluate the topic models in terms of ELBO on the held-out Dev-T set acting as a test set (with the human summaries) for posterior inference and as a prediction set (without human summaries) for BoW summary generation. Multinomial parameters are seeded and Gaussian parameters are randomly initialized. The base measures of $\boldsymbol{\alpha}$ are initialized to 0.1 and normalized while its concentration parameter is set to 10. An issue with the real valued features is the influence of data normalization on the ELBOs from the topic models. We have observed that when the data is not normalized to lie within $[0,1]^P$, the sequence of ELBOs from Corr-MGLDA during EM often indicate suboptimality even during training. The "PDS" suffix (in the table and all other figures) means "Positive Data Scaling" i.e. each real valued vector is sum-normalized to $[0,1]^P$ independently.

Model	**Topic 1**	**Topic 2**	**Topic 3**	
Corr-MMG LDA	wed couple ceremony church ring footage exchange bride groom helicopter	wed ceremony bride groom church flower vow exchange ring walk kiss outdoors	wed Hawai US beach guest place footage scene ring minister lei Kailua	Wedding ceremony
Corr-MG LDA-PDS	wed ceremony couple bride groom church flower exchange vow	wed ceremony bride groom couple flower church man outdoors vow	wed ceremony bride groom couple flower church man outdoors vow	

Table 2: 3 latent topics for an event from proposed 5-topic Corr-MMGLDA and Corr-MGLDA-PDS models. The topics from Corr-MGLDA-PDS are similar as a result of high values of α_k obtained after running Corr-MGLDA-PDS on scaled i.e. normalized data. The topics from Corr-MMGLDA are qualitatively far superior and indicates sub-events of the "Wedding ceremony" event.

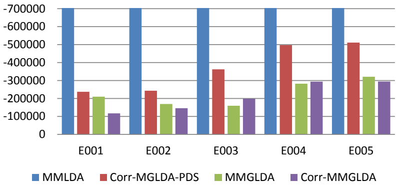


Figure 4: Test ELBOs on events E001-E005 in the Dev-T set. Lower is better.

The PDS normalization fixes this problem and raises ELBOs for Corr-MGLDA significantly but convergence is slower. However in the latter setting, the values of α_k become large which destroys sparsity in topics. This is possibly due to strong overlap of modes within the $[0, 1]^P$ hypercube where one dimension is severely correlated with the others. Examples of such topics on the "Wedding Ceremony" event is given in Table 2 where all topics are almost alike and lose subjective interpretability.

The new topic models with both Multinomial and Gaussian distributions on the video features do not suffer from the data scaling problem. It is possible that the mean parameter space for the tractable distributions over both discrete and real valued observations prevents co-ordinate ascent steps to dwell in suboptimal regions that could arise out of extreme values in the real valued data alone.

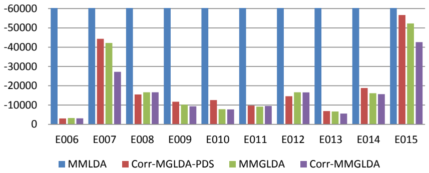


Figure 5: Test ELBOs on events E006-E015 from Dev-T set. Lower is better

Although the Normal-Wishart priors act as regularizers, automatically tuning $\mathbf{W}_0$ using another level of priors or from the data itself is not used here. In general optimization with tractable distributions and parameter constraints (e.g. non-negativity, boundedness and positive definiteness) can be non-convex [31].

Figures 4 and 5 show the test ELBOs of MMGLDA and Corr-MMGLDA versus the MMLDA model and the Corr-MGLDA model with PDS. The ELBOs for MMLDA are off the charts (at least three to four times the cut-off shown in the graphs). For the first 5 events, the videos contain positive instances of the events in Dev-T set. For this subset of events, the MMGLDA family of models outperform the best version of Corr-MGLDA in terms of ELBOs (i.e. with PDS). Figures 4 and 5 are obtained using K=20 topics—K being set through 5-fold cross - validation. For the last 10 events (Fig. 5), the videos contain only related instances of the events in Dev-T set—dissimilar to the training configuration i.e. the annotators are unsure about the relevance of the videos to the event category. In this case, Corr-MGLDA-PDS do not perform worse in general since the GIST features are *global* features [22].

