Project acronym	**AXES**
Project full title	**Access to Audiovisual Archives**
Project No	**269980**
	Large-scale integrating project (IP)

www.axes-project.eu

Deliverable D2.5

Update on final toolbox for specific and category level visual and audio person recognition

September 2014

SEVENTH FRAMEWORK PROGRAMME

Objective ICT- 2009.4.1: Digital Libraries and Digital Preservation

PROJECT DELIVERABLE REPORT

Project	
Grant Agreement number	*269980*
Project acronym:	*AXES*
Project title:	*Access to Audiovisual Archives*
Funding Scheme:	*Large-Scale Integrating project (IP)*
Date of latest version of Annex I against which the assessment will be made:	*30 June 2014*

Document	
Deliverable number:	2.5
Deliverable title	Update on final toolbox for specific and category level visual and audio person recognition
Contractual Date of Delivery:	30/09/2014
Actual Date of Delivery:	30/09/2014
Author (s):	Patrick Pérez
Reviewer (s):	Yann Mombrun
Work package no.:	2
Work package title:	Multimodal analysis of people
Work package leader:	TEC
Version/Revision:	V1.2
Draft/Final:	Final
Total number of pages (including cover):	14

CHANGE LOG

Reason for change	Issue	Revision	Date
Version to reviewers	1	0	28/09/2014
Updated by reviewers	1	1	29/09/2014
Updated by UO	1	1	29/09/2014

DISCLAIMER

This document contains description of the AXES project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AXES project and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

AXES is a project partly funded by the European Union.

TABLE OF CONTENTS

SUMMARY

This document is an update of the toolbox described in deliverable D2.4, describing the modules that constitute the final WP2 toolbox and that are to be integrated either in the web services of the V2 version of the AXES system or within the AXES-x (Pro/Research/Home) interactive platforms. It includes in particular the status of each module with respect to AxesOpen system.

INTRODUCTION

The breakdown of WP2 architecture into ten distinct modules was introduced in the original specification document D2.1. It reads as follows:

- M2.1 ShotExtraction (TEC)
- M2.2 IndividualHumanActionRecognition (UO-Inria)
- M2.3 FaceDetectionAndTracking (UO)
- M2.4 FaceRecognition (UO-Inria)
- M2.5 FaceClustering (TEC)
- M2.6 SpeakerClustering (FhG)
- M2.7 SpeakerRecognition (FhG)
- M2.8 SearchByExample (Inria)
- M2.9 UpperBodyDetectionAndTracking (UO)
- M2.10 VisualPersonCharacterization (UO)

We also remind that these modules are of two types:

- **Annotation modules** (1, 2, 3, 4, 7, 9, 10) output spatial and/or temporal segments of the input audiovisual document with entity labels attached to them; among them, modules 2, 4, 7 and 10 will rely on classifiers trained for given entity thesauri: list of public personalities (e.g., politicians or actors) in modules 4 and 7, list of attributes (e.g., gender or age) in module 10, list of atomic human actions (e.g., smoking) in module 2. In contrast, modules 1, 3 and 9 are each based on a single-class detector (cut, face and upper-body) that will not need retraining throughout the project.
- **Linking modules** (5, 6, 8) group two or more segments of the same type as stemming from the same person. They thus don't assume any pre-defined catalogues of entities and will rely on audio or visual similarity as computed with fixed or adaptive metrics in appropriate feature spaces

In the next section, a high-level algorithmic description is given for each of the modules, along with the following additional information:

- Partner(s) who developed the module;
- Off-line or on-line/on-the-fly nature of the module;
- Development history and current status;
- Current integration status.

MODULE ANALYSIS

M2.1 ShotExtraction (TEC)

Objective and description: The aim of this module is two-fold. It is firstly in charge of detecting *cuts* in an input video document. As a result, it splits the input video into one or several *shots*, which are meaningful and tractable video sub-units. Secondly, it extracts one or several representative *key-frames* per shot. Given a detected shot, the subsequent analysis of its visual content by other modules will either rely on all its frames (person and face tracking, action recognition) or, for computational and storage savings, focus only on extracted key-frames (object/face detection and recognition, place recognition, scene categorization). Key-frames are also useful for visualization in the interactive video search front-end. Although developed in WP2, this module is clearly of generic interest for all visual analysis modules across WPs 2, 3 and 4.

