View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Research Information System University of Turin

UNIVERSITÀ
DEGLI STUDI
DI TORINO

This is the author's final version of the contribution published as:

Lombardo, Vincenzo; Damiano, Rossana. Commonsense knowledge for the collection of ground truth data on semantic descriptors, in: Proceedings - 2012 IEEE International Symposium on Multimedia, ISM 2012, IEEE, 2012, 9780769548753, pp: 78-83.

The publisher's version is available at:
http://ieeexplore.ieee.org/document/6424635/

When citing, please refer to the published version.

Link to this full text:
http://hdl.handle.net/2318/1619171

This full text was downloaded from iris - AperTO: https://iris.unito.it/

iris - AperTO

University of Turin's Institutional Research Information System and Open Access Institutional Repository

Commonsense knowledge for the collection of ground truth data on semantic descriptors

Vincenzo Lombardo and Rossana Damiano
CIRMA and Dipartimento di Informatica
Università di Torino
Torino, Italy
vincenzo, rossana@di.unito.it

***Abstract*—The coverage of the semantic gap in video indexing and retrieval has gone through a continuous increase of the vocabulary of high–level features or semantic descriptors, sometimes organized in light–scale, corpus–specific, computational ontologies. This paper presents a computer–supported manual annotation method that relies on a very large scale, shared, commonsense ontologies for the selection of semantic descriptors. The ontological terms are accessed through a linguistic interface that relies on multi–lingual dictionaries and action/event template structures (or frames). The manual generation or check of annotations provides ground truth data for evaluation purposes and training data for knowledge acquisition. The novelty of the approach relies on the use of widely shared large–scale ontologies, that prevent arbitrariness of annotation and favor interoperability. We test the viability of the approach by carrying out some user studies on the annotation of narrative videos.**

***Keywords*-video annotation, concept ontology, linguistic interface**

I. Introduction

The coverage of the semantic gap in video indexing and retrieval has gone through a continuous increase of the vocabulary of high–level features or semantic descriptors. Starting from a few tens of the first TRECVid conferences, descriptors now amount to a few thousands[1].

As concepts increase in number, the search task benefits from the creation of semantic relations over individual concepts. The incorporation of semantic relations has led to the creation of ontologies, to organize hundreds or thousands of concepts. LSCOM is an ontology of concepts targetedly designed for a corpus of broadcast news [1]; the MediaMill dataset relies on a set of 101 semantic descriptors that are best suited for that repository [2]. As described in [3], the use of rules to define complex semantic concepts from simpler ones allows the acquisition of new rules (and consequently more complex concepts).

In this paper, we introduce an annotation method and an annotation interface to gather reliable semantic descriptors from viewers. The method relies upon two ontologies: one for the structure of the audiovisual addressed (that provides a framework for assigning the semantic descriptors) and one for the vocabulary of the actual annotation terms. This vocabulary relies on a very large scale, shared, commonsense ontology (actually, an integration of ontologies); the structural ontology addresses a video genre that has been quite neglected in multimedia annotation, indexing, and retrieval, namely drama, or narrative at large.

Public video repositories contain very many scenes (or clips) extracted from feature films. Such scenes are freely tagged by users, with different ideas in mind. Consider, for example, in YouTube, the clip from "North by Northwest" (the famous 1959 MGM–Hitchcock's movie) in which Roger (Cary Grant) warns Eve (Eva Marie Saint) that the gangster Vandamm (George Mason) is on to her by writing her a message on a matchbox.

This clip is tagged with the tag string "Alfred Hitchcock North by Northwest matchbox", where almost all tags are media–based (4 out of 5, excluding the function word "by"), and only one tag is content–based, namely a specific object involved in the scene ("matchbox"). This ratio is very common: tags mostly concern the title, the main actors, the director, the production/distribution/publishing company and the genre, all possibly extracted from public databases, such as IMDB. We carried out an informal survey of the user–contributed tags on "North by Northwest" in YouTube (on June 2012). After searching YouTube with the simple keywords "North by northwest", we manually discarded all the results that did not belong to the original movie (59% of the first 100 results consisted of advertising materials, CGI animations inspired by the movie, user–generated editings of the movie, etc.). We restricted our analysis to the Film & Animation category and considered only the first 100 results. So, we collected 183 unique tags and, after a manual, grounded–theory based analysis [4], tags were divided into eleven different categories (Title, Actor, Director, Production, Editing, Publish, Genre, Character, Object, Environment, Action), grouped into two main macro–categories: media-based tags, conveying information about media type, format, etc. and content–based tags. Content based tags are only 32: auction, blonde, boulevard,

[1] http://www.lscom.org

978-0-7695-4875-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ISM.2012.23

IEEE computer society

Figure 1. Three frames of the matchbox scene in North by Northwest.

bourbon, box, city, dress, dritte, drunk, Eva, girl, matchbox, mother, Mount, office, peak, Philip, plane, police, Roger, Rushmore, searchers, secretary, skirt, station, studio, suit, sunset, tunnel, unsichtbare, waterfront, woman. Most tags refer to characters ("Roger", "mother") or their qualities ("blonde", "dress"). According to the structural ontology (see details in [5]), the content tags of the scene above could be referred to:

- actions/events such as "Roger warning Eve", "Roger writing a message on a matchbook to Eve", "a man saving a woman",
- objects such as "matchbook", "warning message",
- characters such as "an elegant man", "a sexy blonde woman", "a gangster",
- environments such as "a living room", "a two–floor villa".

These tags would be useful in searching for contents, even in a cross–media setting, since they describe the narrative features of the content, independently of the specific media involved.

The paper, after providing the necessary background, illustrates the annotation method of narrative audiovisuals, and shows the web application that displays the annotation interface. Finally, we report some preliminary annotation tests.

II. Related Work

In the last years many researches have exploited ontologies to perform semantic annotation and retrieval from video digital libraries. Semantic annotation is generally performed by classifying video elements and/or video documents according to some pre-defined ontology of the video content domain [6], by establishing relationships over terms that specify domain concepts at different abstraction levels [7]. Ballan et al. use the hierarchical linguistic relations within WordNet to learn and refine rules that can detect complex events from simple ones and the participating entities [3]. Beside standard large scale resources, such a WordNet, standardized vocabularies have created for videos, such as the LSCOM initiative [1].

Semantic annotation can be performed manually, by associating the terms of the ontology to the individual elements of the video, or automatically, by exploiting results and developments in pattern recognition and image/video analysis [8]–[10]. However, these approaches generally manage very few concepts, because of the inability to automatically recognize a wide range of elements from videos. In order to permit a wider range of terms to be used within the annotation process, alternative tools (as in [11]) allow the user to manually map a term with a specific ontological concept. This use of large-scale ontologies also introduces a new problem: the access to the data is, for the user, an extremely hard task (both conceptually and computationally), because of the size and the complexity of the considered data (cf. [1] and successive developments). In fact, within these systems, the information available in videos and visual features need to be manually extracted and assigned to concepts, properties, or relationships in the ontology [12].

Another approach to improve the interoperability of the annotations is to constrain the scope of the semantic models: the Lode (meta–)ontology [13] describes the concept of public event (concert, performance, ...), its structure, and properties, by abstracting on the descriptions of several directories. The annotation of entities was addressed by the the Video Event Representation Language (VERL), which models events in the form of changes of states (cf., the Event Calculus), with an annotation framework [14] where primitive events can be composed and sequenced to create complex events. The VERL approach does not refer to large–scale domain ontologies or to acknowledged patterns to provide a structure to the event models.

Though the semantic annotation of videos has been mostly limited to search and navigation systems, such as [15], there is some interest around the systematic annotation for purposes of narrative video indexing [16]. Also, there is a growing interest for the representation of actions carried out by humans in a video (see, e.g., [17]), useful for many practical applications, such as video surveillance.

III. The annotation process

The annotation process consists of three annotation phases and is carried out through a web-based annotation tool. The purpose of the tool is to make the encoding of the annotation in formal languages transparent to the annotator. Since the

annotation process is conducted by filling a set of templates that describe the narrative elements of a unit. The first phase is the *segmentation* into meaningful units: the annotators must be able to identify the subparts within a video, i.e., the boundaries of the narrative units, by identifying the discontinuities in the stream of actions and events of the narrative audiovisual. The second phase is the *annotation of the story elements* (agents, objects, environments, actions, events, states) through the machine–supported multi–lingual access to the vast terminological knowledge base: the linguistic interface suggests the semantic concepts for the annotation, starting from the linguistic terms of the multi–lingual large dictionary and accessing the corresponding concepts in the large–scale ontologies. The third phase concerns the *annotation of the incidents*: such annotation involves the story elements identified in the previous step, which constitute the events and the entities participating in the incidents. This step relies on large–scale knowledge bases of frames, that describe the event as a predicate accompanied by a set of relevant roles, to be identified among the entities in the unit.

