Audio-based Near-Duplicate Video Retrieval with Audio Similarity Learning

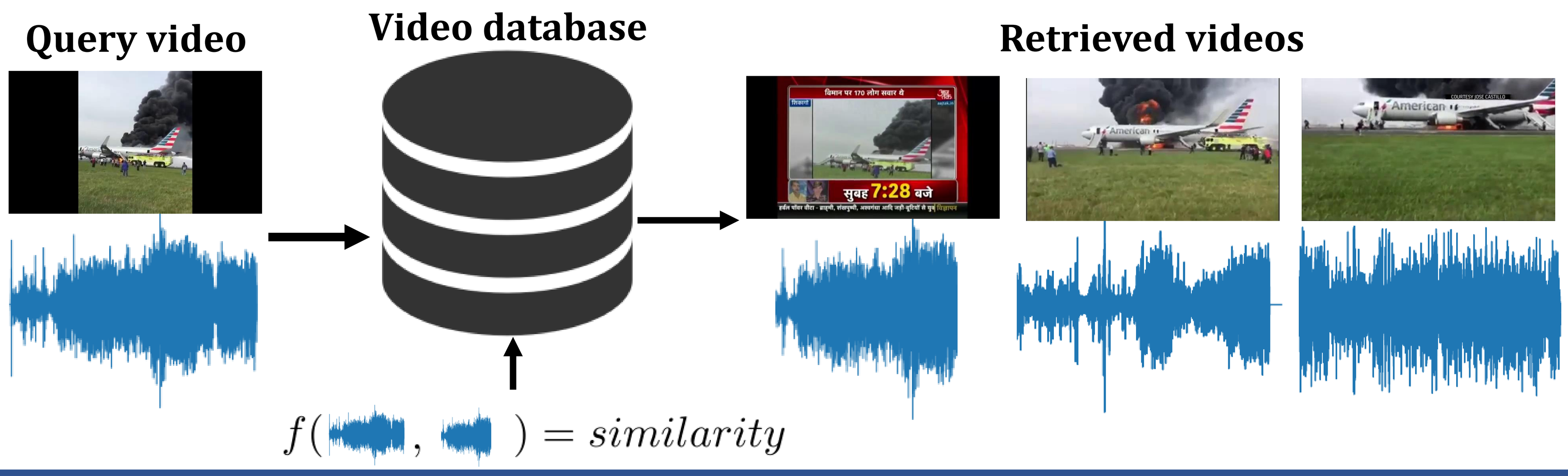
Pavlos Avgoustinakis[1], Giorgos Kordopatis-Zilos[1], Symeon Papadopoulos[1], Andreas L. Symeonidis[2], Ioannis Kompatsiaris[1]

[1]Information Technologies Institute, CERTH [2]Aristotle University of Thessaloniki

Problem statement

Duplicate Audio Video Retrieval (DAVR)

- Given a video query, search a video database and retrieve videos that share the same audio content



Proposed approach

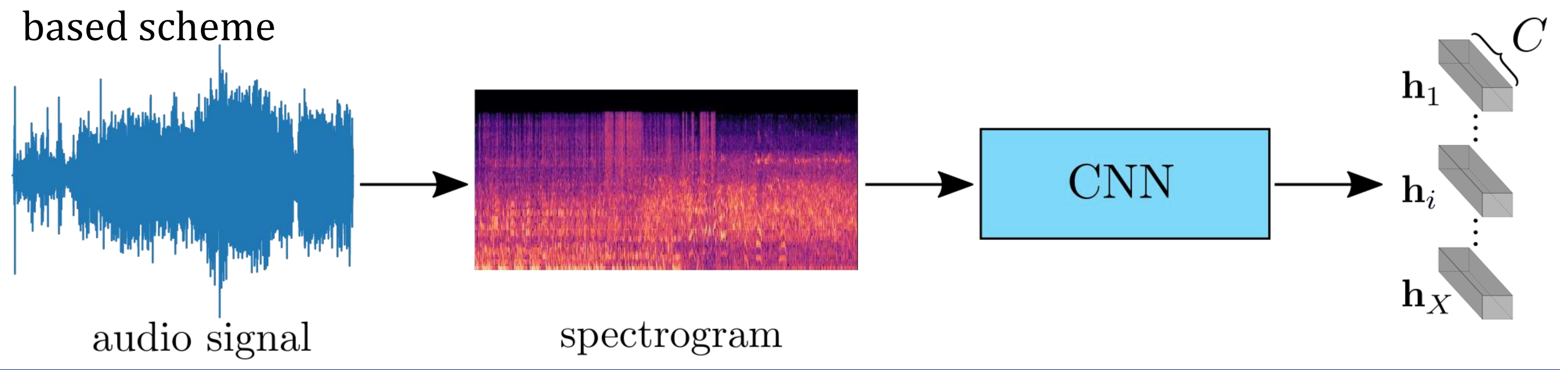
- AuSiL: Audio Similarity Learning network
  - Extract descriptors from the intermediate layers of a pre-trained CNN fed with Mel-spectrograms of videos' audio channels.
  - Generate a similarity matrix containing the pair-wise segment similarities between two compared videos.
  - Robustly capture temporal similarity patterns between videos through a CNN network

