Collaborative Annotation for Person Identification in TV Shows

Matheuz Budnik, Laurent Besacier, Johann Poignant, Hervé Bredin, Claude Barras, Mickael Stefas, Pierrick Bruneau, Thomas Tamisier

Collaborative Annotation for Person Identification in TV Shows

Matheuz Budnik[1], *Laurent Besacier*[1], *Johann Poignant*[2], *Hervé Bredin*[2], *Claude Barras*[2], *Mickael Stefas*[3], *Pierrick Bruneau*[3], *Thomas Tamisier*[3]

[1]Laboratoire d'Informatique de Grenoble (LIG), Univ. Grenoble Alpes, Grenoble, France
[2]LIMSI, CNRS - Orsay, France
[3]LIST, Luxembourg

Mateusz.Budnik@imag.fr

Abstract

This paper presents a collaborative annotation framework for person identification in TV shows. The web annotation front-end will be demonstrated during the *Show and Tell* session. All the code for annotation is made available on *github*. The tool can also be used in a crowd-sourcing environment.

Index Terms: multimodal person identification, collaborative annotation, active learning, data collection.

1. Introduction

1.1. Context - Camomile project

One of the objectives of the Camomile project [1] is to develop a first prototype of a collaborative annotation framework for 3M (Multimodal, Multimedia, Multilingual) data, in which the manual annotation is done remotely on many sites, while the final annotation is localized on the main site.

1.2. Demo Content

The demo presents our annotation interface for person identification in TV shows. Specifically, tracks, i.e. spatio-temporal segments, are annotated with names of people they feature. The tool is supported by a web annotation front end, a server to centralize annotations as well as an active learning backend that are all described in section 2 of this paper. A dry run evaluation (small-scale annotation campaign) is also presented in section 3.

2. Collaborative annotation framework

In this paper, the focus is on manual annotations from multiple users. The proposed collaborative annotation framework follows a client/server architecture (see figure 1).

Figure 1: Overview of the Collaborative annotation using Camomile tools

The frontend and backend parts are coordinated through the server. The data exchanges occur solely via the HTTP protocol, facilitating the design of interoperable software components. The server focuses essentially on data and authentication management tasks, leaving the application logic to both client and backend parts. More details on the framework can be found at [1].

[1]`https://camomile.limsi.fr`

2.1. Camomile server

The server component provides access and basic CRUD operations (create, update, delete) for the resources, which can be any pieces of 3M data (corpus, media, layers and annotations). The web server is built on *node.js* with the *express framework* and *mongodb* as data storage solutions. The latest version of the server is available at [2].

2.2. Web annotation front-end

Figure 2: Overview of the web front-end UI. 1) Video player displaying the track to annotate and the synchronized *context bar*. The red glyph shows the track to annotate, and additional annotations appear in light gray. 2) The *context bar* configuration and the metadata field. The latter displays the video title, and reveals additional details when activated. 3) The textfield to type the annotation. Multiple annotations are supported, and summarized in a table.

An overview of the visual tool is shown in Figure 2. It uses display features provided by HTML5 and D3.js [3]. The *angular.js* framework [4] provides an efficient MVC framework to easily coordinate multiple views. The latest version of the tool is available at [2].

Though there are two main use cases (see 2.2.1 and 2.2.2), components are mostly the same for both: the track or the frame to annotate is displayed in a HTML5 video player and its metadata is shown under the player. The input of multiple annotations is supported by a textfield and a summary table.

2.2.1. Annotating speech

In the first use case, a user has to name the speaker in the track. The video player, restricted to the track, allows to explore it at will.

Owing to the iterative nature of the active learning algorithm, the current speaker might have already been annotated elsewhere in the video. Seeking beyond the current track might reveal such annotations. This obervation led to a *context bar* being proposed, which provides the usual features of a seek bar, while revealing annotations performed in previous steps of the active learning as overlay. A time span can be parametrized around the current track, highlighted in red (see Figure 2). Hovering over contextual annotations displays a tooltip containing a video thumbnail and the associated annotation.

2.2.2. Annotating faces

Annotating people appearing in tracks has also been considered. To facilitate the overlay of bounding boxes on faces to annotate, the display is restricted to a single frame. Doing so also lets the active learning backend require to annotate only a specific person in the frame.

