Next Generation TV through Automatic Multimedia Annotation Systems
A Hybrid Approach

Joël Dumoulin[1], Marco Bertini[2], Alberto Del Bimbo[2], Elena Mugellini[1], Omar Abou Khaled[1] and Maria Sokhn[1]

[1]*Department of Information Technologies, HES-SO, Fribourg, Switzerland*

[2]*Media Integration and Communication Center (MICC), University of Florence, Florence, Italy*

Keywords: Automatic Video Annotation, High Level Content Annotation, Multimedia Information Retrieval, Social Video Retrieval, Tag Suggestion, User-generated Content, Smart TV.

Abstract: After the advent of smartphones, it is time for television to see its next big evolution, to become smart TVs. But to provide a richer television user experience, multimedia content first has to be enriched. In recent years, the evolution of technology has facilitated the way to take and store multimedia assets, like photographs or videos. This causes an increased difficulty in multimedia resources retrieval, mainly because of the lack of methods that handle non-textual features, both in annotation systems and search engines. Moreover, multimedia sharing websites like Flickr or YouTube, in addition to information provided by Wikipedia, offer a tremendous source of knowledge interesting to be explored. In this position paper, we address the automatic multimedia annotation issue, by proposing a hybrid system approach. We want to use unsupervised methods to find relationships between multimedia elements, referred as *hidden topics*, and then take advantage of social knowledge to label these resulting relationships. Resulting enriched multimedia content will allow to bring new user experience possibilities to the next generation television, allowing for instance the creation of recommender systems that merge this information with user profiles and behavior analysis.

1 INTRODUCTION

Technology is changing television practices through integration of computing capabilities and Internet connection. Despite the temporary setback encountered by Google with Google TV[1], effort they put into this direction shows that it is the future of television. Increasing integration of television and computer technology can move traditional television applications towards richer, sociable, computer-mediated user experiences. In particular, we believe that traditional television programs can be enriched by exploiting multimedia content available on the Internet. But this requires to improve such multimedia content, by developing new automatic annotation techniques, to provide richer annotations than existing methods and reduce the so-called “semantic gap”(Smeulders et al., 2000).

In the last few years, the number of video and image archives has dramatically increased. Therefore, it has become really hard to search and retrieve images or videos in an effective way, and one reason is the lack of semantic multimedia content based indexes needed for such retrieval (Tjondronegoro et al., 2005) (Yin et al., 2009). So, the multimedia content management system research became an important research area (Lew et al., 2006). Progress in the field of linked data (Bizer, 2009), with the semantic technologies, have helped a lot to efficiently manage multimedia data, by providing tools for describing and linking concepts and data (Bertini et al., 2010) (Dong and Li, 2006) (Akrivas et al., 2007). In parallel, progress has also been made in computer vision tools, allowing to analyze and process huge amount of multimedia data in a reduced amount of time (Smeaton et al., 2009); most of these solutions have exploited the bag-of-features approach to generate textual labels that represent the categories of the main and easiest to detect entities (objects and persons) in the video sequence keyframes. The TRECVid benchmark has shown an increasing improvement in the performance of appropriately trained classifiers (Snoek et al., 2006)(Hauptmann et al., 2008), but despite these advances, automatic multimedia content annotation is still a prob-

[1] http://www.google.com/tv/

Dumoulin J., Bertini M., Del Bimbo A., Mugellini E., Abou Khaled O. and Sokhn M..
Next Generation TV through Automatic Multimedia Annotation Systems - A Hybrid Approach.
DOI: 10.5220/0004128101920197
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 192-197
ISBN: 978-989-8565-25-9

lematic task. Moreover, the "semantic gap", defined by Smeulders et al. (Smeulders et al., 2000) as the lack of coincidence between the information that can be extracted from a visual data and the interpretation that a user has in a given situation, is still an unsolved problem.

On the one hand, statistical tools based on computer vision techniques provide an automatic mean to process large-scale multimedia databases, but they lack the knowledge that would be required to enrich the resulting elements with additional information and relations, and intervention of people is then required. On the other hand, social networks and also ontologies, through manual annotation and social cooperation, convey a tremendous and quite not used amount of knowledge, although it is still difficult to use it in a profitable way. An idea is to merge both domains, through a hybrid automatic multimedia annotation system, in order to enrich multimedia content, and bring new possibilities to the next generation television, such as interactivity, personalisation, sharing, etc.

