ACTIVE, an Extensible Cataloging Platform for Automatic Indexing of Audiovisual Content

Maurizio Pintus, Maurizio Agelli, Felice Colucci, Nicola Corona, Alessandro Sassu and Federico Santamaria
CRS4, Loc. Piscina Manna, Edificio 1, Pula, Italy

Keywords: Face Detection, Face Recognition, Speaker Recognition, Caption Recognition, Digital Asset Management.

Abstract: The cost of manual metadata production is high, especially for audiovisual content, where a time-consuming inspection is usually required in order to identify the most appropriate annotations. There is a growing need from digital content industries for solutions capable of automating such a process. In this work we present ACTIVE, a platform for indexing and cataloging audiovisual collections through the automatic recognition of faces and speakers. Adopted algorithms are described and our main contributions on people clustering and caption-based people identification are presented. Results of experiments carried out on a set of TV shows and audio files are reported and analyzed. An overview of the whole architecture is presented as well, with a focus on chosen solutions for making the platform easily extensible (plug-ins) and for distributing CPU-intensive calculations across a network of computers.

1 INTRODUCTION

Digital assets have pervaded all segments of modern economies, stimulating the development of a wide range of technological platforms for cataloging, organizing and preserving large digital collections. However, a considerable amount of effort is still taken by metadata production. This is particularly evident with audiovisual content, where a time-consuming visual inspection is usually required for indexing the relevant parts of the video timeline.

The objective of this work is to describe a platform that was created with the express purpose of automating the indexing process of audiovisual material.

The platform was developed with the aim of providing an intelligent cataloging infrastructure capable of adding value to audiovisual archives of the digital content industry, both at production and distribution level.

In order to narrow such a huge scope, the indexing has been circumscribed to the retrieval of people, through the automatic recognition of faces, captions and speakers. However, the paper also describes the adopted approach for allowing the platform to be easily extended to include new indexing algorithms and tools.

An overview of the algorithms (and of how they fit in the whole indexing process) is provided. The main contributions of our work are illustrated, which are: (1) a people clustering method based on face and clothing information; (2) a people recognition method based on extracting names from captions overlaid on the video frames. Results of experiments carried out on a set of TV shows are presented and analyzed.

1.1 Outline

Section 2 will provide an architectural overview of the ACTIVE platform in terms of its main components.The adopted approaches for making the platform extensible through plug-ins and for distributing processing on many workers will be also described in this section.

Section 3 will provide an overview of the platform workflow.

Section 4 will describe the algorithms and techniques used to automatically index video content by recognizing faces, as well as the results of an experimentation carried out on a set of TV programs.

Section 5 will describe the algorithms and techniques used to automatically index audio content by recognizing speakers, as well as the results of an experimentation carried out on a set of audio files.

Finally, section 6 will provide a summary of the results of the project and will highlight the implications that these results may have in terms of various

Pintus, M., Agelli, M., Colucci, F., Corona, N., Sassu, A. and Santamaria, F.
ACTIVE, an Extensible Cataloging Platform for Automatic Indexing of Audiovisual Content.
DOI: 10.5220/0005722205740581
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 574-581
ISBN: 978-989-758-175-5

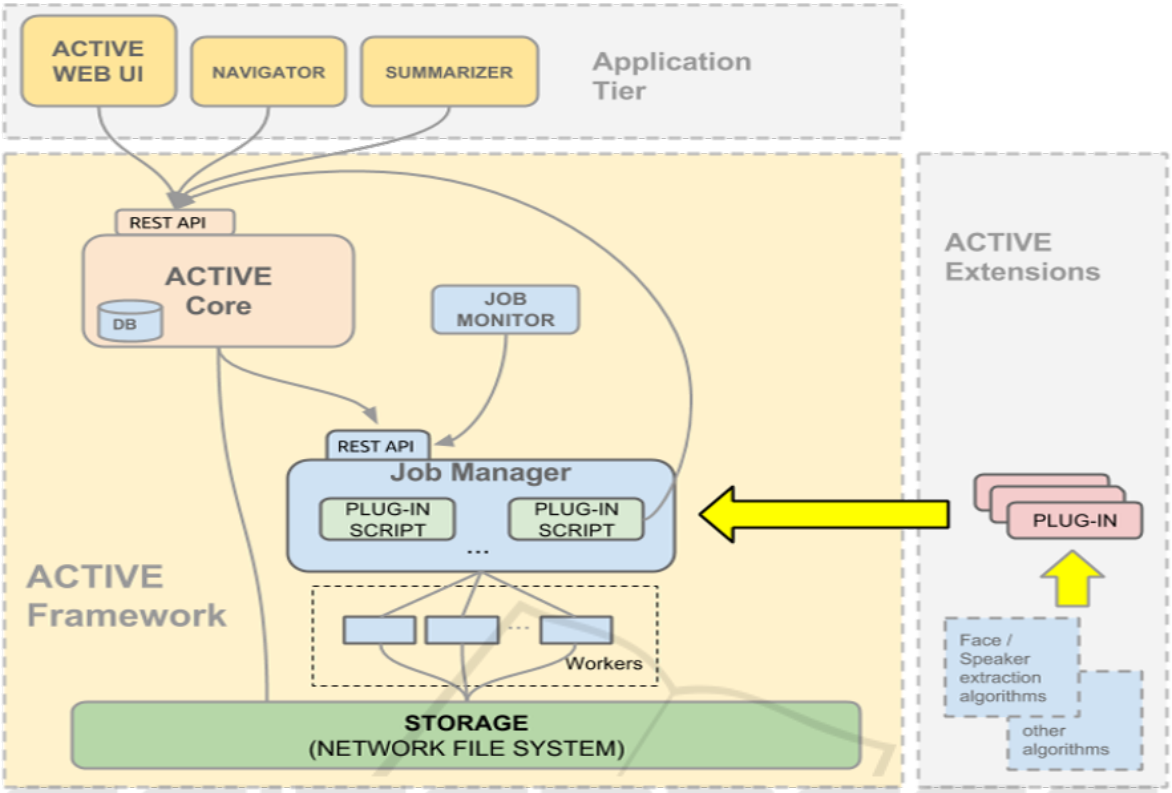


Figure 1: Schema of the ACTIVE System architecture.

application scenarios.

2 ARCHITECTURAL OVERVIEW

The ACTIVE Platform is based on a modular architecture, which was designed with the express intention of enabling a wide range of applications.

As shown in figure 1, the architecture defines three main blocks: the ACTIVE Framework, the Application Tier and the ACTIVE Extensions.

The **ACTIVE Framework** includes a set of reusable components that address the requirements of a variety of applications. These components are:

1. the **ACTIVE Core**, which provides a set of basic features for managing digital assets;
2. the **Job Manager**, which allows to distribute CPU-intensive processing onto several machines;
3. a **Network Storage** for audiovisual resources, based on NFS.

The **ACTIVE Core** implements a data model for users, digital items (internal representation of digital assets and their metadata) and tags (bindings between items and entities, with the latter representing keywords, people, or other kinds of objects which may be defined in future releases). A tag may also be associated to a set of temporal intervals (Dynamic Tags), which specify the time slices of the audiovisual item the entity actually refers to.

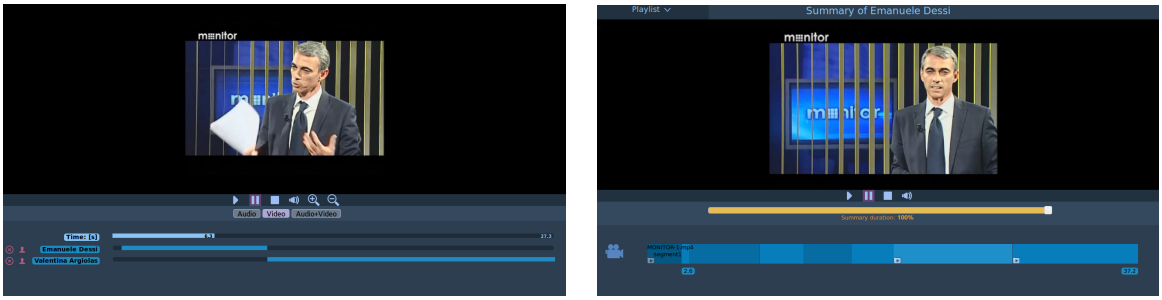
A plug-in system provides a simple mechanism for extending the ACTIVE Core with server-side scripts. The whole of plug-ins builds up a reusable codebase (**ACTIVE Extensions**) that other applications can benefit from.

