☑

www.axes-project.eu

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

Deliverable 3.5

Update on final software toolbox for audio-visual specific and category object/scene 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:	3.5
Deliverable title	Update on final software toolbox for audio-visual specific and category object/scene recognition
Contractual Date of Delivery:	30/09/2014
Actual Date of Delivery:	30/09/2014
Author (s):	Andrew Zisserman
Reviewer (s):	Yann Mombrun
Work package no.:	3
Work package title:	Multimodal analysis of categories and places
Work package leader:	University of Oxford
Version/Revision:	v1.1
Draft/Final:	Final
Total number of pages (including cover):	15

Change Log

Reason for change	Issue	Revision	Date
Initial version		1.1	28/09/2013
Review		1.2	

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 describes the software toolbox for WP3: Visual specific and category object/scene recognition. It focuses on algorithmic details of the provided tools. Details are given in particular on the integration with the V2 version of the AXES system, and within the AXES-x (Pro/Research/Home) interactive platforms. It also describes plans for the remainder of the project.

Introduction

The goal of this work package is to produce annotations and links for specific places/buildings (such as Notre Dame) and for categories of buildings (such as churches) and categories of objects (such as cows and airplanes) in large-scale video collections.

The main modules, as defined in D3.1, D3.2 and D3.3, include:

M3.1 Specific Place Recognition

M3.2 Specific Place/Object Linking

M3.3 Scene Category Recognition

M3.4 Scene Category Linking

M3.5 Object Category Detection

M3.6 Specific Place/Object Clustering

Modules M3.1, M3.3, and M3.5 are concerned with **semantic annotation**.

Several of the modules use on-the-fly training where, for example, the category is specified only at run time. The category is used as a query to a web search engine to retrieve training images, and these images are then used to train a classifier on-the-fly and search the dataset.

Modules 3.2 and 3.4 do not assume any pre-defined places or categories. They amount to exploiting visual similarity to **link** to shots of various places or categories given a query. For these two tasks, there is *a priori* no need for training.

Note, audio-features are included in WP2 (people) and WP4 (events) rather than in this WP, since they bring more benefit there.

All of these modules assume that the video has been partitioned into shots (by one of the services of WP2).

M3.1 Specific Place Recognition

Responsible partner: KUL
Partners involved: KUL, UO

Objective:
This service is in charge of detecting specific places such as landmarks from video content.

Description:

Two methods have been implemented for 'on-the-fly' specific place recognition service: one by UO and one by KUL (see D3.2 for details). Both allow for the use of multiple query images simultaneously. The method from UO has been extended to include `outlier rejection' in the downloaded images. For example, removing a single image of a plan of the building for a text query on `the white house'. This removal is also carried out efficiently on-the-fly by using Hamming embedding of the downloaded images. It leads to an improvement in the retrieved images as an outlier can pollute the high ranked results.

Nature: on-the-fly/off-line.

Status:

- The UO on-the-fly systems are used for all of the AXES-x systems. The on-line system now runs on the BBC server. The off-line components run on the Oxford servers.

Plans:

-
- Distribution of the UO off-line (indexing) code in Y4 depends on the availability of clusters.
- An improved representation of images for instance retrieval is now available, see paper by Jegou 2014.
-
- M3.2 Specific Place/Object Linking

Responsible partner: INRIA
Partners involved: INRIA, KUL, UO

Objective:
This service is in charge of detecting the re-occurrence of the same place or specific object in video content, i.e. deciding whether two video fragments are taken at the same location or contain the same object. Typically, a keyframe or Region of Interest (ROI) of a keyframe will act as the query, and links are then generated to other keyframes/shots that contain the same location or object.

Description:

The method is described in D3.3 and is unchanged.

Nature: on-line

Status:

- Integrated in V2 and the AXES-x services.

Plans:

- Display of matched ROI in AXES-x systems
    -
    -

M3.3 Scene Category Recognition

Responsible partner: UO
Partners involved: UO, KUL

Objective:
This service is in charge of deciding whether keyframes and shots contain a certain visual concept (typically a scene or object category) for example that the scene contains an airplane or is of a city.

