Context-Based Semantic Tagging of Multimedia Data

Nisha Pahal^(✉), Santanu Chaudhury, and Brejesh Lall

Department of Electrical Engineering, Indian Institute of Technology Delhi,
New Delhi, India

nisha23june@gmail.com, {santanuc,brejesh}@ee.iitd.ac.in

Abstract. With the rapid growth of broadcast systems and ease of accessing internet services, lots of information is available and accessible on the web. The information available in multimedia documents may have different *context* and *content*. Since, interpretation of multimedia content cannot be free of context so tagging on the basis on context is indispensable for dealing with this problem. Tagging plays an important role in retrieving multimedia data as now-a-days most of the videos are retrieved based on text describing them and not by the actual context embodied in them. So, in this paper we have proposed a scheme for tagging multimedia data based on the contents and context as identified from web-based resources. The hierarchical LDA (hLDA) is used to model the context information while Correspondence-LDA (Corr-LDA) is used to model the content information of multimedia data. Finally, multimedia data is tagged with the relevant contents and context information on the basis of Context-Matching Algorithm. These tags can then be used by search engines for increasing precision and recall of multimedia search results.

Keywords: Context · Tagging · hLda · SIFT · Corr-LDA

1 Introduction

Due to the availability of high speed digital cameras and smart phones with high-capacity storage, the multimedia data specifically, images and videos have gained importance in everybody's lives. Lots of new videos and images are being uploaded over the internet every second. A major challenge is handling this rapidly growing large volume of multimedia data on the web. This multimedia data encodes information in context to any event and this information needs to be tagged correspondingly. Another issue is that a lot of images and videos on the web are incorrectly tagged. Any description or title tagged with that multimedia data provides information about that data. This information is nothing but meta-data i.e. data about data. The available multimedia search engines rely on this meta-data to search for videos/ images. This meta-data information may be too subjective and may lose its formal semantic meaning. So, it is required to

associate meaningful semantic tags with available multimedia data. This makes the multimedia retrieval process more efficient and accurate.

In this paper we have proposed a tagging framework that tags multimedia data with semantically relevant context and content information. Unlike other approaches, in this approach we have utilized our proposed tagging framework to associate a query video or an image with semantically relevant context and content information, exploiting multi-modal data - visual content of the frames and textual content available on the web. We have established the validity of our approach using experimental results. The remainder of the paper is structured as follows: Sect. 2 discusses the related work. In Sect. 3 the context-based tagging model is introduced. Section 4 contains the Problem Formulation. In this section, initially the problem is defined and then an approach for modeling the context information is discussed. Also, image content classification and merging of context and content information with multimedia data based on contextual meaning is presented. Section 5 contains experimental results and finally the conclusion and references are provided.

2 Related Work

In this section, we discuss some of the approaches that exist today for tagging the multimedia data. Siersdorfer et al. [7, 8] presented the different tag propagation methods for automatically obtaining richer video annotations. Their approach provides the user with additional information about videos which leads to enhanced feature representations for applications like automatic data organization and search. Another approach provided by Nguyen [6] exploits available unstructured data and hidden topic models to infer surrounding contexts for better text and image retrieval. This helps in bridging the semantic gap in Web Mining and Information Retrieval. A description model of video metadata for semantic analysis taking into account various contextual factors was presented by Steinmetz et al. [1]. They considered the contextual information of video metadata for the purpose of Named Entity Recognition (NER) and for calculating a confidence value so as to bring metadata items in a specific order and to use them as context items for the disambiguation process. In another work by Giannakidou et al. [2] the authors provided an unsupervised model for efficient and scalable mining of multimedia social-related data. Their model jointly relies on social, semantic and content knowledge which can be used for tag clusters. And, the tag clusters produced can be used for semantics extraction and knowledge mining required for automated multimedia content analysis. Aytar [3], utilized the semantic word similarity measures for video retrieval. Their approach provided an effective way of using high level semantic relations in video retrieval problem by establishing a bridge between high level visual and semantic relations. Their method for modeling context was trivial as they didn't consider the spatial information due to the complexity of segmentation and localization tasks in TRECVID data sets. So, to bridge the semantic gap between low-level content-based and high-level context-based techniques, hybrid

