CONTENT-BASED ANALYSIS FOR ACCESSING AUDIOVISUAL ARCHIVES: ALTERNATIVES FOR CONCEPT-BASED INDEXING AND SEARCH

Tinne Tuytelaars

ESAT/PSI - IBBT
KU Leuven, Belgium
Tinne.Tuytelaars@esat.kuleuven.be

ABSTRACT

Huge amounts of audiovisual material have been digitized recently, resulting in a great source of information relevant both from a cultural and historical point of view. However, in spite of millions of man hours spent on manual annotation and recent advances in (semi-)automatic metadata generation, accessing these archives and retrieving relevant information from them remains a difficult task. Up to recently, the main paradigm to open up archives by automatic tools for audiovisual analysis has been a concept-based indexing and retrieval oriented approach. However, this approach has its limitations, in that it does not scale well, it requires strong supervision, and does not really match well to the user's needs.

In this paper, we discuss some upcoming alternative approaches that try to overcome or circumvent some of these issues. This includes i) the use of knowledge modeling to bridge the semantic gap; ii) on-the-fly learning of new, user-defined concepts; and iii) weakly supervised methods that learn from associated text data. We also discuss what we consider important open issues at this time that deserve more attention from the research community.

1. INTRODUCTION

Despite a decade of research on multimedia preservation, disclosure and access, methods for understanding multimodal content as well as interactive access tools are still not widely deployed in practice. This especially holds for audiovisual material (still images, video and audio recordings). There already exists an impressive set of methods, tools and good-practices addressing the demands of our contemporary information society. However, scaling up and putting the complex pieces together in service of the various communities of digital archives is still not a fait-accompli. More and more emphasis is now being placed on opening up the traditionally introspective archival practice, allowing innovative tools and practices to benefit improved content exploitation.

At the same time recent advances in semantic understanding of multimodal content have been impressive. Static scenes such as typical landmarks can be recognized and retrieved from an archive containing millions of images in a fraction of a second [1, 2]. Reliable detectors for a variety of object, scene or action classes are readily available (e.g. [3, 4]). Automatic speech recognition has reached acceptable performance levels. Text analysis has moved well beyond keyword search, discovering topics and trends and identifying semantic roles. However, in a more diversified, real world setting, only the most basic tools for content-based audiovisual analysis have really proven their value to date (e.g. shot cut detection, indexing based on ASR output, video copy detection, frontal face detection, or color based similarity search).

Finding relevant entities in existing digital libraries is usually accomplished via keyword-based search. Documents are typically annotated with metadata by the librarian, indexed by the system, and then ready to be searched by matching query terms against these indices. As such, it does not come as a surprise that the main paradigm to leverage the progress in content-based analysis so far has been a concept-based indexing and retrieval approach, that closely resembles current text-based practices. However, unlike for text, we believe this approach, when applied to audiovisual material, has some intrinsic shortcomings that are difficult to overcome.

In this paper, we first summarize the main ideas behind concept-based indexing and retrieval and discuss its limitations (Section 2). Next, we survey some alternative approaches, that still are not that well developed, yet hold good promise in overcoming some of the issues with the concept-based approach (Section 3). This includes the use of knowledge modeling to bridge the semantic gap; on-the-fly learning of new, user-defined concepts; and weakly supervised methods that learn from associated text data. In Section 4, we discuss some open issues and research themes that require further investigation. Finally, Section 5 concludes the paper.

2. CONCEPT-BASED INDEXING AND RETRIEVAL

The concept-based indexing and retrieval paradigm has been advocated, amongst others, by groups at CMU [5] and University of Amsterdam [6]. First, a list of (visual) concepts

that are believed to be relevant for a user is collected. For each of these concepts a classifier is trained, usually starting from a set of low-level features and building on generic machine learning tools. While still far from perfect, these classifiers are usually good enough to retrieve a few relevant images or videos with reasonable precision (albeit poor recall). The whole archive is then processed by each of these classifiers in an offline phase, and the content is indexed accordingly. This allows for fast retrieval of the relevant content whenever a query from the predefined list is entered. This scheme is usually demonstrated with scene-level concepts, like *crowds* or *explosions*, but can equally be applied for people (face recognition) or specific locations.

While this approach has shown top results at benchmark initiatives such as TRECVID [7], it also has some inherent limitations:

- *Mismatch between concepts that can be recognized and concepts users are interested in:* Collecting the list of concepts for which to train classifiers, is not as straightforward as it seems, as it is a balancing act between usefulness and technical feasibility, i.e. what users find relevant (often high-level concepts, like *Perestroika* or *economic crisis*) and what can be learnt by generalizing from a set of example images.

