Collaborative Annotation of Multimedia Resources

Pierrick Bruneau[1], Mickaël Stefas[1], Mateusz Budnik[2], Johann Poignant[2], Hervé Bredin[3], Thomas Tamisier[1], and Benoît Otjacques[1]

[1] Public Research Centre - Gabriel Lippmann, 41 rue du Brill, L-4422 Belvaux
[2] LIG CNRS UMR 5217, BP 53, F-38041 Grenoble Cedex 9
[3] LIMSI-CNRS, BP 133, F-91403 Orsay

Abstract. Reference multimedia corpora for use in automated indexing algorithms require lots of manual work. The Camomile project advocates the joint progress of automated annotation methods and tools for improving the benchmark resources. This paper shows some work in progress in interactive visualization of annotations, and perspectives in harnessing the collaboration between manual annotators, algorithm designers, and benchmark administrators.

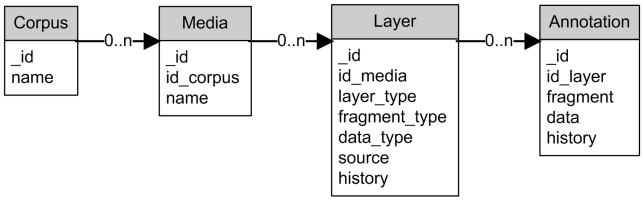
Keywords: multimedia annotation, interactive visualization, collaborative annotation

1 Introduction

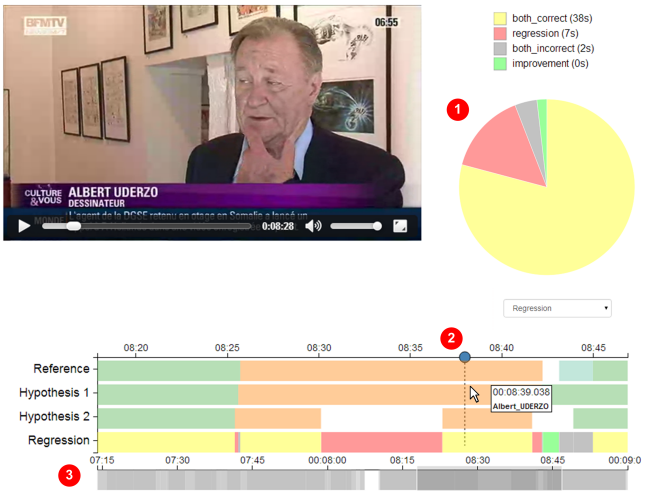
Wide amounts of multimedia data are accessible over the Internet, but for efficient search and retrieval, those have to be properly annotated. Such annotations can also be useful for improving the user experience in the context of multimedia content consumption. However, the growing pace of these data are largely exceeding the abilities of manual annotators: automated means are thus needed.

The Camomile project fosters the development of tools for the Multi-modal and Multi-language processing of Multimedia content (often gathered under the acronym MMM in this domain). This naturally includes contributing to the state-of-the-art of content-based multimedia indexing algorithms. But the project also considers tools for facilitating the constitution of realistic and reliable benchmark resources, thus favoring a general improvement of the contributed literature.

This paper presents on-going work, with a focus on perspectives for the latter aspect. Section 2 shortly introduces the data framework designed for the project, and overviews first results on the use of interactive visualization to support the error analysis of indexing algorithms. Then in Section 3, we consider the perspective of extending this fine-grained task to the context of multiple manual annotators and algorithm designers, that collaborate in the dual objective of improving the quality of both algorithms and the benchmark resources they are evaluated on.



(a) Entity-relation diagram of the annotation data.



(b) Overview of the visual tools for algorithmic results analysis. *1)* Summary statistics displayed in a pie-chart view. *2)* Synchronization of the video playback with an interactive annotation timeline. A tooltip shows the annotation data underlying glyph colors. *3)* A context line shows the summary of all loaded layers, and supports brushing interactions. The time scales are updated interactively.

Fig. 1. Data framework and visual tool in the Camomile project

2 Visual Analysis of Algorithmic Results

Annotations can be summarized as any kind of metadata affecting a fragment (with respect to space or time) in a given medium. For the discussion in this paper, we restrict to speaker (i.e., who speaks) and face (i.e., who is seen) annotations in video media, materialized by temporal fragments.

With a view to scale up the processing of benchmark corpora and associated algorithmic results, a general data framework (see Figure 1(a)) was implemented using NoSQL storage technologies. This design facilitates its further use and extension for a variety of clients, including mere web browsers. It logically stores annotations in a nested structure, gathering annotations in layers, and ultimately referring to *corpora*, i.e. benchmark databases.

Speech and video recognition practitioners are primarily concerned by the design of effective automatic indexing algorithms [2]. They classically assess the quality of their algorithms by aggregated quality metrics w.r.t. a benchmark database.

Taking advantage of the data framework evoked in Figure 1(a), and inspiration from tools such as Advene [1], we recently proposed to use visual and interactive tools to allow a finer analysis of algorithmic results [3] (see Figure 1(b)). With the joint use of timeline and pie-chart views, practitioners can perform the differential analysis of their algorithmic results w.r.t. the benchmark.

In the machine learning literature, algorithms are generally compared to what is often called a *ground truth*, the veracity of which not being questioned. As for automatic annotation, benchmark databases are the result of the intensive manual work of many annotators, and errors are likely to occur in the reference layers used by algorithm designers. Actually, in annotation challenges such as REPERE [5], the *adjudication* step is specially dedicated to debunk potential mistakes highlighted by users of the benchmark [4].

The next section discusses the current perspectives in the project, where multiple manual annotators and algorithms can be considered, with referees filtering the notifications issued by algorithm designers (further known as *adjudicators*).

3 Perspectives in Collaborative Support

In a first stage, the tools presented in the previous section could be extended for the notification of errors in reference annotation layers. Algorithms may occasionally recover the actual ground truth, that was not correctly annotated in the benchmark reference. This capability could be exploited, through the definition of a specialized kind of layer, allowing algorithm designers to highlight the suspected error in the reference. Its visual restitution could emphasize this information, facilitating the decision process by the adjudicator.

We could also add full editing functionalities to the current tool. But taken independently, this feature retains the usual, long, and error-prone annotation sessions, just solving for a more efficient way of distributing the workload. In itself this can be seen as a notable improvement w.r.t. the current, completely manual, procedure, but we decided to aim at a more ambitious target, and harness such distributed editing functionalities towards a greater integration of learning, visualization, and interaction with the data.

Classical automated indexing algorithms proceed in a *supervised* fashion: a predictor is trained to minimize its errors on media where the annotations are known *a priori*. This way of proceeding is demanding in manual annotation effort. Alternatively, *active learning* [6] starts from a completely unlabeled setting, and selects elements to be annotated iteratively, so as to reduce the uncertainty of the model at the current step. We propose to combine this approach with the data framework and visual tools respectively evoked by Figures 1(a) and 1(b). An active learning algorithm could feed a waiting queue of video segments to be annotated in a server. Manual annotators could then process this workload in

a distributed fashion. From the interactive point of view, the required research pertains to providing an adequate context for annotators, so that they make informed decisions, and appropriate controls to reduce the noise of annotations (e.g. dynamic dictionary of named entities in the case of speaker annotations).

Another important direction of research regards findings relevant visualization and interaction metaphors at the corpus scale. Adjudicators could for example inspect and filter the corpus for relevant statistics, navigate in the history of actions taken by all annotators and algorithms, or estimate the annotation coverage currently achieved. Possibilities of integrating the active learning paradigm in the shape of a visual data mining tool, where the algorithm designer and the annotator use a visual abstraction of the currently learned model, both for picking new elements to annotate, and understanding the ongoing learning process, could also be sought.

4 Conclusion

This paper summarized work in progress in the context of the Camomile project. Shortly recalled here, a low-granularity visual tool has already been proposed to support the analysis of errors made by annotation algorithms. We then drew perspectives on scaling up the analysis, and harnessing machine learning and collaborative interactions between involved actors using visual tools.

Acknowledgments

This work was done in the context of the CHIST-ERA CAMOMILE project funded by the ANR (Agence Nationale de la Recherche, France) and the FNR (Fonds National de la Recherche, Luxembourg).

References

1. O. Aubert and Y. Prié. Advene: active reading through hypervideo. *ACM conference on Hypertext and hypermedia*, pages 235–244, 2005.
2. H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. B. Le, T. Napoléon, G. Hua, C. Barras, S. Rosset, L. Besacier, J. Verbeek, G. Quenot, F. Jurie, and E. H. Kemal. Fusion of speech, faces and text for person identification in TV broadcast. *LNCS 7585 (ECCV 2012)*, pages 385–394, 2012.
3. P. Bruneau, M. Stefas, H. Bredin, A.-P. Ta, T. Tamisier, and C. Barras. A web-based tool for the visual analysis of media annotations. In *International Conference on Information Visualisation*, 2014.
4. A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The REPERE Corpus: a Multimodal Corpus for Person Recognition. In *LREC*, 2012.
5. J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly. A Presentation of the REPERE Challenge. In *International Workshop on Content-Based Multimedia Indexing*, pages 1–6, 2012.
6. A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pages 231–238, 1994.