Fig. 1. Tagging framework

approaches have been proposed. A music retrieval system that combines content-based acoustic similarity and context-aware semantic descriptions was presented by Miotto et al. [4]. The system represents a music collection using a hidden Markov model with the purpose to build a music retrieval system that combines content-based acoustic similarity and context-aware semantic descriptions. Various approaches of content-based multimedia retrieval were discussed by Lew et al. [5]. The authors in this paper described various aspects of content-based multimedia information retrieval along with some major challenges for the future.

Although the existing work follows different approaches but none of them aspire to tag multimedia data with the relevant content and context information. Our multimedia tagging approach is different and novel from the work discussed above. It aggregates data from diverse web-based resources and associates a description of the multimedia data with the query video or an image on the basis of contextual meaning.

3 Context-Based Tagging Model

Context refers to some situations or conditions that make the meaning of the content of any web document more clear and specific. Context-based search for multimedia retrieval is an insightful technique which revolutionizes the efficiency of search engines. The multimedia documents with almost the same content may have totally different contexts and thus, may refer to different events. An event is always associated with several inherent contexts like temporal, geographical or entity based. Since context gets defined with respect to an event instance, therefore, identification of context information facilitates in identification of an event that is taking place in multimedia document. The tagging model is depicted in Fig. 1.

The *Event Context* refers to the event related information that is associated with the multimedia data. This includes the time and place related information

of an event which for example help in identifying the Temporal and Geographical context. Some other tags can be identified from the image content using Corr-LDA. On the other hand, the documents from the repository are classified contextually using hLDA approach. The tagging engine makes use of context-matching algorithm to associate the contextually aligned documents and the tags (temporal, spatial and the ones obtained using Corr-LDA) with the query image/ video. The obtained tags are given as input to the tagging engine that facilitates in identifying the contextually relevant documents from the document repository. The documents are considered relevant only if their context matches with the context of the input video or an image.

4 Problem Formulation

Given an image or video, our objective is to annotate the video or an image with the relevant context and content information on the basis of semantic meaning. The contextual information is modeled using hierarchical LDA (hLDA) as explained in Sect. 4.1. For content classification, the Correspondence LDA (hereafter, Corr-LDA) is used to classify videos into different categories corresponding to different events as explained in Sect. 4.2. To associate query image or video with the relevant context and content information on the basis of contextual meaning, Context Matching Algorithm has been devised as described in Sect. 4.3.

4.1 Context Modeling

We have constructed a hierarchical event based context model for an image/ video. In this model, a set of web documents are organised in the form of a hierarchy. Different paths along the hierarchy refer to different sets of contexts. Thus, each path refers to some specific set of web documents. The contexts of the topics of higher levels are more generalised than the lower level ones. The proposed hierarchy consists of *Event-Type* at the lowest level. *Event-Topic* is the super class of the event-type. At the top level we have *Event-Context* class as shown in Fig. 2. For example, *cutting of birthday cake* is an event at the lowest level which belongs to Event-Topic *birthday* event. It further belongs to the Event-Context class *Family*. This hierarchical model is pre-constructed. However, we need to build such hierarchy automatically based on resources available. As an application we have focused on events in the public domain. Accordingly, we chose news websites for building the context models to associate with an image or video of news events. We have built a corpus of news documents classified and clustered at *Google News Website*. On the clustered set of documents we have applied hierarchical LDA (hLDA) [11] to discover a hierarchy of topic models. For example, we have created a repository of all news article pertaining to the domain *entertainment* and *automobile*. Using these documents we get a set of topics.

For obtaining this hierarchical context from text news, hierarchical topic modeling is used which is based on nested Chinese Restaurant Process (CRP) [9].

