[ACM MM 2021] Video Similarity and Alignment Learning on Partial Video Copy Detection

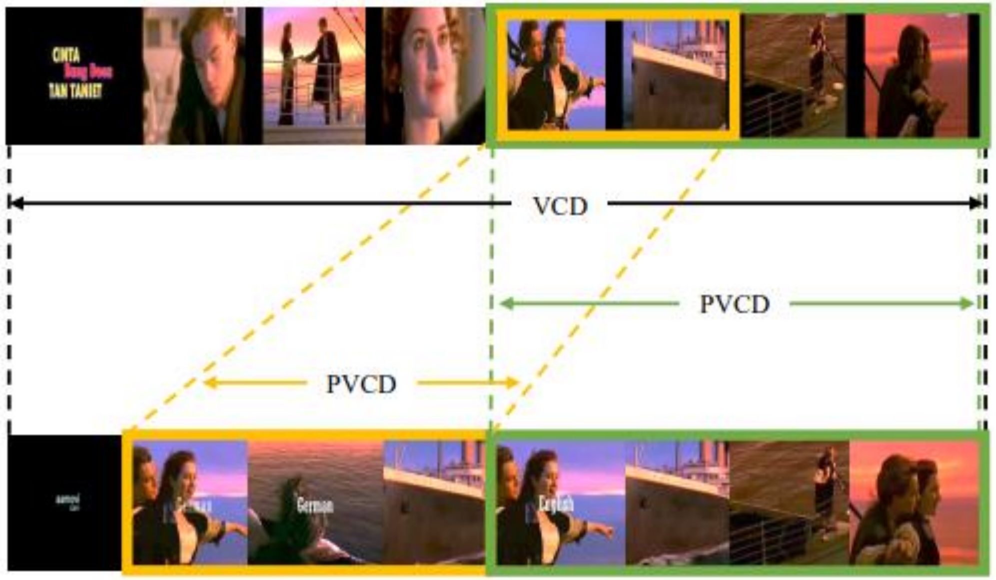
Alibaba

2021.09.13 Won Jo

Overview

VCD (Video-level Copy Detection) vs PVCD (Partial Video Copy Detection)

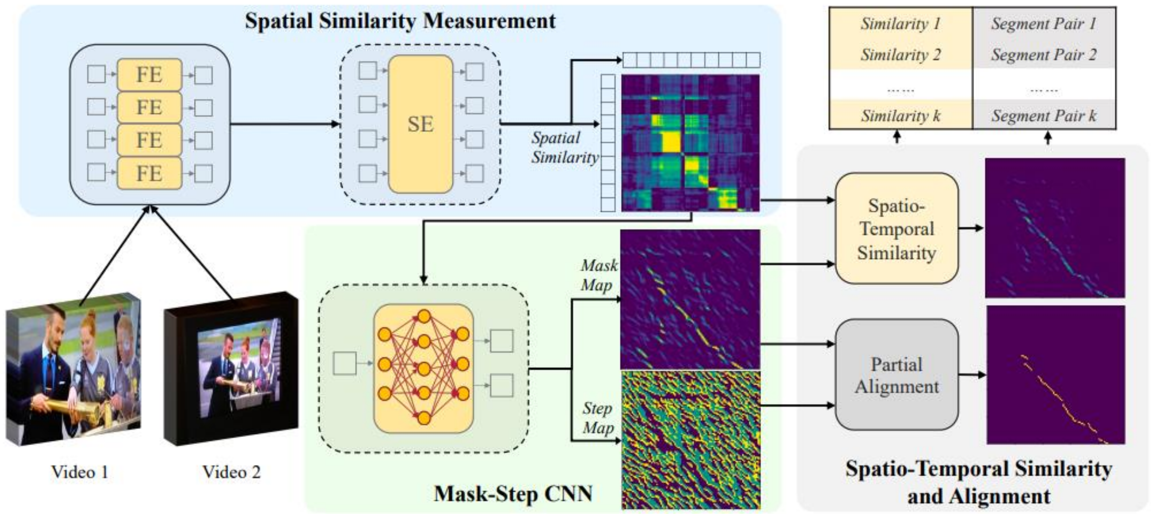
While VCD looks for a copy video in video-level,
PVCD finds the copy video in the video-level and localizes the copied part in that video.



