

Study of Various Video Annotation Techniques

Khushboo Khurana¹, M. B. Chandak²

Student, M.Tech (CSE), RCOEM, Nagpur, India¹

Associate Professor, CSE department, RCOEM, Nagpur, India²

ABSTRACT: Image annotation is an active field of research that serves as a precursor to video annotation. With the increase in the number of videos and important information present in them, there is need to annotate the videos. Annotation improves the efficiency of searching and retrieving the video. Video features are often inspired and sometimes directly borrowed from image techniques and many methods for image indexing are also easily applied to video. A human operator can specify annotations such as time, location and activity. More sophisticated annotations can be provided using automatic or semi-automatic techniques. There are various techniques that can be used for annotation; this paper discusses some of the techniques. Technique which employs the ontology, requires ontology language, web ontology language (OWL) is the ontology language that has gained importance. This paper also discusses ontology and knowledge base editor- Protégé.

Keywords: Video Annotation, Annotation Techniques, Metadata, Ontology, OWL

I. INTRODUCTION

Annotation is a critical or explanatory note or body of notes added to a text, diagram, document, image or video. Annotation implies to attach data to some other piece of data (i.e. add metadata to data) to simplify its access. Video annotation refers to the extraction of the information about video, adding this information to the video which can help in browsing, searching, analysis, retrieval, comparison, and categorization.

There is a lot of video content available. Education, news, medical, surveillance, disaster management videos are some of the categories. As the number of videos is increasing, difficulty in searching desired information is also increasing. Video annotation is done to help semantic retrieval of videos from a large video database. For example, response for a semantic query to search for a video in which suspect is running wearing a red shirt can be very easy if the video is annotated semantically. Video annotation levels can be- complete video, scene, groups, shots or frame level [1]. Video consists of audio, visual and sometimes textual information. In this paper we have presented a study of annotation techniques that use visual or textual information from the video.

Annotation files are generally extensible markup language (XML). These can be referred for the retrieval of desired image or video. Various projects are now focusing on image and video annotation. Datasets for training and testing are also available. LabelMe is a project created by the MIT Computer Science and Artificial Intelligence

Laboratory (CSAIL) which provides a dataset of digital images with annotation [2].

The rest of this paper is organized as follows. Section II describes the metadata that can be attached to videos. Section III presents the video annotation techniques and the related work done. In Section IV we have discussed a general block diagram for ontology based video annotation, available languages and tools for ontology creation. Finally, we conclude in Section V.

II. TYPES OF METADATA FOR VIDEO ANNOTATION

Metadata is attached to video for making it easy to access. Different types of information that can be associated with videos or images are:

- *Content independent metadata* is related to the image or video, but does not describe it directly. For example, name of author, date, location, etc. [3]. It cannot be extracted from the image or video.
- *Content dependent metadata* refers to low-level and intermediate-level features.

Various low-level features can be found from the video and from individual video frames. These features can be used for annotation. Low-level features that can be used are shape, color, texture, edge, motion, etc. MPEG-7 visual descriptors can also be used. MPEG-7 color descriptor [4] and edge descriptor are commonly used.



Low-level features such as SIFT descriptors and Histogram of Oriented Gradients (HoG) can also be used. HoG features have been popular and effective choice for various groups participating in Pascal Visual Object Classes Challenge [5]. Many features can be combined to form a feature vector which can be given as input to machine learning systems to form final annotations.

- *Content descriptive metadata* refers to content semantics. It is related to the content of the image or video, the entities or objects, events, emotions and meaning of scenes.

In the paper we focus on this type of information and study the techniques which annotate the video semantically. The techniques described in this paper may extract some low-level features and do further processing on them to annotate video with content descriptive information.

III. VIDEO ANNOTATION TECHNIQUES

A. Free Text Descriptions

Free textual descriptions may be added to video. There is no pre-defined structure for the annotation [6]. For example, while uploading a video on YouTube the user can add description about the video. Any combination of words or sentences can be used. Such type of annotation helps in accessing the video. Since no structuring exists, annotation is an easy task but, efficient retrieval techniques must be used.

The IAPR-TC12 dataset of 20,000 images [7] contains free text descriptions of each image in English, German and Spanish.