Proposed approach

Feature extraction

Feature vector composition

- Extract Mel-spectrograms from audio signal
  - Divide spectrograms to overlapping time segments
- Employ a pre-trained CNN [1]
  - Apply Maximum Activation of Convolutions (MAC) on intermediate CNN layers
  - Concatenate the K extracted feature vectors.
- Refine vector representations
  - Perform PCA whitening for feature decorrelation.
  - Weigh audio segments based on their captured information through an attention-based scheme

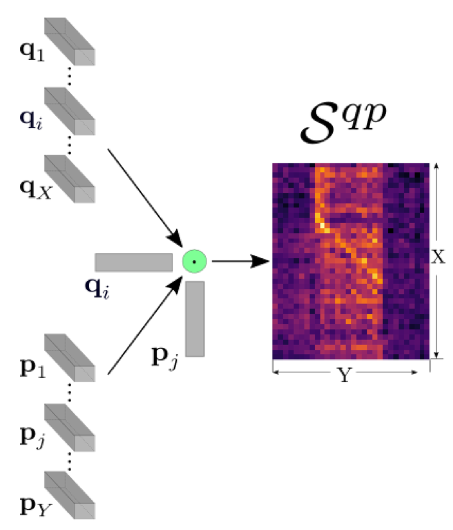


Similarity calculation

Similarity Matrix

- Contains the similarity scores between the audio feature vectors of two compared videos.
  - Let $Q \in \mathbb{R}^{X \times C}$ and $P \in \mathbb{R}^{Y \times C}$ the video tensors of two videos q, p with X and Y audio segments
  - Dot product between pairs of feature vectors.

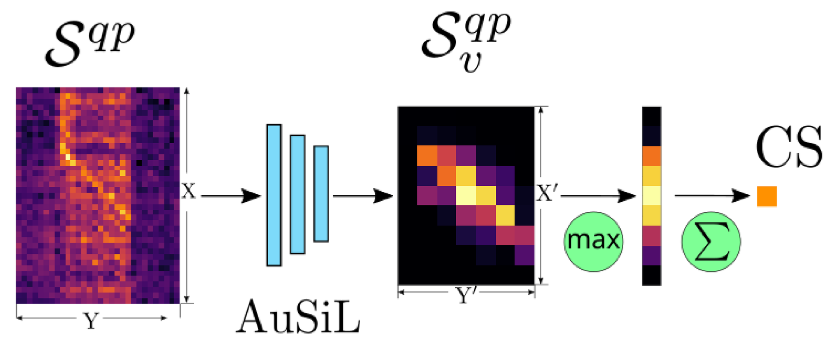
$$S^{qp} = Q \cdot P^T$$



Similarity Calculation

- Feed similarity matrix to AuSiL network [2]
  - Similarity learning four-layer CNN
  - Captures the temporal structures in the similarity matrix
- Chamfer Similarity on the AuSiL output

$$\mathrm{CS}(q,p) = \frac{1}{X'} \sum_{i=1}^{X'} \max_{j \in [1,Y']} \mathrm{Htanh}(\mathcal{S}_v^{qp}(i,j))$$



Training AuSiL

Triplet Loss

- Force network to assign higher similarity scores to relevant video pairs and lower to irrelevant ones.
- Given an anchor, a positive and a negative video (v, v^+, v^-) and a margin parameter γ

$$\mathcal{L}_{tr} = \max\{0, \mathrm{CS}(v, v^-) - \mathrm{CS}(v, v^+) + \gamma\}$$

Similarity Regularization Loss

- Penalizes high values in the input of hard tanh that would lead to saturated outputs

$$\mathcal{L}_{reg} = \sum_{i=1}^{X'} \sum_{j=1}^{Y'} \left|\max\{0, \mathcal{S}_v^{qp} - 1\}\right| + \left|\min\{0, \mathcal{S}_v^{qp} + 1\}\right|$$

Video triplets

- Organize training dataset on triplets of video.
- Generate pairs of video with related audio content.
  - Due to lack of audio level annotation, generate pairs with related visual content
  - Select only the ones that are close in the feature space

$$D(v, v^+) < 0.175$$

- Select hard negatives.
  - Negative videos with anchor-negative distance lower than anchor-positive distance plus a margin value

$$D(v, v^-) < D(v, v^+) + d$$

$D(\cdot,\cdot)$: Euclidean distance of videos' global averaged feature vectors

Evaluation setup

Training dataset

VCDB

- 528 base videos with 9,236 copied segments.
- 100,000 distractor videos.
- 5.8 million generated video triplets.