Overview

VSAL (Video Similarity and Alignment Learning)

A model that learns
Spatial Similarity + Temporal Similarity + Patient Similarity at once for PVCD purposes



Table

1. Problem Formulation
2. Spatial Similarity Measurement
3. Mask-Step CNN
4. Spatio-Temporal Similarity and Alignment

VSAL: Problem Formulation

When two videos u and v with length M and N are inputs, the video similarity consists of three similarities

Sim = F(S, T, P)

(S-Spatial similarity, T-Temporal similarity, P-Partial alignment)

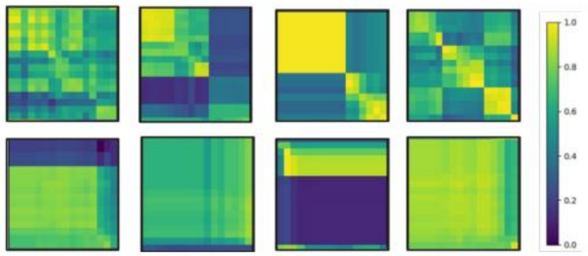
$$Sim_k = \frac{\alpha_k}{|P_k|} \sum_{i,j \in P_k} s_{i,j} t_{i,j}$$

Spatial Similarity Matrix (S)

Feature extracted for each frame is used, and it means a similarity map between frame-level features.

Well-matched

Unmatched



Partial Alignment (S->P)

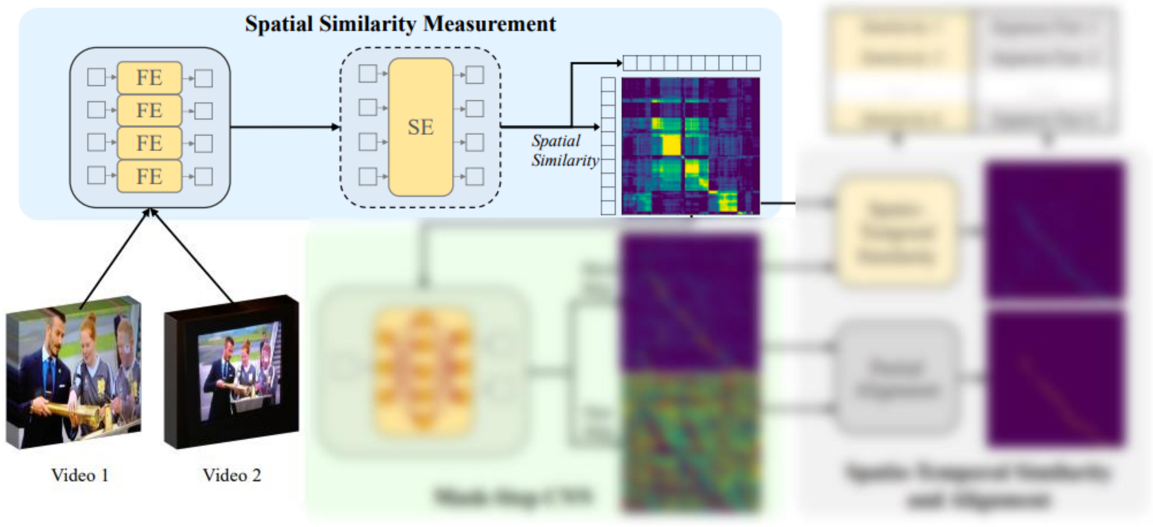
As in the well-matched case of S, the **partial alignment appears as a diagonal path.**
K-th alignment is represented by P_k.
And in this paper, unlike HW (Hard Weight) that drops when len(P_k) is less than a certain threshold, **the following SW (Soft Weight) is used.**

$$\alpha_k = \frac{1}{1 + \gamma e^{-\|P_k\|}}$$

Temporal Similarity Matrix (S->T)

Estimation from S to take advantage of the distinction between Well-matched and Unmatched cases in S.

