ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

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

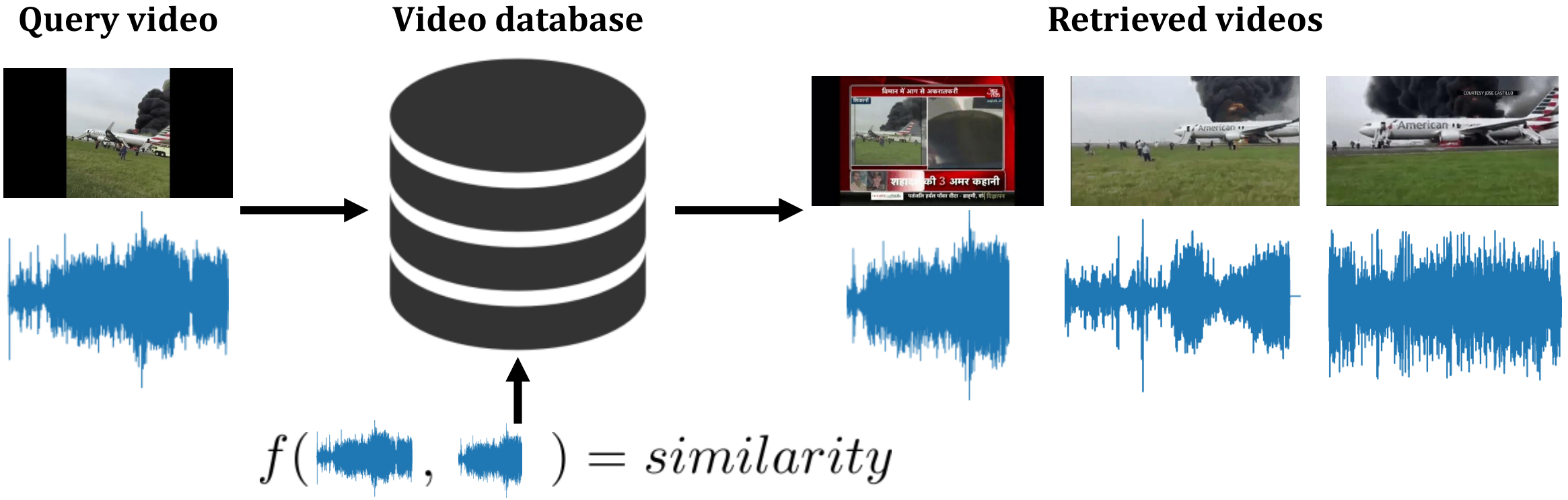
Pavlos Avgoustinakis
Giorgos Kordopatis-Zilos
Symeon Papadopoulos
Andreas L. Symeonidis
Ioannis Kompatsiaris

Information Technologies Institute
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



Motivation

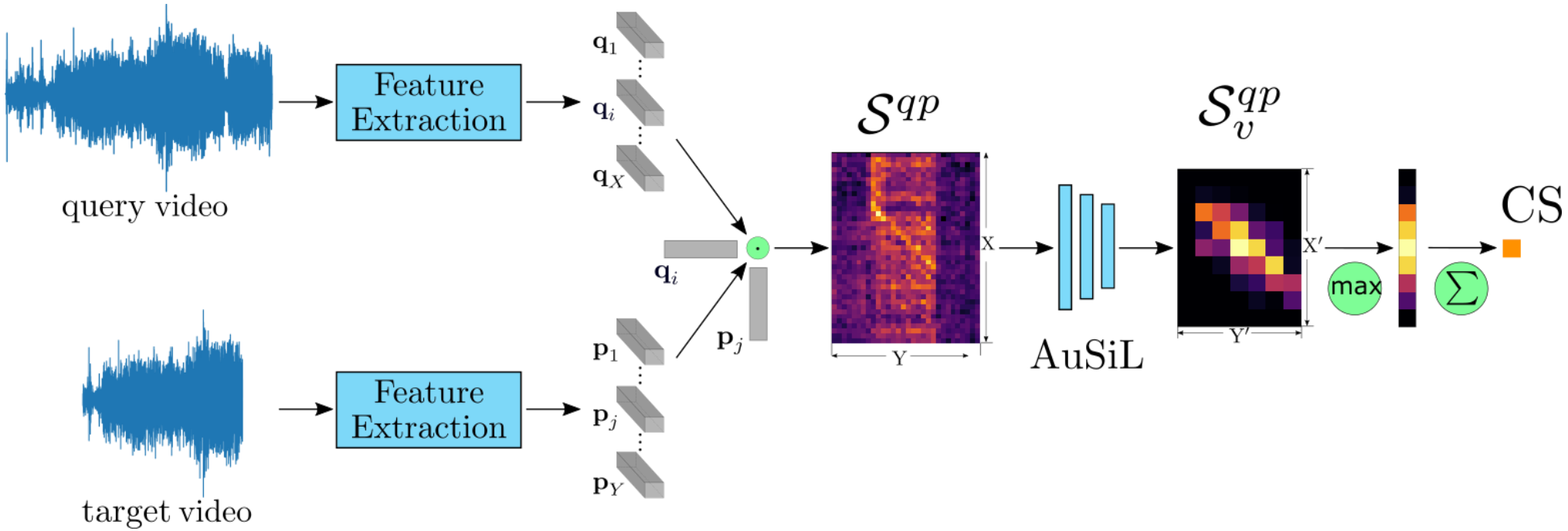
State-of-the-art limitations

- Use of handcrafted approaches for fingerprint extraction
- Rigid aggregation schemes for similarity calculation
- Lack of evaluation benchmarks with user-generated content annotated based on audio duplicity

Our objective

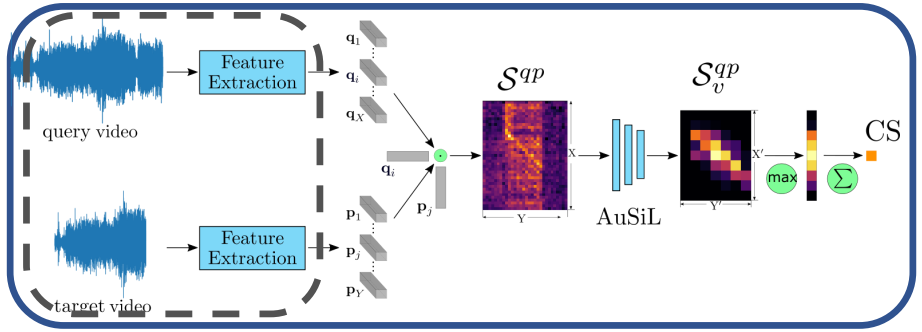
- Leverage of deep learning and transfer learning
- Composition of evaluation benchmarks

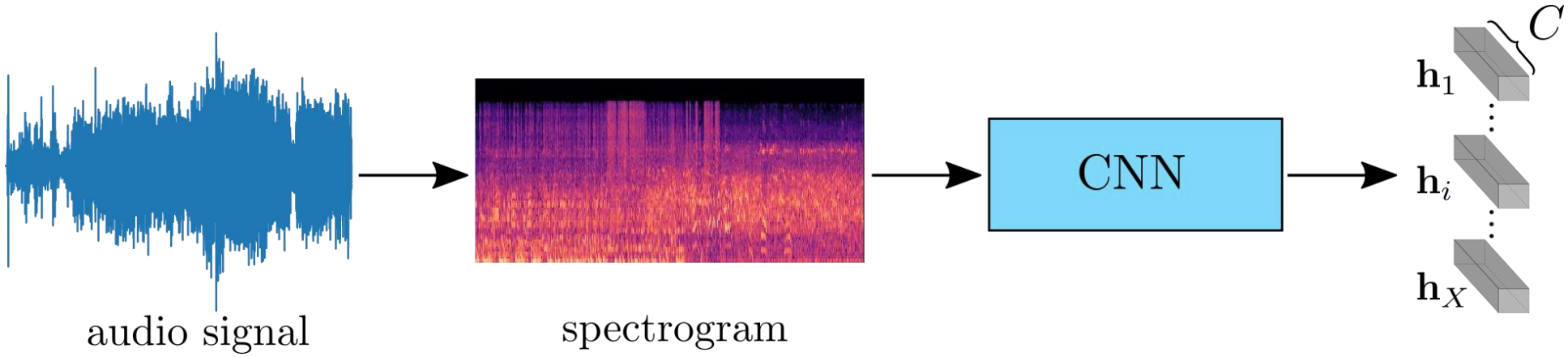
Contribution



- Robust audio-based video similarity calculation
  - Transfer learning from a pre-trained CNN
  - Audio similarity learning
- Two annotated datasets that serve as benchmarks for DAVR

Feature Extraction

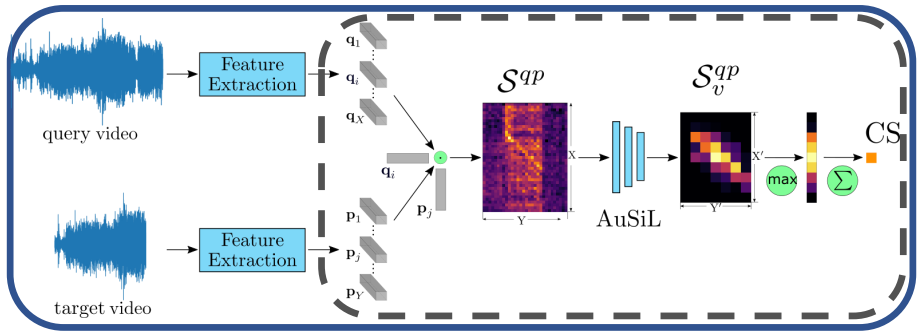


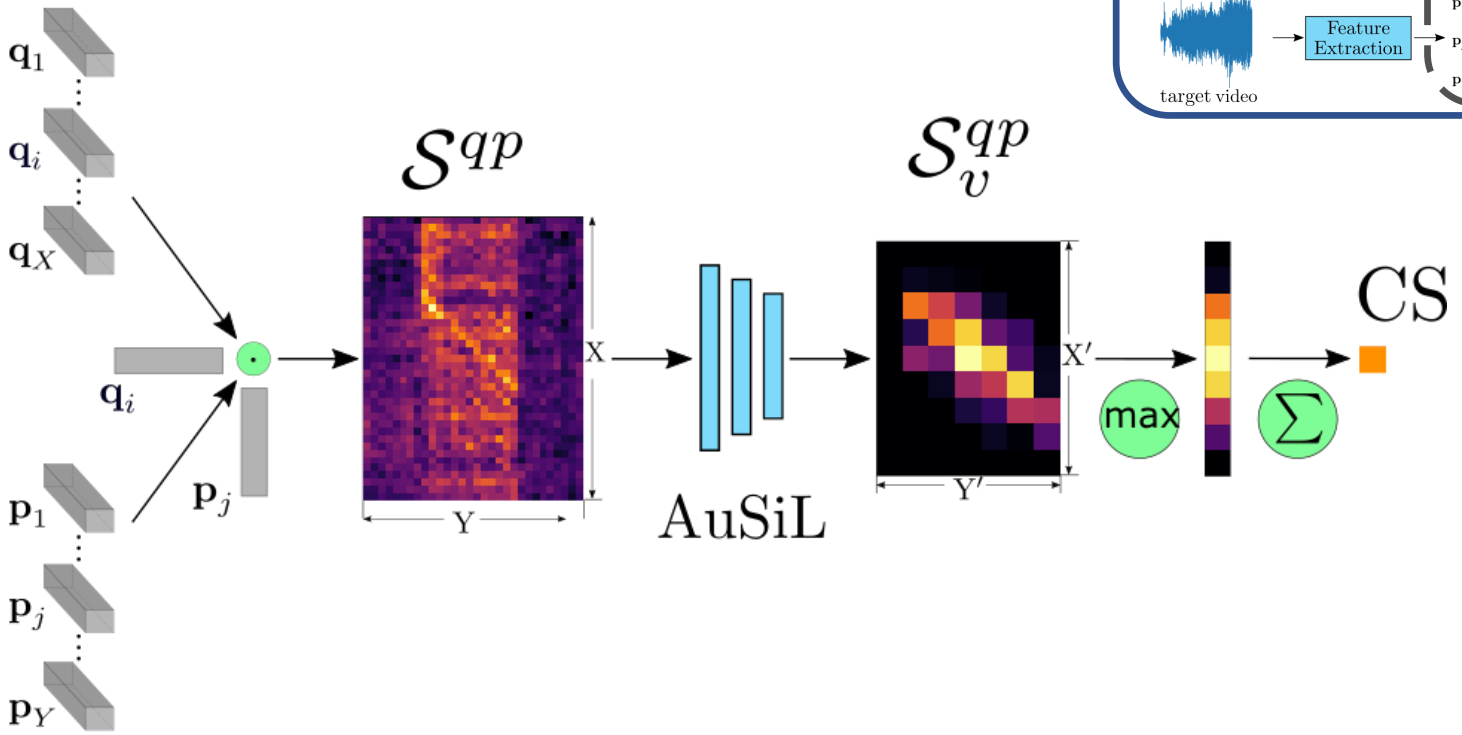


- Generate audio Mel-spectrograms of audio signals
  - Divide into overlapping time segments
- Employ a pre-trained CNN (Kumar et al. 2018)
  - Max Activation of Convolutions (MAC) on intermediate CNN layers
  - Apply PCA whitening and attention-based weighting

A. Kumar, M. Khadkevich, and C. Fugen. “Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes”. ICASSP, 2018.

Similarity calculation





- Generation of the similarity matrix

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

- Audio Similarity Learning network
  - Four-layer CNN
  - Captures the temporal structures

- Chamfer Similarity

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

Evaluation datasets

FIVR-200K$_{\alpha}$

- FIVR-200K (Kordopatis-Zilos et al., 2019) annotated for Fine-grained Incident Video Retrieval
- 76 video queries
- 3,392 audio duplicate pairs

SVD$_{\alpha}$

- SVD (Jiang et al., 2019) annotated for Near-Duplicate Video Retrieval