The ACTIVE Core exposes a REST API which allows applications and plug-in scripts to access the internal data models using CRUD operations.

The **Job Manager** allows to distribute asynchronous jobs across several computers, in order to reduce the execution time of CPU-intensive operations (e.g. transcoding, image processing).

The **Application Tier** contains the modules that provide the overall experience and functionality for any system based on the ACTIVE Framework. These modules can include any kind of applications obeying the requirements of the final system. In the specific case of the platform described in the present paper, the application tier contains three distinct web applications: the main user interface, the navigator and the summarizer.

The **Main User Interface** provides a very basic



(a) (b)

Figure 2: (a) The Navigator. (b) The Summarizer.

set of digital asset management operations, allowing to carry out simple tasks such as uploading new items, inspecting and editing metadata, launching plug-in scripts and applications, performing searches and displaying search results in a thumbnail grid.

The **Navigator** (figure 2) is a tool for visually browsing the Dynamic Tags of an item and for doing some basic editing of the tags associated to people (e.g. to correct the names in case of misrecognition).

The **Summarizer** (figure 2) is a tool for evaluating the results of a search operation based on a specific person. It provides a visual and interactive representation of the occurrences of the query term in different temporal sections of all video items returned as search results. Short samples from these occurrences are concatenated in a unique timeline, so that the user can enjoy an overview of the relevant parts of search results.

3 WORKFLOW OVERVIEW

A short description of the workflow is given, highlighting the aspects related to face / speaker extraction. In order to carry out these tasks, audiovisual resources shall be first imported into the ACTIVE platform. As soon as the import is completed, embedded metadata are extracted, previews are calculated and the newly created items are indexed (so, they may appear in search results).

Face and speaker extraction can be manually launched at any time and their progress can be monitored through the Job Monitor. As soon as the extraction has been completed, a set of tags (with associated dynamic tags) are created and the item is indexed by them. It is also possible to use the Navigator to inspect the face/speaker extraction results, both in terms of clusters (set of time slices where the same person is assumed to be present) and identities (labels assigned to each cluster).

The face/speaker recognition operates on the basis of previously built face/speaker models. In case a detected face (or speaker) does not match any model, it is assumed to belong to a new person and is labeled as "UNKNOWN_XX", where XX is a unique string. Assigned labels can be manually edited, either for assigning an identity to people originally labeled as unknowns, or for correcting the labels assigned to misrecognized people. Although a manual editing is required in case face (or speaker) recognition fails, this editing is automatically applied to the whole cluster, allowing significant time and effort saving.

4 VISUAL INDEXING

In figure 3 the schema of our visual indexing system is shown. On the basis of the results by (Korshunov and Ooi, 2011), all videos are analyzed at a frame rate of about 5 fps.

Considered frames are grouped into shots, using the local thresholding method presented in (Dugad et al., 1998). Histogram calculation is carried out in the HSV space, not considering pixels with high H values and low S and V values. Thereafter frames are analyzed in order to find all faces in them. Face detection is based on the OpenCV (Bradski, 2000) implementation of a method initially proposed by Viola and Jones (Viola and Jones, 2001) (Viola and Jones, 2004) and improved by Rainer Lienhart (Lienhart et al., 2003).

Found faces are then tracked, in order to obtain face tracks belonging to the same person, using a method based on the OpenCV implementation of the Continuously Adaptive Mean Shift (CAMSHIFT) algorithm (Bradski, 1998), and aligned.

In the next step, a people clustering block aggregates face tracks that are likely to belong to the same person by using face and clothing information. After this phase is completed, people recognition is carried out by using face and caption information.

