ITI-CERTH in TRECVID 2016
Ad-hoc Video Search (AVS)

Foteini Markatopoulou, Damianos Galanopoulos, Ioannis Patras, Vasileios Mezaris

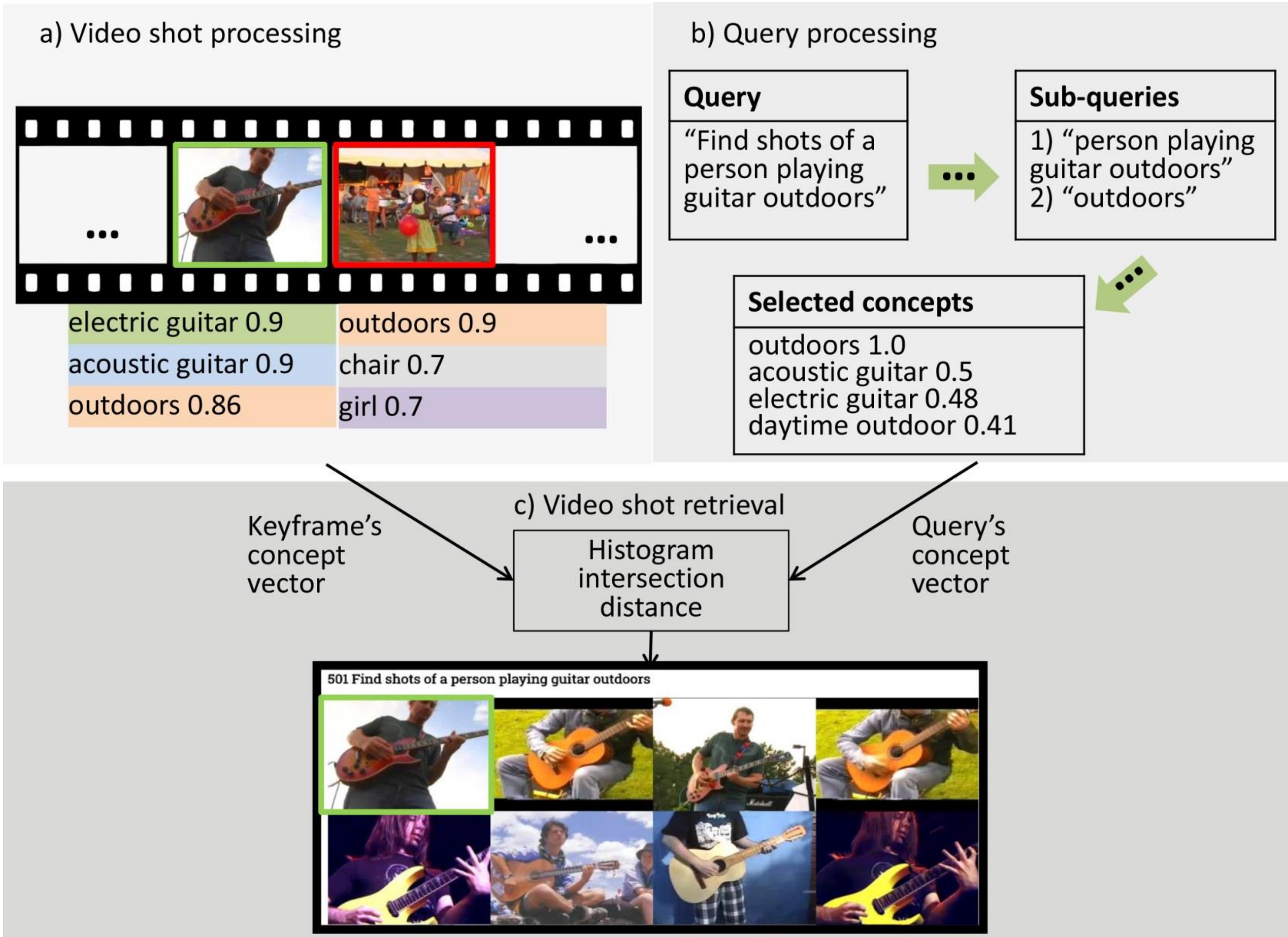
Information Technologies Institute / Centre for Research and Technology Hellas