In this paper we propose a hybrid approach for multimedia annotation systems, by merging automatic processing and data stemming from manual annotation, that could lead to the improvement of the next generation television experience. The rest of the paper is organized as follow. A brief state of the art for the multimedia annotation domain is made in Section 2. The proposed hybrid system approach is discussed in Section 3, and the application to the next generation television domain is proposed in Section 4. Finally, conclusions are drawn in Section 5.

2 MULTIMEDIA ANNOTATION

Two main visual tagging approaches are used to enrich or annotate multimedia content: manual annotation and automatic processing. Manual annotation, done by experts, is the most accurate and provide high level concept annotation. But the time needed to do such task is huge and thus way too expensive. In recent years there has been an extremely common use of social media such as tagging images (e.g. Flickr) and video (e.g. YouTube), resulting in a new human labeling source, which can be more or less structured (ontolgy based approaches such as FOAF[2] vs collaborative systems such as Wikipedia). Human labeling is closer to users, resulting in more precise understanding of the image (or video sequence). But when it comes to process huge quantity of documents, automatic processing techniques are essential.

Image classification methods currently considered state-of-art require extensive training (e.g. SVM (Meyer, 2001), boosting (Schapire, 2003)), and do not scale well with increasing number of concepts that must be recognized. These concepts are also to be determined, typically provided in a standard lexicon as LSCOM[3] or created ad hoc for the problem, while the interest of people may change over time or depending the introduction of some new event or interest.

Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (Hu, 2009) are statistical text modeling techniques allowing automated text documents indexing (Cai et al., 2008) (Nguyen et al., 2009) (Phan et al., 2008). By defining a visual analogy of a word (visual words) (Sivic and Zisserman, 2003), these techniques have recently been successfully applied for image content analysis tasks (both on pictures and videos), such as classification of scenes (Bosch and Zisserman, 2006), multimedia content retrieval (Hörster et al., 2007) (Lienhart and Slaney, 2007) or object localization in images (Andreetto et al., 2008) (Wang and Grimson, 2007). For the multimedia content retrieval, this can add a low-dimensional descriptor, enabling efficient retrieval in very large databases (Hörster et al., 2007).

Several recent approaches try to bring closer both manual annotation and automatic processing, using social tags to help in the creation of supervised classifiers, typically to provide negative examples of concepts to train (Setz and Snoek, 2009) (Li and Snoek, 2009), or to extend more easily the size of the lexicon of concept detectors using user-generated videos as training material (Ulges et al., 2010). Nearest-neighbors methods have received large attention due to the increase of training data available in social media and to the appeal of using an unsupervised approach, e.g. to suggest tags to images (Li et al., 2009) based on their visual similarity with a visual neighborhood. A baseline approach that is based on NN images to annotate new images was proposed in (Makadia et al., 2008), showing how a fusion of different features is fundamental for achieving the best performance, while weighted neighborhood of keywords is used for tag propagation in (Guillaumin et al., 2009). One interesting approach (Tsai et al., 2011) is based on the use of "visual synsets", clusters of visually similar images to whom are associated sets of weighted tags that are semantically similar. (Ballan et al., 2010) have proposed a system for video tag sug-

[2] Friend Of A Friend ontology: http://www.foaf-project.org/

[3] Large Scale Concept Ontology For Multimedia: http://www.lscom.org/

gestion and temporal localization, based on the similarity of keyframes with visual images that have been manually annotated with the same tags on Flickr. In this approach, the tags of a video are used to download images from Flickr (annotated with these words) and these images are then used to evaluate the visual similarity with the keyframe of the video, to locate the visual concepts over time. Most these methods rely on global and compact visual features and CBIR techniques to speed the computation of visual neighborhoods.

3 HYBRID APPROACH

Our goal is to enhance the television user experience, by associating rich annotated multimedia resources, in an interactive and personalized way to the program the user is watching. For this purpose, we propose a hybrid multimedia annotation system. As said before, on the one hand, statistical tools are able to process multimedia contents quite efficiently, but do not scale well in terms of size and changes in the lexicon of concepts to be detected. On the other hand, a huge amount of user generated knowledge is available through social networks, and also ontologies. We want to take advantages of both parts, by merging them together, as schematized in Figure 1.

