ITI-CERTH participation in TRECVID 2018

Konstantinos Avgerinakis, Anastasia Moumtzidou, Damianos Galanopoulos, Georgios Orfanidis, Stelios Andreadis, Foteini Markatopoulou, Elissavet Batziou, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{koafgeri, moumtzid, dgalanop, g.orfanidis, andreadisst, markatopoulou, batziou.el, kioannid, stefanos, bmezaris, ikom} @iti.gr

Ad-hoc Video Search

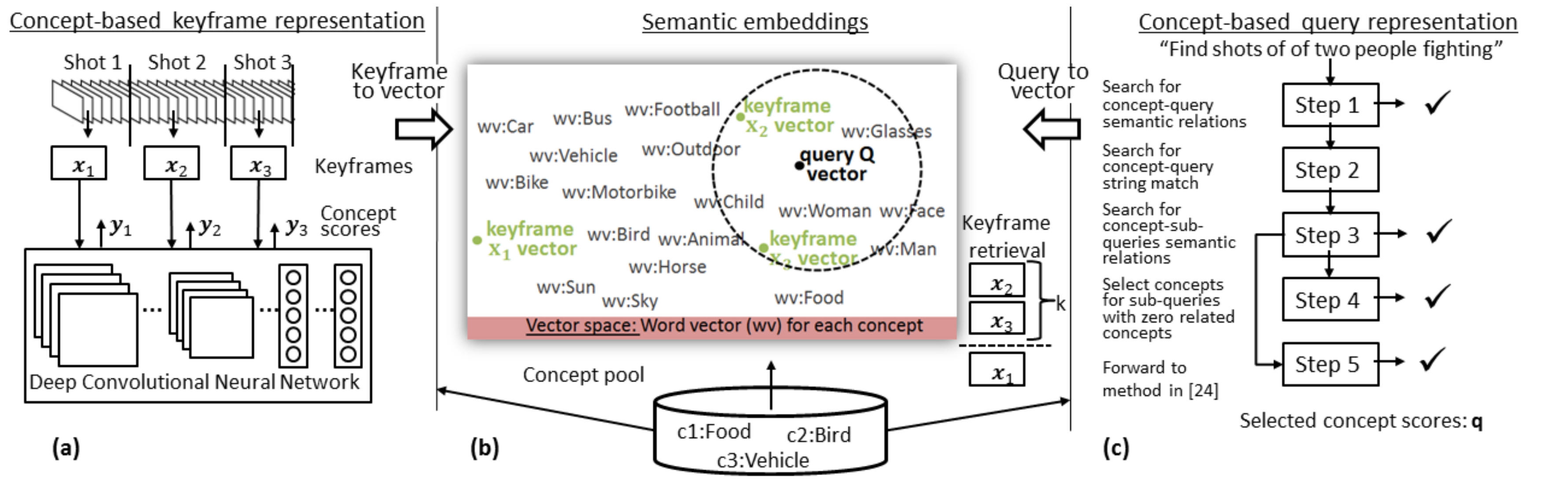
Keyframe Representation: Keyframe annotation using 33142 concept detectors.

- ImageNet 1000 - Late fusion of 5 different DCNNs.
- Pre-trained ResNet ImageNet DCNN; fine-tuned on SIN 345 concepts, plus 3 additional concepts derived from SIN.
- Pre-trained DCNNs for FCVID-239, Places-205, Places-365, Hybrid-1365 & ImageNet 4000, 4437, 8201, 12988.

Query Representation: Cue extraction and query representation as a vector of concepts. Three approaches:

i. A variety of complex NLP rules for cue extraction; multiple stages of matching cues with concepts.
ii. Cue extraction by finding noun phrases; multiple stages of cue-concept matching.
iii. Keywords are extracted by finding nouns; the most similar concept is selected for each noun.

Semantic embeddings: Both query and keyframe concept-based representations are transformed to semantic-embedding representations.