Fig. 2. Hierarchical context classification model

One of the foremost advantages of using hLDA is that it does not rely on a fixed tree structure for the representation of documents. As the documents are added to the model, the structure of the hierarchy keeps on changing to add new levels and topics. Once the hierarchy is obtained, each document in a repository corresponds to a distribution over the L distinct levels and is drawn by choosing an L -level path. The words are drawn from the L topics corresponding to the documents along that path. All the lower-level documents share the topic associated with the highest-level root document.

Assuming a data set consisting of a corpus of documents wherein each document comprises of a collection of words, and a word is considered to be an item in a vocabulary. The fundamental assumption is that the words in a document are generated according to a mixture model where the mixing proportions are random and document specific [9]. A multinomial variable z , and an associated set of distributions over words is represented as $p(w | z, \beta)$, where β is a parameter. The topics, one distribution for each possible value of z , are the basic mixture components in this model. The document specific mixture distribution that temporally assumes K possible topics in the corpus where, z ranges over K possible values and θ is a K -dimensional vector is represented by:

$$p(w | \theta) = \sum_{i=1}^K \theta_i p(w | z = i, \beta_i) \quad (1)$$

where, $\text{Dir}(\alpha)$ is the Dirichlet distribution for parameter α and $\text{multinomial}(\theta)$ is the multinomial distribution for vector θ . The vector θ represents the document specific mixing proportions corresponding to the different components.

The test database consists of a set of 85 documents corresponding to automobile and entertainment category. From these we chose three events, namely *car launch event*, *celebrity event* and *music event*, and pre-processed them to remove stopwords. When we ran hLDA model on this data we obtained a hierarchy of topics (context) as shown in Fig. 3.

Fig. 3. Hierarchy of event context

4.2 Image Content Classification

An image is characterized by temporal and geographic context. Such temporal and geographic context links an image to image content. Hence, for tagging an image we need to formulate a strategy for identification of event context of an image. Image classification plays a key role in many applications such as video surveillance, video tagging, image retrieval and web content analysis. To classify the content information we have used Corr-LDA. First the SIFT features are extracted and then, a dictionary of visual words is obtained. This data is used for image content classification as explained in the next sub-section.

Space-Based SIFT Feature. First the SIFT feature points are obtained for each image. For each SIFT feature point, SIFT descriptors are constructed. The size of each SIFT feature vector is 128. These descriptors are quantized into visterms. In our experiments, we have trained Corr-LDA model using a dataset of 460 images and considered 35 as the total number of clusters. The K -means clustering is then applied and visual vocabulary is constructed which is represented by an $N \times M$ matrix, where each row corresponds to an image and each column represents a visterm. Unlike LDA model that allows documents to be represented as bag-of-words model in text analysis, the bag-of-visterms (BOV) model allows representation of an image as an orderless sequence of visual terms called visterms. The feature descriptors are matched with the vocabulary and a histogram is obtained for each image. Finally, the basic annotations of images are predicted based on image's content using Corr-LDA. We now present a brief introduction of Corr-LDA and how we have used it in our framework.

Table 1. Content classification

Event	Actual No. of Images/ Frames	Relevant Frames Tagged	Accuracy
Music Artist	588	476	.81
Celebrity	601	530	.88
Car Launching	594	495	.83

Correspondence LDA (Corr-LDA). The Corr-LDA model extracts conditional relationships between the set of image regions and set of words. In this, the features of an image are first obtained and then the corresponding words are generated. Let the size of dictionary be M . Assume that an image comprises of N finite regions. While annotating the images, first descriptors for various image regions are generated and then for each of the textual annotations, a region is selected. That is, for each of the M words, one of the regions is chosen from the image and subsequently a word is drawn conditioned on the topic that generates the selected region [10]. The Corr-LDA model was trained using the dataset of the type which one expects to find in entertainment/ automobile video. A total of 16 annotations were used to train the Corr-LDA model. Example of the annotations are guitar, car, person, stage, screen, sky, tree, trophy etc. The test images are then provided as input to predict the basic annotation based on image's content.