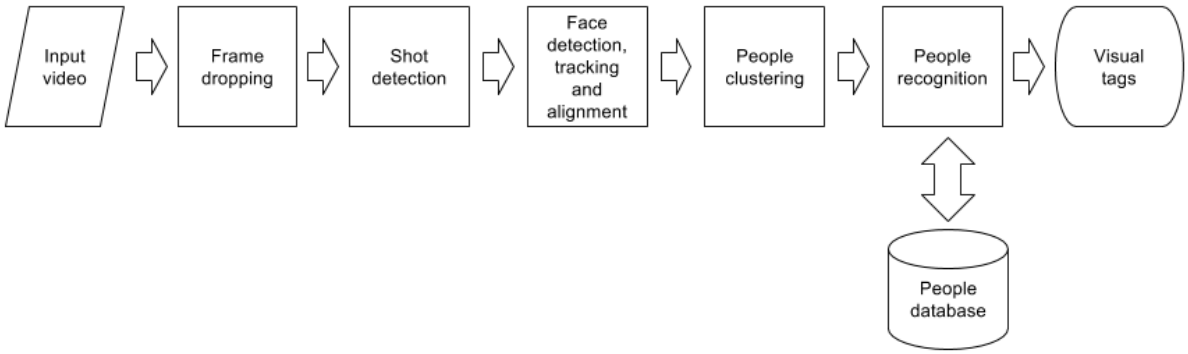


Figure 3: Schema of our visual indexing system.

4.1 People Clustering

People clustering merges found face tracks into clusters in which the same person should be visible: this is useful both for an efficient annotation by users and for improving people recognition. People clustering often relies on features from faces and clothing (Berg et al., 2004) (Sivic et al., 2006) (Everingham et al., 2006) (Maji and Bajcsy, 2007) (Zhang et al., 2009) (El-Khoury et al., 2010).

In our system, firstly face models and clothing models are calculated for each face track, considering only detected faces (i.e. not considering faces located only by tracking). LBP histograms (Ahonen et al., 2004), calculated on the equalized aligned faces, are used as face features. We used the $LBP_{8,1}$ operator, with a 4 x 8 grid (in order to have square cells with the chosen alignment, that produces face images a size of 200 x 400 pixels.). HSV color histograms from a region just below each face detection bounding box are used to represent clothes. Width of clothing bounding box is 2 times the width of the face bounding box, while its height equals the height of the face bounding box. This region permits to include the most significant zone for clothing comparison, e.g., the mask presented in (Sivic et al., 2006), while including little background and few occluding objects as hands or tables.

Face tracks are scanned in a sequential way. In the first scan, the first face track is labeled as "Person 1" and the other N - 1 face tracks are compared to it. If a face track turns out to be similar to the first face track, it is considered belonging to the same person and it is also labeled as "Person 1". After the first face track has been compared to all the other face tracks, the first face track still not labeled is labeled as "Person 2" and the remaining not labeled face tracks are labeled according to it. The process is repeated until all face tracks are labeled. Eventually, the face tracks with the same label are merged into the same cluster. Because faces that are visible at the same time should belong to different people, if two face tracks overlap in time they are considered belonging to different people and are not compared.

When comparing two face tracks, face track 1 and face track 2, firstly the minimum χ^2 distance (d_{face}) between the LBP histograms from the faces in face track 1 and the LBP histograms from the faces in face tracks 2 , is calculated.

Having fixed two thresholds, $th_{face-low}$ and $th_{face-high}$, with $th_{face-low} \leq th_{face-high}$, the two face tracks are considered belonging to the same person if $d_{face} < th_{face-low}$ or if all the following conditions are verified:

1. $th_{face-low} \leq d_{face} < th_{face-high}$;
2. All clothing bounding boxes in the two face tracks are entirely contained by the respective frames;
3. The minimum χ^2 distance between the color histograms from the clothes in the two face tracks is below a local threshold, $th_{clothing}$, calculated from the two face tracks.