The annotation schema includes: *agents* and *objects* (with their properties) in the narrative unit, *goals* (i.e. motivations for actions) of the agents and *actions* observed, (unintentional) *events*, the *environments* in which the incidents take place. Actions, events, and goals are structured according to the role structure defined for some frame.

For the description of the characteristics of the entities involved in the story actions and events, our framework encompasses the YAGO–SUMO ontology [18]. YAGO–SUMO incorporates almost 80 millions of entities from YAGO (which is based on Wikipedia and WordNet, [19])) into SUMO [20], a highly axiomatized formal upper ontology, providing very detailed information about millions of situations, including entities (agents and objects), processes/actions, and events In addition, it provides the integration with FrameNet [21], a linguistic tool where processes and actions are described by a semantic template depicting the situation in terms of roles played by the elements which participate in it.

In order to alleviate the problem of finding the appropriate concept in large scale ontologies, a common approach, adopted by the developers of the ontologies themselves, is to provide a linguistic interface. Taking advantage from the fact that YAGO–SUMO is already accessible through the WordNet lexical data base [22][2]; we have realized an interface for supporting the manual selection of meaning, extending the vocabulary to a multilingual setting (through the lexical data base MultiWordNet [23]), to increase the interoperability of the annotation data across languages. In our framework, the linguistic access to the commonsense knowledge concepts is embedded in the web-based annotation interface. The first part of the negotiation process relies on the lexical knowledge provided by MultiWordNet and can be described as a word sense disambiguation step aimed at associating each inserted term a unique definition which makes it distinguishable from other possible meanings. Then, taking as input the disambiguated word senses, the system searches YAGOSUMO in order to retrieve the most adequate ontological concept, by leveraging several YAGO-SUMO properties (i.e., those created based on the linguistic knowledge provided by WordNet) to efficiently access this knowledge base. Finally, the disambiguated lexical entry is employed to retrieve the relevant frames from FrameNet, based on the mapping between WordNet and FrameNet [24]. The whole process is described in more detail in [5]. In case the linguistic term is not present in MultiWordnet, the annotator is invited to try some synonym (or some other syntactic category) before resorting to the free tags.

The result of the annotation of a video unit consists of an RDF graph that instantiates the structural ontology for the video elements, instantiates well known design patterns for the annotation of stereotypical situations, and instantiates participating characters, objects, and environments with reference to external ontologies, following the paradigm of linked data [25], As an example, we see the annotation of a story incident (see Figure 2), driven by the Time Indexed Situation design pattern developed in the well–known ontology DOLCE [26]. This example represents the segment of “North by Northwest” in which Eve (Eva Marie Saint) shoots Roger (Cary Grant) at a restaurant near Mount Rushmore. This unit, called *#Unit1*, features two agents, *#Roger* and *#Eve* respectively, whose participation to the unit is mediated by the *AgentInUnit* class. The Unit contains a UnitIncident *#UnitIncident1*, which relates the shooting process *#Shooting* (via the *featuresProcess* property) and its participants, Eve and Roger (via the *#incidentFeatures* property). The *ProcessSchema* class (*#ProcessSchema1*) binds the two agents to their respective roles: Eve as the *filler* of the *#AgentRole*, Roger as the filler of the *#TargetRole*. The annotation also includes the intention of Eve, i.e., her goal to shoot Roger (*#shootingRoger*, an instance of the *Goal* class). Here, *#ProcessSchema1* refers to the concept of *shooting* in YAGOSUMO, and describes it through the FrameNet frame (*Hit_Target*). The *Hit_Target* frame has two roles, labeled as *Agent* and *Target*, respectively filled by the two characters, Eve and Roger. Finally, the annotation also includes the qualities of the characters and objects, as deemed relevant by the annotator. For instance, in this example, the character of Eve could described as “blonde” and “charming” (not shown in the figure).

The annotation process is conducted according to the following methodology. An annotation project is created by a Supervisor, who associates the project with at least two annotators and takes care of publishing the approved annotation, possibly comparing the annotators one another.

[2]See the portal http://www.ontologyportal.org/

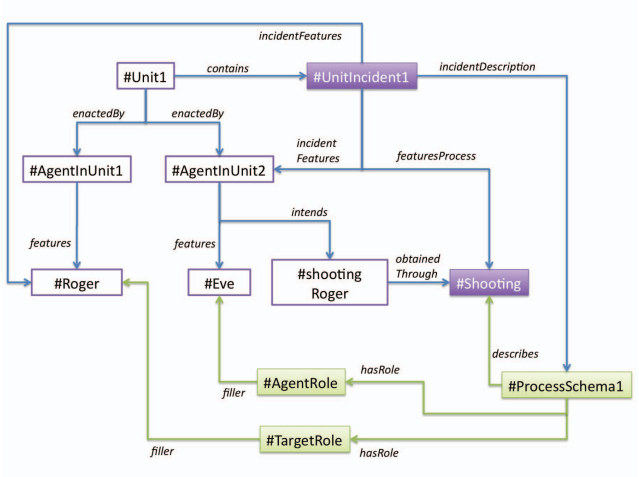

Figure 2. The annotation of a scene of 'North by Northwest" where Eve (Eva Marie Saint) shoots Roger (Cary Grant) at a restaurant near Mount Rushmore.

The guidelines for the annotation process are the following. For the phase of boundary detection, the annotator is invited to identify the onset and offset of the action as she/he perceives it. Annotators are requested to identify both direct observable actions (such as "exiting the train") and narratively meaningful actions (such as "hiding from the detectives"). The latter can sometimes refer to more abstract actions (for example, "hiding from the detectives" can be implemented in many ways). Possible conflicts of interpretation are negotiated by the Supervisor.

IV. Annotation Test

In this example, we describe a preliminary experiment, a proof of concept for the first implementation of the annotation framework, developed in both schema and interface.

Each phase of the annotation process challenges the idea of using manual annotation to create a "golden standard" annotated repository, totally agreed upon by the annotators and interoperable among applications (that can perform search and reasoning on it). In a previous work with students from the cinema programme [16], we had already faced the task of inserting machine readable annotations of narrative units. The preliminary user study concerns the annotation of three different narrative videos: the 2-hour movie "North by northwest" (NbN, from which we have extracted the example above); the multi–prized 2:30 minute animated movie "Oktapodi", about an octopus who tries to save her/his partner from being cooked after having been taken by a vendor from a fish tank; a humorous commercials of the "Zippo" lighter, where a couple of gangsters try to burn a hostage, but waste all the matches they have. The total number of units identified by each annotator was about 100, with differences due to annotators' choices for shot aggregations. Two Italian–speaking annotators annotated the three videos; one English–speaking annotator went through a "North by northwest" scene for comparison's sake.

The first phase, i.e. the detection of boundaries (*segmentation* phase), challenges the unique segmentability of video into units. In our test, there was a significant consistency in the boundaries identification. For the feature film NbN, segmented in about 80 units, we found that 45% of units coincide exactly; of the remaining units, 84% of them were contained in some coincident unit in the other segmentation, and 16% overlapped with the adjacent ones. After the supervised negotiation, almost 90% of units could be considered coincident. These numbers resulted from a tolerance of about 40 seconds on the boundary comparison, a reasonable threshold on a 2-hour feature film. For the short animation Oktapodi, where one annotation has segmented 10 units and the other only 3, the coincidence was of 45%, with a 33% of internal subdivision on the remaining units (boundaries coincidence includes a 5-second tolerance, 78% of coincidence after negotiation). Finally, the 30-second advertisement Zippo was segmented in 3 and 4 units, respectively, with 83% of coincidence, and 100% if we consider inclusion between segments (tolerance 1 second). So, we can conclude that human segmentation on actional/event base can detect similar units, without causing much overload on the Supervisor.

The second phase, i.e. the selection and annotation of the ontological concepts, challenges the interoperability of the such concepts. We started from the inherent ambiguity in the linguistic knowledge bases MultiWordnet and Wordnet, which is less than 2 on average. (i.e., for each linguistic term, the system retrieves in average less than 2 definitions).

Given a total number of 289 requests, we found that the users had to disambiguate in average among 2.83% terms. This means that the annotators tend to use linguistic terms that are more generic than the average. We also ran a qualitative analysis about the difficulty of inputing the appropriate linguistic term and the consequent selection of the adequate definition. We asked the annotators to fill up a questionnaire with the following information:

1) Was it subjectively hard to make a selection from the list of definitions? The answers to this question were: 231 Easy (80%), 39 Medium (13.5%), 19 Hard (6.5%).
2) How many times did you revise your choice by searching for a synonym? The answers were: never 206 times (61%), once 87 times (26%), twice 32 (9%), three times 10 (3%), four times 4 (1%); so 2 or more is about the 13% of cases.
3) How many times did you change your interpretation because of the definitions proposed by the system? This happened 48 times out of 289, 17% of cases.
4) How many times did you resort to free text, giving up the search of an ontological concept? This happened 21 times out of 289, 7% of cases.