2.3. Active learning backend

In order to make the annotations provided by the users more relevant, an active learning system was developed. The approach can be described unsupervised since no biometric models are trained and only speaker clustering is performed (ideally, each cluster corresponds to an individual person). In this method a hierarchical clustering algorithm was used following the approach presented in [5]. The clusters consist of tracks: speech (based on speaker diarization), face (the result of face tracking) or both in the case of multimodal clusters. In the latter case, the distance between tracks from different modalities is based on the output of a multilayer perceptron.

At each step of the algorithm, the user is presented with a set of tracks for annotation. Then, the clustering is refined when new annotations are introduced. The label given to a particular track is propagated to the whole corresponding cluster. Next, a selection strategy is applied, which tries to verify the correctness of the annotated clusters or to label new ones, and feeds a queue of annotations to be processed by annotators using the interface in Figure 2. Already labelled tracks are provided to the queue to enable the context bar (see section 2.2.1). More in-depth description of the method can be found in [6].

3. Dry run evaluation

3.1. Use case : multimodal speaker annotation

A dry run annotation was done to evaluate the efficiency of the whole system. The task consisted of annotating speech tracks extracted automatically following the approach presented in [7]. Each participant was given a video fragment corresponding to the time frame of a speech track and was asked to name the person speaking at the moment. Beyond tuning the context bar, the user could also access to the whole video. Due to the nature of the videos (TV news broadcasts), most people were presented either by an overlaid text or by a spoken name.

3.2. Quantitative analysis

9 users were involved in the dry run. The annotations were done simultaneously and lasted for around 1.5 hours per user. In this run only the speaker annotation scenario was tested. The corpus consisted of 62 videos from the REPERE dataset [8], which included TV debates, news programs and parliamentary broadcasts among others. During the run, a total of 716 speech tracks (81mn) were annotated. Additionally, 654 tracks (68mn) were marked as skipped (tracks which do not contain speech, but music, external noises, etc.). The median annotation time is equal to 10.8s. Additionally, because of the clustering present in the system, the annotations were propagated to the corresponding clusters. This produced a total number of 3504 labeled tracks (including the 716 annotated manually) with the total time equal to 7.81h. As a *by-product*, the use of the multimodal clusters during the dry run enabled to get face annotation (1973 head annotations, for a total duration of 5.47h).

3.3. Qualitative analysis

After the dry run, participants had to fill a feedback questionnaire about the web front-end. While the users were mostly satisfied with the front-end, they pointed out some bugs and lines for improvement. For example, the need for additional tooltips and titles was expressed. Modifications following these suggestions were applied since then. Though the proposed context bar was deemed as an interesting idea, it has not been judged as sufficiently self-explanatory. On the short term, we added a video thumbnail when hovering over the associated annotation, but the chosen visual mapping and layout should be refactored.

4. Supporting content and demo scenario

A video presenting the annotation process was recently published on Youtube [2]. All the code that supports the camomile server, client and active learning is available on the *camomile github* [2].

A poster will be presented with all the latest achievements obtained during the Camomile project. The annotation interface will be also demonstrated while video will be played continuously during the *show and tell* session.

5. Acknowledgements

This work is done within *Camomile* project funded by the French National Research Agency (*Agence Nationale de la Recherche - ANR*) and the Luxembourgish National Research Fund (*Fonds National de la Recherche - FNR*) in the CHIST-ERA international program.

6. References

[1] "Camomile framework," https://camomile.limsi.fr/doku.php?id=framework, 2014.

[2] "CAMOMILE Project," https://github.com/camomile-project, 2014.

[3] M. Bostock, "Data-Driven Documents," http://d3js.org/, 2014.

[4] "Angular.js," http://angularjs.org/, 2014.

[5] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, G. Quénot *et al.*, "Unsupervised speaker identification using overlaid texts in tv broadcast," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[6] M. Budnik, J. Poignant, L. Besacier, and G. Quénot, "Automatic propagation of manual annotations for multimodal person identification in tv shows," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014.

[7] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1505–1512, 2006.

[8] 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, pp. 1102–1107.

[2] https://www.youtube.com/watch?v=PX46s1kcUGY