TRECVID 2016 Workshop, Gaithersburg, MD, USA, November 2016

Highlights

- AVS's task objective is to retrieve a list of the 1000 most related test shots for a specific text query
- Our approach: a fully-automatic system
- The system consists of three components
  - Video shot processing
  - Query processing
  - Video shot retrieval
- Both fully-automatic and manually-assisted (with users just specifying additional cues) runs were submitted

System Overview



Video shot processing

- Extract one keyframe from each video-shot and annotate it using a pool of 1345 concepts:
  - ImageNet 1000
  - TRECVID SIN 345
- A temporal re-ranking method is employed to refine the calculated detection scores
- The final **keyframe's concept vector** in $\mathbb{R}^{1345}$ represents each video shot
- We find all the synonyms of each concept using WordNet; each concept's synonyms are considered as equivalent to the original concept

Video shot processing

ImageNet 1000

- Five pre-trained DCCNs for 1000 concepts
  - AlexNet
  - GoogLeNet
  - ResNet
  - VGG Net
  - GoogLeNet trained on 5055 ImageNet concepts (we only considered the subset of 1000 concepts out of the 5055 ones)
- Late fusion (averaging) on the direct output of the networks to obtain a single score per concept

Video shot processing

TRECVID SIN 345

- Three pre-trained ImageNet networks, fine-tuned (FT; three FT strategies with different parameter instantiations from [1]; in total 51 FT networks) for these concepts
  - AlexNet (1000 ImageNet concepts)
  - GoogLeNet (1000 ImageNet concepts)
  - GoogLeNet originally trained on 5055 ImageNet concepts
- The best performing FT network (as evaluated on the TRECVID SIN 2013 test dataset) is selected
- Examined two approaches for using this for shot annotation
  - Using the direct output of the FT network
  - Linear SVM training with DCNN-based features

[1] N. Pittaras, F. Markatopoulou, V. Mezaris, I. Patras, "Comparison of Fine-tuning and Extension Strategies for Deep Convolutional Neural Networks", at the 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavik, Iceland, 4 January 2017. (accepted for publication)

Query processing

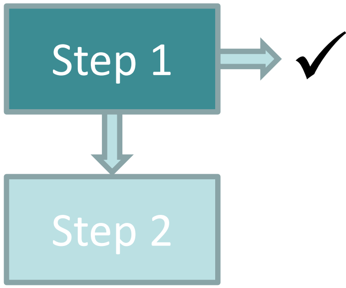
- Each query is represented as a vector of related concepts
  - We select concepts which are most closely related to the query
  - These concepts form the **query's concept vector**
  - Each element of this vector indicates the degree that the corresponding concept is related to the query
- A five-step procedure is used
  - Each step selects concepts, from the concept pool, related to the query

Query processing: Step 1

Motivation: Some concepts are semantically close to input query and they can describe it extremely well

Approach:

- Compare every concept in our pool with the entire input query, using the Explicit Semantic Analysis (ESA) measure
- If the score between the query and a concept is higher than a threshold (0.8) then the concept is selected
- If at least one concept is selected in this way, we assume that the query is very well described and the query processing stops; otherwise the query processing continues in **step 2**

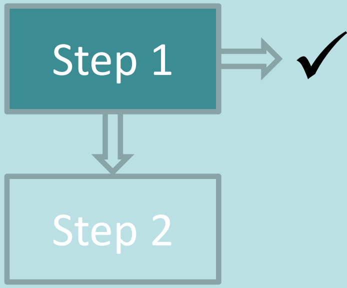


Example: the query ***Find shots of a sewing machine*** and the concept ***sewing machine*** are semantically extremely close

Query processing: Step 1

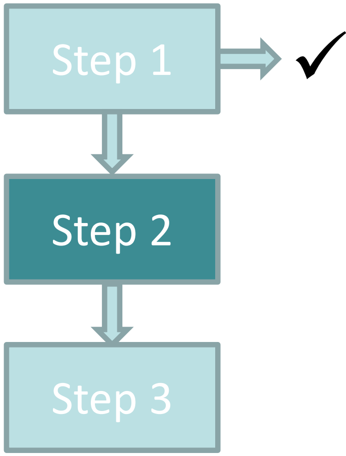
The processing stopped in step 1 for 3 out of the 30 queries:

- For **Find shots of a sewing machine** the concept **sewing machine** was selected
- For **Find shots of a policeman where a police car is visible** the concept **police car** was selected
- For **Find shots of people shopping** the concept **tobacco shop** was selected



Query processing: Step 2

Motivation: Some (complex) concepts may describe the query quite well, but appear in a way that subsequent linguistic analysis to break down the query to sub-queries can make their detection difficult



Approach:

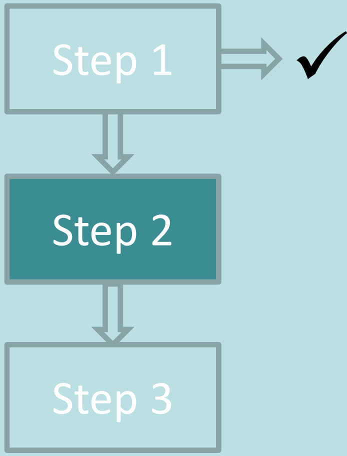
- We search if any of the concepts appear in any part of the query, by string matching
- Any concepts that appear in the query are selected and the query processing continues in **step 3**

Example: For the query ***Find shots of a man with beard and wearing white robe speaking and gesturing to camera*** the concept ***speaking to camera*** was found

Query processing: Step 2

For 5 out of 30 queries concepts were selected through string matching

- For ***Find shots of a man with beard and wearing white robe speaking and gesturing to camera***, the concept ***speaking to camera*** was selected
- For ***Find shots of one or more people opening a door and exiting through it***, the concept ***door opening*** was selected
- For ***Find shots of the 43rd president George W. Bush sitting down talking with people indoors***, the concept ***sitting down*** was selected
- For ***Find shots of military personnel interacting with protesters***, the concept ***military personnel*** was selected
- For ***Find shots of a person sitting down with a laptop visible***, the concept ***sitting down*** was selected

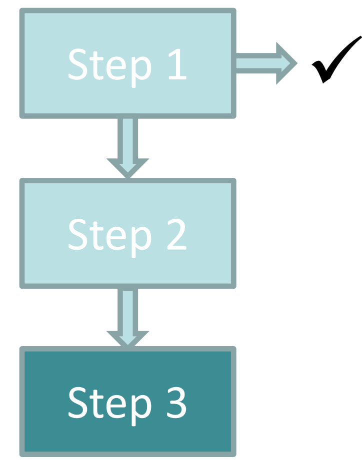


Query processing: Step 3

Motivation: Queries are complex sentences; we decompose queries to understand and process better their parts

Approach:

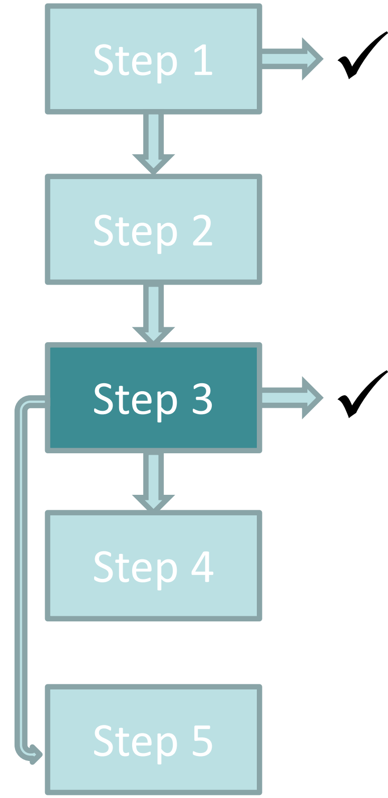
- We define a *sub-query* as a meaningful smaller phrase or term that is included in the original query, and we automatically decompose the query to subqueries
  - NLP procedures (e.g. PoS tagging, stop-word removal) and task-specific NLP rules are used
  - For example the triad **Noun-Verb-Noun** forms a *sub-query*
- The ESA distance is evaluated for every *sub-query* – concept pair
- If the score is higher than our step-1 threshold (0.8), then the concept is selected