Local threshold for clothes comparison is calculated in the following way:

$$th_{clothing} = max\left(\frac{\sum_{i=1}^{l_1}\frac{\sum_{j=i+1}^{l_1} d_{i,j}}{(l_1-i)}}{l_1}, \frac{\sum_{m=1}^{l_2}\frac{\sum_{n=m+1}^{l_2} d_{m,n}}{(l_2-m)}}{l_2}\right) \quad (1)$$

where i and j are indexes of histograms belonging to face track 1, while m and n are indexes of histograms belonging to face track 2. So $d_{i,j}$ represents the χ^2 distance between a histogram i and a histogram j both belonging to face track 1; $d_{m,n}$ represents the χ^2 distance between a histogram m and a histogram n both belonging to face track 2; l_1 and l_2 are the numbers of histograms considered in face track 1 and 2, respectively.

4.2 People Recognition

At this stage, an association of people clusters with real names is carried out. People recognition is usually based on feature extraction from faces. However, other types of features have been used, like subtitles and transcripts (Everingham et al., 2006), film scripts (Zhang et al., 2009), strings extracted from captions and audio tracks (Bertini et al., 2001).

In our system, firstly an attempt to use caption recognition for labeling clusters is made. Secondly, in those cases where this strategy turns out unsuccessful, face recognition is used to label people clusters.

We indicate with the term caption recognition the analysis of text overlaid on video frames in order to find the names of the visible people. A database (tag dictionary) shall be prepared, containing a list of tags identifying the names of the people which may appear in the videos. Words extracted from frames are matched with these tags. A reduced bitrate of 1 frame per second is used in this case and only frames with one face in them are considered. The tool used for OCR is tesseract (Smith, 2007); it is set to recognize only letters, both lowercase and uppercase.

Original frame is binarized by using Otsu's method (Otsu, 1979), then all contours in image are retrieved. Contours that are too small or too big are discarded, the remaining ones are analyzed by the OCR engine in order to recognize single characters.

Found characters are ordered in rows by checking their bounding boxes, discarding characters that are inside other ones. For each row, the portion of original image that contains all characters in the row is binarized by using Otsu's method and found characters are put in another binary image that is analyzed by the OCR engine in order to recognize words in it.

Each found caption block is matched against tags from the tag dictionary by using the Levenshtein distance (Levenshtein, 1966), obtaining, for each tag, a similarity measure between 0 and 1. If the tag that gets the maximum similarity measure is greater than a given threshold, it is assigned to the face. Only tags that are assigned to at least 4 frames in the cluster are considered (usually, a caption is visible for at least 4 seconds). If certain words are identified in a frame, e.g. indicating that the overlaid name refers to a person speaking on the phone, the frame is not considered.

Face comparison is the same used in people clustering, with the difference that faces in the video are compared with previously built face models. In both caption recognition and face recognition, results from single frames are aggregated with a majority rule in order to obtain a final tag for the cluster.

4.3 Experimental Results

Experiments on our visual indexing system were carried out on three full-length episodes of three different TV shows from the Sardinian channel "Videolina", "Facciamo i conti", "Monitor" and "SportClub sugli Spalti", with durations of about 54, 119 and 172 minutes. All used videos have a resolution of 720 x 576 pixels. Only people that could be relevant for a user were considered (e.g., pedestrians were ignored), respectively 6, 10 and 15 for the three videos: start times and durations of video segments in which these people are visible were manually annotated. Annotations extracted by the system were then compared with these ones.

Performance is measured by averaging precision, recall and f-measure calculated on considered people in the video. For each person p we have:

$$precision_p = \frac{T_{CR_p}}{T_{R_p}} \tag{2}$$

$$recall_p = \frac{T_{CR_p}}{T_{T_p}} \tag{3}$$

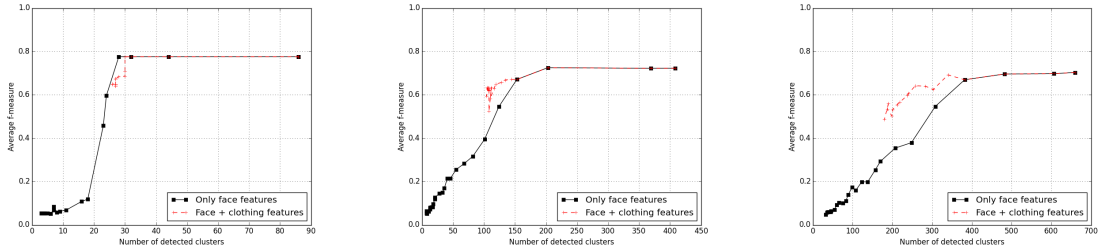
$$F-measure_p = 2 \cdot \frac{precision_p \cdot recall_p}{precision_p + recall_p} \tag{4}$$

where T_{CR_p} is the total duration of video segments correctly assigned to person p, T_{R_p} is the total duration of video segments assigned to person p, T_{T_p} is the total duration of video segments in which that person is actually visible.

Firstly, experiments on people clustering, without the use of people recognition, were carried out. In this case, for each cluster, the central frame of the first face track was selected as a keyframe, highlighting the face of the considered person: annotation was carried out semi-automatically by assigning tags to the clusters according to visual inspection of these keyframes. Figure 4 reports experimental results for several values of threshold for face comparison.

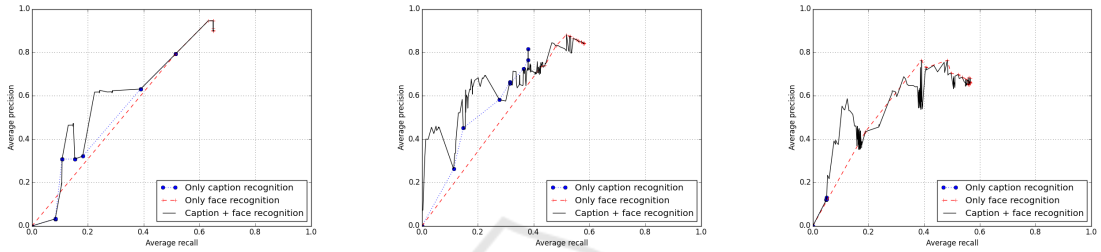
Increasing the thresholds, more face tracks are merged and so the number of detected clusters decreases. Using only face features, average F-measure remains almost constant for low threshold values and exhibits a sharp decrease when the threshold is above a critical value, more or less equal to 8. When clothing features were used, $th_{face-low}$ was fixed to 8 and only $th_{face-high}$ was changed. In the two longest videos, the combined use of face and clothing features outperforms the use of only face features, reducing the decrease of F-measure.

Then, experiments on people recognition were carried out; in this case a fully automatic annotation was used. A threshold just below the critical value



(a) (b) (c)

Figure 4: People clustering results. The mean values of F-measure are reported, with and without the use of clothing features, for three full-length episoded of three TV shows,"Facciamo i conti" (a), "Monitor" (b) and "SportClub sugli Spalti" (c).



(a) (b) (c)

Figure 5: People recognition results. The mean values of precision and recall are reported, using only caption recognition, only face recognition and using both, for three full-length episodes of three TV shows,"Facciamo i conti" (a), "Monitor" (b) and "SportClub sugli Spalti" (c).

was chosen for people clustering. Training set for the creation of the face models used by face recognition was built from frames extracted from 15 videos belonging to the three TV shows (including those used in the experiments). Face detection and face alignment was automatically performed, so these frames were cropped, if necessary, to include only one person. 12 images per person per each of 80 people were considered; similarly to the PIE database (Sim et al., 2002), each person was present under 3 different poses (frontal, turned left and turned right) and with 4 different expressions (neutral, smiling, blinking and talking). The names of these 80 people constituted the tag dictionary used by caption recognition. Figure 5 reports experimental results.

In the first two considered TV shows the use of captions provides good results, while in "SportClub sugli Spalti", where there is a lot of text overlaid on frames (sport results and standings), results using captions are worse. Best results are obtained using thresholds of 0.7-0.9 for caption recognition and 10 for face recognition.