From these data we can conclude that the task of selection of

an ontological concept through linguistic definitions is not very hard and the interface system is adequate for supporting the task.

Finally, the third phase challenges the sharing of templates, that provide a structure for the incidents with participants covering some specific role in a verbal frame. Provided that each linguistic definition is mapped onto one ontological concept, we tested the ambiguity factor in the retrieval of frames, that is in the assignment of a structure to some action/event in a unit. Preliminarily, we measured the amount of mappings that were present between the linguistic knowledge bases, MultiWordnet and Wordnet, with the frame knowledge base FrameNet (VerbNet only indicates generic roles). Numbers are not so nice for frames (22% for the total of English synsets[3] and 32% for the total of Italian synsets, respectively), and this is particularly relevant for verbs. Verbs, though reporting a percentage significantly higher than the other syntactic categories (60% for English and 70% for Italian), require frames for the instantiation of the ontological concept in the situation described by the unit, and this means that the system needs some integration of data in the future.

In the experiment, the average number of frames retrieved per term in MultiWordnet is slightly above one; so, almost no ambiguity (even if the percentage is slightly higher for Italian verbs). Again, we asked the annotators to fill up a questionnaire about the difficulties encountered in annotating the frame, thus providing a structure for the events occurring in the unit. These were the results.

1) How many times did you find the correct frame (exclude the generic frame)? The answer was 151 out of 246 (61%). So, 95 times (39%) the annotators inserted the generic frame.
2) Was it subjectively hard to assign the frame roles to agents and objects? No doubt and immediate selection occurred 106 times out of 175 (61%); hesitant on two entries for a role occurred 53 times out of 175 (30%); mulling over a lot without finding the right assignment and then settled for one occurred 16 times out of 175 (9%).

After the experiment, we measured the total of coincident ontological concepts and frames. Before supervision, coincident concepts were 35% and the coincident frames were 37%. These numbers also depend on the different granularities of unit detection. Also we must notice that the annotators tend to use the same concepts in the annotation of a video, especially in the long case of a feature film, thus increasing the gap. However, percentages doubled after the supervision and the propagation of annotations.

[3]Synsets are groups of words that can be viewed as cognitive synonyms. Each synset expresses a distinct concept.

V. Discussion and conclusions

The annotation framework and the provisional system interface revealed to be effective in the proof of concept experiment, showing the feasibility of the approach. However, for the application of the annotation method in the large, we need to address two quantitative issues. The first is the comparison with the baseline results of the annotation method: what happens if we take two groups of annotators and leave one of them only with free text (so, not relying on an ontology) for term retrieval? Are this control group happier of the annotation produced? The second is the effects of the annotation on some applicative task, such as, e.g., the search for some video fragment in an annotated repository: do users retrieve more relevant fragments (numbers of precision and recall) in case of a free annotation (such as the YouTube example reported in the introduction)? In the next future we are building a prototypical annotated corpus, with examples drawn from cinema studies in order to implement meaningful experiments for tuning the system for some specific application. We have in mind two applications: the first is the annotation of a screenplay and the propagation of the annotation through shooting and editing, in order to test the possible advantages (speed up in realization and less error prone) of having the several media stages annotated; the second is the task of the retrieval of video fragments in the case of the edition of a short movie from annotated stock footage.

In this paper we have presented an approach for the semantic annotation of videos, that relies on very large scale, shared, commonsense ontologies. The ontological terms are accessed through a linguistic interface that relies on multi–lingual dictionaries and action/event template structures (or frames). We have tested the viability of the approach through the application to the annotation of narrative videos and carrying out an experiment as a proof of concept.

The multilingual linguistic interface revealed to be very effective and easy–to–use in the annotation test. The long term goal of this research is to build a gold sample corpus of annotated material, used for the training of machine learning algorithms. A web–based platform that incorporates all the functionalities presented here is ready for deployment, and opens to a large, multi–lingual community of annotators, for the creation of annotated corpora of narrative audiovisuals.