This module relies on a TEC fingerprint technology which amounts to monitoring temporal variations of a global descriptor (RAdon Soft Hashing [RASH]) in order to detect cuts. This module is on the par with previous one (described in D2.3) in terms of accuracy, but it is much faster, with a speed of 10x real time on a single core.

Same global descriptors are also used to defined “stable” (or “representative”) frames of a shot, from which key-frames can be derived: in order to address various requirements from other partners in terms of number and nature of such key-frames, the module outputs, for each shot, the shot middle frame as well as up to *k* (e.g., *k*=10) “ stable” frames and the corresponding closest I-frames.

Nature: off-line.

Development: this module is final and replaces the previous versions described in D2.2 (V0) and D2.3 (V1).

Integration: integrated into V2 system. Used off-line to process corpus indexed in AXES-x platforms (several thousand hours of content).

OpenAxes Status: this component will be part of OpenAxes in the form of a Linux-64 executable. Two technologies it contains are protected by patents, hence this component will only qualify for non-commercial uses. External dependency: ffmpeg.

M2.2 IndividualHumanActionRecognition (UO-Inria)

Due to changes in priorities of concerned partners (more high-added-value work on on-the-fly classification in particular) and partial overlap with action recognition modules developed in WP4 (see D4.4), this module is not part of this toolbox any longer.

M2.3 FaceDetectionAndTracking (TEC)

Objective and description: The objective of this module is two-fold: detecting faces (position and size of rectangular bounding boxes) in video frames and grouping them via visual

tracking. As a result, *face tracks* are output for each input shot. See D2.2 for more details on the motivations and objectives.

Coupling detection and tracking can be done in two ways: either detection is performed independently, typically as a pre-processing step in case of batch analysis, on all frames or on a subset of them, and tracking is subsequently in charge of temporal grouping; or detection is embedded in the tracking system, to start, drive and validate individual tracks. The final implementation of the module is of the former type to fully benefit from its off-line batch nature, i.e. resorting to grouping in both forward and backward temporal directions.

The detection part includes three Viola and Jones face detectors, respectively for the frontal, left profile and right profile faces. The Kanade–Lucas–Tomasi (KLT) tracker is applied between all face detections (within the current shot) lying within a 5 frames horizon and in different frames. This allows evaluating distances between different face pairs based on the proportion of detected interest points within one face detection that after tracking fall onto the other face detection. Based on these distances and so-called “cannot-link” constraints (set for non-overlapping face detections lying within the same frame) a semi-supervised agglomerative hierarchical clustering is performed. Interpolation (to predict punctual missing detections), smoothing (to stabilize face bounding boxes in time) and grouping (to replace multiple detections by different detectors by a unique detection) are all performed jointly via the optimization of a suitable cost function. The last step consists in filtering out very short tracks and the tracks with frontal face detections for which the confidence of facial landmarks detector (Inria component) is very small in average.

Nature: off-line

Development: this version replaces the initial implementation based on sequential detection and tracking described in D2.2. For OpenCV/OS incompatibility issues on server’s side, original video I/O API (OpenCV) has been replaced by a new one that depends on ffmepg only.

Integration: integration in V2 has been done.

OpenAxes Status: this component might be part of OpenAxes in the form of a Linux-64 executable. It includes an instance of the shot detector, hence inheriting the same IP limitations. Its external dependencies are as follows: ffmeg and OpenCV for face detection.

M2.4 FaceRecognition (UO-Inria)

Objective and description: This module aims at analyzing a single face detected in an image or all the faces composing a single face track, to decide whether this face/face-track belongs to a “known” person, such a politician or an actor. Two scenarios are envisaged: (1) an *off-line* automatic annotation scenario where a list of such persons of interest has been established beforehand and face classifiers have been trained off-line in a (semi) supervised fashion for each of them; these classifiers are then run on all faces/face-tracks in the database; (2) a search scenario in which the name of a person, e.g., input by a user, is passed to the system that must first learn, *on-the-fly*, a classifier for this new person, and then return a ranked list of candidate faces/face-tracks from the database for this person.

A module pertaining to the second scenario has been developed by UO and integrated in AXES-x platforms. It allows the user to search for specific people (e.g., George Bush) in the different databases of the system. As with the other on-the-fly systems developed by UO (e.g., in WP3), the method proceeds by first downloading positive training examples for the query from Google images (in this case restricted to faces), via a python Web service.

