Automatic propagation of manual annotations for multimodal person identification in TV shows

Mateusz Budnik, Johann Poignant, Laurent Besacier, Georges Quénot

Automatic propagation of manual annotations for multimodal person identification in TV shows

Mateusz Budnik, Johann Poignant, Laurent Besacier and Georges Quénot
Laboratoire d'Informatique de Grenoble UMR 5217
Université de Grenoble-Alpes, Grenoble, F-38041, France
first.lastname@imag.fr

***Abstract*—In this paper an approach to human annotation propagation for person identification in the multimodal context is proposed. A system is used, which combines speaker diarization and face clustering to produce multimodal clusters. The whole multimodal clusters are later annotated rather than just single tracks, which is done by propagation. Optical character recognition systems provides initial annotation. Four different strategies, which select candidates for annotation, are tested. The initial results of annotation propagation are promising. With the use of a proper active learning selection strategy the human annotator involvement could be reduced even further.**

I. Introduction

The immense quantity of videos, available thanks to the wide spread availability of TV and the Internet, can be a source of very useful and important information. In order to handle such data and be able to utilize it correctly, its indexing and annotation is required. However, because of the complexity and multimodality of the data a human annotator is usually needed. On the other hand, it is not possible to annotate such a large quantity of videos due to the costs of manual intervention. That is why there are techniques being developed that can determine the most suitable instances/tracks for annotation. Active learning is a group of such methods that try to determine the most informative and relevant samples for manual annotation [2].

In this paper an approach reducing human annotator involvement is proposed, namely annotation propagation for multimodal data. This method addresses the problem of how to effectively name numerous persons in a video with the lowest degree of manual involvement. The technique presented could be considered as unsupervised active learning, as opposed to supervised active learning, which can make use of a classifier's output to determine samples for annotation [15].

To propagate labels throughout the dataset, the method finds the most promising cluster to annotate, rather than single tracks. For initial labels the optical character recognition (OCR) approach is used, which utilizes the overlaid text visible in TV broadcasts (e.g. when a person is presented for the first time his or her name is shown at the bottom of the screen).

The main contribution of this paper is an efficient selection strategy for cluster annotation when dealing with the task of multimodal person identification. For this to be possible an unsupervised system for label propagation was developed beforehand [14] and it is shortly described here. However, the application of this system to a simulated manual annotation scenario can be considered as a new insight. The results of this study show the advantages of the use of both the overlaid names (as the source of labels for the cold start) and propagation of annotation within clusters. Additionally, the cross-modal effects are visible when annotation addresses just a single modality.

This work is a part of a project called CAMOMILE[1], which aims at creating a collaborative annotation framework for 3M (multimodal, multimedia and multilingual) data. The rest of the paper is organized as follows. Section 2 describes the recent work related to this study. What follows is the in-detail description of the used system and the contributions (S3). The next section (S4) presents the performance measure, the corpus and the experimental results. Section 5 gives the conclusions and future work.

II. Related work

Concentrating on finding useful and informative samples for labeling is an active field of research. In [8] an unsupervised active learning algorithm is presented. It addresses some of the drawbacks of supervised active learning i.e. the inability of selecting samples which belong to a new category. This problem can also be observed in person identification where there usually is a high number of distinctive classes (each person), but with a small amount of members [5]. In [10] a semantic approach to annotation propagation was proposed. In [11] a person name propagation algorithm in videos is proposed. Due to propagation, the final score for person identification is higher than with the use of SVM trained on the same initial labeled data. Using active learning with clustering was already a subject of research in [9] where the information of the cluster structure (density and distribution) is used alongside the selection of the most representative samples (based on the distance from the classifier decision boundary, as in [16]) to avoid sampling the same cluster.

III. Proposed algorithm

Figure 1 gives a overview of the system used in this study. First, both speaker and face tracks are extracted from the videos. In order to create multimodal cluster, the distance between tracks of different modalities are normalized, so that they can be comparable. The output of a multilayer perceptron (ML) classifier, based on lip activity and other temporal characteristics, is used to establish the association between face and speaker tracks. Using the names obtained by the OCR, the multimodal clusters are initially labeled. Next, the active learning cycle is introduced. Based on the multimodal cluster structure and already available annotation, a given selection strategy chooses a set of unlabeled samples for human annotation. Once the new labels are obtained, cluster recalculation and annotation propagation takes place. This gives rise to a slightly modified cluster structure, which is used for the next iteration of the active learning cycle.