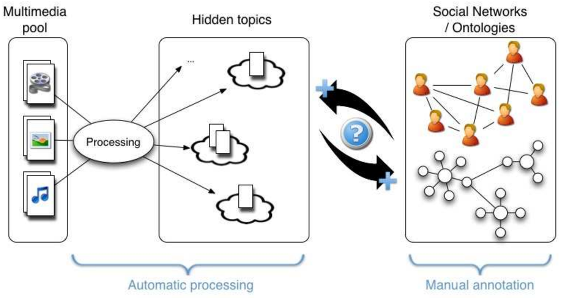

Figure 1: Hybrid annotation system: hidden concepts - social networks and ontologies.

The association between visual topics and textual topics (Monay and Gatica-Perez, 2004) is still an unsolved problem (e.g. due to the sparsity of tags (Qi et al., 2012)), and its resolution could lead to this purpose: take advantage of the knowledge available in both social networks and ontologies, associating the textual latent topics to the visual latent topics; these latter topics could be used also to reduce the dimensionality of the descriptors used in the bag-of-visual-words approach, e.g. describing visual content using an histogram of topics instead of a larger histogram of visual words.

3.1 Unsupervised Techniques for Hidden Topics Analysis

Applied to a pool of multimedia data, the use of unsupervised techniques, such as multi-modal LDA, results in the extraction of relationships between the different multimedia content.

The idea of using such graphical models to discover hidden topics is related to the automatic video analysis based on unsupervised methods domain. For now, these tasks are less accurate than semi-supervised or manually methods. But with the exponential growth of multimedia content (e.g. YouTube reported in May 2012 that more than 72 hours of videos are uploaded every minute), use of unsupervised methods appears essential to us, and therefore enhance these methods is a critical task. Automatic techniques also allow to do tasks much more faster than a person would with a manual system, and this is important in order to one day have real time video analysis systems.

3.2 Social Annotations and Knowledge

In recent years, social sites like Flickr or YouTube have seen a radical increase of the number of users, which also brought an increasing amount of annotated multimedia content. In most cases, people add labels (also called *tags*) to their content, also defined as *folksonomy*, providing a huge amount of contextual and semantic information. This is mainly used by users to organize and access content. But while these tags are often overly personalized, imprecise and ambiguous (Kennedy et al., 2006), they convey a rich information, and they can be very useful. Indeed, as presented in (Ballan et al., 2010), it is possible to use existing social annotations, in order to build tag suggestion systems for video. Moreover, using visual similarity between frames, it is also possible to define temporal localization of the suggested tags.

If social annotations provide a good knowledge base, it lacks in structure. Ontologies could be another way to enrich the manual annotation approach. And semantic relationships represented between concepts can bring new kinds of information. Social annotations and ontologies approach could be both used separately, but perhaps the idea of a *folksontolgy* (Damme et al., 2007) could be an interesting direction.

4 THE NEXT GENERATION TELEVISION

If a huge work has been made to enhance the graphical user interfaces of our computers for the last decades, this has not been the case for our television, so our TV user experience has not really evolved. Indeed, main efforts to image and device quality have been made, until recently. In parallel, the number of available channels and features has increased, making the experience less user friendly. With the advent of good quality Internet connections, we have seen the appearance of connected TV[4], allowing to browse the Web, use applications, or play small embedded games, directly within the TV. But, if it is the main improvement brought to TV, this is not really convenient, mainly because of the use of standard remote control (Cooper, 2008), and not a lot of people use such features. This could be an explanation regarding the limited success encountered by the Google TV or Apple TV. But no doubt it is just a timing question, and other similar products will meet success very soon.

Figure 2 shows a possible application schema, in which our proposed hybrid automatic multimedia annotation system, by enriching multimedia content with additional meta-information (resulting in a "multimedia content ++"), could allow various TV experience improvement ideas.

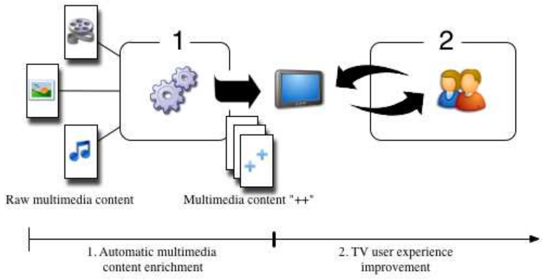
1
2
Raw multimedia content
Multimedia content "++"
1. Automatic multimedia content enrichment
2. TV user experience improvement

Figure 2: Application: 1) Automatic multimedia content enrichment system; 2) TV User interaction.