- *High level of supervision:* Learning a good classifier for a given concept often requires the availability of hundreds of labeled examples. This is a labour intensive process, that has to be repeated for each new concept added to the list.

- *Limited scalability:* Concepts have to be defined manually. Training data has to be collected. Processing the archive has a complexity linear in the number of classifiers (concepts). This limits the number of concepts that can be learnt to a few hundred or a few thousand.

- *Static:* A list of concepts is defined once and for all. Yet audiovisual content and what users find relevant, evolves over time. This is most evident for names of people and events.

- *Limited flexibility for the user:* The user can only select a topic from a predefined list. This is in strong contrast to the free text search people are familiar with when searching the internet, and therefore scores low on user satisfaction.

- *Search oriented:* The indexing structure is especially well suited for a search-based interface. However, for browsing or exploring an archive, it is as of now not clear whether concepts are the way to go. Some concepts may not be relevant for search, but useful nonetheless – think e.g. of face detection.

Various surveys on concept-based video retrieval are already available (e.g. [8]), and the basic methods seem to have become more or less mature. However, while the methods can still be refined, the above characteristics seem to intrinsically limit what can be achieved with a concept-based approach. Therefore, we rather focus in this paper on alternative approaches that have recently emerged and hold good promise to overcome at least some of the issues of concept-based search raised above.

3. ALTERNATIVE APPROACHES

3.1. Use of knowledge modeling to bridge the semantic gap

In [9], Snoek *et al.* propose a way to overcome one of the main limitations of a pure concept-based search: the fact that a user cannot freely decide what to search for. They consider this as part of the semantic gap problem (linking low level descriptors to high-level / semantic information needs from the user), where the concept detectors can be seen as some intermediate representation. To this end, they manually create links between concept detectors and Wordnet synsets. This allows to 'translate' a free-text query into the most related visual concept. This definitely alleviates the problem to some extent and greatly improves the user experience. Moreover, they show that a thesaurus consisting of 100-200 concepts is large enough to obtain state-of-the-art performance.

3.2. On-the-fly learning of new, user-defined concepts

An alternative solution has been explored recently by Parkhi *et al.* in [10]. Instead of restricting the search to a predefined list of concepts, they train new visual concept classifiers on-the-fly while the user is waiting, in particular for specific persons. This is feasible in a reasonable time (a matter of seconds) by appropriate preprocessing of the video corpus (face detection, tracking, and description), such that it becomes searchable for any person. Based on the user query, example images are then collected from the internet using a text-based image search engine, and used for training a discriminative model. Since the model is linear, and the features on which it is applied have been precomputed, it runs over the entire archive in a matter of seconds.

Of course the quality of the final result depends on the quality of the results returned by the text-based search engine. Fortunately, this can leverage the text surrounding images on the web, and therefore is usually quite reliable in terms of precision of the top-ranked results. There is a risk though that the images retrieved from the internet may not be representative for the content found in the archive (e.g. in terms of illumination conditions). Alternatively, one could ask the user to upload some images, but that requires a larger effort from the user, and may therefore not be very attractive.

3.3. Weakly supervised methods learning from associated texts

Finally, several people have looked into another way of exploiting multimodal data: the possibility to learn models from associated text material such as subtitles, transcripts, related websites, etc. This approach has mostly been investigated for identifying people, exploiting the co-occurrence statistics of names and faces (e.g. [11, 12, 13]), but has also been applied to locations [14] and actions [15].

The major challenge for these methods is to scale up from a rather constrained setting (e.g. one TV series) to a more varied data set (e.g. news). To the best of our knowledge, no real large-scale experiments with this type of techniques has been demonstrated so far. Even so, also on a smaller scale, such techniques can already be useful, e.g. to align a transcript to a particular episode, such that the descriptions in the transcript can get appropriate time stamps and the relevant fragment within a video can be retrieved rather than the video as a whole.

4. DISCUSSION

The accuracy of current state-of-the-art video retrieval methods (be it concept-based or not) is still rather low, and this makes archivists reluctant of actually including such noisy, untrustworthy data into their carefully maintained archive. And indeed, wrong or incomprehensible results often put off users more than incomplete results. However, even if not directly used for searching, content-based analysis techniques could be applied in a way that is transparent to the user, e.g. for reranking the retrieved results, for suggesting similar content, and so on. We believe such more creative ways of using the output of the audiovisual analysis should be explored further.

An alternative approach could be interactive schemes that have the archivist-in-the-loop, where he/she can control the quality of the detectors before the results are actually merged into the system. This requires different, dedicated interfaces focussing on annotation rather than retrieval.

Throughout this survey, we focused mostly on the analysis of the visual component of the data. However, there is also a huge potential in exploiting the audio channel to the fullest. This includes not only speech recognition, but also a content-based analysis of the sounds.

More important than advancing the state-of-the art in semantic understanding of multimodal content may be the tackling of an underlying problem, namely the disparity between technology and user needs. If we are to leverage our investments in technology we must foster a stronger connection between the workflow practices and potentially useful technical capabilities. This should not be limited to the traditional ways of accessing an archive, using keyword-based search, but also look beyond that, into really novel ways of interacting with the archive: exploring, browsing, or experiencing the archives in various ways.

Different user groups may also have very different needs when it comes to accessing the archives (see also [16]). Media professionals may be quite familiar with the database and the tools, but often have strong time constraints. Researchers on the other hand are willing to spend some time and effort in retrieving the content, if that brings them to the data they want. Moreover, they do not just want illustrative or qualitative results, but are also interested in quantitative aspects (e.g. *"How often did this politician talk about that topic on national television over the last year ?"*), which puts much stronger constraints on accuracy of the system. Finally, home users may be in it just for fun. They want simple interfaces, and may not have a clue what kind of data is available in a system. These different backgrounds need to be taken into account, and it's unlikely that one paradigm can serve all of them.

Finally, content-based analysis can go hand-in-hand with this other source of information in the form of social tagging. Both are not very reliable, but they seem fairly independent, and therefore one modality could serve to corroborate the results of the other and v.v.

5. CONCLUSIONS

In this paper we have discussed automatic tools for content-based analysis of audiovisual content, with the purpose of opening up large scale multimedia archives. While concept-based search is still the most widely used paradigm to date, we did not give a detailed overview of these methods, as there are already plenty of good surveys out there. Instead, we stressed some fundamental limitations that are intrinsic to this approach, and pointed to some possible alternative schemes to overcome or circumvent some of these. At this point, these alternatives are still in a development phase. Yet they hold good promise to deliver a more enjoyable user experience.

Acknowledgements

We greatly acknowledge support by the EC FP7 Project AXES ICT-269980. We would also like to thank the AXES partners for the fruitful discussions on this and related topics.

6. REFERENCES

[1] S. Gammeter, T. Quack, D. Tingdahl, and Luc van Gool, "Size does matter: improving object recognition and 3d reconstruction with cross-media analysis of image clusters," in *European Conference on Computer Vision (ECCV 2010)*, 2010.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE*

Conference on Computer Vision and Pattern Recognition, 2008.

[3] D. McAllester P. Felzenszwalb, R. Girshick, "Cascade object detection with deformable part models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[4] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[5] A.G. Hauptman, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H.D. Wactlar, "Informedia at trecvid 2003: Analyzing and searching broadcast news video," in *Proceedings TRECVID workshop*, 2003.

[6] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1678–1689, 2006.

[7] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[8] C. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[9] C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.

[10] O. Parkhi, A. Vedaldi, and A. Zisserman, "Learning to recognize people on-the-fly," in *WIAMIS*, 2012.

[11] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. The, E. Learned-Miller, and D.A. Forsyth, "Names and faces in the news," in *Computer Vision and Pattern Recognition (CVPR)*, 2004.

[12] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Face recognition from caption-based supervision," *International Journal on Computer Vision*, vol. 96, no. 1, pp. 64–82, 2012.

[13] P.T. Pham, M.F. Moens, and T. Tuytelaars, "Cross-media alignment of names and faces," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 13–27, 2010.

[14] C. Engels, K. Deschacht, J.H. Becker, T. Tuytelaars, S. Moens, and L. Van Gool, "Automatic annotation of unique locations from video and text," in *British Machine Vision Conference*, 2010.

[15] L. Jie, B. Caputo, and V. Ferrari, "Who's doing what: Joint modelling of names and verbs for simultaneous face and pose annotation," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[16] M. Kemman, M. Kleppe, and H. Beunders, "Classifying users with a profile matrix to support requirements elicitation from a heterogeneous group of users," in *Proceedings WIAMIS*, 2012.