Tags: Car, Screen, Person, Stage

Tags: Person, Mike, Trophy, Screen

Tags: Person, Guitar

Fig. 4. Results of corr-lda

If $r=\{r_1, r_2, \dots, r_N\}$ denotes the set of image features, w denotes the set of associated words, $z=\{z_1, z_2, \dots, z_N\}$ is the set of latent variables, $y=\{y_1, y_2, \dots, y_M\}$ is the set of equiprobable indexing variables and θ is the Dirichlet random variable, then the joint probability distribution is given as follows:

$$p(r, w, \theta, z, y) = p(\theta | \alpha) \left(\prod_{n=1}^N p(z_n | \theta) p(r_n | z_n, \mu, \sigma) \right) \cdot \left(\prod_{m=1}^M (p(y_m) | N) p(w_m | y_m, z, \beta) \right) \quad (2)$$

Conditioned on N and M , Corr-LDA specifies the joint distribution on image regions, caption words and latent variables.

We have sampled videos into frames at the rate of 1 frame/ second to generate the images for our database. The annotations for the various video frames whose $p(\text{annotation} | \text{image})$ is more than a threshold value (0.1 in our case) are selected. Some of the annotations as obtained using Corr-LDA and the classification accuracy are shown in Fig. 4 and Table 1.

4.3 Merging of Context and Content

In this section, the query image or video is tagged with the relevant context and content information on the basis of their contextual meaning. To achieve this objective we assume V to be a set of I video frames or images $V=\{v_1, v_2, v_3, \dots, v_I\}$, C to be a set of J contents $C = \{c_1, c_2, c_3, \dots, c_J\}$ and S be a set of K contexts $S = \{c_1, c_2, c_3, \dots, c_K\}$. The content information $c_J \in C$ along with the tags obtained using Corr-LDA and the temporal and geographical context are associated to an image. This information gets tagged with a query video or an image $v_i \in V$ if their context matches i.e. S_J is equal to S_I . The proposed *Context Matching Algorithm* for associating a query video or an image with context and content information is explained below:

Algorithm 1. Algorithmic Steps

- 1 Input: Query Video or Image (ζ_i)
 - 2 Output: Tagged Video or Image (ζ_t), where (ζ_t) comprises of annotations as resulted from Corr-LDA (ϕ), temporal and geographical context i.e. the information about when and where the multimedia data was uploaded (ρ) and the relevant text news from the corpus of documents (σ).
 - 3 **Description:**
 - 1: The database of text files Γ is created and hLDA is applied over it.
 - 2: The output of hLDA is parsed.
 - 3: Hierarchy of hLDA gives different context along different paths. Let the total no of paths be Π such that $\eta \in \Pi$.
 - 4: Apply the Depth First Traversal to traverse every path of the hierarchy.
 - 5: Map the annotations ϕ (derived from Corr-LDA corresponding to the image/video frames) and tags ρ (temporal and geographical context associated to an image/ video frames) i.e. ($\phi + \rho$) onto η .
 - 6: The path η is selected if the context of ζ_i is matched with the context of the path η .
 - 7: Find the set of documents $\sigma \in \Gamma$ which correspond to the above path η .
 - 8: Tag ζ_i with ζ_t .
-

Table 2. Tagging accuracy

Event	Relavant Videos Tagged	Actual No. of Videos	Accuracy
Music Artist	32	40	.80
Celebrity	34	40	.85
Car Launching	33	40	.82

The basic annotations of images (frames) as obtained after applying Corr-LDA along with the tags (date, location etc.) provided with the video are matched with the topics present in event-context hierarchy as obtained using hLDA. Based on the contextual meaning the content and context information gets tagged to query video or an image. For each context class there is a separate lexicon which contains the frequently used keywords. For example, in weather news, keywords like rain, storm, temperature are common. These contexts are nothing but the topics that we obtained in hLDA, so topic-document probability can be obtained and document based on score of their probability can be linked to a query video or an image. The information like the video broadcast time, source and news event date is tagged with the video, to refine tagging and search processes. The system makes a news video repository, which allows users to search news videos.

Thus, following tags can be associated with a video:

- The date and location of the video
- A list of annotations as obtained using Corr-LDA
- A text summary of the video

Further, key frames can be stored and used to return search results.

