Architecture for semi-automatic multimedia analysis by hypothesis reinforcement

Igor G. Olaizola, Gorka Marcos, Petra Krämer, Julián Flórez and Basilio Sierra

Abstract—
The digitalization of the audiovisual production chain has introduced new opportunities and challenges in the asset management workflow. The huge amount of accesible content requieres new annotation and indexing paradigms that overcome the current limitations in terms of resources and level of detail. A novel approach to improve and automatize professional Media Asset Management systems is proposed in this paper. Our proposed architecture enhances the metadata with new objective concepts that can be ported to the semantic level and can also used through "query by sample" methods. Moreover, the implicit and explicit knowledge about a certain domain can be introduced in the system with a combination of classifiers and a semantic middleware. Last, the system can be replicated in different domains and combined via an initial hypothesis, allowing the scalability of the system to multiple content domains.

Index Terms—Multimedia annotation, architecture, semantic gap

1 INTRODUCTION

In the last years, the digitalization of the content, the increase of bandwidth, and the diversity of networks had a huge impact on the workflows of the broadcasters and the consumption patterns. The way of producing, storing, managing, and distributing the content has radically changed. Moreover, the digitalization has considerably improved the way of searching and retrieving the content. For instance, *Media Asset Management* (MAM)[1] systems have been developed to index the annotations of distributed content, allowing powerful text based retrieval mechanisms.

Consequently, MAM frameworks have a strong dependency on the metadata associated to the media assets. Manual annotations are the main metadata generation method in the current state of the art, but due to the high costs of manual processes and the high subjectivity degree that they introduce, automatic analysis systems are presented as an ancillary method to extract knowledge from the contents. The state-of-the-art in multimedia analysis techniques can be seen as an opportunity to enrich the annotations automatically or with the supervision of an expert. Very promising results have been achieved for the extraction of middle level annotations, especially when the domain of the content is pre-established.

However, even in such conditions, there is still a semantic gap between the information generated by the algorithms and the final constraints of the current systems. Besides, there are enormous opportunities to improve the way of tackling the general problem and the way of combining the different processing modules. In this paper, we propose an architecture to bridge this semantic gap under certain conditions. Its characteristics are the following:

- Annotation process enhancement in professional domains through automatic analysis modules that can add low-level data and convert them into semantic concepts according to the pre-established taxonomy.
- Scalable and modular employment of video analysis, concept detectors, classifiers and semantic tools to extract and infer the meaning of the content of the asset.
- Ability to take advantage of the use of

Igor G. Olaizola, Gorka Marcos Petra Krämer and Julián Flórez are with VICOMTech (www.vicomtech.org, e-mail: iolaizola@vicomtech.org, gmarcos@vicomtech.org, pkraemer@vicomtech.org, jflorez@vicomtech.org)
Basilio Sierra is with University of Basque Country (e-mail: ccpsiarb@si.ehu.es)

the fuzziness of the multimedia processing contents at the semantic level.
- Modularity and scalability to enrich the metadata by plugging further analysis tools or increasing the model with decision rules.

1.1 Manual annotation vs. automatic indexing

Indexing and retrieval processes differ in the professional and non professional domain. General purpose audiovisual content retrieval environments like *Youtube* or *Flickr* (for still images) are based on manual annotations. Basically, those systems offer a free way to annotate the content and text based search engines are used to retrieve the content. This approach is valid for very open content domains where taxonomies would be too big to handle with them and where people who annotate are not expert in indexing processes. In the professional domain the need of accuracy and consistency is much higher and restricted taxonomies are implemented to avoid ambiguities and diffuse concepts. However, this manual process results in very slow and expensive tasks due to the huge amount of content generated by broadcasters and producers. Automatic analysis modules could enrich the metadata and ideally avoid the need of the manual annotations.

The rest of the paper is organized as follows. Section 2 shows our proposed system architecture to enhance the annotation process through automatic analysis modules that can add low-level data and convert them into semantic concepts according to the pre-established taxonomy. Section 3 presents information about the development status of the designed architecture and Section 4 describe last conclusions and the future work planning.

2 System Architecture

2.1 Approach to the problem

From a machine learning point of view, we face an extremely high dimensional problem with a massive quantity of heterogeneous content which cannot be handled by using classical data mining techniques. KDD (Knowledge Discovery in Databases)[2] techniques have to be used instead in order to extract the conceptual information.

