Application of Semantic Web Technologies for Multimedia Interpretation

Ruben Verborgh
Ghent University – IBBT, ELIS – Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium
ruben.verborgh@ugent.be

supervised by
Rik Van de Walle
rik.vandewalle@ugent.be

ABSTRACT

Despite numerous outstanding results, highly complex and specialized multimedia algorithms have not been able to fulfill the promise of fully automated multimedia interpretation. An essential problem is that they are insufficiently aware of the context they operate in. Algorithms that do take a form of context in consideration, often function in a domain-specific environment. The generic framework proposed in this paper stimulates algorithm collaboration on an interpretation task by continuously actualizing the context of the multimedia item under interpretation. Semantic Web knowledge, combined with reasoning methods, forms the corner stone of the integration of these various interacting agents. We believe that this framework will enable an advanced interpretation of multimedia data that goes beyond the capabilities of individual algorithms. A basic platform implementation already indicates the potential of the concept, clearing the path for even more complex interpretation scenarios.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*Semantic Web*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*; I.4.8 [**Image processing and computer vision**]: Scene Analysis

General Terms

Theory

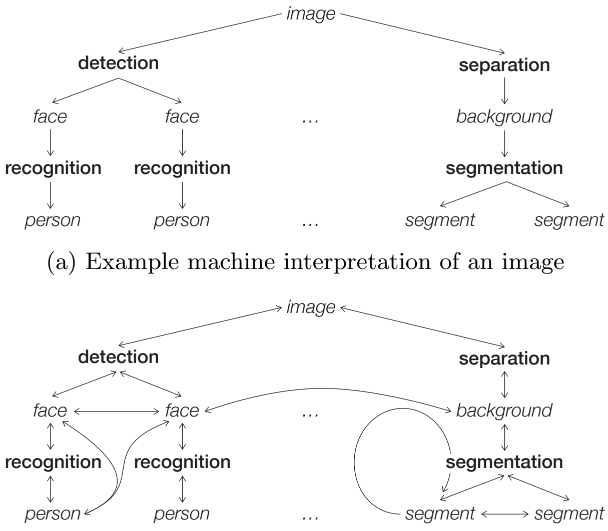
Keywords

feature extraction, multimedia annotation, reasoning, Semantic Web, service composition

1. INTRODUCTION

1.1 Human interpretation of multimedia

At the dawn of the 21st century, a large gap still exists between human and machine interpretation of multimedia data [1, 24]. Machines decompose an interpretation task recursively into smaller subtasks, each of which is to be solved independently by a highly specialized algorithm. This decomposition results in a hierarchical task model, as depicted in Figure 1(a). Humans, on the other hand, seemingly employ a far more complex interpretation model (Figure 1(b)), as based on various natural observations [10].



(a) Example machine interpretation of an image

(b) Example human interpretation of an image

Figure 1: Human versus human interpretation

Some characteristic differences include:

a. **vertical feedback** from each phase to its predecessor, indicated by bidirectional arrowheads;

b. **horizontal feedback** between phases in various sections of the original hierarchy;

c. **circular feedback** that iteratively passes through different phases.

In concrete examples, this could come down to:

a. validating the correctness of a face detection depending on the successfulness of subsequent face recognition;

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

b. recognizing the face of a person by the presence of other persons in the same image;

c. detecting and recognizing several parts of the same object.

The above three examples show frequent interpretation patterns that – while completely natural to humans – do not fit into a simple hierarchical task model. Therefore, machines experience serious difficulties with common situations that we consider straightforward to interpret.

1.2 Multimedia interpretation needs context

The interesting part of this problem is that these difficulties are *not* due to possible inadequacies of the individual subtask handlers. For example, it would not matter if an algorithm performed as good as humans on face recognition in a certain region. Quite the contrary: it is exactly the algorithm's high degree of specialization that causes its isolation and therefore failure when its normal input parameters are insufficient to provide meaningful results. The real issue is that, because of missing feedback, algorithms are unaware of the *context* they operate in and therefore cannot extract contextual information to adapt their inner workings to.

This conclusion, together with recent insights in human information processing mechanisms, make it fair to state that no algorithm will ever succeed in interpreting an image on its own. The key lies in intelligent integration of different algorithms with a constantly actualized context. This is where semantics come into play. The contextual information should be stored in a semantic format that associates it with a meaning, which can be related to external concepts. This insight immediately draws our attention towards the Semantic Web, which at present is arguably the richest interconnected knowledge base containing the foundations for many different contexts we will have to express.

If we want to use Semantic Web technologies to provide a context during a multimedia interpretation process, we face research challenges consisting of diverse topics. Some important open questions are listed below.

- How can algorithms interact in a uniform way with their environment?
- How to create communication channels between several algorithms to account for all feedback types?
- How to create, actualize and maintain the context based on algorithm output?
- How to choose which algorithms should be applied in a certain context?
- How to represent and deal with uncertainties and inconsistencies inherent to the interpretation process?
- How to compare different possible interpretations of the exact same situation?
- How to formulate an answer after the process finishes?

These questions form the core problem I want to address during my PhD research. Section 3 describes how they could be answered in a coordinating platform that integrates multimedia algorithms and Semantic Web knowledge, while Section 4 outlines my methodology. I describe current results in Section 5 and future work in Section 6.

2. RELATED WORK

[8] provides a comprehensive overview of the current burdens in multimedia information retrieval. My work focuses on retrieval through semantic indexing, which can then be combined with techniques such as those described in [9] to offer interactive retrieval. Such semantic annotations have received attention in many publications [2, 7].

On the algorithm side, much work has already been done [18, 27], providing us with many powerful techniques for the individual subtasks. Also, the omnipresence of frameworks such as *OpenCV* [19] facilitate custom implementations of common algorithms to serve a specific purpose. For other multimedia data such as audio, similar techniques and tools are available [3].

Many standards for metadata storage and exchange already exist [25]. One such standard is MPEG-7 [17], which can be serialized to RDF for Semantic Web applications [5]. However, there is a lack of algorithms that output in a standard metadata format; instead, they mostly rely on proprietary input and output schemes. A relevant and complex issue with RDF, however, is the representation of uncertainty [13].

The communication and integration of several independent agents belongs to the Semantic Web services research domain. A first component is service description, which is offered by methods such as WSDL [4], OWL-S [16], and WSMO [14]. The later two include possibilities to semantically describe the capabilities and requirements of services by the use of rule languages (e.g., KIF [6] and SWRL [11]). The other component consists of matching and composition, which can be performed both statically [23] and dynamically [12]. Composition techniques can be used to determine subtask decompositions for a given interpretation task.

The fusion of multimodal multimedia analysis techniques is discussed in [1], covering the diverse associated issues such as methods, timing and selection. This article also indicates the important relationship with artificial intelligence techniques. Furthermore, it tackles topics such as the inclusion of confidence levels and contextual information, which are very relevant to my research, yet it omits the techniques and practices required to describe that context.

From another perspective, several authors look at improving existing metadata with information on the Web. Overell *et al.* classify user-supplied annotations ("tags") using open content resources [20]. Troncy *et al.* associate events with multimedia fragments using Linked Open Data [26]. Other directions include tag recommendation systems [15]. Rahurkar and Dagli demonstrate that even Web resources targeting humans, such as *Wikipedia*, can provide valuable information to represent high-level world knowledge in images [22].

However, the whole idea of integrating multimedia feature extraction algorithms by using Semantic Web knowledge is relatively novel. This is indicated by the long list of related work on both sides but the absence of literature on their integration. We should therefore carefully validate whether this approach can be ported to different practical situations, as the developed techniques should not solely constitute a theoretical exercise.

3. PROPOSED APPROACH

I am developing a generic semantic problem solving platform with a blackboard architecture in which services and knowledge can be plugged (Figure 2, [29]). The central com-

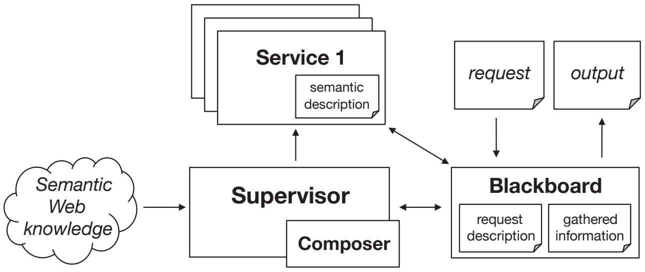


Figure 2: Semantic problem solving platform

ponent is the supervisor, which governs the interpretation process. The services perform specialized subtasks as indicated by their semantic description. Semantic Web knowledge is used to connect concepts and services, and to expand information on the blackboard.

Multimedia processing algorithms are treated as Semantic Web services and described as such [28]. Using service composition, the supervisor creates a task structure, which it subsequently executes. Communication from and to the blackboard takes place using RDF, maintaining formal semantics throughout the entire process.

Algorithms interact through a blackboard, by which the *context* of the item under annotation is created and constantly updated. In this way, algorithms have access to more information than usual, allowing complex interaction patterns such as the *horizontal feedback* depicted in Figure 1(b). The generated solution is formed by a symbiosis of different algorithms that are executed in an iterative fashion. The supervisor adjusts the composition dynamically, based on the output of algorithms and the associated certainty. This ultimately leads to the flow of *vertical* and *circular feedback* necessary for realistic interpretation scenarios.

The novelty of this approach lies in the *cooperation* of several algorithms that work on a common context, instead of operating in isolation. As indicated before, this uncovers new possibilities for interpretation that go far beyond the individual capabilities of each algorithm. The important challenge here is to maintain a generic platform, ignorant of the application domain, while retaining the ability to act on a wide variety of specific situations. Consequently, we have to balance the amount of required knowledge against the generalization possibilities of that knowledge. The same holds for algorithms: the relatively high success rate of very specialized algorithms conflicts with their relatively low application rate.

Semantic Web *reasoning* plays a prominent role in the interpretation process. To start with, advanced reasoning is necessary to devise whether service descriptions match. This becomes increasingly complex as the effects of the combination of several services can differ from the combined effects of the individual services. Reasoning also takes place on the blackboard contents to discover and interpret relations between concepts and to derive related information.

4. METHODOLOGY

We will focus our research on the interpretation of images because of low computational complexity, enabling fast results, and simple solution visualization, enabling straightforward verification by humans. However, we need to keep the expansion to other multimedia types in mind to assure generic applicability of our framework.

To qualify the added value of Semantic Web technologies for image interpretation, we must validate the results of our platform with image sets against both human-annotated and machine-annotated images. We could for example compare the results of automatic annotation generation against that of user-generated tags that have been automatically classified afterwards. Therefore, we need to develop several quality metrics, such as:

- **correctness:** the extent to which the returned annotations are accurate and, if applicable, relevant to the original request *(cf. precision of search engines)*;
- **completeness:** the amount of information found compared to the requested or required amount *(cf. recall)*;
- **performance:** how many resources (time, memory, budget, etc.) were consumed;
- **certainty:** how (un)sure the platform is about the generated annotations;
- **consistency:** whether the annotations are compatible and form a coherent whole.

It is immediately clear that these metrics entail conflicting interests. It should thus be apparent that an ideal solution does not exist, but rather is a delicate compromise between different aspects. For example, it is hard to compare a bad annotation with a low probability to a good annotation with an equally low probability. The first annotation is wrong but improbable, whereas the second is right but also improbable. This illustrates only one of the many tensions that can appear between the above quality metrics.

Another aspect is the required degree of automation. The platform can be used in standalone or user-assisting mode. In the second use case, the platform helps a human in annotating footage, for example in a live context with strict time constraints. In this case, we could express the utility of the platform as the amount by which it decreases the user's efforts in the annotation task.

Finally, it could prove interesting to quantify the amount of specific knowledge required to annotate images in a concrete domain such as news, sport or theater. This will consist of the amount of readily available knowledge on the Web, the amount of available knowledge that needs some adaptation (ontology mapping etc.) and the amount of specifically created knowledge. Is is of utmost importance that the benefits of having a generic platform do not outweigh the adaptation cost of specific situations. The same principle holds for the transition to other multimedia types such as sound and video. We could even test whether the capabilities of the platform extend beyond multimedia interpretation purposes.

5. RESULTS

I have developed a basic version of the platform and its elementary components [29]. I implemented a reasoner-based composition algorithm that is capable of combining services with complex constraints, and a supervisor that is able to recover from certain common types of failure. This platform version is already capable of demonstrating the potential of multimedia interpretation using Semantic Web technologies.

My initial aim was to present a use case that indicates the added value of Semantic Web knowledge for the multimedia interpretation process. Some existing feature extraction algorithms were encapsulated as web services with a semantic description [28]. Next, I formalized basic knowledge about people and photographs, and plugged this in, together with Linked Open Data endpoints. The platform was then asked to annotate pictures depicting, amongst others, people that could not be recognized individually by the algorithm. Based on the formalized knowledge and Linked Open Data on subjects that did pass recognition, it was indeed able to identify unrecognized people with satisfying certainty.

This encouraging success convinced me that further research will eventually lead to more advanced applications of the outlined principles and that it is therefore meaningful to continue in this direction.

6. CONCLUSIONS AND FUTURE WORK

While the first results look promising, various research possibilities emerge, some of which are discussed below. The listed topics are those most likely to become subjects of my future PhD research.

6.1 Handling of Imperfect Information

Imperfections are inherent to the domain of multimedia interpretation, so an adequate way of dealing with them urges itself upon us. [21] distinguishes between several orthogonal dimensions of imperfection: *inaccuracy, vagueness, uncertainty, incompleteness,* and *inconsistency.*

Their successful incorporation into the platform requires that at least some algorithms somehow know how to interpret or generate imperfections qualitatively. It will prove challenging to determine the contribution of each dimension, if possible and applicable at all. Furthermore, the reasoning mechanisms should be able to deal with imperfections and their propagation during the interpretation process.

6.2 Knowledge Modeling

The first experiments with knowledge integration have indicated that interpretation requires a vast amount of every-day knowledge, which unfortunately cannot always be expressed trivially. For example, our intuitive notion that a single person cannot appear multiple times in the same photograph, does not lend itself to simple ontological expressions in OWL.

In general, relationships with an arity greater than 2 are difficult to express using current Semantic Web ontological mechanisms. An additional difficulty is that each rule leads to exceptional situations, which is even harder to express and reason with, given the open world assumption.

6.3 Quality Metrics

As indicated in Section 4, the platform's solutions need to be measured against several – and sometimes conflicting – quality metrics. In the end, a unique solution will almost never exist. Therefore, several metrics must be combined to determine one or several optimal solutions. The definition of "optimal" will vary from one application domain to another and will probably contain some subjective criteria.

Furthermore, we could evaluate the result quality of certain (combinations of) algorithms. We could then create statistics indicating which of them work best given certain circumstances. Those could be taken into account when choosing between different composition alternatives for future interpretation tasks.

6.4 Evaluating Different Configurations

Crucial in the proposed approach is the reuse of existing algorithms and knowledge. I will do so by providing an open architecture with few dependencies between components. This enables the platform to function in a research environment where the influence of specific component implementations can be measured. It will prove interesting to compare different composition algorithms with different knowledge sources and algorithm implementations.

Combined with quality metrics for algorithms, this could happen in a (semi)-automated fashion. Ultimately, we could arrive at self-organizing configurations that adapt to a specific application domain. We should of course keep a realistic attitude towards knowledge source selection, which will probably always require human intervention.

6.5 Comparison to Other Platforms

After optimal configurations have been evaluated, we should compare them to other platforms in order to quantify how this platform improves current techniques and technology. Because of the substantially different possible approaches, this task will be difficult. For example, it is hard to express the amount of knowledge used by a certain platform, all the more because specialized platforms often contain vast amounts of implicit knowledge in their mechanisms.

6.6 Integration in Production Environments

Finally, an important point is the integration of the platform into production environments where image interpretation or the creation of annotations is at the center. Some interesting application domains include surveillance and annotation of sports events because of their rule-based nature.

We should distinguish between fully automated environments and human-assisting tools, the main difference being who determines or selects the final solution. In the latter case, a handful of probable alternatives – between which a machine cannot decide – could be helpful to users.

Either way, the transition from a research situation to a practical context will reveal the real-world opportunities for multimedia interpretation guided by Semantic Web technologies. It is my conviction that this concept can be of fundamental value to bridge the semantic gap between feature extraction algorithms and multimedia data interpretation.

7. ACKNOWLEDGMENTS

The research activities as described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO-Flanders), and the European Union.

8. REFERENCES

[1] P. Atrey, M. Hossain, A. El Saddik, and M. S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, Jan 2010.

[2] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. Strintzis. Semantic

annotation of images and videos for multimedia analysis. *The Semantic Web: Research and Applications*, pages 592–607, 2005.

[3] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416 –430, May 2008.

[4] E. Christensen, F. Curbera, M. Greg, and S. Weerawarana. Web Services Description Language (WSDL) 1.1. W3C Member Submission, Mar. 2001. Accessed 28 October 2010.

[5] R. García and O. Celma. Semantic Integration and Retrieval of Multimedia Metadata. In *Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005)*, 2005.

[6] M. R. Genereth. Knowledge Interchange Format. Draft Proposed American National Standard. http://logic.stanford.edu/kif/dpans.html.

[7] J. Geurts, J. Van Ossenbruggen, and L. Hardman. Requirements for practical multimedia annotation. *Workshop on Multimedia and the Semantic Web*, pages 4–11, 2005.

[8] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R Smith. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? *Proceedings of the IEEE*, Jan 2008.

[9] A. G. Hauptmann, M. G. Christel, and R. Yan. Video Retrieval Based on Semantic Concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[10] J. Hawkins, G. Dileep, and J. Niemasik. Sequence memory for prediction, inference and behaviour. *Philosophical Transactions of the Royal Society*, Jan 2009.

[11] I. Horrocks, P. F. Patel-Schneider, H. Boley, and S. Tabet. SWRL: A Semantic Web Rule Language combining OWL and RuleML. W3C Member Submission, May 2004. http://www.w3.org/Submission/SWRL/.

[12] A. Hristoskova, B. Volckaert, and F. De Turck. Dynamic Composition of Semantically Annotated Web Services through QoS-Aware HTN Planning Algorithms. In *ICIW '09: Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services*, pages 377–382, Washington, DC, USA, 2009. IEEE Computer Society.

[13] K. J. Laskey, K. B. Laskey, P. C. G. Costa, M. M. Kokar, T. Martin, and T. Lukasiewicz. Uncertainty Reasoning for the World Wide Web. W3C Incubator Group Report, Mar. 2008. http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/.

[14] H. Lausen, A. Polleres, and D. Roman. Web Service Modeling Ontology (WSMO). W3C Member Submission, June 2005. http://www.w3.org/Submission/WSMO/.

[15] S. Lee, W. De Neve, K. N Plataniotis, and Y. Man Ro. MAP-based image tag recommendation using a visual folksonomy. *Pattern Recognition Letters*, 31(9):976–982, Jan 2010.

[16] D. Martin, M. Burstein, J. Hobbs, and O. Lassila. OWL-S: Semantic Markup for Web Services, Nov. 2004. W3C Member Submission.

[17] MPEG-7 Overview, Oct. 2004. http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm.

[18] M. Nixon and A. S. Aguado. *Feature Extraction & Image Processing, Second Edition*. Academic Press, 2 edition, January 2008.

[19] OpenCV. http://opencv.willowgarage.com/.

[20] S. Overell, B. Sigurbjörnsson, and R. Van Zwol. Classifying tags using open content resources. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 64–73, 2009.

[21] S. Parsons. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. *IEEE Transactions on Knowledge and Data Engineering*, 8(3):353–372, 1996.

[22] M. Rahurkar, S. Tsai, C. Dagli, and T. Huang. Image Interpretation Using Large Corpus: Wikipedia. *Proceedings of the IEEE*, 98(8):1509 – 1525, 2010.

[23] D. Redavid, L. Iannone, and T. Payne. OWL-S Atomic services composition with SWRL rules. In *Proceedings of the 4th Italian Semantic Web Workshop*, Dec. 2007. http://eprints.ecs.soton.ac.uk/15658/.

[24] M. Riesenhuber and T. Poggio. How the Visual Cortex Recognizes Objects: The Tale of the Standard Model. *Citeseer*, Jan 2002.

[25] J. Smith and P. Schirling. Metadata standards roundup. *IEEE MultiMedia*, Jan 2006.

[26] R. Troncy, B. Malocha, and A. Fialho. Linking events with media. *Proceedings of the 6th International Conference on Semantic Systems*, Jan 2010.

[27] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundation and Trends in Computer Graphics and Vision*, 3:177–280, July 2008.

[28] R. Verborgh, D. Van Deursen, J. De Roo, E. Mannens, and R. Van de Walle. SPARQL Endpoints as Front-end for Multimedia Processing Algorithms. *Proceedings of the Fourth International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web (SMR2 2010)*, 2010.

[29] R. Verborgh, D. Van Deursen, E. Mannens, C. Poppe, and R. Van de Walle. Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. *Multimedia Tools and Applications special issue on Multimedia and Semantic Technologies for Future Computing Environments*, 2011.