Using a Lightweight Multimedia Content Model for Semantic annotation

Smitashree Choudhury, John G. Breslin

DERI, National University of Ireland, Galway
IDA Business Park, Dangan, Galway, Ireland
smitashree.choudhury@deri.org

DERI, National University of Ireland, Galway
IDA Business Park, Dangan, Galway, Ireland
john.breslin@deri.org

Abstract

In this paper we discuss the use of a multimedia content model for automatic extraction of semantic metadata from multimedia content. We developed a modular and extensible framework to model the content feature of multimedia data and also describe the way it can be integrated with other existing vocabularies. The goal of this model is to generate sufficient understanding of media content, its context and its relation to domain knowledge in order to perform multimedia reasoning. We implemented a tool that analyzes and links low-level descriptions to higher-level domain specific semantic concepts by means of statistical learning and clustering analysis. Experimental result shows the approach performs well in visual concept prediction in the image which can be further augmented with other information sources such as context text and or audio source.

Keywords: Ontology, multimedia annotation, semantic web, image annotation, Concept learning

1 Introduction

With the explosive growth of the web in terms of image and videos in recent years it makes imperative to filter, process and personalize the media as much as possible in order to save precious time from being wasted in the ongoing process of searching and browsing. To enable multimedia content to be discovered and reused by services and applications it needs to be semantically described in machine readable form. Automatic generation of multimedia content description is inherently complex due to its complex dynamics and subjective interpretation by users. Significant progress has been made in content based feature extraction and media segmentation. Semantic Web effort can make use of these developments to describe the media structure and content in an effort to bridge the semantic gap with reasoning over low level and contextual features, for e.g. we can extract the image content features and compute the probability of the image content to the higher semantic classes such as building, people, vegetation etc.
Recently efforts have increased manifold to tap these non textual mediums for various possible knowledge based application in biomedicine, security, e-science in the entertainment industry.

In understanding multimedia content, combination of ontological infrastructures and statistical learning theory with background knowledge has attracted large research interest recently. As part of the present study we will discuss those papers which closely resemble our work. Most of the multimedia ontology developed so far has tried to extend some subset of the MPEG7 [8] specification. One of the initial

attempts made in [9] to convert MPEG-7 specification into machine readable format. A Visual Description Ontology is described in [1] modeled to represent the visual part of the MPEG 7. Multimedia Ontology (COMM) is described in [2] for describing the multimedia data on web. Visual information Object modeled in [3] is also extended from Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) as in [2]. We have adopted very light weight yet holistic approach of modeling multimedia data by integrating multiple vocabularies yet keeping it simple and lean for obvious reason.

For the concept learning part we refer to some of the recent works on automatic image annotation and concept detection in video data. Most of the studies in computer vision and video retrieval research concept detection in videos are based on classification of each and every single visual concept with uni-modal or multimodal approach and adopting some kind of fusion method. Co-occurrence model described in [4] is based on the co-occurrence of visual features and words. Duygulu et al. [5] used machine translation approach to establish one to one correspondence between image region and visual keywords. Jeon et al. [6] attempted to compute the joint probability of blobs and concepts in order to predict the concepts in the image. Fan et al. [7] proposed hierarchical classification framework for automatic image annotation.

Since we have a light weight video model based on which we want to annotate the multimedia data (both image and video) we have deliberately avoided the single concept detector approach for each visual concept adopted by many video retrieval studies. Instead our work is more aligned to the approach adopted by Image annotation researchers to identify the concept in an image. Our approach in concept learning also counts the joint probability of semantic concepts and multiple visual feature space but at the same time it also depends on the concept co-occurrence space in order to refine the final prediction.

The rest of this paper organized as follows: section 2 provides the multimedia object model and its class structure description section 3 illustrates the Concept learning Framework (CLF). Section 4 discusses the training and evaluation phase followed by conclusion and future work in section 5.

