www.axes-project.eu

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

Deliverable D4.5

Update on final software toolbox for audio-visual specific and category event 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:	D4.5
Deliverable title	Update on final software toolbox for audio-visual specific and category event recognition
Contractual Date of Delivery:	30/09/2014
Actual Date of Delivery:	30/09/2014
Author (s):	J. Verbeek, C. Schmid
Reviewer (s):	Y. Mombrun
Work package no.:	4
Work package title:	Multimodal analysis of events
Work package leader:	INRIA
Version/Revision:	1/1
Draft/Final:	Final
Total number of pages (including cover):	12

CHANGE LOG

Reason for change	Issue	Revision	Date
Initial draft, INRIA	1	0	01/09/2014
Review, Cassidian	1	1	03/09/2014
Final version INRIA	2	0	04/09/2014

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 report gives an update on the WP4 final software toolbox for audio-visual specific and category event recognition. For each component we give a brief description of the functionality and the integration status in the AXES platform. For details on the underlying technology we refer to the corresponding technical publications.

INTRODUCTION

Work package 4 is focused on recognition of events in video. In contrast to WP2 and WP3, events are "things that happen", and not entities such as people, objects, or places. So, for example, where WP2 and WP3 might detect "people" or "airplanes" in video, WP4 will focus on detecting demonstrations or an airplane taking off. We focus on developing approaches that characterize the temporally dynamic nature of events in videos. We differentiate between recognition of event categories and recognition of specific events.

Some of the techniques developed in this work package are related to those in WP2. For example, human actions can be seen as an event, but this also fits the human-centred analysis of WP2. However, WP4 targets techniques that are generic and can be applied to any type of event, such as event classification based on "bags of spatio-temporal features". Techniques that are specifically tailored to detection of human actions (e.g. using body part localization) are the responsibility of WP2. Furthermore, WP4 also includes techniques describing activities (as opposed to simple actions). These techniques allow capturing the multiple aspects of temporally dynamic events, such as dynamic interaction with objects, scenes and between multiple humans.

This report presents an update on the components of the final software toolbox for audio-visual specific and category event recognition, as outlined in the earlier report D4.1. The description below follows the WP4 services of the AXES system. Earlier versions of the software toolbox are described in reports D4.2, D4.3, and D4.4.

The main modules of WP4 are:

M4.1 Event category recognition

M4.2 Event clustering and segmentation

M4.3 Link structure generation

M4.1: Event category recognition

Responsible partner: INRIA

Partners involved: INRIA, UO, FHG

Objective

The goal of this service is to recognize event categories, such as “soccer games”, or “rock concerts”. Generic recognition techniques are used, e.g. not techniques specifically tuned to human (inter)actions, which are addressed in WP2.

Status

Two recognition systems are currently available. The first, is based on visual features extracted from single frames. The second system uses additional features, most importantly those extracted from motion patterns. Both systems are integrated in the AXES platform.

The first system, called VISOR and developed by UO, is based on key frame classification techniques. It is suitable for actions and events that can be identified in still images, such as demonstrations and concerts. An "on-the-fly" learning mechanism is available that given a textual category description fetches relevant images from an external image search engine (such as Google images). These images are processed to extract features, and a classifier is learned to separate these images from a fixed set of generic "background" images. The classifier is then applied to key frames from videos in the archive. This service is integrated in the AXES platform with both the off-line and on-line code fully distributed. This component will be part of Open AXES.

The second system in the final toolbox, developed by INRIA, combines static image features extracted from frames with motion, audio, OCR, and ASR features. With respect to the advanced toolbox, this system has been extended to include improved motion features, color features, and ASR (contributed by FHG). This system was used in the winning AXES submission to the TRECVID 2013 Multimedia Event Detection (MED) task. To validate the new tools internally we used the TRECVID MED 2011 data set, which shows a clear improvement in performance as compared to the versions of the system included in the baseline and advanced toolboxes. Currently an indexing and retrieval component are integrated in the AXES PRO system. The first component scans videos in the archive for 30 event categories of TRECVID MED 2013, and stores the shot classification scores in an index file. The currently indexing component relies only on motion pattern features. The second component allows the AXES platform to interact with the stored index. See [Wang & Schmid, 2013] and [Oneata et al, 2013] for details.