5 Experimental Results

Initially, news videos were downloaded from YouTube website and frames were extracted from them. We have trained the Corr-LDA using 460 images. The training images were manually annotated using a list of 16 annotations (like person, car, sky, guitar, grass, tree, screen, trophy etc.). The calculated histogram vectors corresponding to each image feature along with the manual annotations are given as an input for training of the Corr-LDA. Approximately, 60 frames per video were extracted (typical video duration was 1-2 minutes). The histogram vectors for the test images have been calculated similarly. These histogram vectors are given as test input to the Corr-LDA. Using the image-annotation probabilities for the test images, only those annotations are considered which have a significant probability i.e., have image-annotation probability greater than a threshold value. In our case, we have considered the threshold value as 0.1. The key frames and hence the videos are classified into three given categories *viz* music artist event, celebrity event and car launch event. The category into

Tags: February, 2014, New Delhi, Car, BMW, i8, Car Launch Event
 BMW India presented the new generation of individual mobility by unveiling the BMW i8, the all-new BMW X5 and the all-new BMW M6 Gran Coupe at the 12th Auto Expo 2014...

Fig. 5. Car launch event february 2014

Tags: January, 2014, Mumbai, Person, Trophy, Celebrity Event
 Lifetime Achievement Award to Amitabh Bachchan, was presented by Vidhu Vinod Chopra and Shatrughan Sinha...

Fig. 6. Celebrity event january 2014

which most of the frames extracted from the video are classified is chosen as the video category. For extracting context for news videos, our system crawled GoogleNews website to retrieve text news related to the three categories mentioned above. Then we obtained a hierarchy of news context using hLDA on text news. The query video or image is tagged with the context and content information on the basis of contextual meaning using Context Matching Algorithm. The context and content information thus gets tagged to a query video or an image.

We have considered 40 videos per category for experimentation out of which in the paper 2 illustrations have been presented as depicted in Figs. 5 and 6. The tagging accuracy is shown in Table 2.

6 Conclusion

In this paper, we have proposed an approach for tagging multimedia data based on its context and content information. The query video or an image is tagged automatically without any manual intervention on the basis of contextual meaning. We have discussed an application for tagging multimedia data that utilizes context matching algorithm for associating a resultant video or an image with the appropriate content and context information.

References

1. Steinmetz, N., Sack, H.: Semantic multimedia information retrieval based on contextual descriptions. In: Philipp, P., Oscar, C., Valentina, P., Laura, H., Sebastian, R. (eds.) ESWC 2013. LNCS, pp. 382–396. Springer, Heidelberg (2013)
2. Giannakidou, E., Kompatsiaris, I., Vakali, A.: Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems. In: 2008 IEEE International Conference on Semantic Computing, pp. 128–135, IEEE (2008)

3. Aytar, Y.: Semantic video retrieval using high level context. In: *Semantic Video Retrieval Using High Level Context* (Doctoral dissertation, University of Central Florida Orlando, Florida) (2008)
4. Miotto, R., Orio, N.: A Probabilistic approach to merge context and content information for music retrieval. In: *Proceedings of ISMIR*, pp. 15–20 (2010)
5. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimed. Comput. Commun. Appl.* **2**(1), 1–19 (2006)
6. Nguyen, C.T.: Bridging semantic gaps in information retrieval: context-based approaches. In: *ACM VLDB*, p. 10 (2010)
7. Siersdorfer, S., San Pedro, J., Sanderson, M.: Automatic video tagging using content redundancy. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 395–402, ACM (2009)
8. Pedro, J.S., Siersdorfer, S., Sanderson, M.: Content redundancy in YouTube and its application to video tagging. *ACM Trans. Inf. Syst. (TOIS)* **29**(3), 47–58 (2011). Article No. 13
9. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: *NIPS*, Vol. 16 (2004)
10. Blei, D. M., Jordan, M. I.: Modeling annotated data. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 127–134, ACM (2003)
11. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM (JACM)* **57**(2), 7 (2010)