2. Ontology model

The model under discussion is a light weight multimedia ontology which satisfies all the basic requirements of a multimedia ontology and gives sufficient scope for fully integration with domain knowledge and other existing vocabularies such as Dublin core, Simple Knowledge Organization System (SKOS) [12] and Friend of a Friend (FOAF) [13] as and where needed.
This section describes the underlying model used in the concept learning framework. The model concepts encapsulate the multimodal nature of multimedia data for data representation. Requirements for a multimedia ontology are modularity, interoperability, extensibility and separation of domain knowledge from multimedia document depicting the content. We describe both the model and its requirement satisfaction below.

2.1 Class Structure:

In this section we will discuss the model and its design issues in brief. The model based on the concept of *MultimediaObject* which (which is also subclass of a FOAF document class) has five subclasses, *Video, Audio, Image, Segment, Feature*. Video, Audio and Image are different classes used to represent media content where as segment is the super class of the classes results from structural decomposition of media objects such as Shot, Frame and Still Region. Each media object is described with its global properties such as *language, duration, copyright, creator, genre* and contextual properties such as available web *text, subtitles ,reviews, awards*, source, date, ratings etc. the prime focus of our model is to facilitate content modeling. The class structure of the model is described in Figure 1 below.

Multimedia Object Description ontology is aimed to integrate the multimedia objects on the web with other information objects in order to give an interlinked and integrated view of the user information needs. This is only possible when we model not only the media object used to represent the content but also the content abstracting the higher semantic concepts of specific domain. The model is aimed not only helps to retrieve and browse the media objects based on its properties but also to filter and segment it with its semantic description.

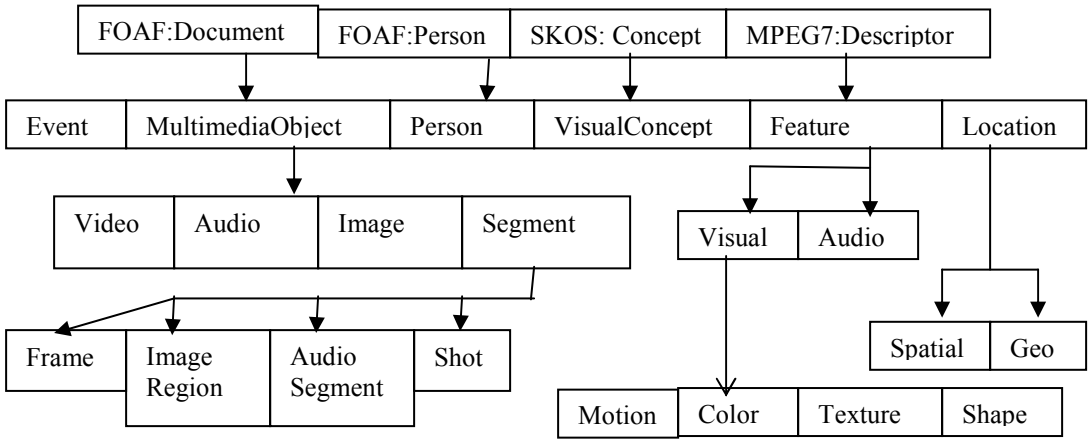

Figure 1: Multimedia Ontology structure

MultimediaObject is the main class which has four subclasses such as *Video, Audio, Image, Segment. Feature* class is an abstract class to describe the low level features of different objects.

Feature class can be integrated with MPEG-7 Descriptors. This class subsumes *VisualFeature* and *AudioFeature*. *VisualFeature* can be further specialized into classes such as *ColorFeature, TextureFeature, ShapeFeature and MotionFeature* where as *AudioFeature* can be further specialized into *SpectralFeature, CepstralFeatures* as and when necessary to describe the audio content features.