In order to convert the information in knowledge, dimensionality reduction methods like (PCA [3] LDA [4], MFA [5], etc.) have to be used. The resulting features must have a high correlation degree with the concepts contained within the video content. However, this methods are only suitable for expert system where the knowledge can be modeled with an affordable number of features (e.g: face recognition systems). For more general domains, existing automatic dimensionality reduction methods do not provide the needed information to bridge the semantic gap and convert the information in knowledge.

On the semantic part of the problem, the knowledge can be provided explicitly[6] by using ontologies that represent the domain with its elements and the relationships among them, but again, complex or big domains cannot be defined manually.

We propose a combined solution where the explicit knowledge is introduced to a semantic middleware and the implicit knowledge is acquired by classifiers. This combination allows the modeling of bigger and more complex domains[7] and reduces the semantic gap by connecting low-level features with high-level hypothesis and reinforcement factors. The reinforcement factors allow to extend the dimensionality of the domain and provide the framework for specific analysis methods.

2.2 Architecture

The proposed architecture (Figure 1) combines low-level feature analyzers with data mining techniques and a semantic middleware. Thus, the analysis process consists of three iterated steps: Domain pre-establishment, hypothesis generation, and hypothesis reinforcement.

The domain of the content is fixed during the ingesting process in order to define the initial configuration of the system. This configuration avoids ambiguities that appear in the content and applies specific modules related to that domain. In the further steps, low-level analysis modules are applied to segment the frames into regions and extract low-level data for each of them.

