Efficient Media Exploitation towards Collective Intelligence

Phivos Mylonas[1], Vassilios Solachidis[2], Andreas Geyer-Schulz[3], Bettina Hoser[3], Sam Chapman[4], Fabio Ciravegna[4], Steffen Staab[5], Pavel Smrz[6], Yiannis Kompatsiaris[2], and Yannis Avrithis[1]

[1] National Technical University of Athens,
Image, Video and Multimedia Systems Laboratory,
Iroon Polytechneiou 9, Zographou Campus, Athens, GR 157 80, Greece
{fmylonas, iavr}@image.ntua.gr
[2] Centre of Research and Technology Hellas
Informatics and Telematics Institute
1st Km Thermi-Panorama Road, Thermi-Thessaloniki, GR 570 01, Greece
{vsol, ikom}@iti.gr
[3] Department of Economics and Business Engineering,
Information Service and Electronic Markets,
Kaiserstraße 12, Karlsruhe 76128, Germany
{andreas.geyer-schulz, bettina.hoser}@kit.edu
[4] University of Sheffield,
Department of Computer Science,
Regent Court, 211 Portobello Street, S1 4DP, Sheffield, UK
{s.chapman, fabio}@dcs.shef.ac.uk
[5] Universität Koblenz-Landau,
Information Systems and Semantic Web,
Universitätsstraße 1, 57070 Koblenz, Germany
staab@uni-koblenz.de
[6] Brno University of Technology,
Faculty of Information Technology,
Bozetechova 2, CZ-61266 Brno, Czech Republic
smrz@fit.vutbr.cz

Summary. In this work we propose intelligent, automated content analysis techniques for different media to extract knowledge from the multimedia content. Information derived from different sources/modalities will be analyzed and fused, in terms of spatiotemporal, personal and even social contextual information. In order to achieve this goal, semantic analysis will be applied to the content items, taking into account the content itself (e.g. text, images and video), as well as existing personal, social and contextual information (e.g. semantic and machine-processable metadata and tags). The above process exploits the so-called "Media Intelligence" towards the ultimate goal of identifying "Collective Intelligence", emerging from the collaboration and competition among people, empowering innovative services

and user interactions. The utilization of "Media Intelligence" constitutes a departure from traditional methods for information sharing, since semantic multimedia analysis has to fuse information from both the content itself and the social context, while at the same time the social dynamics have to be taken into account. Such intelligence provides added-value to the available multimedia content and renders existing procedures and research efforts more efficient.

Key words: media intelligence, collective intelligence, social media

1 Introduction

It is rather true that nowadays most of community-related knowledge and information originates from raw content, be it in the form of e.g. text, images, video, or speech. Human annotation or tagging used in social networks is a way to represent or handle the underlying knowledge, yet despite the human intervention, content remains highly unstructured and it is quite difficult to extract semantics and correlate to other sources of information. The term "Media Intelligence" is introduced in this work and aims at the development of intelligent, automated content analysis techniques for different media to extract knowledge from the content itself. It contributes to the current state-of-the-art techniques in single modality content processing, and at the same time makes a significant step in proposing novel research methods of fusing information from different sources/modalities, contextual information (e.g. time, location, acquisition metadata), personal context (e.g. profile or preferences) and social context (tagging, ratings, group profiles, relevant content collections etc.).

2 Progress over Related Scientific Work

The main drawback of current multimedia content is the fact that it remains highly unstructured and it is quite difficult to extract semantics from it and correlate them to other sources of information. Consequently, we may identify a number of principal challenges to tackle within our work:

- **Semantic Gap**. Although description of multimedia information has seen significant progress, the pace of automatic extraction of such a description, and especially of its semantic part, is rather slow, due to the limitations of state of the art multimedia analysis systems. A "semantic gap" has been acknowledged between current multimedia analysis methods and tools on the one hand and semantic description and annotation methods and tools on the other [17].
- **Scope/domain generalization**. Due to the above challenges, multimedia content analysis techniques are mostly being developed and tuned to a narrow application scope and are not extendible to other application

contexts [11]. In some cases, the methods do not even work on a test set covering multiple aspects of one and the same target application scope.

- **Heterogeneity of modalities**. Multimodal processing [13] and the use of contextual information [3] have been recognized as key in dealing with the above challenges. However, text, image, video, speech processing and analysis are so diverse in nature that experts in different disciplines often have difficulties in reaching a common methodology when dealing with multiple modalities at the same time.
- **Unstructured data**. The Web has emerged as a massive source of multimedia content and automated methods have appeared that collect and organize such content into ground truth data for research, vocabularies and so on [16]. On the other hand, information extraction over such sparse, distributed, unstructured sources, integrating information from content with information from metadata and Web 2.0 tags as well is another challenge on its own.

Going a step further into the detail and regarding text analysis, a number of challenges exist when dealing with documents originating from user input and existing sources:

1. Ability to adapt large scale tasks and domains using limited user input; current methods designed for large scale mining are not applicable to the task, as they require redundancy of information [6], [8] that is not present in many cases (e.g. in an "Emergency Response" case) where information may be scarce and not repeated.
2. Exploitation of contextual information to limit the complexity of extraction, for example to help disambiguate information; in this work we will go beyond the current methods in that we will extend the concept of contextual information beyond the use of simple lists or gazetteers (as in [18]) or user interaction [4]. The use of contextual information will be considered across media (where information in one medium helps disambiguating information in another medium) or by reusing existing meta-information about documents (creator, time, etc.) or the information looked for (background information).

In this work we shall focus text analysis on user input coming from messages (e.g. multimedia messages, emails, etc.) and on the analysis of existing documents (web pages, blogs, RSS feeds and material at the user sites, e.g. PowerPoint presentations, HTML documents, etc.). Technologies suitable for large scale processing of text in knowledge management environments (e.g., [12], [5]) will be enhanced and adapted to the social web requirements of the current era.

On the other hand, regarding image and video analysis, state-of-the-art techniques are frequently used for content-based retrieval by use of low-level features [11], [10], [9] in conjunction with high-level human understandable concepts [1] for multimodal analysis. For instance, Blinkx (`www.blinkx.com`)

is a web platform that enables multimodal video retrieval based on embedded metadata, audio and video cues. Our work will extend such technologies and exploit state of the art techniques in single modality content processing and will make a significant step in researching novel methods of fusing information from diverse modalities [13], contextual information [3], personal context [19] and social context.

It will also utilize existing multimedia analysis approaches based on explicit knowledge that model visual features/structure, algorithms, domain concepts and rules guiding the analysis process [7] and will advance the state of the art by providing intelligent extensions to knowledge-assisted analysis [15], [2] based on visual and complementary contextual information [14] derived from other modalities, metadata, personal and social context. It will extend them to better adapt to and benefit from the remarkable success of Web 2.0 applications by:

1. Formalisation of multimedia analysis based on user context (e.g. physical location, media acquisition conditions).
2. Fusion of knowledge originating from different modalities and different analysis processes.
3. Exploitation of the intelligence emerging from user groups in order to boost the performance of multimedia analysis and support novel content access and delivery mechanisms.
4. Adaptation and combination of vocabulary-based telephone speech recognition techniques and utilization of phonetic search that enables identification of proper names and usually unrecognized words, especially in the context of an emergency response case study.

Regarding speech analysis, today's standard vocabulary-based speech recognition technology cannot provide sufficient accuracy for such cases; the word-error rate can be as high as 40% for noisy environments. Moreover, the tools can only recognize a given set of words (from their limited vocabularies). They cannot deal with new names of persons, places etc. that are crucial for real-life use cases. To address this situation, we will combine a vocabulary-based speech recognizer with a keyword spotting module implementing the functionality of the phonetic search, also supporting detection of detection of OOV (out-of-vocabulary) words. The phonetic recognition can be more or less language independent but can also benefit from an identification of the language-specific set of phonemes.

The proposed framework will provide a unique opportunity in exploiting challenging research directions such as using multimodal processing, contextual information, personal and social context, tags and other information to improve the performance of existing multimedia analysis methods. Most importantly, such analysis is not currently used in social network environments mainly due to its unreliability. The greatest challenge will be to demonstrate that combined with metadata, tags and other information currently used, content analysis can provide a more powerful experience for the user. Success

will highlight the value of content analysis in such environments, generate awareness and open the way to a number of future applications.

3 Intelligent media analysis

The main step towards efficient "Media Intelligence" concerns automated analysis and semantic extraction from raw visual, textual or audio content and associated metadata. Analysis focuses on each medium in isolation and without taking into account any contextual information or the social environment. However, it does take into account prior knowledge, either implicit, in the form of supervised learning from training data, or explicit, in the form of knowledge driven approaches.

Extracting knowledge from raw data is a huge research problem on its own so work in this field is expected to advance current existing state-of-the-art techniques for each medium, while a significant effort will be devoted on

1. adapting to the individual domains of interest and intelligence methodologies,
2. handling heterogeneity of unstructured user-contributed content and
3. supporting interoperability with contextual information.

3.1 Text analysis

Textual information is of fundamental importance in every scenario where humans are involved as it it the most common medium of communication. Unfortunately the ability humans have to understand and generate language is not matched by machines and therefore information in textual documents is generally unavailable for automatic processing. Textual information is pervasive and - in the digital era - its availability is increasing. Intelligent techniques are required to enable automatic Information Extraction (IE) from text and make this information available for further processing, e.g. to integrate with other sources or to proactively execute tasks.

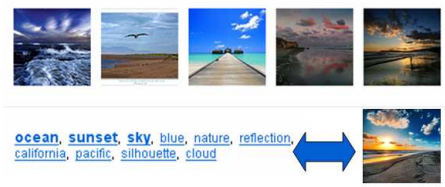

Fig. 1. Intelligent text analysis paradigm

The current research challenges for text analysis are:

1. information extraction over sparse distributed documents (e.g. the Web), integrating information from documents with information from metadata and Web 2.0 tags as well. Technologies for large scale processing of text in knowledge management environments (e.g., [12], [5]) will be enhanced and adapted to the current social web requirements.
2. Analysis of time and spatial information: Modelling information with a strong spatial and temporal connotation over a large scale is a complex and challenging problem and - to our knowledge - has been only partially coped with previously.
3. User profiling and monitoring as a way to empower and direct text extraction, i.e. strategies for focusing extraction and making sense of facts based on user and task profiling or in other words take personal context into account for multimodal media analysis.

3.2 Visual information analysis

Visual information, that is, still images and especially video, tend to impose huge requirements on current repositories or social networks in terms of storage or transmission due to the size of the data involved, yet its contribution to the knowledge and intelligence of related applications remains insignificant. Research in disciplines like image processing, pattern recognition and computer vision has been ongoing for decades but satisfactory performance can usually only be achieved in constrained domains, scales and environments.

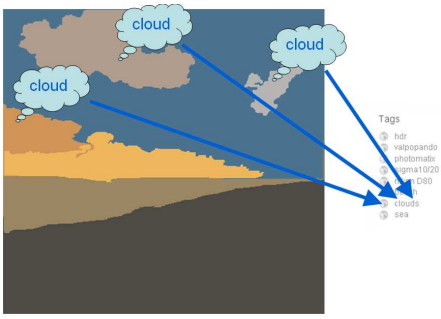

Fig. 2. Intelligent visual analysis paradigm

Our work leverages existing tools to handle well-specified problems like spatial/temporal decomposition and structuring, object/event detection, recognition and tracking, investigating appropriate visual features, metrics and supervised learning approaches using annotated training data. It expands also existing approaches based on explicit knowledge that models visual features/structure, algorithms, domain concepts and rules guiding the analysis process [7]. In doing so, it employs state-of-the-art knowledge-assisted analysis tools and advances them by providing extensions to

1. support processing of incomplete, uncertain, partial or conflicting information,
2. exploit visual context emerging from interaction between local and global processing, and
3. support use of complementary contextual information derived from other modalities, metadata, user and social context.

3.3 Speech analysis

Speech is considered to be an important modality, e.g. in the case of emergencies, where due to the number available resources, handling by humans is clearly inadequate and automated extraction of semantics is necessary. In this manner, a vocabulary-based speech recognizer will be combined with a keyword spotting module implementing the functionality of the phonetic search. The phonetic recognition can be more or less language independent (but can also benefit from an identification of the language-specific set of phonemes). The research agenda will focus on the detection of OOV (out-of-vocabulary) words in the output of the large vocabulary speech recognition module and on the advantages the phonetic recognizer can provide for their search. The development part of the task will concentrate on the generality of the designed interfaces to enable easy updates of the background recognition modules.