Complemented with a generic set of negative face images, this training set is used to trained a classifier which is then used to rank the pre-computed face tracks of the video dataset using their visual descriptors. The classifier is a linear SVM efficiently trained with LibLinea and operating on face description vectors built as follows: gradient-based local descriptors are computed for each of 13 facial landmarks previously detected and concatenated into a high dimensional descriptor for each face. Note that this face descriptor is likely to be replaced in Y4 by an even more powerful one, following recent research done by UO (with no impact on component API).
Interestingly, the same training-recognition pipeline can also be used in a completely off-line way to train, and run over the database, classifiers for popular personalities.
Nature: on-the-fly/off-line

Development: components are final, except for face descriptor which will be replace byd Fisher Vector face and face track description (see details in P3 activity reports).

Integration: On-the-fly part by UO is part of AXES-x platforms and is installed on BBC server.
OpenAxes Status: these components will not be part of OpenAxes.

M2.5 FaceClustering (TEC)

Objective and description: This module is in charge of deciding whether two or more face tracks originating from one or more video documents belong to a single person, with no prior knowledge (*fully unsupervised*). More generally, it allows, in a batch mode, to spot faces that reoccur either within a given video document or across a video corpus. The output of the module is composed of groups or *clusters* of face tracks. Face-based linking thus produced can serve on its own as a means to navigate and search. It will also be useful to expedite subsequent annotation of specific people. For instance, manual or automatic identification of one face track will be instantaneously propagated to all face tracks in the same cluster.

Based on most recent advances in face authentication, the implementation of this module relies on a low-complexity, yet powerful, description/comparison pipeline: face bounding box is divided into cells (e.g., 4x5 tilling) and each cell is described by a histogram of local quantized patterns (LBP, LTP or LHS), properly normalized and scaled. The similarity between two faces is obtained by averaging cell-wise cosine similarity over this tiling. Similarity between two face tracks is readily derived from the latter through pair-wise comparisons between all faces in tracks, or only subsets (e.g., key-faces extracted in the same manner as key-frames in a shot). Classic clustering such as spectral clustering is finally applied to the large inter-track similarity matrix that is thus computed over a given document or set of documents.

Nature: off-line.

Development: for reasons exposed in Y1 and Y2 activity reports, the development of this module has been substantially delayed. It is now planned to be completed by end of Y3. Present specification replaces the one presented in D2.3, that appeared as more complex while likely to produce under optimal results.

Integration: not integrated.

OpenAxes Status: this component will not be part of OpenAxes.

M2.6 SpeakerClustering (FhG)

Objective and description: This module is the audio counterpart of FaceClustering. It is in charge of deciding whether two or more speech audio samples have been uttered by the same person, with no prior knowledge. More generally, it aims at spotting voices that reoccur either within a given audio document or across an audio corpus (“audio diarization”).

To achieve these goals, the module implements the following processing chain. First the audio is segmented into homogeneous parts with consistent audio environments. The underlying assumption is that no speaker changes occur inside these segments, but rather at the boundaries. Given these segments, GMM-based classifiers are used to first determine whether a segment contains speech and, if this is the case, whether the speaker is female or male. The last step is then a gender-dependent clustering, which is implemented as hierarchical clustering using the Bayesian Information Criterion (BIC) as similarity measure. The final output of the module is then the segmentation and for each segment a local speaker information using labels that are unique for the given file.

Nature: off-line

Development: Currently the module is not capable of clustering reoccurring speakers in a set of files; instead each file is processed separately. As this extension is expected to be very useful in the usage scenarios we are envisioning, we plan to implement this as an optional post-processing step for future versions. Since clustering complete collection with sizes above 100h would require very high processing power, we foresee a selection of relevant files by metadata.

Integration: Integrated in AXES V1 and V2 as part of a web service shared by M2.7 module.

OpenAxes Status: this component will not be part of OpenAxes

M2.7 SpeakerRecognition (FhG)

Objective and description: This module is the *supervised* counterpart of SpeakerClustering. It is in charge of deciding whether a given speaker run belongs to a given person within a predefined list of persons, such as politicians or actors, for which audio classifiers are pre-trained and passed to the service.

The module for speaker recognition is using GMM-based classifiers that are trained in a supervised manner on manually selected material. Given a set of possible speakers, the module selects the most probable person or, in case the corresponding probability is too low, hypothesizes an unknown speaker.