Description:
The UO **method is described in** D3.3. As planned in D3.3, the encoding method is updated as new methods become available. In D3.4 the encoding was updated from BoW to VLAD leading to improved performance (though no change to the API). Following the recent progress in using Convolutional Neural Network (CNN) descriptors for retrieval (see paper by Chatfield et al 2014), the encoding will next be updated from VLAD to CNNs.

KUL has extended its set of pretrained classifiers ("über-classifiers") to 1537 categories. These are trained based on the Imagenet dataset. Apart from the 1000 categories used in the ILSVRC challenge, this includes an extra 537 popular queries. Whenever a user poses a query, it's first checked whether it corresponds to one of these 1537 pretrained categories. If not, the UO on-the-fly service is called.

Nature: on-the-fly/off-line

Status:
- Both the on-line and off-line (index computation) have been integrated into V2 and run on the AXES servers. This module is used with all the AXES-x services.
- Code to index an archive, using a fixed set of über-classifiers, has been integrated in V2 and installed on the AXES servers.
-

Plans:
-
- The encoding method will be updated again in Y4 to use CNNs (this does not require any changes to the API)
- Adding visual attributes, e.g. colour

Dependencies:
- The offline index computation ("über-classifiers") is integrated into AXES OPEN as a binary, compiled from matlab. It uses/depends on:

- VLFeat (for code for SIFT features, GMMs, Fisher Vectors) (BSD license)
- liblinear-1.93 (for code for training the support vector machine classifiers) (BSD license)
- binary compiler MCR matlab
- data: imagenet (use of database limited to non-commercial use)

Use of this tool as part of AXES OPEN for commercial use may be limited due to patents related to the SIFT descriptor and the Fisher Vector encoding. These could be avoided by switching, e.g., to the CNN features provided by UO. Even then, it's unclear whether the models trained based on ImageNet data can be used for commercial use. For non-commercial use, there seem to be no issues.

M3.4 Scene Category Linking

Responsible partner: UO

Partners involved: UO, INRIA

Objective:

This service is in charge of detecting the re-occurrence of the same scene category in video content, i.e. deciding whether two video fragments contain the same content. Typically, one video or keyframe will act as the query, and links are then generated to other keyframe/shots which contain the same category.

Nature: on-the-fly

Status:

- This service is now provided by the UO M3.3 API

Plans:

- It will be integrated into the AXES-x systems (the interface is equivalent to providing relevance feedback).

M3.5 Object Category Detection

Responsible partner: KUL

Partners involved: KUL, UO

Objective:

This service is in charge of deciding whether keyframes and shots contain a certain visual object category, for example that the scene contains an elephant, and if so outputting its (spatio-temporal) location (i.e. a bounding box in space within each frame and possibly over time).

Description:

The sliding window based detector has been replaced by methods that only sample a reduced set of candidate windows or segments, either based on objectness or object proposals. Both INRIA as well as KUL have worked on methods for object category detection starting from weakly supervised data (no bounding boxes available at training time). UO has a new method of detecting object categories on-the-fly. See paper by Aytar 2014.

Nature: offline / on-the-fly

Status:

Being able to localize the objects in the image seems of little added value, so this feature has not been developed in any of the AXES-x systems

Plans:

-
-

M3.6 Specific Place/Object Clustering

Responsible partner: KUL
Partners involved: KUL, UO

Status: Since it was not clear how this component would be of use for the different demonstrators, this direction has not been further explored.

REFERENCES

Y. Aytar, A. Zisserman
Immediate, scalable object category detection
IEEE Conference on Computer Vision and Pattern Recognition, 2014

K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman
Return of the Devil in the Details: Delving Deep into Convolutional Nets
British Machine Vision Conference, 2014

H. Jégou, A. Zisserman
Triangulation embedding and democratic aggregation for image search
IEEE Conference on Computer Vision and Pattern Recognition, 2014