An updated version of the event recognition system has been developed, and is used in the AXES submission to the TRECVID 2014 Multimedia Event Detection (MED) evaluation benchmark. The update includes new visual features based on convolutional networks [Jia et al, 2014], and a novel audio feature based on the scatter transform [Andén & Mallat, 2011]. Some of the existing features have been improved by using larger dictionary sizes for the Fisher Vector calculation. In addition the code has been optimized in various places.

Plans

An updated version of the event recognition indexing service is expected to be fully integrated by the end of 2014. The update will include besides the motion pattern features, also audio features and static image features. The indexing component will be extended by adding another 20 categories, bringing the total to 50 event categories. This component is planned for integration in Open AXES as an add-on module. Since it is computationally quite heavy integration in the default package of Open AXES is not desirable.

M4.2: Event clustering and segmentation

Responsible partner: TEC

Partners involved: KUL, INRIA, TEC

Objective

The goal of this service is to recognize specific events, such as a particular speech of a politician, despite changes in viewpoint (different cameras filming the same event), cropping of the video (spatially or temporally), or differences due to compression. Videos that contain the same event can be clustered together, and the matching segments in the videos can be identified.

Status

The final software toolbox contains three methods for specific event recognition.

The first system, developed by INRIA and TEC [Douze2010], is an extension of techniques used for matching still images. Matching local descriptors provide "votes" on geometric and temporal deformation between the query and target. Consistency of local matches is used to obtain a robust matching.

A second system, developed by TEC [Bagri et al, 2013], exploits an audio-based technique to cluster and synchronize videos of a same event.

The third system, developed by INRIA [Revaud et al, 2013], is based on video-level descriptors and exploits the properties of circulant matrices to efficiently compare the videos in the frequency domain. This offers a significant gain in complexity and accurately localizes and synchronizes the matching parts of videos.

These modules are not integrated in the AXES platform.

Plans

Integration of these components is not planned; they will not be part of Open AXES.

M4.3: Link structure generation

Responsible partner: UO

Partners involved: INRIA, UO, KUL

Description

The goal of this service is to provide search by example functionality. It will allow end-users and the link management system to create links between similar videos. The functionality is similar to that of the "Event clustering and segmentation" service M4.2, however, the matching will be looser in the sense that links will also be generated between video fragments that are similar in terms of the events (specific or categories) that are detected, and/or based on similar textual descriptions obtained by speech recognition, video OCR, and other existing meta-data.

Status

No modules for this service are included in the final software toolbox.

Link generation has been deferred to WP6 based on the automatic annotation of categories of events M4.1. For instance, videos that contains similar events might be proposed to the user. To this end the M4.1 components are able to communcate their results to WP6 components, which use this input for link generation.

Plans

No integration is planned.

CONCLUSION

This report gives an update on the components of the final software toolbox D4.4 delivered in M33. Two systems for event category recognition developed by UO and INRIA are included (both are integrated). Three systems for recognition of specific events developed by INRIA and TEC are included, but not integrated. Link structure generation has been deferred to WP6, and is not included in the final software toolbox.

REFERENCES

[Douze2010]
An image-based approach to video copy detection with spatio-temporal post-filtering
Douze Matthijs; Jégou Hervé; Schmid Cordelia
IEEE Transactions on Multimedia, IEEE, 2010, 12 (4), pp. 257-266

[Revaud et al, 2013]
J. Revaud, M. Douze, C. Schmid, H. Jégou.
Event retrieval in large video collections with circulant temporal encoding.
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[Wang & Schmid 2013]
H. Wang and C. Schmid
Action recognition with improved trajectories
IEEE International Conference on Computer Vision (ICCV), 2013.

[Oneata et al, 2013]
D. Oneata, J. Verbeek, and C. Schmid
Action and Event Recognition with Fisher vectors on a Compact Feature Set
IEEE International Conference on Computer Vision (ICCV), 2013.

[Bagri et al, 2013]
A. Bagri, F. Thudor, A. Ozerov, and P. Hellier,
Scalable framework for joint clustering and synchronizing multi-camera videos
EUSIPCO 2013

[Jia et al, 2014]
Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell
Caffe: Convolutional Architecture
for Fast Feature Embedding
arXiv:1408.5093

[Andén & Mallat, 2011]
J. Andén and S. Mallat.
Multiscale Scattering for Audio Classification
Proceedings of the ISMIR conference, 2011.