- 167 video queries
- 1,492 audio duplicate pairs

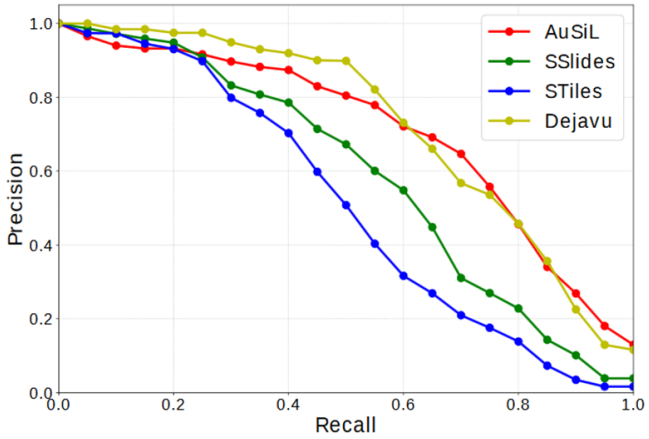
G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I.Kompatsiaris. “Fine-grained Incident Video Retrieval”. IEEE TMM, 2019.
Q. Y. Jiang, Y. He, G. Li, J. Lin, L. Li, and W. J. Li. “SVD: A Large-Scale Short Video Dataset for Near Duplicate Video Retrieval”. ICCV, 2019.

Experimental results

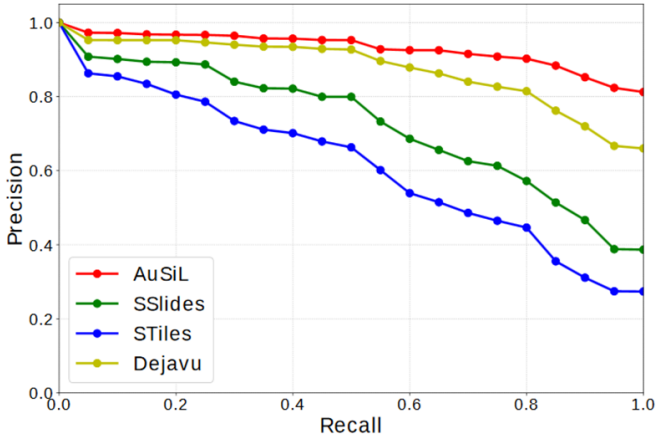
- Duplicate Audio Video Retrieval (DAVR)

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**

FIVR-200K$_\alpha$



SVD$_\alpha$



- Audio speed transformations

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

Experimental results

- Visual-based video retrieval

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

Thank you!

Get in touch:

Giorgos Kordopatis-Zilos: georgekordopatis@iti.gr / @g_kordo

MeVer team:

https://mever.iti.gr/web/ / @meverteam

Code available in:

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

With the support of: