Supplementary material of ViSiL: Fine-grained Spatio-Temporal Video Similarity Learning

Giorgos Kordopatis-Zilos[1,2], Symeon Papadopoulos[1], Ioannis Patras[2], Ioannis Kompatsiaris[1]

[1]Information Technologies Institute, CERTH, Thessaloniki, Greece
[2]Queen Mary University of London, Mile End road, E1 4NS London, UK

A. Additional Results

A.1. Different similarity calculation functions

In this section, we compare the impact of different functions, other than CS, on the frame-to-frame (F2F) and video-to-video (V2V) similarity calculation. In general, CS can be considered to be equivalent to a Max-Pooling (MP) function followed by Average-Pooling (AP). A different combination could be the application of two AP functions. Table 1 illustrates the results for different combinations of the core similarity functions of the proposed system on FIVR-5K. It is evident that the use of two AP functions for V2V does not work at all. The run with the two AP for F2F and CS for V2V achieves competitive mAP, but still lower than the run with CS in both functions as proposed.

F2F	V2V	DSVR	CSVR	ISVR
MP-AP	MP-AP	0.880	0.869	0.777
AP-AP	MP-AP	0.769	0.748	0.682
MP-AP	AP-AP	0.640	0.652	0.623
AP-AP	AP-AP	0.439	0.436	0.341

Table 1. mAP comparison of four pooling combinations for frame-to-frame and video-to-video similarity calculation on FIVR-5K. **MP** stands for Max-Pooling and **AP** for Average-Pooling.

A.2. Impact of hyperparameter values

In this section, we compare the impact of different values of hyperparameter γ, r and W, on the performance of the proposed system. As default values, we use the values reported in the original paper, i.e. $\gamma = 0.5$, $r = 0.1$ and $W = 64$, and change one at a time.

We first assess the impact of the margin parameter γ on the retrieval performance of the proposed approach. Figure 1(a) illustrates the performance of the method trained with different margins on the three tasks of FIVR-5K. Regarding the DSVR task, one may notice that that the performance of the model improves as the margin parameter increases. However, this is not the case for the ISVR task. The approach reports high performance (mAP greater than 0.775) for small values of γ, i.e. within range [0.25, 0.5], but performance drops as γ increases.

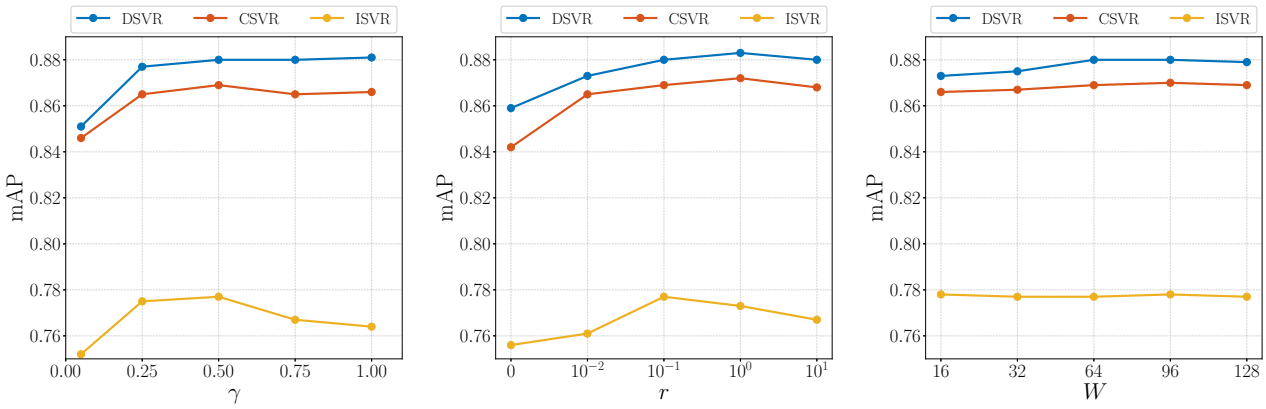
Additionally, we assess the impact of the regularization parameter r on the retrieval performance of the proposed approach. Figure 1(b) illustrates the performance of the method trained with different regularization parameters on the three tasks of FIVR-5K. On DSVR and CSVR tasks, the proposed approach achieves the best results for $r = 1.0$ with considerable margin from the second best, approximately 0.003 mAP. However, on the ISVR task, the performance significantly dropped in comparison to the default value ($r = 0.1$). For values lower than the default, the proposed approach does not report competitive results on any evaluation task.