5 AUDIO INDEXING

The audio indexing system performs speaker recognition on audiovisual items. The system, whose schematic diagram is outlined in figure 6, is based on the LIUM_Speaker_Diarization framework (Meignier and Merlin, 2010) (Rouvier et al., 2013).

The key concepts of the system are summarized below:

1. the extracted audio is classified to obtain segments containing either music, silence or speech (Ajmera et al., 2004);
2. the speech parts are segmented on the basis of speaker changes;
3. segments are grouped into clusters, each one containing speech from the same speaker;
4. the speaker recognition process associates to each cluster an audio tag, representing the name of the speaker.

A two-step speaker diarization has been used (Barras et al., 2006). First, in order to produce homogeneous speech segments, an acoustic BIC (Bayesian Information Criterion) segmentation (Chen and Gopalakrishnan, 1998) is carried out, using generalized likelihood ratio as metric (Meignier and Merlin, 2010) to determine the similarity over an audio segment, followed by a BIC hierarchical clustering, which groups similar segments. Next, a Viterbi resegmentation is applied, which produces a new set of segment clusters.

The next step is the the Speaker Recognition process, which is based on the extraction of a set of fea-

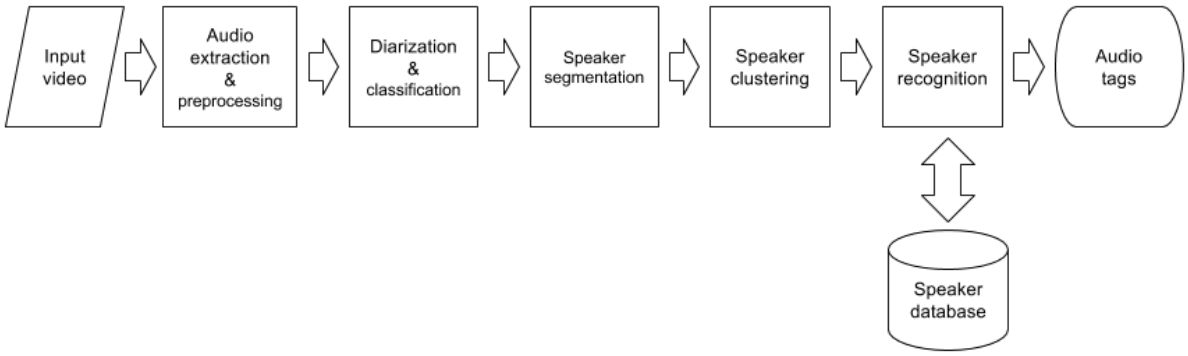


Figure 6: Schema of our audio indexing system.

tures (Khan et al., 2008).

The voice of an unknown speaker is analyzed and modeled as a random Gaussian process from which its corresponding sequence of MFCC vectors is extracted. The parameters of the Gaussians are computed from MFCC vectors, using a Maximum Likelihood (ML) method. The model of the unknown speaker is compared with the model of known speakers.

For accomplishing this task our system uses a *statistical background model* from which the models of each speaker are adapted (Reynolds et al., 2000). The specific models of each speaker are adapted from the UBM using the maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994).

5.1 Experimental Results

Experiments were carried out on a set of audio files with the following characteristics:

- high-quality audio (40 files of around 20 minutes each, one speaker for each file, 5 total speakers);
- average-quality audio (40 files of around 2 minutes each, unknown speaker number);
- average-quality audio, close dialogue, i.e. where each speaker speaks for no longer than 20 seconds (10 files of around 10 minutes each, unknown speaker number)

The database of known models includes 80 models (i.e. 80 different speakers). For each person, the audio file used for the test and the audio file used for the enrollment process have been obteined in similar condition (ambient, noise level, etc.).

The performance of a speaker recognition system is based on FAR ("False Acceptance Rate, which measures how many speakers are falsely recognized) and FRR ("False Rejection Rate", which identifies the probability that the system fails to identify a speaker whose model is already present in the database) (Martin et al., 1997). Table 1 shows the results.

Table 1: Speaker recognition results.

