ITI-CERTH participation in TRECVID 2016

Foteini Markatopoulou[1,2], Anastasia Moumtzidou[1], Damianos Galanopoulos[1], Theodoros Mironidis[1], Vagia Kaltsa[1], Anastasia Ioannidou[1], Spyridon Symeonidis[1], Konstantinos Avgerinakis[1], Stelios Andreadis[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], Alexia Briassouli[1], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], Ioannis Patras[2]

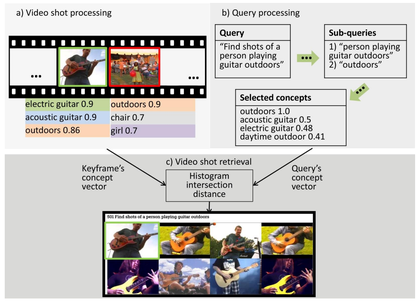
[1]Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{markatopoulou, moumtzid, dgalanop, mironidis, vagiakal, ioananas, spyridons, koafgeri, andreadisst, heliasgj, stefanos, abria, bmezaris, ikom} @iti.gr

[2]Queen Mary University of London, Mile end Campus, UK
i.patras@qmul.ac.uk

Ad-hoc Video Search

Video shot processing: Video shot annotation with concepts from two concept pools:
A) 1000 ImageNet concepts - Late fusion of 5 different pre-trained Deep Convolutional Neural Nets (DCNNs)
B) 345 TRECVID SIN concepts - 3 pre-trained ImageNet networks fine-tuned (FT) on these concepts

- The best performing FT network was chosen
- Two different approaches for video shot annotation
  - Direct output of the FT network
  - Linear SVM training with DCNN-based features



AVS system overview

Query processing: Selecting concepts which are the most related to a query, exploiting NLP procedures

- Concept selection using the ESA distance between the query and each of the concepts
- String matching between concepts and any part of the query
- Query's linguistic analysis for **sub-queries** creation
- Concept selection using the ESA distance between sub-queries and each of the concepts
- Our MED16 000Ex pipeline is used, if none of the previous steps is able to select concepts

Runs

ITI-CERTH 1: Annotation using DCNN outputs for 1000 ImageNet concepts and SVM-based scores for 345 TRECVID SIN concepts
ITI-CERTH 2: Annotation using DCNN outputs for 1000 ImageNet concepts and the direct output of the FT network for 345 SIN concepts
ITI-CERTH 3: ITI-CERTH 1 ignoring sub-queries which do not present high semantic relatedness with any of the concepts
ITI-CERTH 4: ITI-CERTH 1 without string matching

Submitted run:	ITI-CERTH 1	ITI-CERTH 2	ITI-CERTH 3	ITI-CERTH 4
Fully-automatic	**0.051**	0.042	**0.051**	**0.051**

Multimedia Event Detection

010Ex and 100Ex

- Kernel Subclass Discriminant Analysis for dimensionality reduction and fast linear SVM for training (**KSDA+LSVM**)
- Both DCNN-based and motion features

DCNN-based features

- GoogLeNet trained with 12988 ImageNet concepts
- The best FT network on the 345 TRECVID SIN concepts (the same used for the AVS task)
- GoogLeNet trained with 500 EventNet concepts
- GoogLeNet trained with 5055 ImageNet concepts and FT on 487 Sports-1M concepts
- GoogLeNet trained with Places 205 concepts

Motion features

- HOG, HOF, MBHx, MBHy

Video Representation

- Concatenate motion and DCNN-based feature vectors
- New feature vector in $\mathbb{R}^{153781}$
- KSDA drastically reduces the feature vector dimensionality

Run ID	MAP%	mInfAP@200%
010Ex p-1KDALSVM	31.8	34.2
100Ex p-1KDALSVM	46.2	47.5

Conclusion

- Plentiful DCNN-based features lead to better results

000Ex

- An improved version of our MED15 zero-example event detection framework is used
- Extensive visual concept pool: up to 14k concepts
- Pseudo-relevance feedback using KSDA+LSVM and retrieved videos from the MED16–EvalSub set

DCNN-based features

- Direct output of 5 DCNNs (from 010Ex and 100Ex)
- Direct output of GoogLeNet trained with 5055 ImageNet concepts

Run ID	MAP%	mInfAP@200%
p-1DCNN13K_1	14.6	12.2
c-1DCNN14K_1	14.5	11.9
c-1DCNN05K_1	02.5	01.4
c-3Train_1	**16.2**	**14.2**

Runs

p-1DCNN13K_1
- ImageNet 12988 concepts
- EventNet 500 concepts

c-1DCNN14K_1
- ImageNet 12988 concepts
- EventNet 500 concepts
- TRECVID SIN 345 concepts
- Sports-1M 478 concepts
- Places 205 concepts

c-1DCNN05K_1
- ImageNet 5055 concepts
- EventNet 500 concepts

c-3Train_1
- KSDA+LSVM using as positive samples the 10 top retrieved videos of p-1DCNN13K_1 run

Conclusion

- Pseudo-relevance feedback has a significant impact to our performance (the relative improvement is 16.39%)
- Using a large number of visual concepts gives a boost to the performance compared with our MED15 submission

Instance Search

VERGE video search engine

VERGE retrieval and presentation modules

High Level Visual Concept Retrieval

- 346 TRECVID SIN concepts using pre-trained DCNNs & training with Linear SVMs
- 205 Places scene concepts using GoogLeNet for landscape recognition

Visual Similarity Search module

- Use of pre-trained DCNN on 5055 ImageNet categories & selection of last pooling layer for keyframe representation
- Nearest Neighbour search realized using Asymmetric Distance Computation

Face detection and Face Retrieval Module

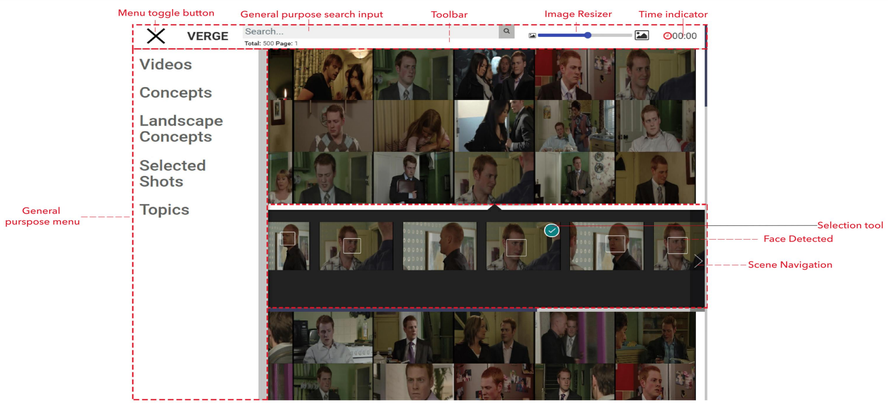
- Face detection: detect facial landmarks using cascade of DCNNs
- Face feature extraction: extract DCNN descriptors using VGG-Very-Deep-16 architecture & use of last FC layer as feature vector
- Construction of an IVFADC index for fast face retrieval

Run ID	MAP	Recall
Run 1	0.114	1000/11197

Evaluation of INS results

Future work

- Combination of the different modules for improving the video search results



VERGE GUI
http://mklab-services.iti.gr/trec2015/

KSDA+LSVM software (executables and sources) available at: http://mklab.iti.gr/project/aksda

Supported by: