# ITI-CERTH participation to TRECVID 2015

Foteini Markatopoulou[1,2], Anastasia Ioannidou[1], Christos Tzelepis[1,2], Theodoros Mironidis[1], Damianos Galanopoulos[1], Stavros Arestis-Chartampilas[1], Nikiforos Pittaras[1], Konstantinos Avgerinakis[1], Nikolaos Gkalelis[1], Anastasia Moumtzidou[1], Stefanos Vrochidis[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, ioananas, tzelepis, mironidis, dgalanop, stav_ares, npittaras, koafgeri, gkalelis, moumtzid, stefanos, bmezaris, ikom} @iti.gr

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

## Semantic Indexing

**System overview:** 1) Features based on pre-trained and fine-tuned deep convolutional neural networks (DCNN); 2) Dimensionality reduction using Kernel Subclass Discriminant Analysis (KSDA); 3) Cascade of classifiers using local and DCNN-based descriptors; 4) Multi-task learning (MTL) using the logistic-lasso algorithm; 5) Two-layer stacking architecture

| | MXinfAP |
|---|---|
| **System A** | 0.239 |
| **System B** | **0.258** |
| **System C** | 0.219 |
| **System D** | 0.232 |

| ITI-CERTH-Runs | MXinfAP |
|---|---|
| **Run1** | **0.263** |
| **Run2** | 0.26 |
| **Run3** | 0.25 |
| **Run4** | 0.252 |

**System A**: Cascade of four stages; 11 visual descriptors, one LSVM per descriptor (3xORB, 3xSIFT, 3xSURF, 2xDCNN-based)
**System B**: Cascade as in **A**, with features reduced to a much lower dimension using KSDA
**System C**: One MTL model for each of the reduced descriptors in **B**; per-concept late fusion of the MTL models' scores
**System D**: Late fusion of 7 DCNN-based descriptors, using 5 LR models per descriptor; resulting in 35 LR models per concept

**ITI-CERTH-Run1**: "Combination". Combination of **A** through **D**
**ITI-CERTH-Run2**: "MTL-LogLasso". Combination of **C** and **D**
**ITI-CERTH-Run3**: "Cascade KSDA". Combination of **B** and **D**
**ITI-CERTH-Run4**: "Cascade". Combination of **A** and **D**

- **Runs 3** and **4**: Similar accuracy; the cascades offered ≈37% less classifier evaluations compared to the late fusion alternative
- **Run 2**: Slight improvement due to Multi-task learning
- **Run 1**: Marginal improvement over previous runs

## Multimedia Event Detection

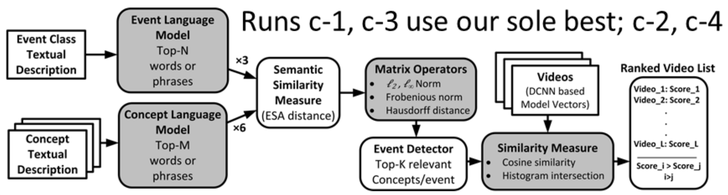
**In 010Ex and 100Ex**:

- **KSDA+LSVM** (c-1KDALSVM used visual, motion, fc7+fc8 DCNN descriptors); achieved better results; 2nd best among all contestants on 100Ex
- **RDKSVM** (c-2RDKSVM used fc8, c-3RDKSVM used fc7+fc8); exploits near-miss videos; performance limited by descriptors

| | 010Ex | 100Ex |
|---|---|---|
| **Run ID** | **mInfAP@200** | **mInfAP@200** |
| **p-1Fusion** | 0.211 | 0.3649 |
| **c-1KDALSVM** | **0.2493** | **0.4111** |
| **c-2RDKSVM** | 0.1588 | 0.2894 |
| **c-3RDKSVM** | 0.2026 | 0.2367 |

**In 000Ex**:

3 ELMs*6 CLMs*4 mat. op. = 72 detectors;
Runs c-1, c-3 use our sole best; c-2, c-4 use top-10 average.




| | 000Ex |
|---|---|
| **Run ID** | **mInfAP@200** |
| **p-1Fusion** | **0.0617** |
| **c-1oneCosine** | 0.0478 |
| **c-2avgCosine** | 0.0473 |
| **c-3oneHist** | 0.0474 |
| **c-4avgHist** | 0.0592 |

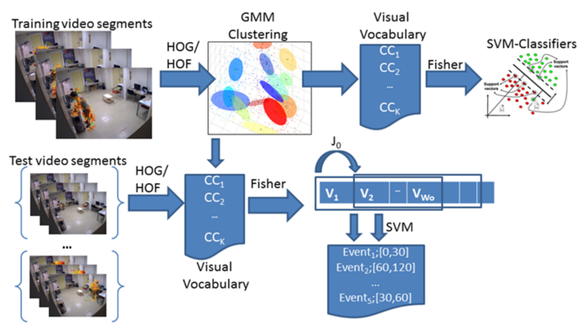
Fusion of multiple detectors leads to performance increase.

## Surveillance Event Detection

Activity detection system stages:

- Offline Stage: Extraction of Dense Trajectories + HOGHOF descriptors; Fisher encoding; LSVM model training
- Online Stage: a 50-frame sliding window along with pre-trained models for localizing the desired events inside test videos

Steps of the Interactive SED system

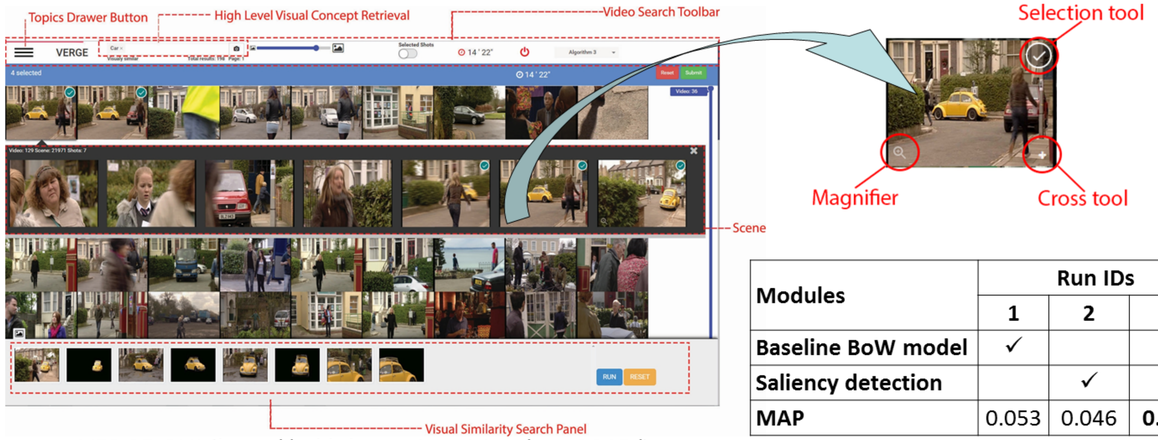



**5 Events of interest**: Embrace, PeopleMeet, Pointing, PersonRuns, and PeopleSplitUp

| Metric \ Event | Embrace | PeopleMeet | Pointing | PersonRuns | PeopleSplitUp |
|---|---|---|---|---|---|
| ADCR | 0.9855 | 0.9990 | 1.0054 | 0.9834 | 0.9868 |
| MDCR | 0.9855 | 0.9984 | 1.006 | 0.9823 | 0.9868 |

## Instance Search

- Participation with VERGE interactive video search engine
- VERGE retrieval and presentation modules
  - High Level Visual Concept Retrieval (Semantic Indexing)
  - Object-based Visual Search Module
    - Fast Hessian detector + SIFT descriptor for local feature extraction
    - BoW model: 2-layer Visual Vocabulary (100K) construction using Repeated Bisecting K-Means
    - Inverted Index implementation using Apache Lucene search engine library
    - Capability of querying the system either with whole frames or any cropped part of them
    - Support of Multiple Queries
    - Similarity Score: Borda Count based on tf-idf weights and ranking positions in the retrieved list
    - Best performance with fusion of baseline BoW + Saliency-based BoW models




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

| Modules | Run IDs | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| **Baseline BoW model** | ✓ | | ✓ |
| **Saliency detection** | | ✓ | ✓ |
| **MAP** | 0.053 | 0.046 | **0.064** |

Evaluation of INS results

KSDA demo software (for GPU): http://mklab.iti.gr/project/gpu-agsda
Online demo for video concept detection: http://multimedia2.iti.gr/onlinevideoanalysis/service/start.html

Supported by: ForgetIT, MultiSensor, HOMER (WWW.HOMER-PROJECT.EU)