Finally, we assess the impact of the size of video snippet W on the retrieval performance of the proposed approach. Figure 1(c) depicts the mAP of the method with different values of W on the three tasks of FIVR-5K dataset. Regarding the DSVR and CSVR tasks, it is evident that the larger the size of video snippets W the better the performance of the proposed methods. The run with $W = 96$ yields the best results on both tasks with 0.880 and 0.870 mAP, respectively. However, the system's performance on the ISVR task is independent of the size of video snippets used for training, since all runs report approximately the same mAP.

A.3. Computational complexity

In this section, we compare the computational complexity of different setups of the proposed approach. The proposed method can be split in two distinct processes, an offline and an online. The offline process comprises the feature extraction from video frames, whereas the online one the similarity calculation between two videos.

In Table 2, we compare the MAC and iMAC runs (cf. Table 2 of the paper) with the ViSiL$_f$ and ViSiL$_v$ in terms of execution time and performance. In that way, we assess the trade-off between the performance gain from the introduc-



(a) (b) (c)

Figure 1. Impact of the margin hyperparameter γ, the regularization parameter r and video snippet size W on the performance of the proposed method on FIVR-5K.

Run	Comp. Time		FIVR-5K		
	Offline	Online	DSVR	CSVR	ISVR
MAC	0.95s	2.0ms	0.747	0.730	0.684
iMAC	0.95s	2.3ms	0.755	0.749	0.689
ViSiL$_f$	0.96s	6.0ms	0.838	0.832	0.739
ViSiL$_v$	1.04s	9.5ms	0.880	0.869	0.777

Table 2. mAP and execution time comparison of four versions of the proposed approach on FIVR-5K. The execution time of the offline process refers to the average feature extraction time per video. The execution time of the online process refers to the average time for the calculation of video similarity of video pairs.

tion of each component of the method, and the associated computational cost. The average length of videos in FIVR-5K is 103 seconds. All the experiments were executed on a machine with an Intel i7-4770K CPU and a GTX1070 GPU.

For the offline process, all runs need approximately the same time to extract frame features. The use of intermediate convolutional layer does not slow down the feature extraction process, since both MAC and iMAC needs 950 ms for feature extraction. The extraction of regional vectors (ViSiL$_f$) has minor impact on the speed, approximately 1% increase of the total extraction time. Also, the application of whitening and attention-based weighting does not significantly increases the extraction time; ViSiL$_v$ needs 80 ms more than ViSiL$_f$ per video.

Regarding the online process, the complexity of calculating the frame-to-frame similarity matrix between videos of M frames each, is $O(M^2N^2)$, where N is the number of regions per frame. This is to be compared to $O(M^2)$ of frame-to-frame methods such as iMAC (where $N = 1$). Based on our experiments, the MAC and iMAC runs need less than 2.5 ms to calculate video similarity. The computation of the proposed frame-to-frame similarity matrix increases the execution time by 3.7 ms, which is more than a 150% increase (comparing iMAC and ViSiL$_f$). Finally, in ViSiL$_v$, the second-stage CNN on the frame-to-frame similarity matrix takes 40% of the execution time, and further increasing it approximately by 3.5 ms but for a significant performance gain.

B. Visual Examples

This section presents some visual examples of the outputs of the system components.

Figure 2 illustrates three visual examples of video frames coloured based on the attention weights of their regions vectors. Apparently, the proposed attention mechanism weights the frame regions independently based on their saliency. It assigns high weight values on the information-rich regions (e.g. the concert stage, the Mandalay Bay building); whereas, it assigns low values on regions that contain no meaningful object (e.g. solid dark regions).

Additionally, Figure 3 illustrates examples of the input frame-to-frame similarity matrix, the network output and the calculated video similarity of two compared videos for three video categories. The network is able to extract temporal patterns from the input frame-to-frame similarity matrices (e.g. strong diagonals, consistent parts with high similarity) and suppress the noisy (i.e. small inconsistent parts with high similarity values), in order to calculate the final video-to-video similarity precisely. Also, sampled frames from the compared videos are depicted for the better understanding of the different video relation types.

Figure 2. Examples of the attention weighting on arbitrary video frames: sampled video frames from the same video (top), attention maps of the corresponding frames (bottom). Red colour indicates high attention weights, whereas blue indicates low ones.