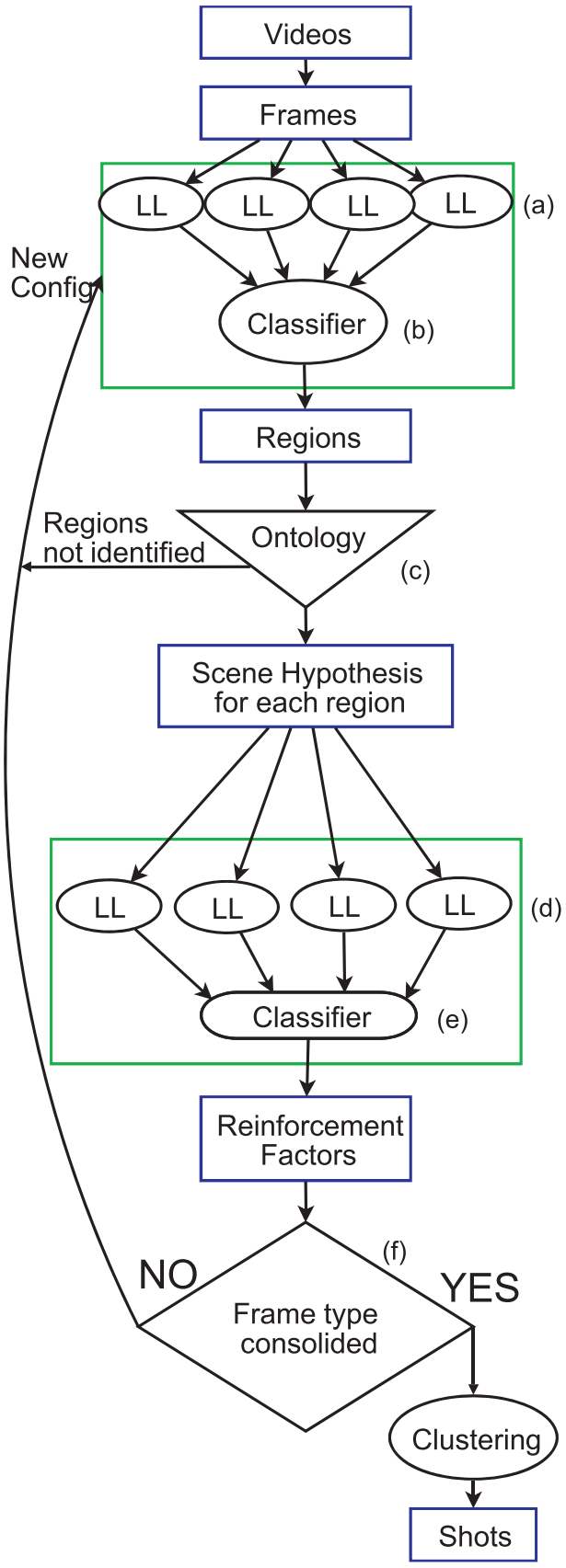

Fig. 1. System Architecture

Fig. 2. Original frame

2.2.1 *First low-level feature extraction layer*

The first module in the architecture extracts region-based features which will be processed by classifiers for first frame categorization (labeled as 'a' in Figure 1). These low level features are domain dependent and have mostly a strong relation with edges and shapes or color and texture features[8], [9], [10]. Figure 2 and 3 show an example of the process of simplifying regions based on color distribution. Then the main regions and their properties are submitend to the classifier. Time related features like movement vectors or similarity among different frames can be another type of low-level features that in some cases can provide meaningful information to the next layer. Thus, the outcome of this module will be a simplified description of the image based on a set of features.

2.2.2 *Classifiers*

The implicit knowledge is extracted by using a combination of automatic classifiers (labeled as 'b' in Figure 1). These classifiers obtain the first middle-level classes[11] which will be used as the basis for the hypothesis assumption. The classifiers are selected according to the experimental results obtained in a testbench where all low-level features, middle-level classes and classification algorithms are evaluated. For the experimental evaluation, a testbenching application has been developed in Matlab™. This

Fig. 3. Color based regioned frame

application can access all the classification algorithms of Weka Data Mining Software[1] and make an automatic ranking of the results.

2.2.3 Semantic middleware and hypothesis generation

The semantic middleware applies fuzzy reasoning rules to the knowledge gathered from middle-level features by a set of ontologies to select the next analysis modules that will provide more specific information about each labeled region[12](‘c’ in Figure 1). The outcome of these modules is considered as a set of “reinforcement factors”. For instance, in a nature footage domain a “big green area” with specific texture properties would be assumed as “grass” by classifiers. Then then semantic middleware has to establish the “meadow” hypothesis. According to this hypothesis, specific analysis modules have to executed in order to demonstrate or rule out the hypothesis through the reinforcement factors.

2.2.4 Reinforcement factors and hypothesis validation

A similar architecture to the presented subsystem (‘d’,‘e’,‘f’ in Figure 1) is initiated to analyze the reinforcement factors. Low-level analysis modules are executed to find specific features to prove hypothesis and classifiers provide the middle-level features to be processed by the semantic middleware. If these factors confirm the expected features, the hypothesis of an asset will be considered as true and these factors will be the characteristics of this frame. Otherwise, if these factors do not satisfy the expectations of the semantic middleware, the hypothesis will be dismissed and the frame will be considered without a specific semantic meaning inside this domain and the process is reiterated with different initial conditions.

1. http://www.cs.waikato.ac.nz/ml/weka/

2.3 Scalability to other domains

The presented architecture can be extended to any new domain and all these new extensions can be integrated in a single framework. The only common part among different domains is the region labeling part, carried out by the classifiers. Once the hypothesis is estimated, only one of the ontologies will be used. This fact improves dramatically the scalability of the system and allows the independent behavior of each semantic definition. Moreover, this architecture based on consecutive layers of data mining modules and ontologies allows the combination of automatically extracted features and manual annotations. A coherent combination of these two sources improves the final result of the content retrieval process.

3 Implementation

The modules presented in this paper have been developed and tested in a professional broadcasting context. The selected domain has been “landscape documentaries” and the testing classes are the following ones:

- Water
  - Sea
  - Reservoir/Lake
  - River
- Buildings
  - Village
  - City
- Land
  - Forest
  - Grass
  - Wheat

  - Soil
- Sky
  - Clouds
  - Clear Sky

Very promising preliminary results have been obtained by the different modules of the architecture.

4 Conclusions and Future Work

An architecture to avoid the dimensionality problem and to minimize the semantic gap in specific audiovisual content domains has been presented in this paper. The implicit and explicit knowledge are integrated with a combined solution of classifiers network and a semantic middleware. Further work is focused on the appliance of the architecture within the entire workflow of an testing and validation processes within the selected domain.

Acknowledgments

This work has been supported by the public Basque Radio and TV Broadcaster EiTB (Euskal Irrati Telebista, http://www.eitb.com). The authors would like to thank Mikel Agirre, Mikel Frutos and Leticia Fuentes for letting all the audiovisual material used in this project and for their advice.