B. Based on Text in the Video

The textual information that exists in images and video sequences are called collateral text. This is used so that keywords, and potentially richer representations, are extracted from text fragments. Consider, for example, the text of news, documentary programs, movies and even newspaper film reviews [8]. Textual data is a source of highly semantic information and thus, if available, would allow the filtering and searching of video data by users in a more intuitive and natural way. Text embedded in images and video, especially captions provide brief and important content information, such as the name of players or speakers, the title, location, date of an event, etc. [9].

Rainer Lienhart and Wolfgang Effelsberg[10], proposed text extraction method based on segmentation. Color is used to form homogeneous regions, texture and motion analysis is done for automatic segmentation of text in digital videos. The output is directly passed to a standard optical character

recognition (OCR) software package in order to translate the segmented text into ASCII.

In [11], segmentation method based on Markov random field to extract more accurate text characters is used. This methodology allows handling background gray-scale multi-modality and unknown text gray-scale values. Support vector machine (SVM) is used for text verification followed by traditional OCR algorithm.

A hybrid wavelet/neural network are used to locate text in videos; high-frequency coefficients are the input to the network [12]. Annotation of images and videos without OCR is proposed in [13] where word-images extracted from visual documents are matched with keyword images. The text equivalent of the keywords is used to annotate the video, clustering is used to identify words with the same stem.

The semantics can be extracted from video text. Sunitha Abburu [6] used super imposed text to extract semantics of the video which increased the efficiency of retrieval system. A semiautomatic method to generate annotation for cricket videos has been proposed, semantics are extracted by analyzing the text content using the rule based approach.

C. Based on Machine Learning

Low-level features can be extracted from the video or image. Various machine learning techniques such as support vector machine (SVM), Bayesian networks, Clustering, similarity and metric learning can be used.

In [14], a framework for semantic video event annotation is presented, which exploits global feature, local feature and motion feature. Using these features, video clip can be encoded as a set of feature vectors. Then according to different features, SVM classifiers are trained, and a bi-coded chromosome based genetic algorithm is performed to obtain optimal classifiers and relevant optimal weights based on training stage. With the optimal classifiers set and optimal weights, the maximum similarity between video clip in original database and unlabeled video clip is considered to be the final label result.

In [15], annotation is a supervised learning problem under Multiple-Instance Learning (MIL) framework. A novel Asymmetrical Support Vector Machine-based MIL algorithm is proposed, which extends the conventional Support Vector Machine. By maximizing the pattern margins subject to the MIL constraints, ASVM-MIL converts the MIL problem to a traditional supervised learning problem.

Sabine Barrat and Salvatore Tabbone [16], classifying weakly-annotated images, where just a small subset of the database is annotated with keywords. In this paper a new method by integrating semantic concepts extracted from text and by automatically extending annotations to the images



with missing keywords is proposed. The model is inspired from the probabilistic graphical model theory. Bayesian networks are used.

D. Based on Rule Learning

Visual features can be directly extracted from the video or images. These low-level features can be used for annotation but gap exists between the information that can be extracted automatically from visual data and the interpretation that the same data has for a user in a given situation: *the semantic gap* [17]. Rules are built to infer a set of high-level concepts from low-level descriptors.

Jardon *et al.* introduced a rule-based approach for the generation of inference rules using fuzzy logic [18]. However; the used knowledge representations are predefined and static, limiting the adaptability to different contexts.

A rule based video annotation system is proposed in [19]. The proposed system annotates video sequences automatically using knowledge from a pre-annotated dataset. It creates representations from a set of low-level video features and infers the association rules between them and high-level concepts from a predefined lexicon.

In [20], learning by means of Fuzzy Decision Trees (FDT), automatic rules based on a limited set of examples is proposed. Rules intended, in an exploitation step, to reduce the need of human usage in the process of indexation.

Occurrence of some audiovisual features demonstrates remarkable patterns for detection of semantic events. Monireh-Sadat Hosseini, Amir-Masoud Eftekhari Moghadam [21] present an approach for event detection and annotation of broadcast soccer video. A fuzzy rule-based reasoning system is designed as a classifier which adopts statistical information from a set of audiovisual features as its crisp input values and produces semantic concepts corresponding to the occurred events. A set of tuples is created by discretization and fuzzification of continuous feature vectors derived from the training data. We extract the hidden knowledge among the tuples and correlation between the features and related events by constructing a decision tree (DT).