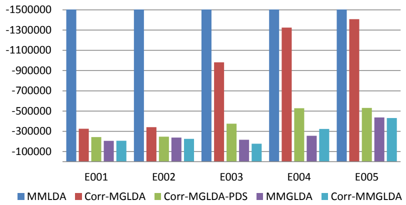


Figure 6: Prediction ELBOs on first 5 event for Dev-T set. Lower is better.

The prediction performance on the first 5 events is shown in terms of ELBO in Fig. 6 for the same value of K. Fig. 6 shows that MMLDA does not perform well in terms of word prediction ELBO measure. We can also see the effects of sub-optimality when PDS is suppressed for Corr-MGLDA (Corr-MGLDA in Fig. 6). MMGLDA and Corr-MMGLDA again perform comparably and outperforms Corr-MGLDA-PDS on the first 5 events except E002—"feeding an animal"—a very complex event for computer vision.

For events 6 through 15, the prediction ELBO graphs also look very similar to that in Fig. 5 (not shown here due to space constraints). PDS on our proposed MMGLDA family shows even better ELBOs, but topic sparsity problems mitigate only a little and we do not report those here. All these experiments are run using $\mathbf{m}_0$ set to $\mathbf{0}$, $\mathbf{W}_0$ set to a broader prior $\mathbf{I}$, the identity matrix, ν_0 set to P and κ_0 set to 1. Normalizing the data to lie in $[0,1]^P$ with $\mathbf{I}$ as priors for $\mathbf{\Lambda}_k$s leads to sharing of topic responsibilities of the real valued data by only a few Gaussians thereby contributing much less to the overall log-likelihood.

It is also observed that the means of the ELBOs of our proposed models are significantly less negative (i.e. better) at 95% confidence level (using paired t-test) than the existing topic models during cross-validation on the training set. For most events, ELBOs for proposed models with K=10 are not statistically worse either and show slightly higher ROUGE-1 scores for some events. Figure 7 shows the macro average of test ELBOs across all the 15 events in the Dev-T set. We omit the line graph for MMLDA as it is out of axis limits. The graphs confirm the superior fit of our proposed models to a natural representation of multimedia (test) data.

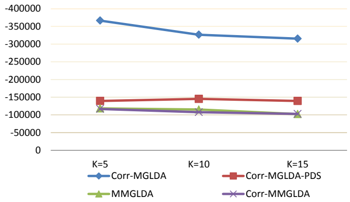


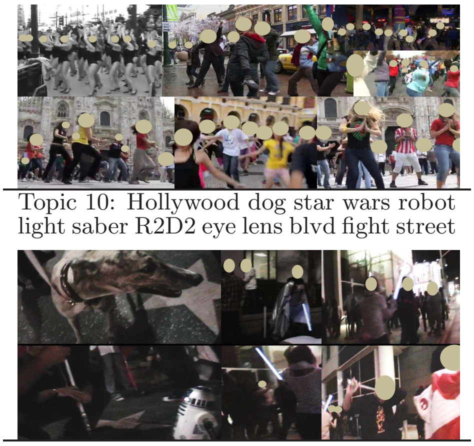
Figure 7: Average test ELBOs on all events in the Dev-T set for different topics. Lower is better.

4.2 Translating Related Words to Videos

Table 3 shows how latent topics can first be used to discover most probable related words from unstructured text which can then be translated to most probable frames from one or more videos (and hence the videos themselves). The frames correspond to $\mathbf{w}_o$s in Fig. 3 and Table 1.

We observe from Table 3 how topics 6 and 10 decompose the "Flash mob gathering" event into its constituent subthemes. While topic 6 describes flash mob dances in outdoors and near plazas, topic 10 focuses on a flash mob in Hollywood posing in Star Wars costumes and light sabers along with the famous miniature robot R2D2. Due to space constraints we cannot show more samples of numerous such examples.