Runs:

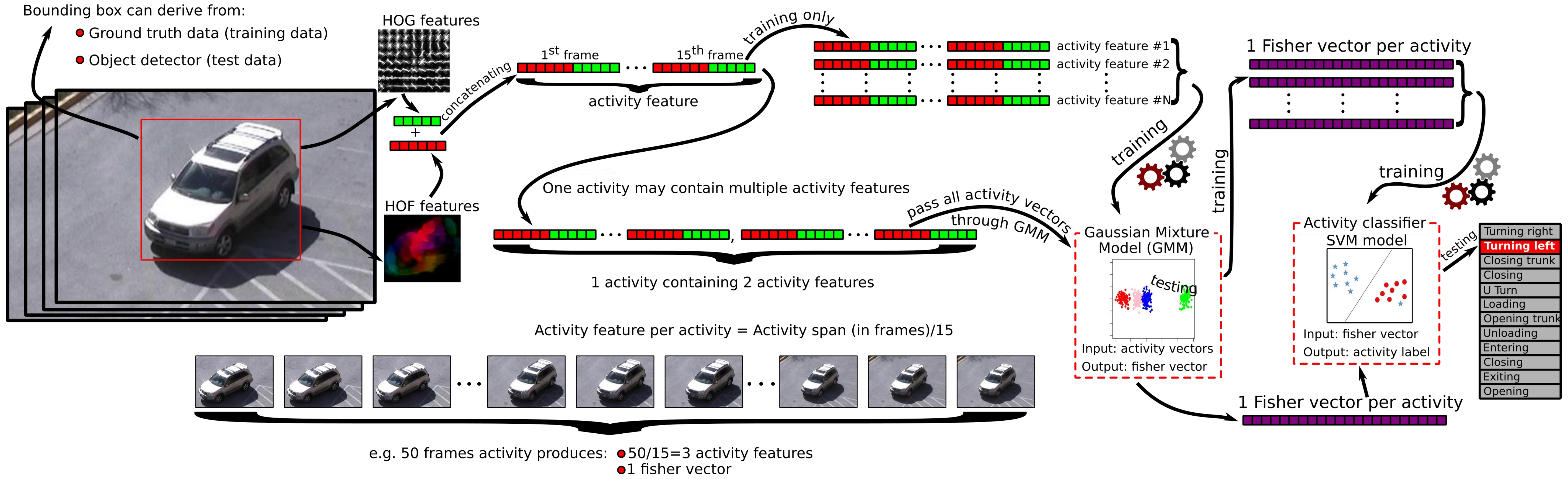
ITI-CERTH 1: Runs 2 & 4 combination (late fusion)
ITI-CERTH 2: Approach (iii) for query representation
ITI-CERTH 3: Approach (ii) for query representation
ITI-CERTH 4: Approach (i) for query representation

Results:

Run	ITI-CERTH 1	ITI-CERTH 2	ITI-CERTH 3	ITI-CERTH 4
MXinfAP	0.043	**0.047**	0.040	0.034

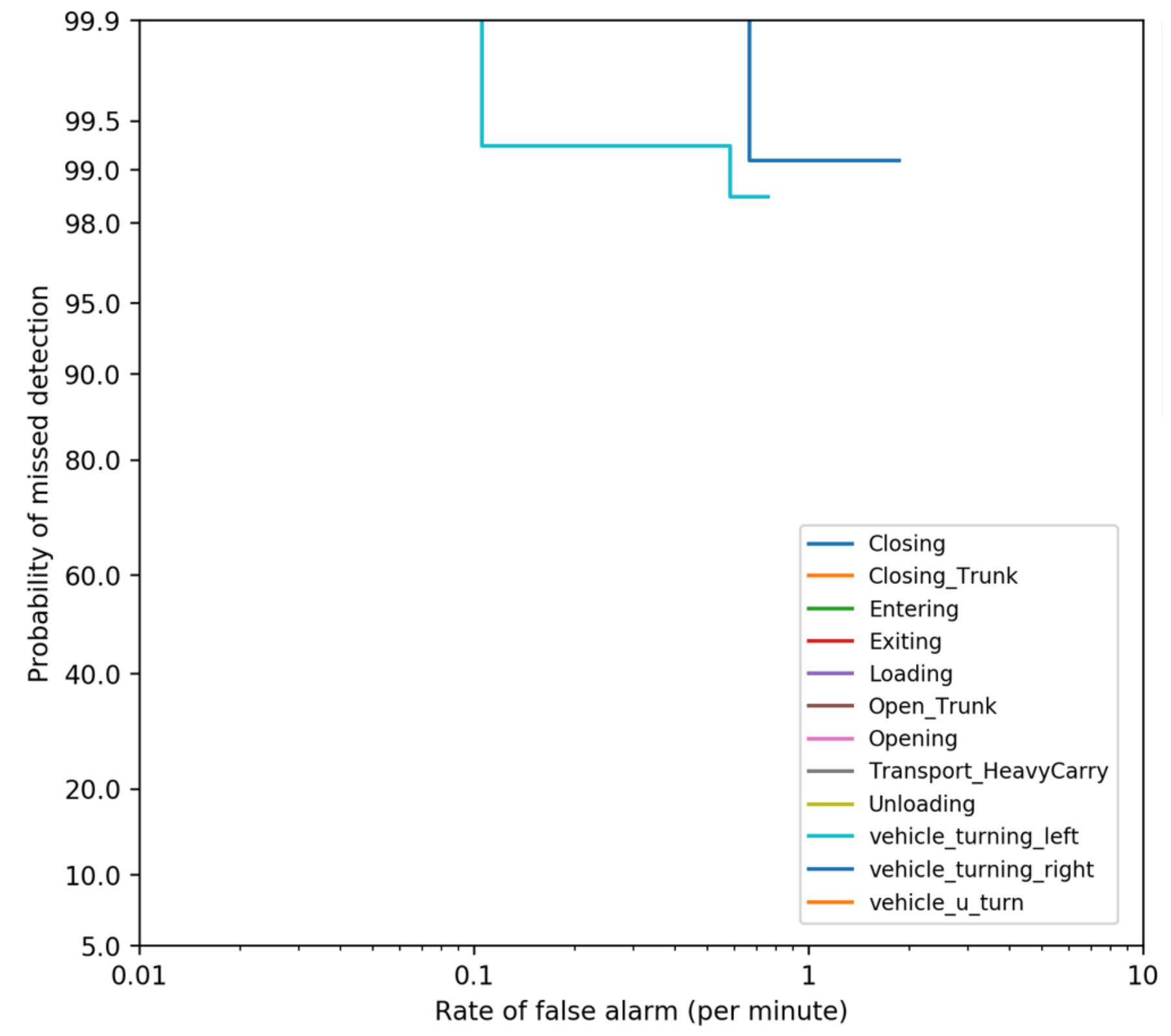
Activities in Extended Video

Activity Recognition Pipeline:


Bounding box can derive from:
● Ground truth data (training data)
● Object detector (test data)
HOG features
HOF features
+
concatenating
1st frame
15th frame
activity feature
training only
activity feature #1
activity feature #2
activity feature #N
One activity may contain multiple activity features
1 activity containing 2 activity features
pass all activity vectors through GMM
Activity feature per activity = Activity span (in frames)/15
e.g. 50 frames activity produces: ●50/15=3 activity features ●1 fisher vector
training
Gaussian Mixture Model (GMM)
testing
Input: activity vectors
Output: fisher vector
training
1 Fisher vector per activity
training
Activity classifier SVM model
Input: fisher vector
Output: activity label
1 Fisher vector per activity
testing
Turning right
Turning left
Closing trunk
Closing
U Turn
Loading
Opening trunk
Unloading
Entering
Closing
Exiting
Opening

Results:

min max	min	max
mean $P_{miss}@Rate_{FA=1AD}$	0.4536572	0.99807
mean $P_{miss}@Rate_{FA=1AOD}$	0.5576526	0.9994005



Conclusions:

- Combination of DCNNs with traditional approaches (HOG-HOF and SVM classification).
- Accuracy above 99% in training, indicates that the contestants activity features are linearly separable.
- The performance of the object detection module is vital, which in our case was not secured. For this, many activities weren't evaluated.
- An improved object detection model, as well as, an activity filter for irrelevant activities, will improve the performance.

Instance Search

VERGE video search engine modules:

High Level Visual Concept Retrieval:

- 346 TRECVID concepts using pre-trained DCNNs & Training with Linear SVMs.
- GoogLeNet CNN network for landscape recognition using the Places-205 scene categories.

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 faces using Tiny face detection algorithm.
- Face feature extraction: extract CNN descriptors using VGG-Very-Deep-16 CNN architecture & use of last FC layer as feature vector.
- Construction of an IVFADC index for fast face retrieval.

Scene Similarity Search Module:

- Use the feature vector of the last fully connected layer and the output of the softmax layer of the VGG16 DCNN pre-trained on Places-365 dataset.