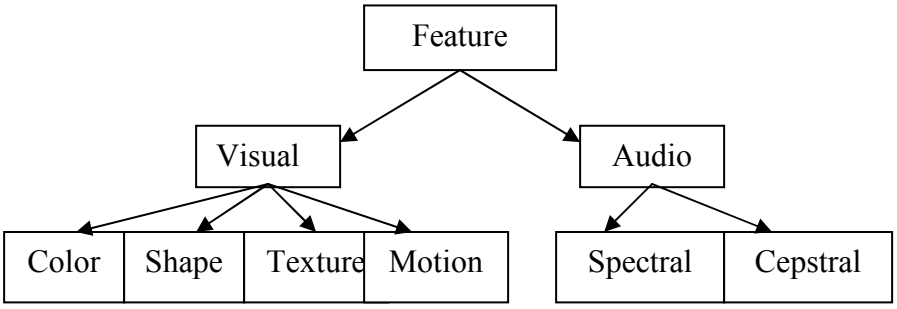

Figure 2: Feature Class structure

VisualConcept class is the entry point for describing the higher level semantic concepts in a multimedia object. A visual concept unlike a linguistic concept may be simple (mountain, boat, car, flower) or complex where many sub concept form a visual concept such as 'explosion', 'earthquake', 'meeting'. A VisualConcept instance can be described in terms of SKOS Concept class to integrate with domain knowledge. However VisualConcept class also carries attributes such as *co_depictsWith* to describe what the other visual concepts are appears in time space dimension.

VisualConcept:
subClassOf: ***skos: Concept***

hasSemanticLabel : Literal
co_depictsWith : VisualConcept
hasSubConcepts :VisualConcept

Event class of the model attempts to describe the semantic content in terms of events which is interplay between objects and actions or other sub-events. Event detection is a challenging issue for multimedia processing and retrieval community. Any attempt to model events has to consider complex interaction between objects and agents in time and space. An *Event* has temporal attributes such as *startTime* and *endTime, hasDuration occursAt, occursAfter, occursBefore.*

Event:

hasSemanticLabel: Literal
hasStartTime :Time
hasEndTime :Time
involvesObject:VisualConcept
hasSubEvents :Event

Location class is another major concept in the model which has 2 subclasses called *SpatialLocation, GeoLocation*. There are two type of location information, one is the *GeoLocation* as depicted in the media such as a *City , Street* or *Country* and the other is *SpatialLocation* which describes the spatial properties of a region in terms of geometric coordinates. *SpatialLocation* has attributes such as *xCoordinates* and *yCoordinates* and other spatial attributes specifying directional semantics.

Segment class is the broader class for both spatial segments and temporal segments such as Image region, Frame and Shots, Audio segments.

2.2 Interoperability and Integration

Present model proposes clear but simple distinction between the domain ontological concept and the media structure components for e.g. a still region of a frame is an instance of the media segment which depicts a *VisualConcept* of type *SKOS:Concept* linking to a particular domain. Below are some of the examples of interlinking properties of the model.

1. Video, Image, Audio are subclass of *MultimediaObject* which is a subclass of *FOAF: Document*.
2. Creator, Director, User class of the model are subclass of *FOAF:Person* who can be further linked with other works of the same Person.
3. *ImageRegion :depictsConcept* property takes the value in the range of *VisualConcept* which is a subclass of *SKOS:Concept* in order to integrate with domain knowledge
4. *ImageRegion* can be described with *"sameAs*" class of StillRegion of MPEG -7.

2.3 Multimedia Reasoning

Proposed model can help us to reason about the semantic concepts depicted in the media as follows:

1. Mapping of low level features of an image region or any media segment helps to classify and predict the color, shape or category of an underlying semantic object or event.
2. Spatiotemporal information helps further for detection and interaction of different objects in the frame and scene.

3. Concept Learning Framework (CLF)

In order to prepare a use case for the proposed model we have implemented a simple java based Concept Learning Framework. Concept learning can be considered as a classification problem where the input is classified into most likely group and the predictions are made on the basis of fuzzy class membership. CLF extracts the low level visual content of an image or Frame and predicts the possible depicted concepts with certain degree of relevance.