Query processing: Step 3

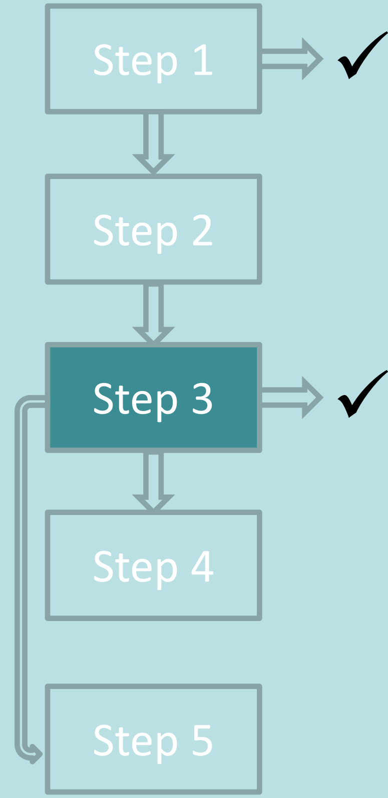
Example: the query ***Find shots of a diver wearing diving suit and swimming under water*** is split into the following four *sub-queries*: ***diver wearing diving suit, swimming, water***

- If for every sub-query at least one concept is selected we consider the query completely analyzed and we proceed to **video shot retrieval** component
- If for a subset of the *sub-queries* no concepts have been selected we continue to **step 4**
- If for all of the of the *sub-queries* no concepts have been selected we continue to **step 5**



Query processing: Step 3

- On average, a query was broken down to 3.7 sub-queries
- For none of the test queries there was at least one concept from our pool matched to each sub-query
- For 17 out of 27 queries, concepts were matched to a subset of the sub-queries, thus the processing continued to **step 4**
- For the remaining 10 queries, no concept was matched to any of their sub-queries, thus the processing continued to **step 5**

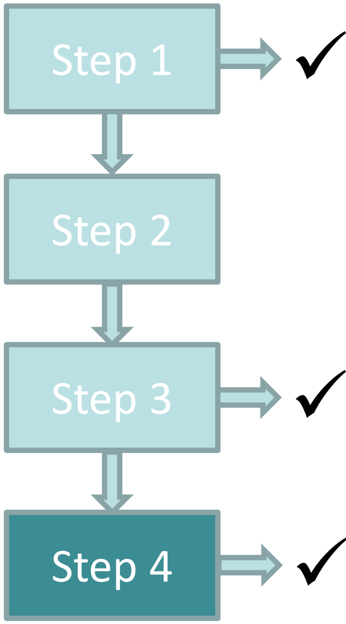


Query processing: Step 4

Motivation: For a subset of the *sub-queries* no concepts were selected due to their small semantic relatedness (i.e., in terms of ESA measure their relatedness is lower than the 0.8 threshold)

Approach:

- For these *sub-queries* the concept with the higher value of ESA measure is selected, and the we proceed to **video shot retrieval**



Example:

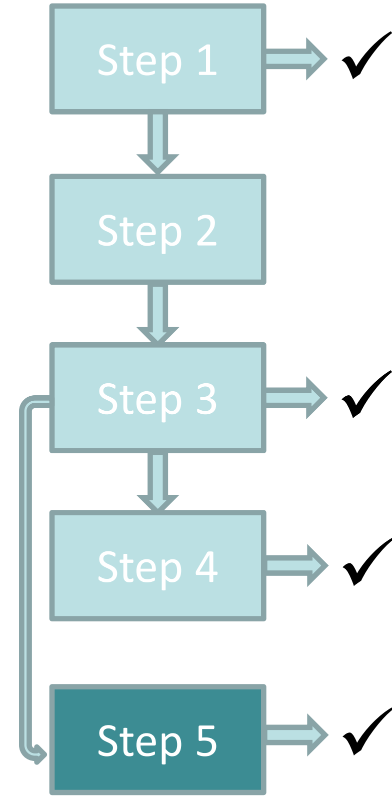
Query: Find shots of one or more people walking or bicycling on a bridge during daytime		
	Sub-queries	Selected concepts (ESA score)
Steps 2,3	• people walking • bicycling • bridge	• walking (1.0) • bicycle-built-for-two (1.0) • suspension bridge (1.0) • bicycles (0.85) • bridges (0.84) • bicycling (0.84)
Step 4	• daytime	• daytime outdoor (0.74)