VSAL: Spatial Similarity Measurement

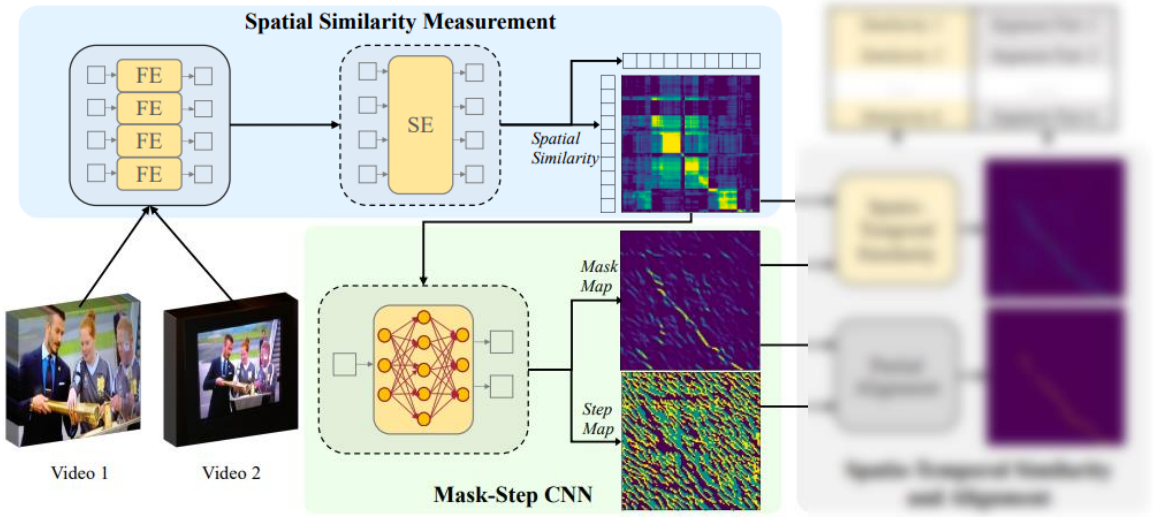


Spatial Similarity Measurement

- It consists of FE (Feature Encoder) and SE (Sequence Encoder)
- After frame feature is created with FE, SE is passed for interaction between individual spatial information
- FE is the frame feature encoder SVRTN_f, SE is the self-attention layer of Transformer
- When the frame feature of video u and v is U and V, create a Spatial Similarity Map S in the following equation, f_θ is L2 norm

$$S = f_\theta(U) f_\theta(V)^{\mathsf{T}}$$

VSAL: Mask-Step CNN

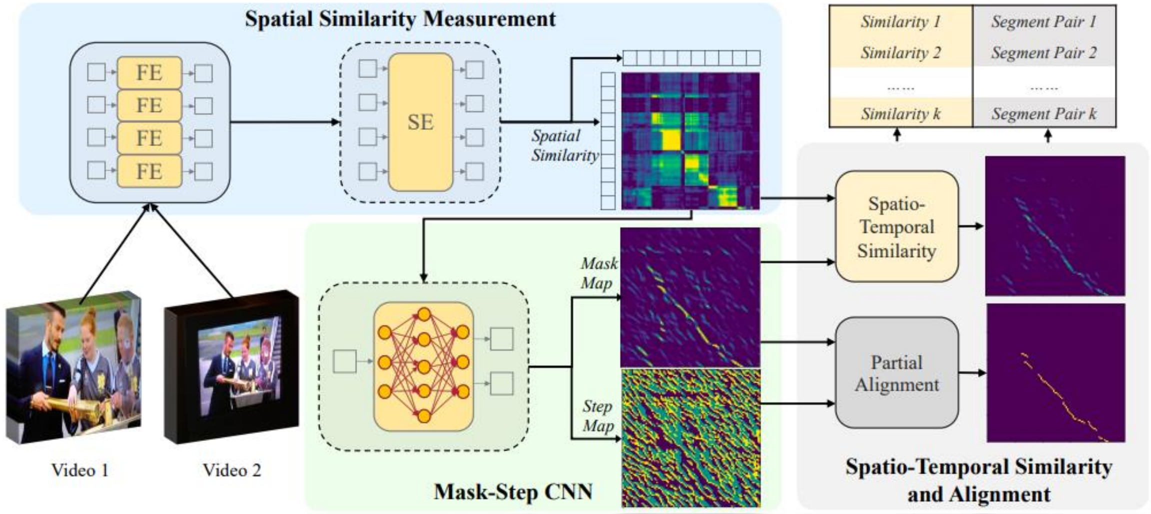


Layer	Kernel size/ Padding	Output size	Activation
Conv-1	$3 \times 3/1$	$M \times N \times 8$	ReLU
Conv-2	$3 \times 3/1$	$M \times N \times 16$	ReLU
Conv-3	$3 \times 3/1$	$M \times N \times 32$	ReLU
Mask	$3 \times 3/1$	$M \times N \times 2$	Softmax
Step	$2 \times 2/0$	$(M-1) \times (N-1) \times 3$	Softmax

Mask-Step CNN

- It is divided into a mask branch and a step branch with S as input
- Mask Branch -> MM(Mask Map)
  - The output value of the mask branch is two channels and serves to **determine the probability of the partial alignment**
  - The output value of the mask branch can also represent how well S aligned along their temporal direction, so that is a representation of the temporal similarity T
- Step Branch -> SM(Step Map)
  - It is convolution layer for step predictor **making a direction classification on each position**, and the categories indicate directions to step next from current position to continue alignment path
  - Categories: "stepping right-down", "stepping right" and "stepping down"
- Learn by applying spatial & temporal transform to one video in a self-supervised manner
  - Mask Loss = BCE, Step Loss = CE

VSAL: Spatio-Temporal Similarity and Alignment



Spatio-Temporal Similarity and Alignment

- Partial Alignment
  - Referring to S and T by element, the below pseudo-algorithm is used to a partial alignment
- Spatio-Temporal Similarity
  - S and T are elementally multiplied by weight for each partial alignment position

Algorithm 1 Partial Alignment

Input: Spatial similarity $S = (s_{i,j}) \in \mathbb{R}^{M\times N}$; Temporal similarity $T = (t_{i,j}) \in \mathbb{R}^{M\times N}$; Step map $D = (d_{i,j}) \in \mathbb{N}^{(M-1)\times(N-1)}$; Threshold to find start points τ; Similarity threshold σ.