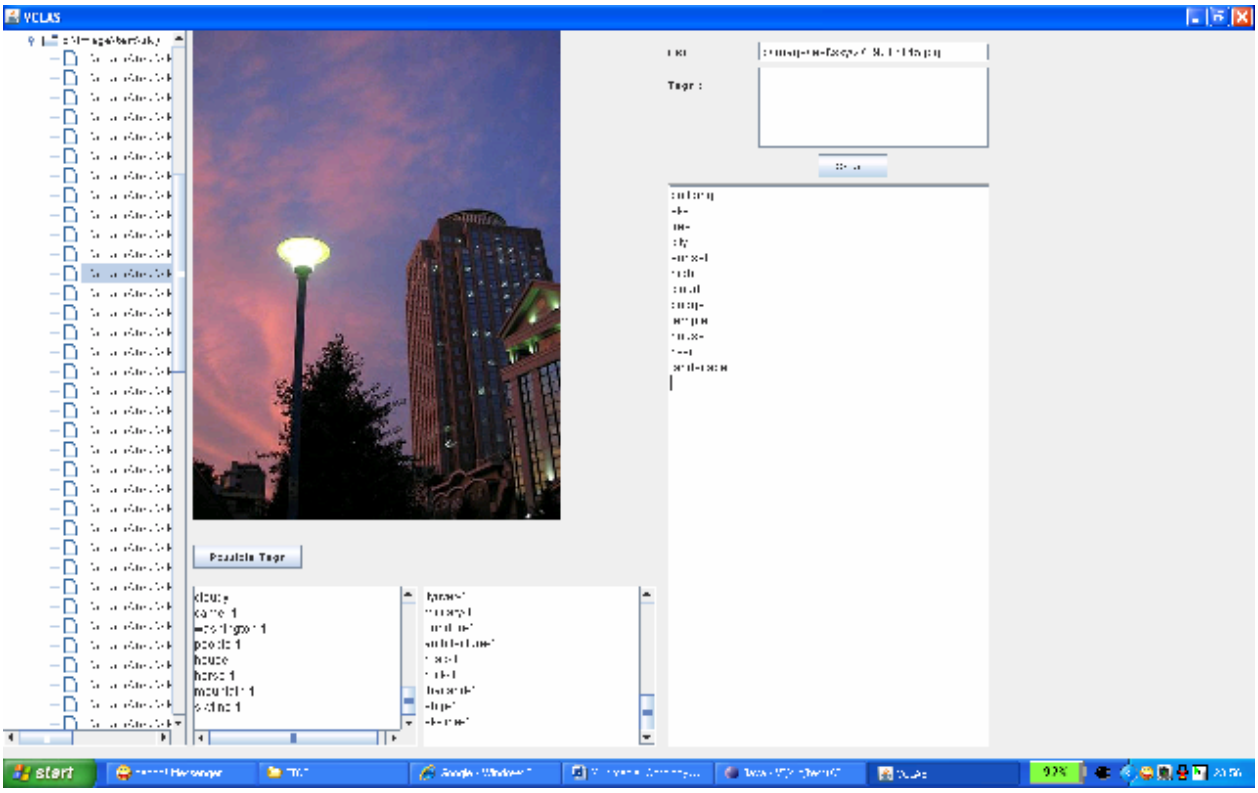

Figure 3: Visual Concept Learning UI

In the present form the tool extracts and uses the different global features of the image to suggest the labels without any prior segmentation. Localizations of object in the image and linking to specific domain concept will be the part of our future work.
The suggested labels can be added to values of the property "*depictsConcepts*" with the degree of probability.

3.1 Visual Concept

Our set of visual concepts is data driven and built incrementally. There are 520 unique keywords labels in our training data. However for more precise evaluation we grouped them under 20 broader visual concepts (animal,building,city/urban,car,bus,mountain,snow,boat/ship,airplane,people,desert,garden/park,beach,road,bridge,flower,explosion,waterfront) manually, for e.g. keywords such as church, house, skyscraper, temple, apartment, office are grouped under the concept 'building'.

3.2 Frame feature and visual concept association:

To describe the content of the frame / image we have selected two feature spaces to represent because of their good and reliable performance and wide use in distance matching. We have used only global feature distribution to learn the presence or absence of the concepts in the frame for simplicity reason. The two feature spaces are MPEG7 edgehistogram [14] and color-corellogram [15] used to create the feature vector. Edge histogram gives good result for object centered frames where as color corellogram which takes into account the spatial correlation values of the color has proved to be good for the nature based frames and images. When the input frame comes with these feature vectors the system computes the Euclidian distance (equation 1) between the queried frame and the training sets in order to retrieve the more similar frames and their labeled concepts.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \quad (1)$$

We retrieve two result sets from two feature space of the image. Two result sets are fused and the stronger cluster gets higher ranking. The final visual concepts are given as part of the frame content with the degree of confidence which is described as the probability value of the domain concepts. Thus we have combined both the low level features and the contextual properties of a visual concept in order to learn the gap between content and visual concepts from the frame. We have only implemented the global feature extraction and concept association. Object localization in the frame can be our future work which includes adaptive segmentation and concept association.