	High quality	**Average quality**	**Average quality, close dialogue**
FAR	$< 2\%$	12 %	18 %
FRR	$< 1\%$	10 %	15 %

6 CONCLUSIONS

An extensible framework for managing audiovisual assets has been presented. A specific application of this framework in the field of automatic audiovisual indexing has been thoroughly described. Experimental results on a set of TV programs and audio files have been presented, showing a quite good behavior.

Further testing will be carried out, aimed at evaluating the practical advantages of the proposed solution in a real-world operating environment.

Future work may take several directions: (1) developing new indexing algorithms, e.g. for recognizing specific classes of objects; (2) improving the ACTIVE framework with additional cataloging features; (3) developing a set of general-purpose plugins (e.g. for transcoding, watermarking, batch editing, extracting embedded metadata, evaluating user profiles, etc.); (4) implementing a full digital asset management application on top of the ACTIVE framework, in order to better exploit the work carried out on the automatic indexing tools.

REFERENCES

Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face Recognition with Local Binary Patterns. In *Proc. ECCV*, pages 469–481.

Ajmera, J., McCowan, I., and Bourlard, H. (2004). Robust speaker change detection. *IEEE Signal Processing Letters*, 11(8):649–651.

Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512.

Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y.-W., Learned-Miller, E., and Forsyth, D. (2004). Names and Faces in the News. In *Proc. CVPR*, pages II–848–II–854 Vol.2.

Bertini, M., Bimbo, A. D., and Pala, P. (2001). Content-based indexing and retrieval of tv news. *Pattern Recognition Letters*, 22(5):503–516.

Bradski, G. R. (1998). Real Time Face and Object Tracking as a Component of a Perceptual User Interface. In *Proc. WACV*, pages 214–219.

Bradski, G. R. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 25(11):120, 122–125.

Chen, S. S. and Gopalakrishnan, P. S. (1998). Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132.

Dugad, R., Ratakonda, K., and Ahuja, N. (1998). Robust Video Shot Change Detection. In *Proc. MMSP*, pages 376–381.

El-Khoury, E., Senac, C., and Joly, P. (2010). Face-and-clothing based people clustering in video content. In *Proc. MIR*, pages 295–304.

Everingham, M. R., Sivic, J., and Zisserman, A. (2006). "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In *Proc. BMVC*, pages 92.1–92.10.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Khan, S., Rafibullslam, M., Faizul, M., and Doll, D. (2008). Speaker recognition using mfcc. *International Journal of Computer Science and Engineering System*, 2(1).

Korshunov, P. and Ooi, W. T. (2011). Video quality for face detection, recognition, and tracking. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7(3):14:1–14:21.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and control theory*, 10(8):707–710.

Lienhart, R., Kuranov, A., and Pisarevsky, V. (2003). Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. In *Proc. DAGM*, pages 297–304.

Maji, S. and Bajcsy, R. (2007). Fast Unsupervised Alignment of Video and Text for Indexing/Names and Faces. In *Proc. MM*, pages 57–64.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. EUROSPEECH*, pages 1895–1898.

Meignier, S. and Merlin, T. (2010). LIUM SpkDiarization: An Open Source Toolkit For Diarization. In *Proc. CMU SPUD Workshp*.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. In *Proc. INTERSPEECH*.

Sim, T., Baker, S., and Bsat, M. (2002). The CMU Pose, Illumination, and Expression (PIE) database. In *Proc. FG*, pages 46–51.

Sivic, J., Zitnick, C. L., and Szeliski, R. (2006). Finding people in repeated shots of the same scene. In *Proc. BMVC*, pages 93.1–93.10.

Smith, R. (2007). An overview of the Tesseract OCR Engine. In *Proc. ICDAR*, pages 629–633.

Viola, P. and Jones, M. J. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proc. CVPR*, pages I–511–I–518 Vol.1.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Zhang, Y.-F., Xu, C., Lu, H., and Huang, Y.-M. (2009). Character identification in feature-length films using global face-name matching. *IEEE Transactions on Multimedia*, 11(7):1276–1288.