Collaborative Annotation for Person Identification in TV Shows

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

To cite this version:

Matheuz Budnik, Laurent Besacier, Johann Poignant, Hervé Bredin, Claude Barras, et al.. Collaborative Annotation for Person Identification in TV Shows. Interspeech 2015 (short demo paper), Sep 2015, Dresden, Germany. hal-01170513

HAL Id: hal-01170513
https://hal.archives-ouvertes.fr/hal-01170513
Submitted on 15 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Collaborative Annotation for Person Identification in TV Shows

Matheuz Budnik¹, Laurent Besacier¹, Johann Poignant², Hervé Bredin², Claude Barras², Mickael Stefas³, Pierrick Bruneau³, Thomas Tamisier³

¹Laboratoire d’Informatique de Grenoble (LIG), Univ. Grenoble Alpes, Grenoble, France
²LIMSI, CNRS - Orsay, France
³LIST, 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 project¹ 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).

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

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].

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,
In-depth description of the method can be found in [6].

Interface in Figure 2. Already labelled tracks are provided to a queue of annotations to be processed by annotators using the selection strategy is applied, which tries to verify the correct-set of tracks for annotation. Then, the clustering is refined when 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 Luxembourg National Research Fund (Fonds National de la Recherche - FNR) in the CHISTERA international program.

6. References