Acknowledgment

The work presented here is part of project CADMOS, funded by Regione Piemonte, Innovation Hub for in Multimedia and Digital Creativity, 2010-2012, POR-FESR 07-13. We thank Carmi Terzulli, David Pugliese, and Renata Sheppard for their annotation work. We also thank Antonio Pizzo and Mario Cataldi for their support in the framework definition and experiment.

REFERENCES

[1] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, pp. 86–91, July 2006.

[2] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of ACM Multimedia,*, Santa Barbara, USA, October 2006, pp. 421–430.

[3] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, "Video annotation and retrieval using ontologies and rule learning," *IEEE MultiMedia*, pp. 80–88, October-December 2010.

[4] A. Strauss and J. Corbin, *Basics of qualitative research: grounded theory procedures and techniques*. Newbury Park, Calif.: Sage Publications, 1990.

[5] M. Cataldi, R. Damiano, V. Lombardo, and A. Pizzo, "Lexical mediation for ontology-based annotation of multimedia," in *New Trends of Research in Ontologies and Lexical Resources*, A. Oltramari, P. Vossen, L. Qin, and E. H. (Eds.), Eds. Springer, 2012, to appear.

[6] M. Bertini, R. Cucchiara, A. Del Bimbo, and C. Torniai, "Video annotation with pictorially enriched ontologies," in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, Amsterdam, NL, July 2005. [Online]. Available: http://www.micc.unifi.it/publications/2005/BCDT05a

[7] A. Hauptman, "How many high-level concepts will fill the semantic gap in video retrieval?" in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2007.

[8] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE MultiMedia*, vol. 9, pp. 44–51, April 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=614674.615110

[9] A. Ekin and A. M. Tekalp, "Automatic soccer video analysis and summarization," in *Storage and Retrieval for Media Databases*, 2003, pp. 339–350.

[10] M. Bertini, A. Del Bimbo, and G. Serra, "Learning ontology rules for semantic video annotation," in *Proceedings of the 2nd ACM workshop on Multimedia semantics*, ser. MS '08. New York, NY, USA: ACM, 2008, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/1460676.1460678

[11] C. Saathoff, S. Schenk, and A. Scherp, "Kat: The k-space annotation tool," in *SAMT 2008, Demo Session Proceedings*, 2008.

[12] A. Jaimes and J. R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," in *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2*, ser. ICME '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 781–784. [Online]. Available: http://dl.acm.org/citation.cfm?id=1170745.1171482

[13] X. Liu, R. Troncy, and B. Huet, "Finding media illustrating events," in *1st ACM International Conference on Multimedia Retrieval (ICMR'11)*, Trento, Italy, 2011.

[14] A. R. François, R. Nevatia, J. Hobbs, and R. C. Bolles, "Verl: An ontology framework for representing and annotating video events," *IEEE MultiMedia*, vol. 5, pp. 76–86, 2005.

[15] B. Richard, Y. Prié, and S. Calabretto, "Towards a unified model for audiovisual active reading," in *Tenth IEEE International Symposium on Multimedia.*, Dec. 2008, pp. 673–678. [Online]. Available: http://liris.cnrs.fr/publis/?id=3719

[16] V. Lombardo and R. Damiano, "Semantic annotation of narrative media objects," *Multimedia Tools and Applications*, pp. 1–33, 2011, 10.1007/s11042-011-0813-2. [Online]. Available: http://dx.doi.org/10.1007/s11042-011-0813-2

[17] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor," in *Proceedings of the ACM Multimedia Conference*, 2009, pp. 165–174.

[18] G. De Melo, F. Suchanek, and A. Pease, "Integrating yago into the suggested upper merged ontology," in *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, vol. 1. IEEE, 2008, pp. 190–193.

[19] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[20] A. Pease, I. Niles, and J. Li, "The suggested upper merged ontology: A large ontology for the semantic web and its applications," in *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002.

[21] C. Baker, C. Fillmore, and J. Lowe, "The berkeley framenet project," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998, pp. 86–90.

[22] G. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[23] L. Bentivogli, E. Pianta, and C. Girardi, "Multiwordnet: developing an aligned multilingual database," in *First International Conference on Global WordNet, Mysore, India*, 2002.

[24] S. Tonelli and D. Pighin, "New features for framenet - wordnet mapping," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, CO, USA, 2009.

[25] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, pp. 1–136, 2011.

[26] A. Gangemi and V. Presutti, "Ontology design patterns," *Handbook on Ontologies*, pp. 221–243, 2009.