Topic 6: mob flash dance people mall outdoors large gather woman public plaza



Topic 10: Hollywood dog star wars robot light saber R2D2 eye lens blvd fight street

Table 3: 2 latent topics for "Flash mob" event from a 10-topic Corr-MMGLDA

Table 4 shows the inter-translation of modalities for topic 10 from MMGLDA corresponding to that in Table 3 from Corr-MMGLDA. Note how the topic loses specificity (e.g. misses "R2D2", "light," "saber" within the top few words) and focuses on generality (e.g. flash mob). Topic 6 for MMGLDA is exactly the same as that for Corr-MMGLDA. Table 5 shows log of the ratio: $\frac{\boldsymbol{\alpha}_k}{|\mathbf{\Lambda}_k|}$ for the two proposed

Topic 10: place fight fake public event gather flash mob music star outdoors war

Table 4: Topic 10 for the "Flash mob" event from a 10-topic MMGLDA

models and gives us a hint on how “broad” a topic k may be vs. how much variance in the visual summary is it able to capture. A relatively higher value of the ratio means that a topic captures more variance and hence the volume captured by the determinant of the inverse covariance matrix $\mathbf{\Lambda}_k$, i.e. $|\mathbf{\Lambda}_k|$, through its spanning eigenvectors is proportionally less. For MMGLDA, this ratio is always relatively lower in our setting and this means that the model captures more generic patterns first giving rise to a lower $|\mathbf{\Lambda}_k|^{-1}$.

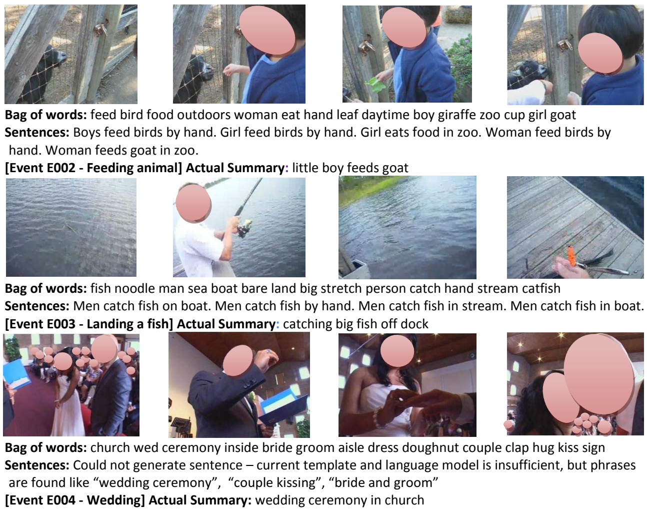
The last column in Table 5 is the average of the ratios for the other topics from the two models for event eight. The ratios can be large for a more general topic (e.g. topic 6 in Table 3) owing to a higher $\boldsymbol{\alpha}_k$ too. Although, all values are close due to the use of the broader prior, $\mathbf{I}$, it is observed that Corr-MMGLDA discovers related words which are qualitatively superior. However, the corresponding most probable frames are almost similar for both models in most cases. Translating related words to frames is best judged manually, but, the ratio we use here can be a viable alternative.

Model	k=6	k=10	avg$_{k\neq\{6,10\}}$
Corr-MMGLDA	104.628	8.164	40.2398
MMGLDA	104.623	8.102	40.0702

Table 5: log $\frac{\boldsymbol{\alpha}_k}{|\mathbf{\Lambda}_k|}$ values for topics in event 8

4.3 Translating/Summarizing Videos To Text

Table 6 reports the ROUGE-1 (henceforth R-1) scores of the predicted 5 to 10 keyword summaries from the different models when compared with the corresponding short human synopses. Sometimes full sentences cannot be generated from the predicted words due to a deficient language model. This and the short nature of human synopses are the primary reasons why we perform only R-1 evaluation. Some examples of the sentences/phrases are shown in Fig. 8.



Bag of words: feed bird food outdoors woman eat hand leaf daytime boy giraffe zoo cup girl goat
Sentences: Boys feed birds by hand. Girl feed birds by hand. Girl eats food in zoo. Woman feed birds by hand. Woman feeds goat in zoo.
[Event E002 - Feeding animal] Actual Summary: little boy feeds goat

Bag of words: fish noodle man sea boat bare land big stretch person catch hand stream catfish
Sentences: Men catch fish on boat. Men catch fish by hand. Men catch fish in stream. Men catch fish in boat.
[Event E003 - Landing a fish] Actual Summary: catching big fish off dock

Bag of words: church wed ceremony inside bride groom aisle dress doughnut couple clap hug kiss sign
Sentences: Could not generate sentence – current template and language model is insufficient, but phrases are found like “wedding ceremony”, “couple kissing”, “bride and groom”
[Event E004 - Wedding] Actual Summary: wedding ceremony in church

Figure 8: Bag of keywords and sentence translations from our proposed MMGLDA ($K = 20$) for some clips from events 2, 3 and 4 from the Dev-T set. Best viewed in color and magnification.