Indeed, we plan to take advantage of this "multimedia content ++", for example to improve navigation through TV Guide, but also in the program content itself. Additional information could be made available to the user, according to what he is watching, taking in account his interests. By providing new ways of community interactions, we think that social aspects could take new dimensions, also helping to build recommender or collaborative filtering systems for example.

Actually, a recommender system, based on social information, could represent an interesting starting point in our vision of TV user experience enhancement. If recommender systems are becoming more and more popular (i.e. Last.fm or Pandora for music, MovieLens or Netflix for movies, etc.), mainly because of the ever increasing user generated content on multimedia sharing websites like YouTube, they are generally not able to perfectly meet the user needs. We think that building a system able to dynamically improve the comprehension of the user needs, while he is using the system through the time, and requires the least amount of explicit user interaction, eliciting information from other sources or analyzing the user behavior, could give better results. Nowadays, people do not like to spend time in system configurations. So, using existing information defining the user interests, like his Facebook profile, is a good way to create an initial user profile. This could be done by analyzing his profile information (interests, etc.), but also his generated content (tags, links, comments, etc.). As users evolve, so should evolve their profiles. In addition to a textual user profile builder, we would like to integrate a behavioral user profile builder. It could use standard information, like the way the user use the system (zapping, the user stops a programme before it ends, etc.), but more interesting would be to take advantage of more human behavioral activity, by defining his attention: is the user talking with other people? Does he look attentively at the TV, or is he sleeping? The inclusion of visual sensors on TV sets, as the gesture based control system in the Samsung Smart TV, make this type of analysis possible. Finally, the whole system, as schematized in Figure 3, could propose interesting content to the user. It will be able to efficiently select the more interesting videos, thanks to their annotations, automatically added by the first part of the system.

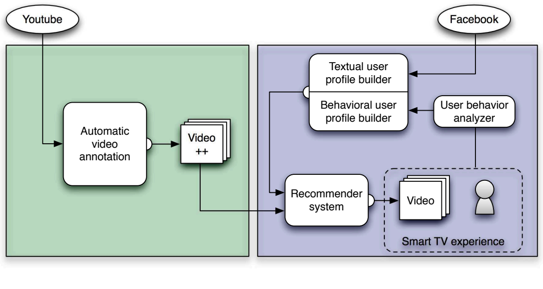

Figure 3: The proposed recommender system: social media and video annotation (left), user profile and behavior analysis (right) contribute to the improvement of the TV user experience.

[4] http://en.wikipedia.org/wiki/Smart_TV, Intel & Smart TV: http://www.intel.com/content/www/us/en/smart-tv/smart-tv-with-intel-inside.html

5 CONCLUSIONS

As the images and videos databases become huge, it is essential to find efficient, fast, but also accurate techniques to annotate and enrich multimedia contents in an automatic way. Both automatic processing and manual annotation methods have advantages, and we propose to merge them in a hybrid system that links user-generated annotations with visual content analysis techniques. Resulting enriched multimedia content will allow us to improve current television experience (interaction, personalization, sharing, etc.), in order to meet expectations of the users of the next generation television systems. As a starting point in this vision, we would like to build a dynamic recommender system, that exploits both video content annotations and a user profile builder based on behavioral analysis.

REFERENCES

Akrivas, G., Papadopoulos, G., Douze, M., Heinecke, J., O'Connor, N., Saathoff, C., and Waddington, S. (2007). Knowledge-based semantic annotation and retrieval of multimedia content. In *Proc. of 2nd International Conference on Semantic and Digital Media Technologies*, pages 5–6, Genoa, Italy.

Andreetto, M., Zelnik-Manor, L., and Perona, P. (2008). Unsupervised learning of categorical segments in image collections. *Proc. of Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08.*, pages 1–8.

Ballan, L., Bertini, M., Bimbo, A. D., Meoni, M., and Serra, G. (2010). Tag suggestion and localization in user-generated videos based on social knowledge. In *Proc. of second ACM SIGMM Workshop on Social Media (WSM)*, pages 3–8.

Bertini, M., Amico, G. D., Ferracani, A., Meoni, M., and Serra, G. (2010). Web-based Semantic Browsing of Video Collections using Multimedia Ontologies. In *Proceedings of the international conference on Multimedia - MM'10*, pages 1629–1632, Firenze, Italy. ACM.

Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE Intelligent Systems*, 24(5):87–92.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bosch, A. and Zisserman, A. (2006). Scene classification via pLSA. *Proc. of ECCV*.

Cai, D., Mei, Q., Han, J., and Zhai, C. (2008). Modeling hidden topics on document manifold. In *Proc. of the 17th ACM conference on Information and knowledge management*, pages 911–920.

Cooper, W. (2008). The interactive television user experience so far. *Proc. of the 1st international conference on Designing interactive user experiences for TV and video (UXTV)*, 44:133.

Damme, C. V., Hepp, M., and Siorpaes, K. (2007). FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. *Social Networks*, 2:57–70.

Dong, A. and Li, H. (2006). Multi-ontology Based Multimedia Annotation for Domain-specific Information Retrieval. *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing - Vol 2 - Workshops*, 2:158–165.

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309 –316.

Hauptmann, A. G., Christel, M. G., and Yan, R. (2008). Video retrieval based on semantic concepts. In *Proceedings of the IEEE*, volume 96, pages 602–622.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.*, pages 177–196.

Hörster, E., Lienhart, R., and Slaney, M. (2007). Image retrieval on large-scale image databases. *Proc. of Conference on Image and video retrieval*, pages 17–24.

Hu, D. (2009). Latent Dirichlet Allocation for Text, Images, and Music. *cseweb.ucsd.edu*, pages 1–19.

Kennedy, L. S., Chang, S. F., and Kozintsev, I. V. (2006). To search or to label?: predicting the performance of search-based automatic image classifiers. *Proc. of the 8th ACM international workshop on Multimedia Information Retrieval (MIR)*, pages 249–258.

Lew, M. S., Sebe, N., Djereba, C., and Jain, R. (2006). Content-Based Multimedia Information Retrieval : State of the Art and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19.

Li, X. and Snoek, C. (2009). Visual categorization with negative examples for free. In *Proc.s of ACM Multimedia*, pages 661–664.

Li, X., Snoek, C. G. M., and Worring, M. (2009). Learning Social Tag Relevance by Neighbor Voting. *IEEE Transactions on Multimedia*, 11:1310–1322.

Lienhart, R. and Slaney, M. (2007). pLSA on large scale image databases. *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 1217–1220.

Makadia, A., Pavlovic, V., and Kumar, S. (2008). A new baseline for image annotation. In *Proc. ECCV*, pages 316–329.

Meyer, D. (2001). Support Vector Machines. *R News*, 2(2):23–26.

Monay, F. and Gatica-Perez, D. (2004). PLSA-based image auto-annotation: constraining the latent space. In *Proc. of ACM Multimedia*, pages 348–351.

Nguyen, C., Phan, X., and Horiguchi, S. (2009). Web Search Clustering and Labeling with Hidden Topics. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(3).

Phan, X., Nguyen, L., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proc. of the 17th international conference on World Wide Web*, pages 91–100.

Qi, G.-J., Aggarwal, C., Tian, Q., Ji, H., and Huang, T. (2012). Exploring context and content links in social media: A latent space method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):850 –862.

Schapire, R. E. (2003). The Boosting Approach to Machine Learning An Overview. *MSRI Workshop on Nonlinear Estimation and Classification*, 7(4):1–23.

Setz, A. and Snoek, C. (2009). Can social tagged images aid concept-based video search? In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pages 1460–1463.

Sivic, J. and Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2.

Smeaton, A., Over, P., and Kraaij, W. (2009). High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. *Multimedia Content Analysis*, pages 1–24.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. of ACM Multimedia*.

Tjondronegoro, D., Chen, Y.-P. P., and Pham, B. (2005). Content-based video indexing for sports applications using integrated multi-modal approach. *Proc. of ACM Multimedia*, page 1035.

Tsai, D., Jing, Y., Liu, Y., Rowley, H., Ioffe, S., and Rehg, J. (2011). Large-Scale Image Annotation using Visual Synset. *Proc. of ICCV*.

Ulges, A., Schulze, C., Koch, M., and Breuel, T. M. (2010). Learning automatic concept detectors from online video. *Computer Vision and Image Understanding*, 114(4):429–438.

Wang, X. and Grimson, E. (2007). Spatial latent dirichlet allocation. *Proc. of Neural Information Processing Systems Conference*, pages 1–8.

Yin, Z., Li, R., Mei, Q., and Han, J. (2009). Exploring social tagging graph for web object classification. *Proc. of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, page 957.