Evaluation datasets

Datasets with audio annotation

FIVR-200K$_\alpha$

- Derived from visual annotations of FIVR-200K
- Manual annotation for DAVR problem
- 76 video queries and 3,392 audio duplicate pairs
- FIVR-5K$_\alpha$: a subset of FIVR-200K$_\alpha$ with 50 randomly selected queries intended for quick comparisons

SVD$_\alpha$

- Derived from visual annotations of SVD
- Manual annotation for DAVR problem
- 167 queries and 1,492 audio duplicate pairs

Datasets with visual annotation

FIVR-200K

- Fine-grained incident video retrieval
- 225,960 videos, 100 queries and three retrieval tasks

SVD

- Near-duplicate video retrieval
- 1,206 video queries, 34,020 labeled and 526,787 unlabeled videos

EVVE

- Event-based video retrieval
- 620 queries and 2,375 database videos

Experiments

Ablation Study

- Timestep impact

Time step	FIVR-5K$_\alpha$	SVD$_\alpha$
1000	**0.794**	0.903
500	0.789	0.915
250	0.787	0.928
125	0.790	**0.940**

- Contribution of each component
  MAC: Extracted CNN features
  W: PCA whitening
  A: Attention layer
  AuSiL: Proposed approach

Network Components	FIVR-5K$_\alpha$	SVD$_\alpha$
MAC	0.656	0.891
MAC + W	0.740	0.932
MAC + W + A	0.742	0.934
AuSiL	**0.794**	**0.940**

- Transfer learning settings

Transfer	Update	FIVR-5K$_\alpha$	SVD$_\alpha$
✓	×	**0.794**	**0.940**
✓	✓	0.588	0.857
×	✓	0.445	0.764

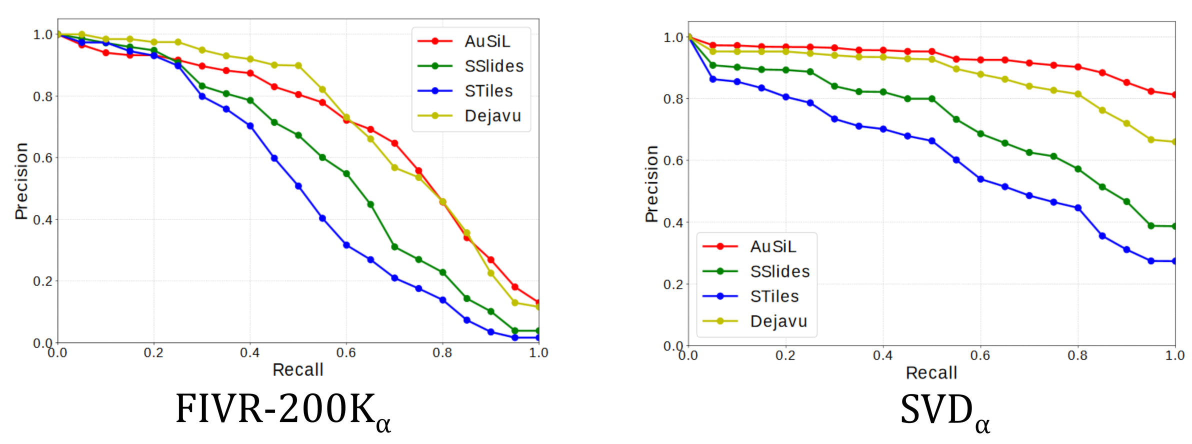
- Impact of margin hyperparameter γ

gamma (γ)	0.4	0.6	0.8	1.0	1.2
FIVR-5K$_\alpha$	0.764	0.761	0.786	**0.794**	0.767
SVD$_\alpha$	0.903	0.895	0.937	**0.940**	0.919

Comparison against state-of-the-art

- Duplicate audio video retrieval (DAVR)

Method	FIVR-200K$_\alpha$	SVD$_\alpha$
Dejavu [17]	**0.726**	0.874
Spectro Slides [14]	0.588	0.716
Spectro Tiles [16]	0.510	0.605
AuSiL (ours)	0.701	**0.940**



- Evaluation on audio speed transformations
  - Generate duplicates by altering the speed of audio signals
  - Exclude original audio duplicates from the dataset

Method	FIVR-200K$_\alpha^T$	SVD$_\alpha^T$
Dejavu [17]	0.443	0.741
AuSiL (ours)	**0.865**	**0.923**

- Visual-based video retrieval tasks

Method	FIVR-200K			SVD	EVVE
	DSVR	CSVR	ISVR		
Dejavu [17]	**0.352**	**0.324**	0.230	0.477	0.160
Spectro Slides [14]	0.288	0.269	0.189	0.406	0.146
Spectro Tiles [16]	0.249	0.228	0.159	0.323	0.144
AuSiL (ours)	0.327	0.310	**0.232**	**0.516**	**0.288**
Best visual	0.892	0.841	0.702	0.785	0.631

References

[1] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in ICASSP, 2018.

[2] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I.Kompatsiaris,"ViSiL: Fine-grained spatio-temporal video similarity learning," in ICCV, 2019

https://github.com/mever-team/ausil

FIVR http://ndd.iti.gr/fivr/