References

[1] J. Van Tassel, *Digital content management, creating and distributing media assets by broadcasters*. NAB, National Association ofd Broadcasters, 2001.
[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.
[3] I. Jolliffe, *Principal Component Analysis*. Springer - Verlag, 2002.
[4] H. Kim, B. L. Drake, and H. Park, "Multiclass classifiers based on dimension reduction with generalized lda," *Pattern Recogn.*, vol. 40, no. 11, pp. 2939–2945, 2007.
[5] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," vol. 29, no. 1, pp. 40–51, Jan. 2007.
[6] S. Sav, N. E. O'Connor, A. F. Smeaton, and N. Murphy, "Associating low-level features with semantic concepts using video objects and relevance feedback," in *WIAMIS 2005 - 6th International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.
[7] N. Simou, T. Athanasiadis, G. Stoilos, and S. Kollias, "Image indexing and retrieval using expressive fuzzy description logics," *Signal, Image and Video Processing*, vol. 2, pp. 321–335, 2008.
[8] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," vol. 29, no. 1, pp. 167–172, Jan. 2007.
[9] F. Li, Q. Dai, W. Xu, and G. Er, "Multilabel neighborhood propagation for region-based image retrieval," vol. 10, no. 8, pp. 1592–1604, Dec. 2008.
[10] C. R. Jung, "Unsupervised multiscale segmentation of color images," *Pattern Recogn. Lett.*, vol. 28, no. 4, pp. 523–533, 2007.
[11] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu, "Multi-layer multi-instance learning for video concept detection," vol. 10, no. 8, pp. 1605–1616, Dec. 2008.
[12] R. C. F. Wong and C. H. C. Leung, "Automatic semantic annotation of real-world web images," vol. 30, no. 11, pp. 1933–1944, Nov. 2008.

Igor García Olaizola is the head of Digital TV and Multimedia Services Department in VICOMTech (http://www.vicomtech.org). He received his MEng degree in Electronic and Control Engineering from the University of Navarra, Spain (2001). He developed his Master thesis at Fraunhofer Institut f'ur Integrierte Schaltungen (IIS), Erlangen -Germany- 2001 and currently he is preparing his PhD in Computing Science and Artificial Intelligence at University of Basque Country. He has participated in many industrial projects related with Digital TV as well as several European research projects in the area of audiovisual content management. His current research interests include multimedia content analysis frameworks and techniques to decrease the semantic gap.

Gorka Marcos received the BSc degree from the Telecommunications Enginneering Faculty of the Basque Country University in 2001. Since them he is a researcher in VICOMTech, where he has been actively participating in diverse european and national research projects related with his main research interest: the improvement of the management of multimedia workflows by employing content-aware techniques. In the last 8 years he has been working closely with international researchers and companies from the basque country in this field. He has actively published his results and reviewed third party papers in several conferences and journals.

Dr. Petra Krämer is postdoctoral researcher at the Biomedical Applications and Digital TV & Multimedia Services departments of VICOMTech (Spain). She received MSc degree in 2004 and PhD degree in 2007, both in computer science from the University of Bordeaux 1 (France). Her main research interests are biomedical image processing, as well as video processing and indexing.

Julián Flórez Esnal got his PhD in University of Manchester Institute of Science and Technology (UMIST) in Manchester, United Kingdom. He is currently the General Manager of VICOMTech Research Center and since 1994 a Professor in the University of Navarra. From 1983 to 1994 he was an associate professor in the same university. Dr. Julián Flórez- Esnal has participated in several industrial and European research projects and has written more than 60 technical research papers in different areas of Information Systems, Electronics, Control and Electrical Engineering and holds several international awards. Dr Flórez holds nine industrial patents and has directed 16 Doctoral Thesis.

Basilio Sierra is an Assistant Professor in the Computer Sciences and Artificial Intelligence Department at the University of the Basque Country. He received his BSc in Computer Sciences in 1990, MSc in Computer Science and Architecture in 1992 and PhD in Computer Sciences in 2000 at the University of the Basque Country. He is the director of the Robotics and Autonomous Systems Group in Donostia-San Sebastian. Dr. Sierra is presently a researcher in the fields of Robotics, Computer Vision and Machine Learning, and he is working on the use of different paradigms to improve behaviors; he has written more than 15 journal papers in those fields, as well as 100 conference contributions and more than 30 book chapters.