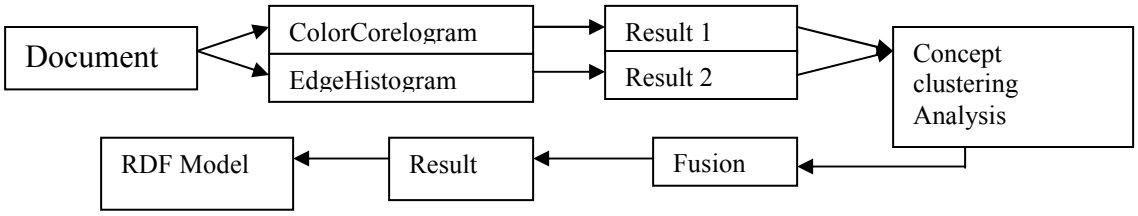

Figure 4: Concept Learning Architecture

4. Experimental Setup

4.1 Training

Training data set consist of 5000 training images collected from varied sources including video frames and Flickr images covering 20 plus broad visual concepts and most of the visual concepts also includes sub concepts**.** Two step training process includes signature extraction and manual concept labeling.

1. Extract a signature for each image *i*, $i \in \{1, 2, ..., N\}$. The signature consists of two discrete distributions, one of 256 bin color corellogram features, and the other is 80 bin edgehistogram features.

Edges in images constitute an important feature to represent their content. Also, human eyes are sensitive to edge features for image perception. One way of representing such an important edge feature is to use a histogram. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. *Color correlogram* combines color information with spatial layout while retaining the advantages of histograms. It computes the spatial correlation of pairs of colors as a function of the distance between pixels [5].

2. Each frame / image is manually annotated with 4-8 keywords

We have decided to take global feature of the image instead of region based local feature because we need to predict the probability of the presence of the concept in the image rather than localizing the object. It can be considered as image classification task rather than object recognition. In this way we avoided the segmentation of image.

4.2 Testing and Evaluation

In our work we used two different sets of data for training and testing. Although several automatic and semi-automatic image annotation studies carried out recently but a comprehensive quantitative evaluation is absent in public domain. The online demo from Penn State University is ALIPR [16], can be used for a comparative evaluation. Initial evaluation for different visual concept categories shows that our system predicts most of the relevant labels within top 10 ranked labels as compared to 15 of ALIPR, but a more comprehensive evaluation is required.

In the present study we have adopted a simple evaluation approach to test the framework. The evaluation aims to see the performance of the CLF in terms of given an image *I* what are the top 10 suggested labels .For simplistic purpose we designed 3 scale evaluation 'Good' indicates that the expected label appears within top 5 labels, 'poor' indicates that the concept label appears within 5-10 ranks and the third measure is absent of the concept from the predicted list. Even though the result set retrieves 4-5 relevant concepts for each image, present evaluation is only restricted to the major visual concept appear in the image.