Nature: off-line

Development: current version uses only very few speakers. We plan to expand this number to achieve reasonable coverage of the persons present in the corpora used for evaluations. Furthermore we plan to implement a system that can be used with lighter supervision and especially detect additional persons without having to reprocess the whole corpus with the new model. Since the external interface will not change, integrating these updated versions into the processing chain will not require changes to other components.

Integration: integrated in AXES V1 and V2 as part of the same web service as SpeakerClustering module M2.7.

OpenAxes Status: this component will not be part of OpenAxes.

M2.8 SearchByExample (Inria)

Objective and description: This module is in charge of providing a face-based search by example functionality and of creating face-based links interactively between various videos.

This module can take as input an image or video frame with a bounding box of a face (if no bounding box is provided a face detector automatically selects the bounding box of the largest detected face). Given the bounding box, a set of 9 facial landmark points are detected, located on the eyes, nose, and mouth. The face image is then warped using a similarity transform (translation, rotation, and scaling) so that the landmark points are mapped as close as possible to a canonical configuration of a frontal pose. For each of the 9 landmarks, a signature is extracted by means of histograms of oriented gradients (HOG). These 9 signatures are concatenated to form the face signature. The high-dimensional face signature is then compressed into a lower dimensional signature by means of a linear projection. The projection matrix has been obtained by an off-line metric learning algorithm so that the L2 distance between signatures after projection is small for face signatures of the same person (despite pose and expression changes), and large for signatures of different people (despite similarities in pose and expression).The compressed face signature of the query image is then matched to face signatures in the database to find similar faces across other videos, which are returned by the service.

Nature: on-line (search part) / off-line (database face description and indexing)

Development: An additional command-line interface to this component is being developed to facilitate off-line processing of large video collections.

Integration: Integrated in AXES V1 and V2 as well as AXES-x platforms.

OpenAxes Status: this component will not be part of OpenAxes.

M2.9 UpperBodyDetectionAndTracking (UO)

Objective and description: this module is similar to M2.3, but with a larger visual focus on whole upper body of visible persons. Similarly to the face detection and tracking pipeline, is combines a pre-trained detector and grouping of detections via tracking. The human upper-body (UB) detector is based on Felzenswalb et al. deformable part model. Each video frame is independently processed with this detector and resulting detections are then grouped over time into UB tracks when they are connected by many KLT point tracks. Next, false-positive tracks are discarded based on track length and the score of the detections they contain. A track is discarded if it contains fewer than 20 detections, or if the sum of detection scores over it is below a threshold.
Nature: off-line
Development: this module is final
Integration: a preliminary version of this module is now installed on the BBC server.
This will be updated and tuned further in the final part of the project.
OpenAxes Status: this component will not be part of OpenAxes.

M2.10 VisualPersonCharacterization (UO)

Objective and description: This module aims at analyzing a single face detected in an image or all the faces composing a single face track, to decide whether this face/face-track exhibits certain characteristics such as age category, gender, presence of beard/moustache or speaking action. As for face recognition, two scenarios are envisaged: (1) an off-line automatic annotation scenario where a list of characteristics of potential interest and with visual distinctiveness has been established beforehand and associated classifiers have been trained off-line in a (semi) supervised fashion for each of them; these classifiers are then run on all faces/face-tracks in the database; (2) a search scenario in which a certain characteristic of interest, e.g., input by a user, is passed to the system that must first learn, on-the-fly, a classifier for this attribute, and then returns a ranked list of candidate faces/face-tracks from the database.

As for FaceRecognition M2.4, the search version of the module described above has been developed. It allows the user to search for specific visual facial attributes (e.g., has beard, has glasses, is a baby) in the different databases of the system. If follows exactly the same on-the-fly learning workflow and ranking as described in M2.4.

The same system can know be used in a completely off-line way, to pre-train and run in a batch mode classifiers for common attributes.

Nature: on-the-fly/off-line

Development: components are final, except for face descriptor which is likely to be modified in Y4 (see M2.4).

Integration: This is now available on the BBC server as part of the on-the-fly face retrieval installed there.

OpenAxes Status: this component will not be part of OpenAxes.

CONCLUSION

This report describes all WP2 modules. They provide AXES systems and platforms with a range of person-based analysis and search functions, respectively focused on visual face analysis (face grouping, description, recognition, search and characterization) and on speech analysis (speaker diarization and recognition).