Results:

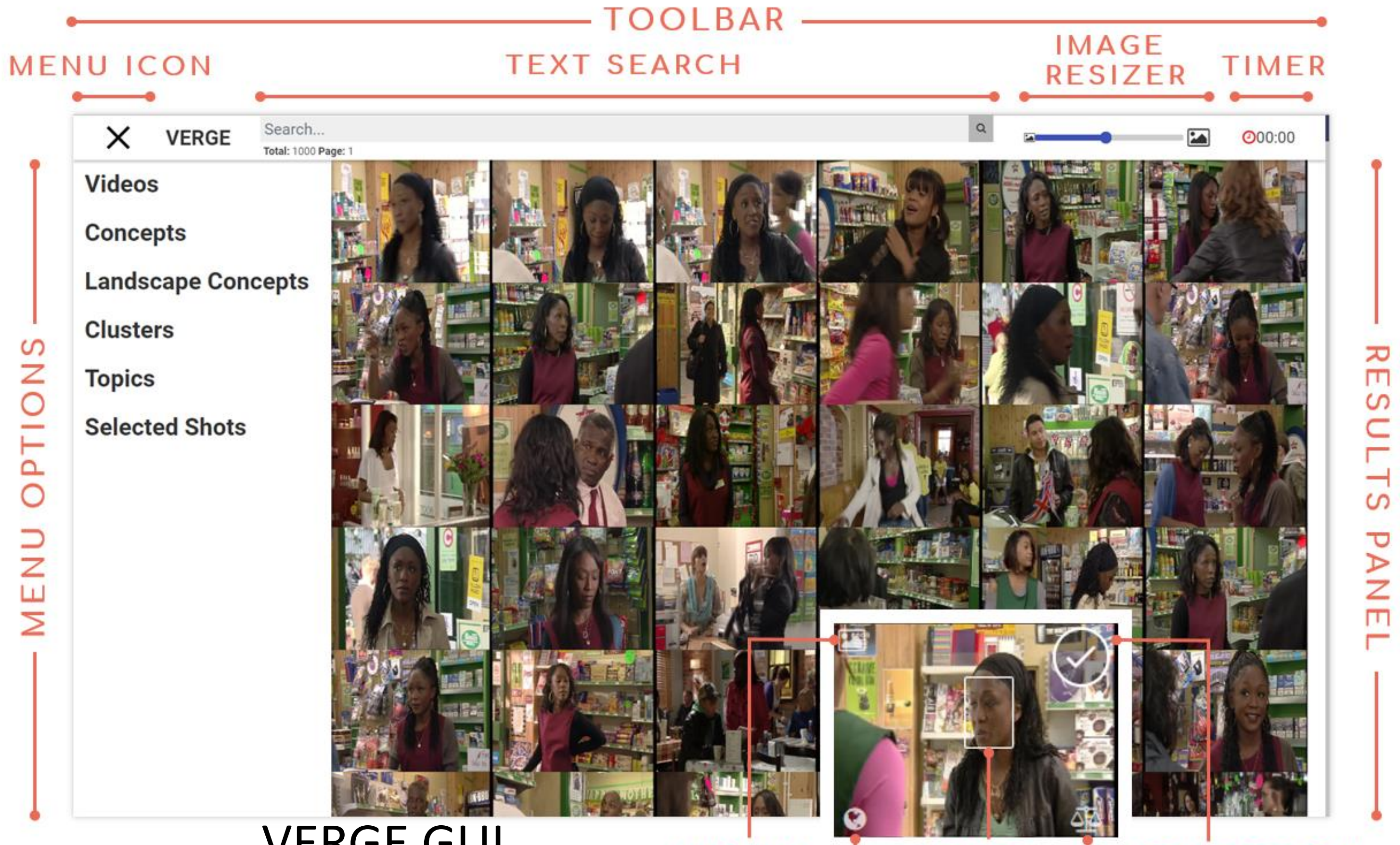
Run ID	MAP	Recall
Run 1	0.114	478/11717

Conclusion:

- High level features are extracted and combined.

VERGE graphical user interface:

- Friendly and efficient navigation.
- Various retrieval utilities, displayed in a sliding menu or as buttons.
- Shot-based representation of results.



VERGE GUI

http://mklab-services.iti.gr/verge/trec2018

Supported by: MOVING ROBORDER beAWARE