In principle, response to emergency cases will benefit from speaker identification technology (and verification of the speaker identity). State of the art speaker identification technologies will be examined and employed; the latter are mainly based on advanced feature extraction and machine learning techniques. Requirements for speech analysis related to visual and text analysis will also be investigated.

4 Contextual media analysis and fusion

The aim of this section is to combine the semantics extracted from different modalities in a structured way, along with contextual information like time, location, or acquisition metadata, and personal context like user profile and semantic preferences. Fusion of heterogeneous information derived from different sources and modalities is a problem that has been long studied but only partially dealt with (e.g. still images with associated metadata, or video sequences with associated audio), mainly due its multidisciplinary nature and the different requirements of each study. An integrated theoretical model will be developed in this task to handle fusion of textual, visual and speech semantics, coupled with contextual and personal information. Contextual information typically refers to metadata automatically created and stored with the content, like acquisition time, location (mobile cell, GPS coordinates) and parameters, e.g. device or camera metadata (aperture, focal

length, lighting conditions, flash use) for the still image case [3]. Such information may be available with the content itself (e.g. EXIF metadata for images) or in separate resources (e.g. MPEG-7 metadata for video). Such metadata can be valuable when combined in the analysis process and will be exploited to disambiguate, resolve inconsistencies or complement missing information in simple tasks like indoor/outdoor classification.

On the other hand, personal context [14], like the profile and the personal preferences of the user contributing a specific resource, provides additional evidence to assist content-based analysis of the resource, e.g. types of places one visits, style of writing and so on. Such contextual extraction mechanisms will be investigated, that are necessary for the creation and exploitation of personal context. All evidence will be taken into account both during multimodal media analysis (early fusion) and as a post-analysis integration step (late fusion). To support the required information fusion processes, existing knowledge representation formalisms and reasoning tools will be extended to support temporal/spatial analysis, media interpretation and information fusion under uncertainty, and incomplete or contradicting evidence.

5 Social Media Intelligence

It is expected that actionable knowledge can be extracted by analysing how multimedia content is shared, accessed, annotated and otherwise used by communities. The aim of this task is to exploit workspace statistics (e.g. access/usage history) and social content (tags, ratings, related content items, related users, group profiles etc) in order to improve the intelligent media analysis and use. This includes e.g. analysing how images are stored or how frequently they are accessed in the context of a given task in order to learn what they have in common in terms of content.

Knowledge about how users are interacting in a shared community and semantics of social interaction extracted from system usage will be investigated in this task. This kind of information will be exploited in the content analysis process; such an approach has not been explored yet, to our knowledge. This would permit e.g. to analyze an image given the profile of the contributing user, other pictures in the same collection (e.g. from vacation), comments from his/her friends, or related pictures within the community. This is yet another information fusion task that will be carried out, where contextual information extends to include personal or social context. Collaborative computing techniques will be employed to enable collaborative media tagging, manipulation, or search over a social network.

Knowledge extracted from the social content will be represented in a machine understandable way in order to achieve interoperability and knowledge sharing. Semantic Web technologies will provide the formal framework for the representation and processing of syntax and semantics of the extracted social knowledge.

Existing reasoning engines will be investigated and used to process the extracted knowledge. Reasoning services are essential to check the consistency and the validity of instances, to extract implicit knowledge using subsumption and equivalence relations defined in the ontologies and, finally to take decisions using inference rules. Existing work and standards will be examined in order to construct the social content ontology. For instance, Friend-Of-A-Friend (FOAF) is a simple technology that makes easier sharing and using information about people and their activities (e.g. photos, calendars, weblogs), to transfer information between Web sites, and to automatically extend, merge and re-use it online. FOAF can be used as a starting point and extended where needed in order to support representation and exchange of knowledge with respect to content, metadata, users and community interactions.