In Table 6, OB is the Object Bank baseline (see just ahead of Section 3.1)—it confirms the difficulty of detecting objects on this dataset. The quantized OB responses perform poorly since a-priori it is hard to know which object detectors will be needed and existing but irrelevant object detectors can produce an unpredictable pattern of false positives. Creating object models for every genre of data requires expensive annotation efforts. Even if there is a 100% overlap between our training vocabulary and object models, the R-1 scores for OB *may* only increase by 10-folds which is still low.

Purely multinomial topic models showing lower ELBOs can perform quite well in BoW summarization. MMLDA assigns likelihoods based on success and failure of *independent* events and failures contribute highly negative terms to the log likelihoods but this does not indicate the model’s summarization performance where low probability terms are pruned out. Gaussian components can partially remove the *independence* through covariance modeling and fit the data better at the cost of higher time and space complexity. The R-1 scores from MM(G)LDAs are comparable for 5 and 10 keywords with no statistical difference, however, a possible reason for lower R-1 scores for Corr-MMGLDA model is that due to better correspondence to the topic of the GIST energy in the scene, when a topically relevant but non-summary word is chosen upfront, more related but non-summary words are also drawn in. As future research, we also like to do a principled initialization of Gaussian parameter priors as in [19].

	Model	n=5	n=10	OB
E001	MMLDA	0.182	0.248	0.0*
	Corr-MGLDA-PDS	0.187	0.257	
	MMGLDA	0.179	0.245	
	Corr-MMGLDA	0.139*	0.192*	
E002	MMLDA	0.186	0.249	0.0*
	Corr-MGLDA-PDS	0.182	0.242	
	MMGLDA	0.186	0.237	
	Corr-MMGLDA	0.143*	0.176*	
E003	MMLDA	0.221	0.265	0.012* =1%
	Corr-MGLDA-PDS	0.233	0.263	
	MMGLDA	0.228	0.267	
	Corr-MMGLDA	0.171*	0.230	
E004	MMLDA	0.265	0.302	0.0*
	Corr-MGLDA-PDS	0.263	0.292	
	MMGLDA	0.264	0.321	
	Corr-MMGLDA	0.221	0.247*	
E005	MMLDA	0.167	0.213	0.005* =0.5%
	Corr-MGLDA-PDS	0.180	0.208	
	MMGLDA	0.165	0.205	
	Corr-MMGLDA	0.129*	0.142*	
6-15 Avg.	MMLDA	0.216	0.252	0.001* =0.1%
	Corr-MGLDA-PDS	0.211	0.258	
	MMGLDA	0.210	0.243	
	Corr-MMGLDA	0.179*	0.221	

Table 6: Individual and average ROUGE-1 scores on the events—best results from 10/20 latent topics are shown. The value of n represents the top-n most probable keywords. A (*) means significantly **worse** performance at 95% confidence to {MM,MMG}LDAs. These results are only reported for the same hyperparameter settings.

However, the high scores with Corr-MGLDA-PDS is entirely co-incidental—the topics are more or less uniform and each one covers parts of the sub-events equally. Further, each $\mathbf{w}_o$’s density over those topics is uniform enough to not achieve a reasonable permutation. The same thing happens when PDS is used for our MMGLDA family of models and the summaries completely lose subjective appeal although R-1 scores improve considerably. This is similar to the *qualitatively degenerate approach*—taking the top n frequent words from the event vocabulary and using those as summaries for **every** test video. The scores for MMLDA and MMGLDA are also comparable to this setting. Quantification of the permutation quality has not been done.

Scores in Table 6 need to be multiplied by the event classification accuracies to obtain lower bounds for clips having no event labels. The scores become competitive for larger n and much larger K if we topic model on the entire corpus.