[1] http://camomile.limsi.fr

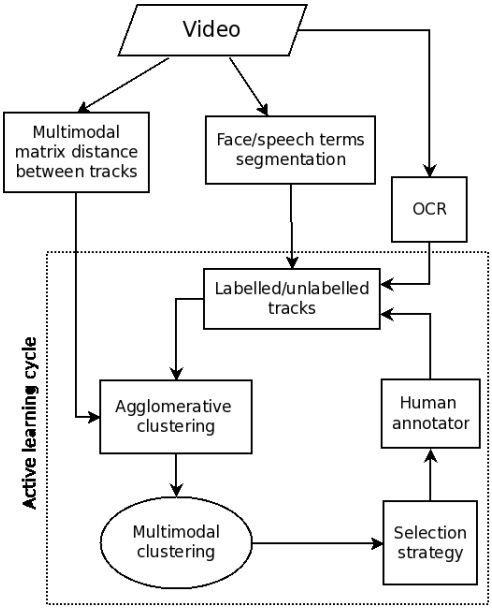


Fig. 1: System overview.

A. Text detection and recognition

Optical character recognition (OCR) system is following the design proposed in [13]. Often, guests and speakers, when introduced on a given television show to the viewer, are presented alongside overlaid text containing their name. In the context of this paper, the OCR system is used to generate automatic initial annotation, which would be later improved and expanded by human annotators.

This module is composed of two parts: text detection and text recognition. For text detection a two step approach following [1] is adopted. The coarse detection is obtained through a Sobel filter and dilatation/erosion. Additionally, to overcome the shortcomings of binarization, several binarized images are extracted of the same text, but temporally shifted. This is done to filter out false positive text boxes. For the text recognition part a publicly available OCR system from Google called `Tesseract`[2] was used.

B. Multimodal cluster structure

The structure of this module is presented in detail in [14] and [12]. In this paper just an overview is given. The module consists of three parts: speaker diarization, face clustering and the unsupervised fusion. In the case of speaker diarization, after splitting the signal into acoustically homogeneous segments, the calculation of the similarity score matrix between each pair of speech tracks is done with the use of the BIC criterion [4] with single full-covariance Gaussians. Next, the distances are normalized to have values between 0 and 1.

Face detection and tracking follows the particle-filter framework with detector-based face tracker introduced in [3]. The initialization of the face tracks is done by scanning the first and the fifth frame of every shot. Three detectors are included: frontal, half-profile and profile, which makes the face detector independent on the initial pose. Tracking is done on-line, i.e. with the use of the information from the previous frame, the location and head pose of the current frame are established.

Afterwards, a 9-point mesh is imposed on the image of the face (2 point per eye, 3 for the nose and 2 for the lips). A confidence score is obtained, helps to determine if a given face can be successfully used. If so, a HOG descriptor with 490 dimensions is calculated on that face image. After such a descriptor is made for every suitable image in the sequence, an average descriptor is then established for the whole sequence. This is then projected to a reduced space of 200 dimensions thanks to the LDML approach [7]. Next, the Euclidean distance is computed between each track. Finally, the output is normalized (values between 0 and 1).

After a selection step where faces that are deemed too small are eliminated, normalization is performed on the distances of the outputs of speaker diarization and face detection to make them comparable. In order to construct multimodal clusters, an association score between face and speaker is necessary. To that end the ML classifier is used. Color histograms calculated on the region around the speaker's lips serve as the main input. Additionally, the size of the head, its proximity to the center of the screen is used among others. The output of the classifier (between 0 and 1) indicates the distance between tracks of different modalities. ML is the only supervised step in the method. However, due to the universality of its task (association between voice and face), the model could be trained only once and be applied to different video corpora without additional involvement from the annotator.