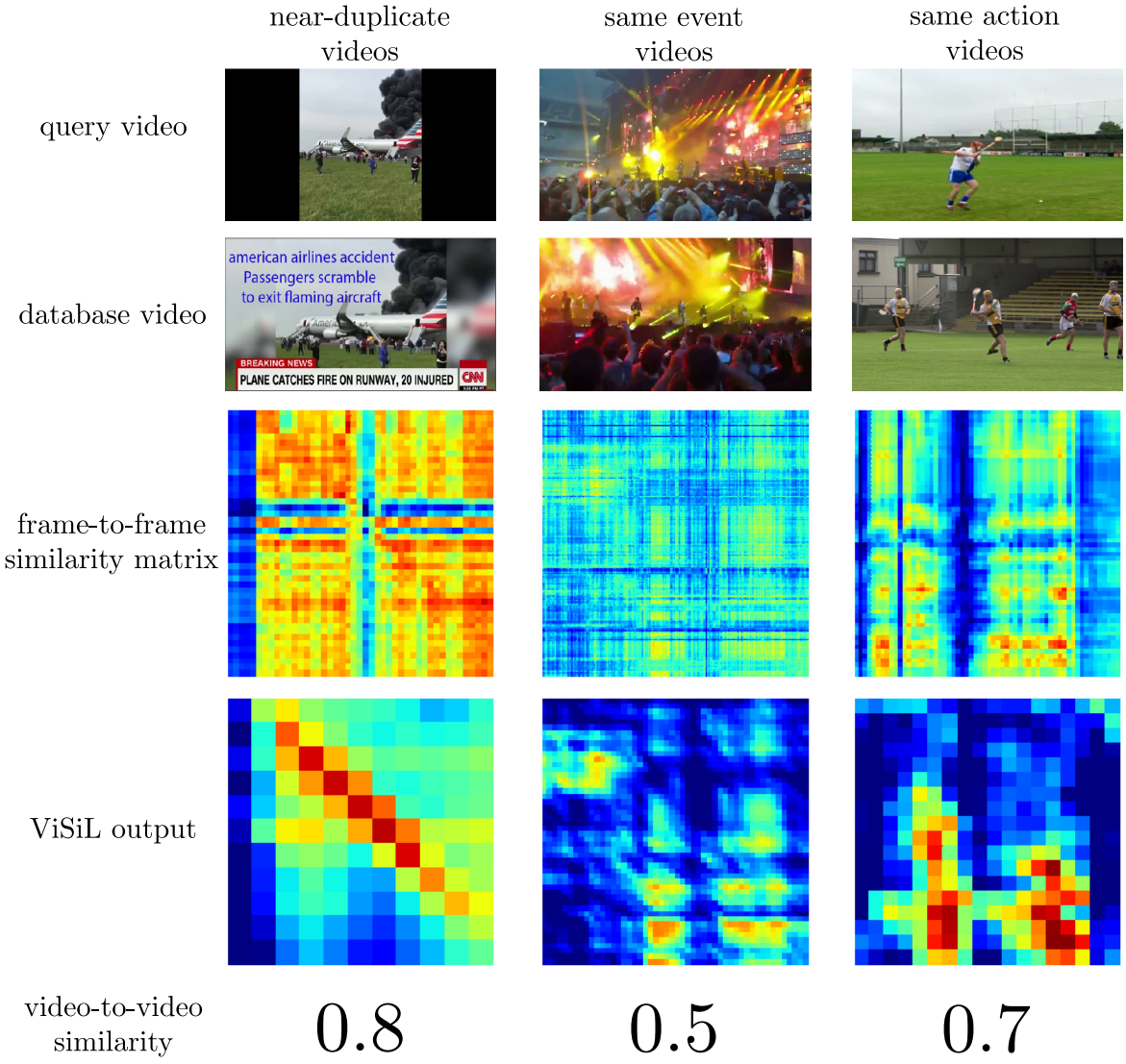


Figure 3. Visual examples of the input and output of ViSiL for three different video relation types. Two sampled frames of the compared videos are depicted on top, then the input frame-to-frame similarity matrix and the ViSiL output are displayed, and the final video-to-video similarity is reported. In the similarity matrices, red colour indicates a high similarity score, whereas blue indicates low similarity.