E. Based on Graph

Graph-based learning is a semi-supervised method. Graph with labelled and unlabeled vertices are used. These vertices are samples; the edges reflect the similarities between sample pairs. A function is estimated on the graph based on a label smoothness assumption. These methods have already been successfully applied in image and video content analysis [22], [23].

In [24], Weng *et al.* proposed a method to learn the inter-concept relationships and then used graphical model to improve the concept annotation results.

Video annotation mainly aim at the assignment of single or multiple concept labels to a target data set, where the assignment is often done independently without considering the inter-concept relationship. Due to the fact that concepts do not occur in isolation (e.g., smoke and explosion); context-based video annotation with graph diffusion process is used [25].

A graph reinforcement method driven by a particular modality (e.g., visual) is used to determine the contribution of a similar document to the annotation target. The graph supplies possible annotations of a different modality (e.g., text) that can be mined for annotations of the target [26]. Filtered tags are used; they are superior to a state-of-the-art semi-supervised technique for graph reinforcement learning on the initial user-supplied annotations.

In [27], a multi-graph based semi-supervised learning method is proposed. This framework amounts to fusing graphs and then conducting semi-supervised learning on fused graph.

F. Based on Ontology

Ontology is defined as an explicit specification of a conceptualization [28]. It is a large classification system that classifies different aspects of life into hierarchical categories. This is similar to classification by keywords, but the fact that the keywords belong to a hierarchy enriches the annotations [29]. For example, it can be found that a “room” is a subclass of the class “house”. Ontology consists of entities and their relationships, which may be organized as classes and subclasses, each class may also consist of one or more instances.

In [30], a framework for ontology enriched semantic annotation of CCTV video is proposed. Visual and text semantics are linked with appropriate keywords provided by domain experts. Video segmentation is done to find moving objects, which are classified as agent, action and recipient. These visual semantics are annotated by keywords of CCTV ontology.

Video annotation based on ontology can also certain rules and/or machine learning. Semantic concept detectors can be linked to corresponding concepts in the ontology [31]. A rule-based method for automatic semantic annotation is used; rule learning is built in SWRL. Concepts’ relationship of co-occurrence and temporal consistency of video are used to improve performance of individual concept detectors.

The system in [32] detects objects in video images captured in vehicular traffic situations, maps them to Open



Cyc ontology and generates descriptions of traffic scene in CycL language. Spatio-temporal rules are used for object classification. The objects are classified into four categories: car, pedestrians, poles and other objects.

Jin-Woo Jeong, Hyun-Ki Hong, and Dong-Ho Lee [33], have presented an automatic video annotation technique which employ ontology to facilitate video retrieval and sharing process in smart TV environment. Multimedia contents are represented through a well-structured ontology. The visual features are extracted from MPEG-7 visual descriptors. These features are mapped to semi concept values. Semantic inference rules and support vector machine are used to detect the high level concepts. In [34], method proposed consists of ontology based on image domain and low-level features extracted from the images. The image annotation problem is transformed into an image retrieval one for annotation purpose. In [35], events and objects in the video are used for annotation.

IV. ONTOLOGY BASED VIDEO ANNOTATION TECHNIQUE

A. Block Schematic

Annotation system based on ontology is shown in fig.1, the video is provided as input. Key-frames are extracted from the video. Key-frames are the still frames from the video that contain the important information of video. Feature extraction is extraction of various low-level features from the still image frames; a feature vector is the output of this module, which is given to the annotation module. The annotation module provides annotation based on ontology. Machine learning technique and rules can also be used. Ontology can be constructed or existing ontology can be used. One or more ontologies can be employed.

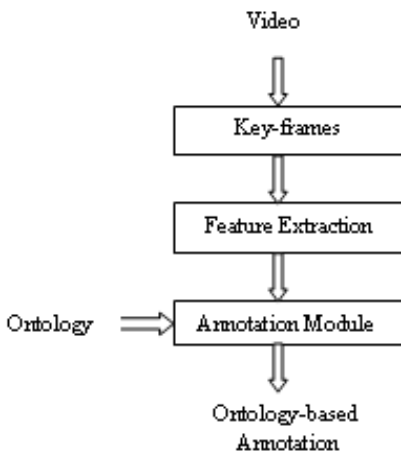


Fig.1. Block Schematic for Ontology based Video Annotation

B. Ontology Languages

Ontology languages are formal languages used to construct ontologies. Several ontology languages have been developed in the last few years. Various ontology languages are DAML+OIL, Ontology Interface Layer (OIL), Resource Description Framework (RDF), RDF Schema (RDFS), SHOE [36]. Web ontology language (OWL) [37] is a new ontology language for the Semantic Web, developed by the World Wide Web Consortium (W3C) Web Ontology Working Group. OWL language is the advanced version of DAML+OIL (DARPA agent mark-up language). These languages use a markup scheme to encode knowledge, most commonly with XML.

OWL ontologies may be categorised into three sub-languages: OWL-Lite, OWL-DL and OWL-Full. A defining feature of each sub-language is its expressiveness. OWL-Lite is the least expressive, OWL-Full is the most expressive. The expressiveness of OWL-DL falls between that of OWL-Lite and OWL-Full. OWL-DL may be considered as an extension of OWL-Lite and OWL-Full an extension of OWL-DL.

Every individual in the OWL world is a member of the class owl:Thing. Thus each user-defined class is implicitly a subclass of owl:Thing. Classes, subclasses, individuals can exist. These may also have properties. OWL language guide [37] shows how ontologies can be created and its properties. For example, to define PotableLiquid (liquids suitable for drinking) to be subclass of ConsumableThing it will be written as follows:

```

    <owl:Class rdf:ID="PotableLiquid">
        <rdfs:subClassOf rdf:resource="#ConsumableThing"
    />
    ...
    </owl:Class>
    
```

In addition to classes, their members can be defined these are individuals. An individual is minimally introduced by declaring it to be a member of a class.

```

    <owl:Thing rdf:ID="CentralCoastRegion" />
    <owl:Thing rdf:about="#CentralCoastRegion">
        <rdf:type rdf:resource="#Region"/>
    </owl:Thing>
    
```

rdf:type is an RDF property that ties an individual to a class of which it is a member. In addition to these many properties and constraints can be added [37].



C. Available Ontologies

There are a number of ontologies available that can be downloaded and used for research purpose. The LSCOM Large scale Concept Ontology for Broadcast Video effort has produced an ontology of 1,000 concepts that's proving to be a valuable resource for the multimedia research community. The teams have used approximately 450 concepts to manually annotate a large corpus of 80 hours of broadcast news video [38]. Concept definitions/concept detectors can also be downloaded and used to train the machine learning algorithms. WordNet [39], Cyc or Open-Cyc [40], SUMO [41], and domain-specific ontologies and taxonomies such as the GeneOntology (<http://www.geneontology.org>) also exist.

D. Protégé Ontology Tool

Available ontology can be used; Protégé is an ontology and knowledge base editor produced by Stanford University. Protégé is a tool that enables the construction of domain ontologies, customized data entry forms to enter data. Protégé is based on Java, is extensible, and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development.

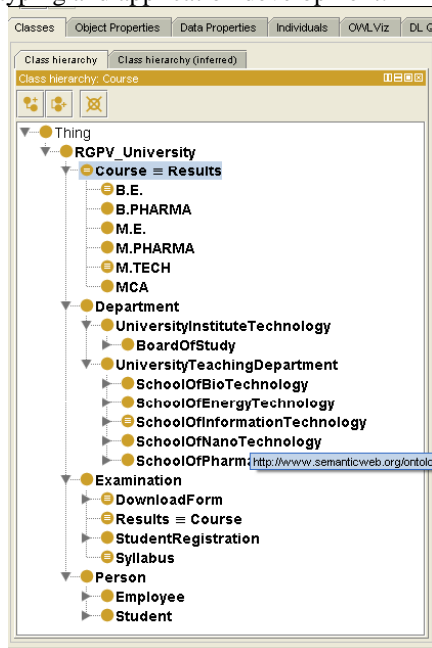


Fig. 2 Ontology classes for university (from [43])

Protégé allows the definition of classes, class hierarchies, variables, variable-value restrictions, and the relationships between classes and the properties of these relationships. Protégé is free and can be downloaded from

[42]. Protégé comes with visualization packages, which help the user visualize ontology with the help of diagrams. The Protégé-OWL editor enables users to load and save OWL and RDF ontologies, edit and visualize classes, properties, and SWRL rules, define logical class characteristics as OWL expressions, execute reasoner such as description logic classifiers.

Available ontologies can be loaded, integrated in this tool. Protégé plug-in is also available for download. Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema. Ontology classes for university are show in Figure 2, created using Protégé tool.

V. CONCLUSION

We have studied various video annotation techniques. These annotation techniques improve the access to the videos and are more efficient compared to manual annotation. Ontologies developed and Protégé tool for ontology creation, visualization is discussed. We have also presented a general block schematic for ontology based video annotation. Using ontology based technique help in semantic annotation. Machine learning or rule learning along with ontology can give better annotations compared to other techniques.

REFERENCES

- [1] X. Zhu, J. Fan, X. Xue, L. Wu and A. K. Elmagarmid, "Semi-automatic video content annotation", Proceeding of Third IEEE Pacific Rim Conference on Multimedia, pp. 37-52, 2002.
- [2] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, 2008.
- [3] A.del Bimbo, "Visual Information Retrieval," Morgan Kaufmann Publishers, Inc., 1999.
- [4] L Cieplinski, "MPEG-7 color descriptors and their applications", 9th International conference, CAIP, 2001.
- [5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008)Results," <http://www.pascalnetwork.org/challenges/VOC/voc2008/workshop/index.html>
- [6] S. Abburu, "Multi Level Semantic Extraction for Cricket Video by Text Processing," International Journal of Engineering Science and Technology, Vol. 2(10), pp. 5377-5384, 2010.
- [7] M. Grubinger, P. Clough, H. M'uller, T. Deselaers, "The IAPR TC-12 benchmark - a new evaluation resource for visual information systems," in: Proceedings of the International Workshop OntoImage, pp. 13-23, 2006.
- [8] A. Salway and E. Tomadaki, "Temporal Information in Collateral Texts for Indexing Video," Procs. LREC Workshop on Annotation Standards for Temporal Information in Natural Language, pp. 36-43, 2002.
- [9] D. Palma, J. Ascenso, F. Pereira, "Automatic Text Extraction in Digital Video based on Motion Analysis," Int. Conf. on Image Analysis and Recognition (ICIAR'2004), Porto, 2004.



- [10] R. Lienhart e W. Effelsberg, "Automatic Text Segmentation and Text Recognition for Video Indexing", *Multimedia Systems*, Vol. 8, No. 1, 69 – 81, 2000.
- [11] D. Chen, J. Odobez, H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition* 37, pp. 595 – 608, 2004.
- [12] H. Li, D. Doermann, O. Kia, "Automatic Text Detection and Tracking in Digital Video", *IEEE Transactions on Image Processing*, Vol. 1, No 1, pp. 147 – 156, 2000.
- [13] P. Sankar , M. Meshesha, C. Jawahar, "Annotation of Images and Videos based on Textual Content without OCR," In: *Proc. ECCV Workshop on Computation Intensive Methods in Computer Vision*, 2006.
- [14] J. LU, Y. Tian, Y. Li, Y. Zhang, Z. Lu, "A Framework for Video Event Detection Using Weighted SVM Classifiers," *Artificial Intelligence and Computational Intelligence, AICI '09 International Conference*, Vol. No.4, pp.255 – 259, 2009.
- [15] C. Yang, M. Dong, "Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning," In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2, June 17 - 22 , 2006.
- [16] S. Barrat, S. Tabbone, "Classification and automatic annotation extension of images using Bayesian network," in *S+SSPR*, 2008.
- [17] A. W. M. Smoulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, Dec. 2000.
- [18] R. S. Jardon, S. Chaudhury, and K. K. Biswas, "Generic video classification: An evolutionary learning based fuzzy theoretic approach," in *Proc. Indian Conf. Computer Vision Graphics and Image Processing*, pp. 79–91, 2002.
- [19] A. Dorado, J. Calic, E. Izquierdo, "A Rule-Based Video Annotation System," *IEEE Transactions on Circuits and Systems for Video Technology* , Vol. 14, No. 5, May 2004.
- [20] M. Detyniecki, C. Marsala, "Automatic Video Annotation with Forests of Fuzzy Decision Trees," *Mathware & Soft Computing* 7 , 2000.
- [21] M.-S. Hosseini, A.-M.E. Moghadam, "Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video," *Appl. Soft Comput.* J. (2012), <http://dx.doi.org/10.1016/j.asoc.2012.10.007> :in press
- [22] J. R. He, M. J. Li, H. J. Zhang, H. H. Tong, and C. S. Zhang, "Manifoldranking based image retrieval," in *Proc. ACM Multimedia*, New York, NY, pp. 9–16, 2004.
- [23] H. Tong, J. R. He, M. J. Li, C. S. Zhang, and W. Y. Ma, "Graph-based multi-modality learning," in *Proc. ACM Multimedia*, Singapore, pp. 862–871, 2005.
- [24] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," In *ACM Multimedia*, 2008.
- [25] Y-G Jiang, J Wang, S-F Chang, C-W Ngo, "Domain Adaptive Semantic Diffusion for Large Scale Context-Based Video Annotation," in: *Proc. IEEE International Conf. Computer Vision* , 2009
- [26] E. Moxley, T. Mei, B. S. Manjunath, "Video Annotation Through Search and Graph Reinforcement Mining," *IEEE Transactions On Multimedia*, Vol. 12, No. 3, April 2010.
- [27] M Wang, X-S Hua, R Hong, J Tang, G-J Qi, Y Song, "Unified Video Annotation via Multigraph Learning," *IEEE Trans. Circuits Syst. Video Technol* , Vol. 19, No. 5, May 2009.
- [28] T. R. Gruber. "A translation approach to portable ontology specifications", *Knowledge Acquisition*, vol. 5, no. 2, pp.199-220, 1993.
- [29] A. Hanbury, "A Survey of Methods for Image Annotation," *Journal of Visual Languages and Computing*, Volume 19 Issue 5, pp.617-627, 2008.
- [30] B. Vrusias, D. Makris, J. Renno , " A Framework For Ontology Enriched Semantic Annotation of CCTV Video", *Eight International workshop on image analysis for multimedia interactive Services*, IEEE, 2007.
- [31] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, " Video Annotation and Retrieval using ontologies and rule learning", *IEEE computer society*, 2010.
- [32] R. Brehar, C. Fortuna, et al., "Spatio-temporal Reasoning for Traffic Scene Understanding", *Intelligent Computer Communication and Processing IEEE*, pp. 377 – 384, 2011.
- [33] J. Jeong, H.Hong, and D. Lee, "Ontology-based Automatic Video Annotation Technique In Smart TV Environment", *IEEE Transaction on consumer Electronics*, Vol. 57, No. 4, November 2011.
- [34] P. Koletsis, G. M. Petrakis, "SIA: Semantic Image Annotation using Ontologies and Image Content Analysis," *Image Analysis and Recognition*, 7th International Conference, ICIAR, pp. 374-383,2010. http://dx.doi.org/10.1007/978-3-642-13772-3_38
- [35] A R J Francois, R Nevatia, J Hobbs, R C Bolles, "Verl: An ontology framework for representing and annotating video events", *IEEE MultiMedia Mag* , 2005.
- [36] R. Miroguchi, "part 2: ontology development tools and languages"
- [37] M. K. Smith, C. Welty, and D. McGuinness, "OWL Web Ontology Language Guide," *W3C Recommendation*, 2004, <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>
- [38] M Naphade, J Smith, J Tesic, S-F Chang, W Hsu, L Kennedy, A Hauptmann, J Curtis, "Large-Scale Concept Ontology for Multimedia," *IEEE Multimedia* , 2006.
- [39] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.
- [40] Cynthia Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira. An introduction to the syntax and content of Cyc. In *AAAI Spring Symposium*, 2006.
- [41] Ian Niles and Adam Pease, "Towards a standard upper ontology," In *FOIS*, 2001
- [42] Protégé. The Protégé project, [Online] Available: <http://protege.stanford.edu>
- [43] N. Malviya, N. Mishra, S. Sahu, "Developing University Ontology using protégé OWL Tool: Process and Reasoning," *International Journal of Scientific & Engineering Research* Volume 2, Issue 9, September-2011.

Biography



Khushboo Khurana received her B.E. degree in computer science and engineering from RTM Nagpur University, Nagpur, in 2010. She is currently pursuing her M.Tech. (CSE) from Ramdeobaba College of Engineering and Management (Autonomous), Nagpur. Her interests include image and video processing.