4.3.1 Natural Language Generation

To translate a video into multiple sentences from predicted keywords, we use an ordered-sequence template as <subject, verb, object, preposition, scene-noun>. Language models from the data at hand is used to prune impossible sequences. The subjects, objects, verbs and nouns extracted from the training synopses using dependency grammars and POS models appear in each generated sentence only once.

We use the parser in [12] to score the sequence of words following the template. The sentences are ordered according to bigram and parse tree scores. When complete sentences cannot be generated due to a deficient language model, we output possible bigrams and trigrams (see E004 in Fig. 8). Similar corpus based sentence generation techniques can be found in [34, 9] but NLG is a research topic in its own right.

4.4 Event Classification

Fig. 9 shows 5-fold cross-validation on the 15-event training set and also the test accuracies on Dev-T set for event classification. A c-SVM classifier from the libSVM [6] package is used with default settings for multiclass classification. Although around 50% classification accuracy can be easily achieved using the discrete visual features that we use, higher accuracies can be obtained using better kernels, fusion of classifiers and optimizing Detection-Error-Tradeoff curves while cross validating [20]. However, these discussions are outside the scope of this paper.

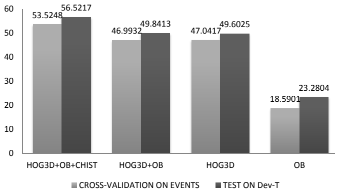


Figure 9: Event detection accuracies for cross-validation (light gray bars) and test (dark gray bars) with different features

5. CONCLUSION

Our new topic models show better fits to multimedia data representation consisting of discrete and real valued features from videos as well as accompanying short textual synopses. In general Corr-MMGLDA improves on text to video translation while the non-correspondence versions perform better in video to text summarization. Video summarization through topic models significantly out-perform that through state-of-the-art object detectors and thus can be used as new baselines. Our NLG component has suffered from severe data sparsity and impoverished language models and we wish to overcome these using external knowledge bases.

ACKNOWLEDGEMENTS This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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. We thank Honeywell ACS Labs, Simon Fraser Univ. & Kitware Inc. for support in feature extraction and the anonymous reviewers for their comments.

6. REFERENCES

[1] A Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *SIGKDD*, 2009.

[2] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, UCL, 2003.

[3] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *ICCV*, 2011.

[4] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, 2003.

[5] D. M. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[7] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR*, pages 19:1–19:8, 2009.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[9] A. Gupta, Y. Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[10] T. M. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, 2009.

[11] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[12] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL*, 2003.

[13] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*. 2004.

[14] L. Li, K. Zhou, G. Xue, H. Zha, and Y. Yu. Video summarization via transferrable structured learning. In *WWW*, 2011.

[15] L-J. Li, H. Su, E. P. Xing, and L. Fei-fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[16] C-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL HLT*, 2003.

[17] L. V. D. Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, pages 316–329, 2008.

[19] N. Nasios and A.G. Bors. Variational learning for gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(4):849 –862, 2006.

[20] P. Natarajan et al. BBN VISER. In *TRECVID MED*, 2011.

[21] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, 2004.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[23] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 155(1):23–36, 2006.

[24] G. Parisi. *Statistical Field Theory.* Addison-Wesley, 1988.

[25] D. Putthividhya, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010.

[26] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *WSDM*, 2009.

[27] R. K. Srihari. Piction: A system that uses captions to label human faces in newspaper photographs. In *AAAI*, 1991.

[28] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[29] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces. In *ECCV*, 2010.

[30] Yang W. and Greg M. Human action recognition by semi-latent topic models. *IEEE PAMI*, 31(10):1762–1774, 2009.

[31] M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference.* Now Publishers Inc., Hanover, MA, USA, 2008.

[32] J. Wanke, A. Ulges, C. Lampert, and T. Breuel. Topic models for semantics-preserving video compression. In *MIR*, 2010.

[33] M. J. Welch, J. Cho, and W. Chang. Generating advertising keywords from video content. In *CIKM*, 2010.

[34] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.

[35] H. J. Zhang, C. Low, S. Smoliar, and J. Wu. Video parsing, retrieval and browsing: an integrated and content-based solution. In *ACM MM*, 1995.