6 Conclusions

In this work we proposed our initial research efforts and ideas in designing and implementing efficient, automated content analysis techniques for different media to extract knowledge from the multimedia content. Contextual information, in terms of spatial, temporal, semantic, personal and social information constitute the added value to current traditional media analysis approaches. Applying and exploiting this kind of additional information forms the way to efficiently advance the so-called "Media Intelligence" towards the ultimate goal of identifying "Collective Intelligence", emerging from the collaboration among large communities and empowering innovative services and people interactions. Such combined intelligence provides added value to the available multimedia content and results into a more efficient rendering of procedures and research.

7 Acknowledgement

The research leading to these results has received funding from the European Community's 7th Framework Programme FP7/2007-2013 under grant agreement n.215453 - WeKnowIt.

References

1. W. Al-Khatib, Y.F. Day, A. Ghafoor, P.B. Berra. Semantic Annotation of Images and Videos for Multimedia Analysis. 2nd European Semantic Web Conference (ESWC), Greece, 2005.
2. Th. Athanasiadis, Ph. Mylonas, Y. Avrithis, S. Kollias. Semantic Image Segmentation and Object Labeling. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, Issue 3, March 2007.

3. M. Boutell. Exploiting Context for Semantic Scene Classification. Technical Report 894 (Ph.D. Thesis), University of Rochester, 2006.
4. S. Chakrabarti, K. Puniyani, S. Das. Optimizing scoring functions and indexes for proximity search in type-annotated corpora. WWW '06: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006.
5. F. Ciravegna, A. Lavelli. Learning Pinocchio: Adaptive Information Extraction for Real World Applications Journal of Natural Lang. Engineering, 10 (2), 2004.
6. F. Ciravegna, S. Chapman, A. Dingli, Y. Wilks. Learning to Harvest Information for the Semantic Web. Proceedings of the 1st European Semantic Web Symposium (ESWS), Heraklion, Greece, May 2004.
7. S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, M. G. Strintzis. Knowledge-Assisted Semantic Video Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Analysis and Understanding for Video Adaptation, 15(10) 1210-1224, October 2005.
8. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Web-scale information extraction in KnowItAll. ACM, New York, May 2004.
9. J. Foote. An Overview of Audio Information Retrieval. ACM Multimedia Systems 7(1), 42-51, 1999.
10. M. Haas, M.S. Lew, D.P. Huijsmanns. A New Method for Key Frame based Video Content Representation. Image Databases and Multimedia Search
11. R.M. Haralick, L.G. Shapiro. Computer and Robot Vision. Addison-Wesley, New York, USA, 1993.
12. J. Iria and F. Ciravegna. A Methodology and Tool for Representing Language Resources for Information Extraction. In 5th International Conference on Language Resources and Evaluation (LREC 2007), Genoa, 24-26 May 2006.
13. P. Maragos. Cross-Modal Integration for Performance Improving in Multimedia: State of the Art Report. MUSCLE NoE Deliverable D.6.1, September 2004.
14. Ph. Mylonas and Y. Avrithis. Using Multiple Domain Visual Context in Image Analysis. 8th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2007), Santorini, Greece, 6-8 June 2007.
15. G. Th. Papadopoulos, Ph. Mylonas, V. Mezaris, Y. Avrithis and I. Kompatsiaris. Knowledge-Assisted Image Analysis Based on Context and Spatial Optimisation. International Journal on Semantic Web and Information Systems, Vol. 2, no. 3, pp. 17-36, July-September 2006.
16. A. Popescu, G. Grefenstette, C. Millet, P.-A. Moellic, P. Hede. Imaging Words - Wording Images. 1st International Conference on Semantic And digital Media Technologies (SAMT 2006), Athens, Greece, 6-8 December 2006.
17. A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans Pattern Anal Mach Intell 22(12), pp.1349-80, 2000.
18. M. Stevenson, M. Greenwood. Comparing Information Extraction Pattern Models. Proceedings of the Workshop on Information Extraction Beyond The Document, pages 12-19, ACL, 2006.
19. D. Vallet, P. Castells, M. Fernandez, Ph. Mylonas and Y. Avrithis. Personalised Content Retrieval in Context Using Ontological Knowledge. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, Issue 3, March 2007.