C. Selection strategies

In this work four different annotation selection strategies are explored and evaluated:

- Random – the basic baseline, which chooses random annotation for every show.
- Chronological – chooses the annotation according to its time of appearance in a given show starting from the beginning.
- Biggest cluster first – this strategy makes use of the multimodal cluster structure of the tracks (both the face tracks and speaker tracks). Let n_t be the number of tracks within a given cluster and a_t be the number of annotations already assigned to that cluster. The score S_c is calculated as:

$$S_c = \frac{a_t}{n_t} \tag{1}$$

  and the cluster with the minimum score is selected. Afterwards, a track for manual annotation from the cluster is chosen in a chronological manner.
- Biggest cluster probability – a modification of the previous algorithm, which rather than selecting the cluster with the lowest S_c score, assigns a probability to be chosen for annotation, which is proportional to the score at a given step.

IV. Experimental evaluation

A. REPERE corpus

The REPERE challenge [6] was designed to help evaluate person identification in videos. This evaluation also provides the participants with the data, which consists of a series of shows from the French TV channels BFM TV and LCP. For this study, the test set, with the running time of around 3 hours, is used for evaluation. Not all frames of the video are annotated, but rather one every 10 seconds. There are 7 different types of shows and 28 separate videos in total. The series differ in length (from around 3 minutes to half an hour)

[2] http://code.google.com/p/tesseract-ocr/

and therefore, also in number of annotation available for each (from around 20 to more than 100). For the test set there are 1229 annotated frames in total.

B. Experimental settings

In this study four different selection strategies were evaluated. The experiment was a simulated active learning scenario where all the labels provided by human annotators are initially unknown and are reviled for a given track when an algorithm selects it. At each step of the simulation (consisting of 20 steps in total) a single track is selected for labeling for every show. The whole experiment is repeated 10 times, at each time 80 % of the corpus is randomly selected, while the rest is not used in any way. There are two types of annotations available corresponding to two modalities, i.e. the head annotation and the speaker annotation, and thus two types of corresponding tasks. Therefore, two separate experiments were performed to assess the influence of each of those on different modalities.

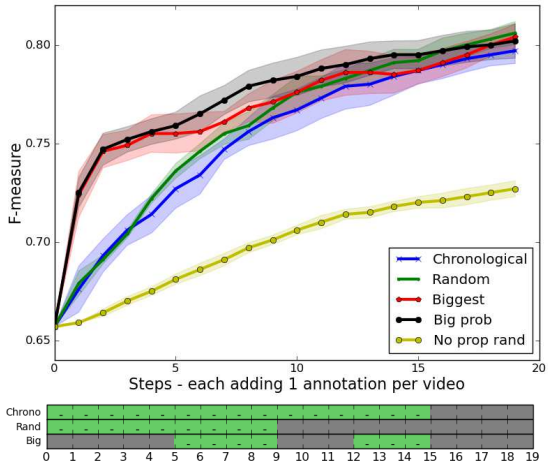
The standard F-measure was adopted as an evaluation criteria. When calculating the metric, faces with less than 2000 pixels are not considered. Also for the evaluation purpose, persons, which were not identified in the corpus, are not included in the score. In other words, a given face may be annotated correctly by the algorithm, but will not be counted as such due to the lack of reference in the corpus.

