Query and Keyframe Representations for Ad-hoc Video Search

Foteini Markatopoulou\textsuperscript{1,2}, Damianos Galanopoulos\textsuperscript{3}, Vasilieos Mezaris\textsuperscript{1}, Ioannis Patras\textsuperscript{2}

\textsuperscript{1}Information Technologies Institute (ITI), CERTH, Thessaloniki, Greece \textsuperscript{2}Queen Mary University of London, London, UK

**Problem and motivation**

- **Ad-hoc video search**: retrieving, from a large video collection, video fragments (keyframes) that are related to a given query
- **Typical solution**: treat the query as a set of simple terms
- **Motivation**:
  - Detecting the most useful parts of the query, e.g., subsequences that contain the main content that the user asks for retrieval
  - Combining two different measures for the distance between the video shots and the target query, calculated on concept-based and semantic embedding representations

**Background**

- Each concept is enriched with additional information captured by Google or Wikipedia \cite{20}
- An inverted index structure is used in order to associate the query with the concepts \cite{4}
- A semi-automatic system \cite{21}, where the user is asked to choose keywords given a test query

**Proposed Method**

**Method outline**

- (a) Concept-based keyframe representation: apply a DCNN in every keyframe
- (c) Concept-based query representation: translate the query in a set of related concepts using NLP
- (b) Semantic embeddings for concept-based query and keyframe representations: project both into a given semantic embedding space

**Proposed solution**

- Two different distances are combined in terms of arithmetic mean

**Experimental results**

- **Datasets**: TRECVID AVS 2016, TRECVID Video Search 2008
  - Test set: 600 and 100 hours, respectively
  - Evaluated queries: 30 and 48, respectively

**Evaluation measure**: MXInfAP (%)

| Steps | All | Excluding one step |
|-------|-----|--------------------|
|       | step 1 | step 2 | step 3 | step 4 | step 5 |
| (a) Concept-based representation | 5.94 | 5.92 | 5.74 | 3.96 | 5.95 | 4.53 |
| (b) Semantic embeddings | 3.77 | 3.86 | 2.98 | 3.22 | 3.75 | 2.80 |
| (c) Combination | 6.35 | 6.51 | 5.77 | 4.37 | 6.27 | 4.99 |

- **Semantic embeddings**: pre-trained Google News Corpus word2vec model
- **Keyframe representation**: 1346 concepts
- 1000 Imagenet concepts extracted using 5 pre-trained ImageNet DCNNs; fused in terms of arithmetic mean
- 346 TRECVID SIN concepts extracted using 2 fine-tuned DCNNs, again fused

**Comparisons**

| Methods | AVS16 | V508 |
|---------|-------|------|
| (a) Literature methods |
| Tzelepis et al. \cite{20} | 4.16 | 8.27 |
| Ueki et al. \cite{21} | 5.65 | 7.24 |
| Norouzi et al. \cite{15} | 3.14 | 7.30 |
| (b) Top-4 TRECVID finalists |
| Top-1 Le et al. \cite{4} | 5.4 | 6.7 |
| Top-2 Markat. et al. \cite{13} | 5.1 | 5.4 |
| Top-3 Liang et al. \cite{6} | 4.0 | 4.2 |
| Top-4 Zhangy et al. \cite{23} | 3.8 | 4.1 |
| Proposed | 6.35 | 9.11 |

Supported by: MOVING InVID