Query processing: Step 5

Motivation: For some queries none of the above steps is able to select concepts

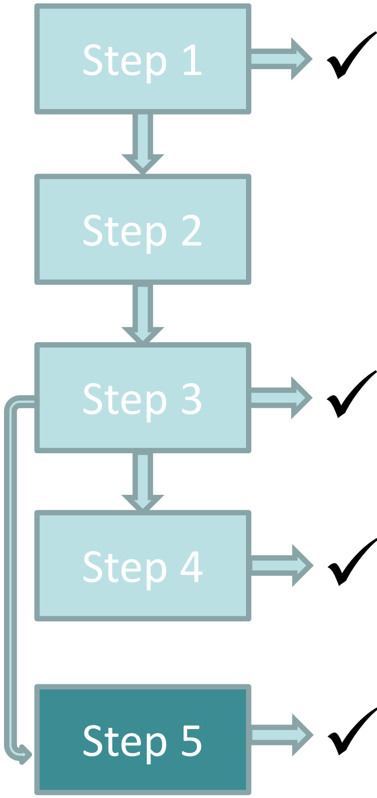
Approach:

- Our MED16 000Ex framework is used
- The query title and its sub-queries form an Event Language Model
- A Concept Language Model is formed for every concept using retrieved articles from Wikipedia
- A ranked list of the most relevant concepts and the corresponding scores (semantic correlation between each query-concept pair) is returned
- We proceed to **video shot retrieval** component



Query processing: Step 5

Example: For the query **Find shots of a person playing guitar outdoors** the framework returns the following concepts: **outdoor**, **acoustic guitar**, **electric guitar** and **daytime outdoor**



Video shot retrieval

- The query’s concept vector is formed by the corresponding scores of the selected concepts
- If a concept has been selected in steps 1, 3, 4 or 5 the corresponding vector’s element is assigned with the relatedness score (calculated using the ESA measure) and if it has been selected in step 2 it is set equal to 1
- Histogram intersection calculates the distance between **query’s concept vector** and **keyframe’s concept vector** for each of the test keyframes
- The 1000 keyframes with the smallest distance from query’s concept vector are retrieved

Submitted Runs

- We submitted both fully-automatic and manually-assisted runs
- For the manually-assisted ones
  - We used the same fully-automatic system, but
  - A member of our team that was not involved in the development of our AVS system took a look at each query and manually suggested *sub-queries* for it, without knowledge of the automatically-generated ones
  - The manually defined *sub-queries* were added to the automatically-generated ones, and our automatic AVS system was applied

Submitted Runs

ITI-CERTH 1:

- Late fusion of the direct output from 5 DCNNs for ImageNet 1000 concepts
- SVM-based concepts detectors for 345 TRECVID SIN concepts

ITI-CERTH 2:

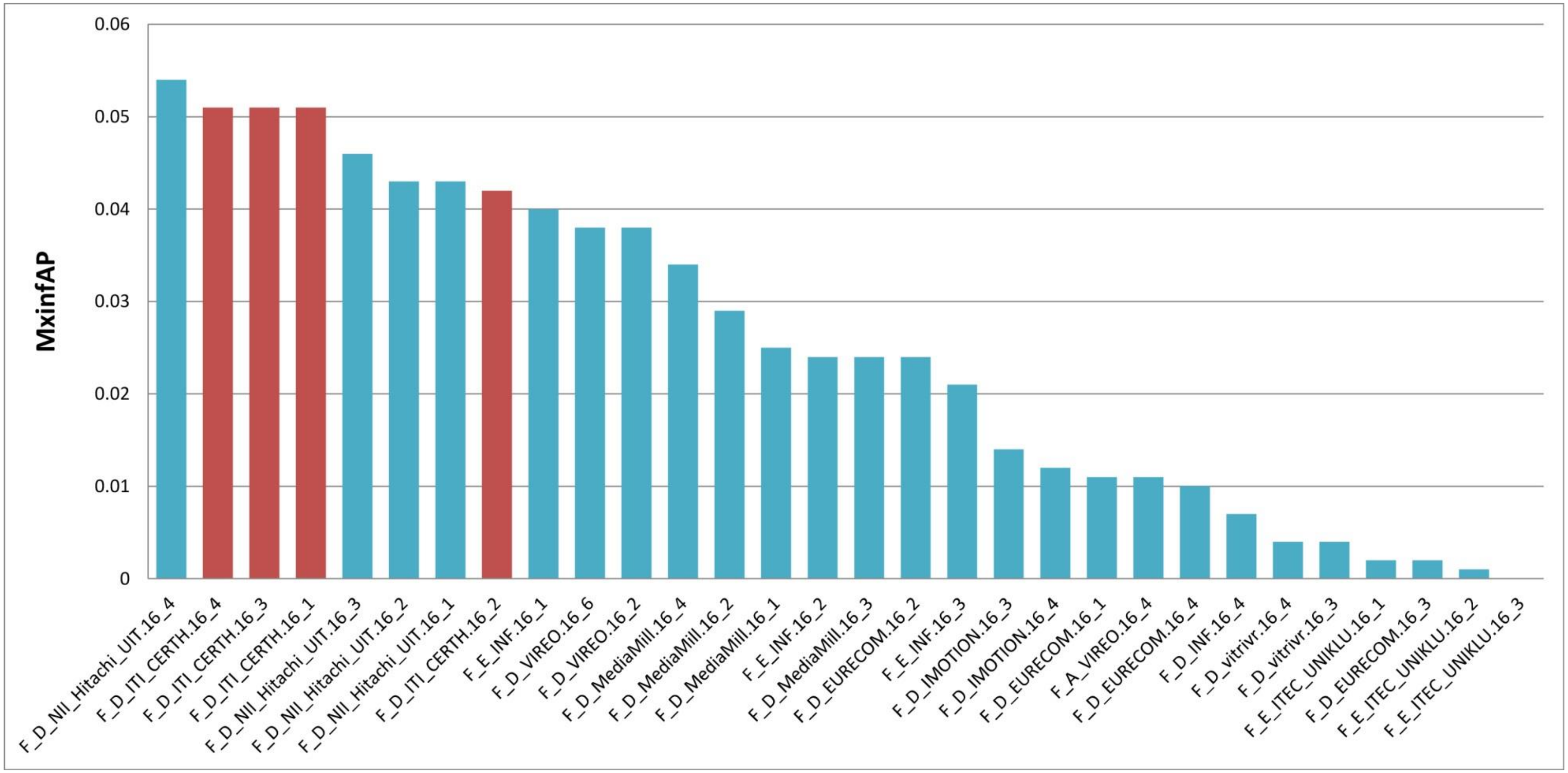
- Late fusion of the direct output from 5 DCNNs for ImageNet 1000 concepts
- The direct output of the FT network for 345 TRECVID SIN concepts

ITI-CERTH 3: ITI-CERTH 1 run without step 4

ITI-CERTH 4: ITI-CERTH 1 run without step 2

Submitted run:	ITI-CERTH 1	ITI-CERTH 2	ITI-CERTH 3	ITI-CERTH 4
MXinfAP (fully-automatic)	**0.051**	0.042	**0.051**	**0.051**
MXinfAP (manually-assisted)	0.043	0.037	0.037	0.043

Results (fully-automatic runs)



Results and conclusions

- Training SVMs on DCNN-based features instead of using the direct output of the DCNNs, for the 345 TRECVID SIN concepts, improves the accuracy (i.e., run ITI-CERTH 1 outperforms ITI-CERTH 2)
- In the AVS 2016 dataset
  - Step 4 could be omitted for the fully-automatic runs
    - Sub-queries without high semantic relatedness can be ignored; ITI-CERTH 1 & ITI-CERTH 3 achieve the same results
  - Step 2 could be omitted
    - String matching between the test query and concepts does not improve the accuracy; semantic relatedness makes the difference
- Fully-automatic runs outperformed the manually-assisted ones
- Our best fully-automatic run was ranked 2nd-best in the fully-automatic run category; it also outperformed the runs of all but one participant in the manually-assisted run category

Questions?

More information and contact:
Vasileios Mezaris, http://www.iti.gr/~bmezaris, bmezaris@iti.gr

TRECVID 2016 paper:
F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, T. Mironidis, V. Kaltsa, A. Ioannidou, S. Symeonidis, K. Avgerinakis, S. Andreadis, I. Gialampoukidis, S. Vrochidis, A. Briassouli, V. Mezaris, I. Kompatsiaris, I. Patras, "ITI-CERTH participation in TRECVID 2016", Proc. TRECVID 2016 Workshop, Gaithersburg, MD USA, November 2016.