C. Results and discussion

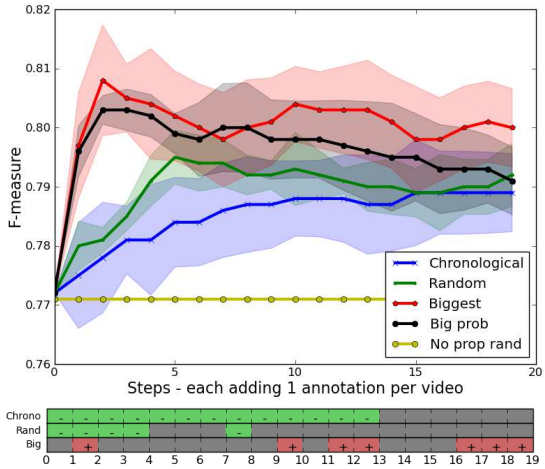
Figure 2 shows the results when using head annotation. Plots (a) and (b) give the F-measure score for different modalities, i.e. face and speaker. In Figure 3 a corresponding set of plots can be found, but using only speaker annotation. The standard deviation at each point is also visualized. As in many other studies concerning active learning, the first few steps of the algorithm are the most important. For that reason the statistical significance results are not aggravated, but rather can be seen underneath the corresponding plots where each colored block represents a consecutive step of the simulation. The outcomes of the standard Student t–test with $p = 0.05$ are calculated at each step of the simulation, excluding the starting point, which gives 19 points per plot. The statistical significance results are color coded in the following way: green ('-') – the BigProb is significantly better, gray – there is no significant difference in performance, red ('+') – the BigProb's performance is worse. As an additional experiment the random selection strategy was launched without the use of the annotation propagation after every step (called 'No prop rand' on the plots). This was done to show the effectiveness of the propagation mechanism. When applying annotation from a different modality the score of 'No prop rand' does not change, due to the lack of cross–modal effect.

A promising strategy is to pick the biggest clusters for annotation first. It is very effective at the beginning, but the increase in performance at the later steps is not that satisfactory. When using head annotation for the face error it tends to be at times significantly worse than random. Amongst the four tested strategies the BigProb tends to display the best performance overall. It is also the most consistent. It is significantly better than Random and Chronological strategies at the beginning of the simulation, but also manages to keep this advantage in the following steps.

The increase of performance can be observed on one modality when using the annotation from the other. This is



(a) Face score and t-test results.

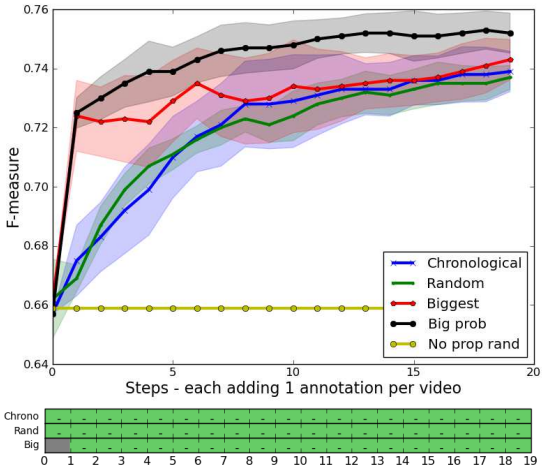


(b) Speaker score and t-test results.

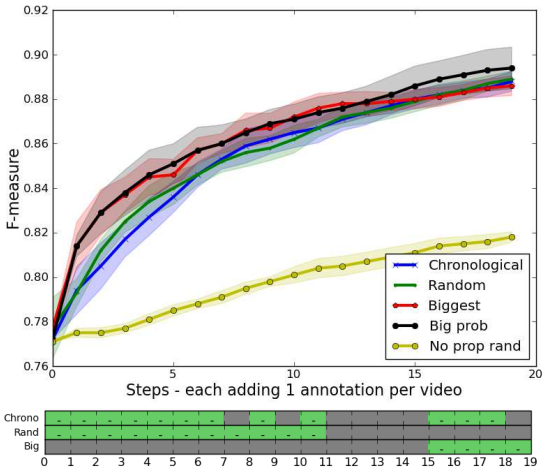
Fig. 2: F score for different modalities with the use of head annotation.

due to the use of multimodal clusters and the annotation propagation within them. In other words, while annotating speakers one can also significantly increase the performance of face annotation. This could be used in a practical active learning scenario, where annotation of different modalities has a different cost (time, difficulty for the human annotator) associated with them. In this case some of the modalities may be preferable for annotation, while the labeling of others can be reduced without a major drop in performance.

In a case where there is no prior labels available the behaviour of the strategies changes significantly. Figure 4 shows the result of the simulation using the head annotation without the OCR initialization, the annotation is propagated at each step and the score is calculated for the face. Without any additional prior knowledge the multimodal clusters are constructed based on the distances alone. This means that there is a higher probability that they will have lower purity, i.e. containing tracks from different people, compared to their counterparts with the OCR labels. Therefore, strategies that make use of the cluster structure of the data and emphasize larger clusters (which probably have lower purity) perform worse than those that ignore this information.



(a) Face score and t-test results.



(b) Speaker score and t-test results.

Fig. 3: F score for different modalities with the use of speaker annotation.

V. Conclusion and Future Work

In this paper a method for annotation propagation using multimodal clusters for person identification is proposed. It aims at reducing the human annotator involvement, which directly decreases the overall cost of producing a labeled dataset. With the application of an active learning selection strategy the costs can be reduced even further. The multimodal clusters with automatic pre-annotation produced by the OCR seem to be a very good starting point for manual annotation that builds upon it. Adding just a few additional labels can significantly increases the overall F-measure and thus improve the annotation coverage.

As for future work, apart from more complex selection strategies, a batch selection methods could be developed, so that the clusters need not be recalculated after every annotation. A real life active learning experiment, which would include real annotation costs, could help to evaluate the applicability of this system.

Acknowledgment

This work was conducted as a part of the CHIST-ERA CAMOMILE project, which was funded by the ANR (Agence Nationale de la Recherche, France).

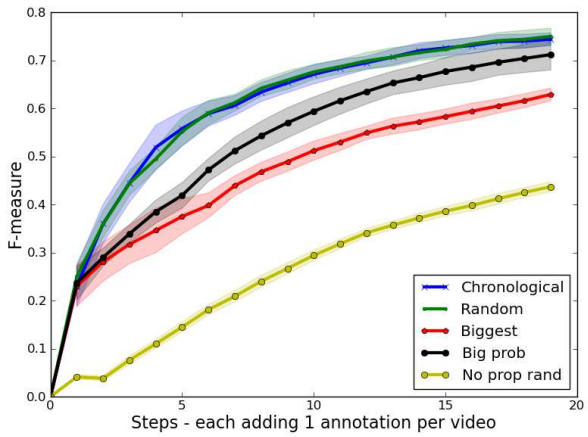


Fig. 4: The F-measure score without the initialization done by the OCR labels. Face score using the head annotation.

References

[1] Marios Anthimopoulos, Basilis Gatos, and Ioannis Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, 28(9):1413–1426, 2010.

[2] Stephane Ayache and Georges Quenot. Evaluation of active learning strategies for video indexing. *Signal Processing: Image Communication*, 22(7):692–704, 2007.

[3] Martin Bauml et al. Multi-pose face recognition for person retrieval in camera networks. In *AVSS*, pages 441–447, 2010.

[4] Scott Chen and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, page 8, 1998.

[5] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. Mulitmodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pages 176–181, 1999.

[6] Aude Giraudel et al. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107, 2012.

[7] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 96(1):64–82, 2012.

[8] Weiming Hu, Wei Hu, Nianhua Xie, and Steve Maybank. Unsupervised active learning based on hierarchical graph-theoretic clustering. *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, 39(5):1147–1161, 2009.

[9] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *ICML*, page 79. ACM, 2004.

[10] Gilberto Zonta Pastorello, Jaudete Daltio, and Claudia Bauzer Medeiros. Multimedia semantic annotation propagation. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 509–514, 2008.

[11] Phi The Pham, Tinne Tuytelaars, and Marie-Francine Moens. Naming people in news videos with label propagation. *IEEE Multimedia*, 18(3):44–55, 2011.

[12] Johann Poignant. *Identification non-supervisée de personnes dans les flux télévisés*. PhD thesis, Université de Grenoble, 2013.

[13] Johann Poignant, Laurent Besacier, Georges Quénot, and Franck Thollard. From text detection in videos to person identification. In *ICME*, pages 854–859. IEEE, 2012.

[14] Johann Poignant et al. Towards, a better integration of written names for unsupervised speakers identification in videos. In *SLAM*, 2013.

[15] Bahjat Safadi and Georges Quénot. Active learning with multiple classifiers for multimedia indexing. *Multimedia Tools and Applications*, 60(2):403–417, 2012.

[16] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846, 2000.