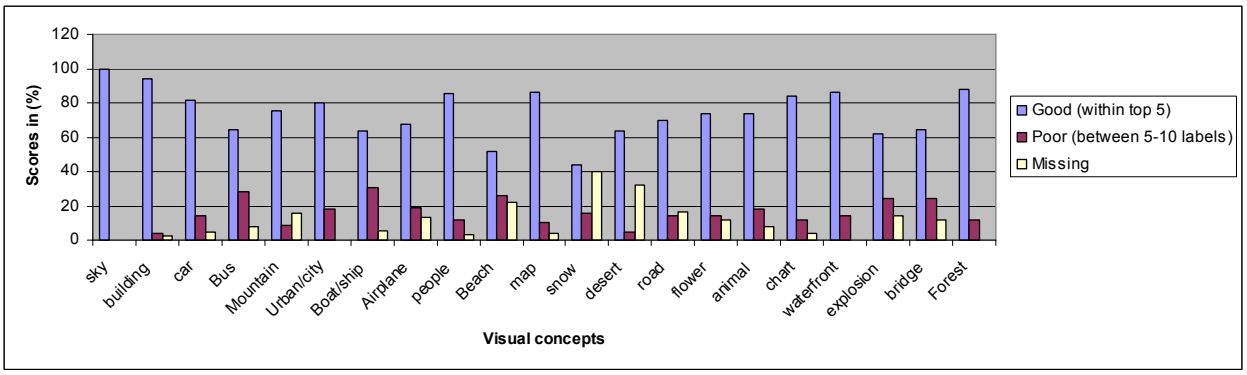

Figure 5: Visual concepts and their evaluation score (%).

4.3 Analysis of results

The result showed some visual concepts such as Sky, building, city/urban, people, map, chart, waterfront give very good result where more than 80 % of cases predict the correct label within top 5 results. The poor results for many concepts such as desert, snow/glacier, bridge and flower may be due to the intra class visual variations. Where the intra class variations are great we can explore the hierarchical image classification approach as adopted in [7] in our future work.

5. Conclusion and Future work

In this paper we presented a multimedia concept annotation system based on light weight multimedia object ontology to describe the content and its interlinking properties to domain specific visual concepts. We have used some of the visual concepts aligned with concepts developed as part of the LSCOM project [11] to annotate the visual images and videos. This work greatly alleviates the time consuming process of

manual annotation of multimedia data by extracting the low level global features and linking them to the visual concepts.
Our future works will be mainly focused few areas such as efficient integration of spatial-temporal attributes in our model and multi modal approach for robust concept detection.

References

[1] Simou, N.(2005). A Visual Descriptor Ontology for Multimedia Reasoning.
[2] Arndt, R.(2007). COMM: Designing a Well-Founded Multimedia Ontology for the Web. In Proc. of the 6th International Semantic Web Conference (ISWC'2007), Busan, Korea.
[3] Bertini, M.(2005). Ontologies Enriched with Visual Information for Video Annotation.In proc.of the
[4] Mori, Y., Takahashi, H. and Oka, R. (1999). Image-to-word Transformation Based on Dividing and Vector Quantizing Images with Words.. *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*.
[5] Duygulu, P., Barnard, K., Fretias, N. and Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Proceedings of the European Conference on Computer Vision*: 97-112.
[6] Jeon, J., Lavrenko, V. and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*: 119-126.
[7] Fan, J. (2008). Hierarchical classification for automatic image annotation. in Proc.of the International Conference on Image and Video Retrieval (CIVR,08).
[8] MPEG-7: Multimedia Content Description Interface. ISO/IEC 15938 (2001)
[9] Hunter, J.(2001). Adding multimedia to the semantic web -building an mpeg-7 ontology. In International Semantic Web Working Symposium.
[10] Jeon, J. and Manmatha, R.(2004). Using maximum entropy for automatic image annotation. CIVR, pp. 24-32, 2004
[11] Hauptmann,A.G.(2004).Towards a large scale concept ontology for broadcast video. In Proc. of the 3rd int conf on Image and Video Retrieval (CIVR'04).
[12] www.w3.org/TR/swbp-skos-core-guide
[13] Brickley, D., Miller, L.: FOAF Vocabulary Specification: Working Draft (2005), http://xmlns.com/foaf/0.1/.
[14] Manjunath, B.S. (2001). Color and Texture Descriptors .IEEE Transactions on Circuits and Systems for Video Technology: 11.
[15]Huang, J. (1998). Combining Color and Spatial Information for Content-based Image Retrieval. Available online at http://www.cs.cornell.edu/rdz/Papers/ecdl2/spatial.htm.
[16] Li. J., and Wang. J.Z., (2008). "Real-time Computerized Annotation of Pictures". *IEEE Trans. on Pattern Analysis and Machine Intelligence*.