Output: Partial alignments P.

```
Φ = {(i, j) : t_{i,j} > τ}; k = 0.
while |Φ| > 0 do
    k = k + 1.
    Set P_k = ∅; g = 0.
    Select (i, j) from Φ with smallest i + j value.
    while i < M and j < N and g < 3 do
        st = s_{i,j} t_{i,j}.
        if st < σ then
            g = g + 1.
        else
            Add (i, j) to P_k.
        end if
        Remove (i, j) and its 8 neighborhoods from Φ.
        (i, j) = C_{d_{i,j}}(i, j).
    end while
end while
return P.
```

VSAL: Experiments - VCDB

Comparison of segment-level performance between VSAL and other state-of-the-art methods on VCDB core dataset

Methods	SP	SR	F_1-score
ATN[11]	0.7050	0.5220	0.5956
CNN[11]	-	-	0.6503
SNN[11]	-	-	0.6317
CNN+SNN[11]	-	-	0.6454
TH+CC+ORB[6]	0.5052	0.9294	0.6546
LAMV[1]	-	-	0.6870
CNN+SC[23] (1fps)	-	-	0.6995
CNN+SC[23] (all frames)	-	-	0.7038
BTA[26]	0.7600	0.7500	0.7549
Q-Learning[7]	0.8829	0.7355	0.8025
FPVCD[24]	-	-	0.8613
VSAL	**0.8971**	**0.8462**	**0.8709**

SP: Segment-level Precision, SR: Segment-level Recall

Ablation studies on VCDB core dataset

Methods	SP	SR	F_1-score
HV (baseline)	0.8513	0.6912	0.7629
HV+SE	0.8607	0.6936	0.7682
HV+SE+SW	0.7686	0.7887	0.7785
SM+SE+SW	0.8447	0.8047	0.8242
SM+SE+SW+MM	**0.8971**	**0.8462**	**0.8709**

VSAL: Experiments - VCDB

FIVR-200K-PVCD

ABOUT RESULTS DOWNLOADS CODES

To evaluating the performances of PVCD systems on more complicated spatial and temporal situations, we add annotation of the segment pairs for DSVR subset of FIVR-200k to construct the new partial video copy detection benchmark, called FIVR-200k-PVCD. Original FIVR-200k is a Fine-grained instance video retrieval dataset consisting of 225,960 videos and 100 queries, including three retrieval tasks namely Duplicate Scene Video Retrieval (DSVR), Complementary Scene Video Retrieval (CSVR) and Incident Scene Video Retrieval (ISVR). Here we only focus on the annotations DSVR videos.

Newly Added Segment-level Annotations

Overall FIVR-200k-PVCD contains 10870 annotated copy segment pairs involving 5935 different video pairs. Many partial copy segments are more challenging with abundant temporal and spatial editing.

Performance comparison on FIVR-200k-PVCD

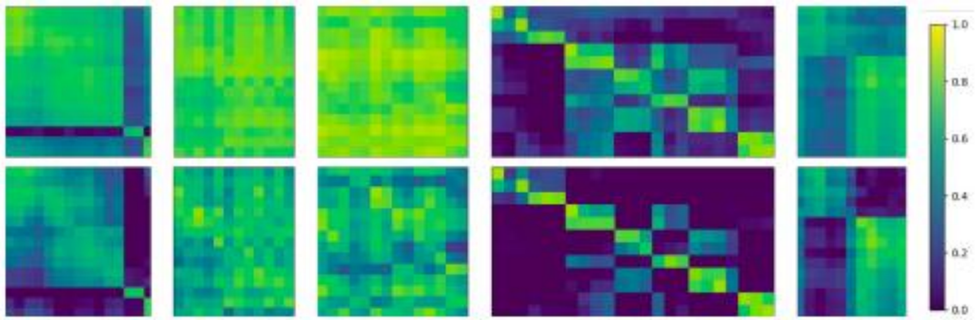
	IoU>0			IoU>0.3			IoU>0.5			IoU>0.7		
Methods	SP	SR	F_1-score	SP	SR	F_1-score	SP	SR	F_1-score	SP	SR	F_1-score
HV(baseline)	0.4350	0.5911	0.5012	0.6069	0.3491	0.4433	0.5501	0.3142	0.4000	0.4778	0.2708	0.3457
HV+SE	0.4579	0.5936	0.5170	0.5827	0.3794	0.4596	0.5281	0.3439	0.4165	0.5164	0.2755	0.3593
HV+SE+SW	0.5730	0.5255	0.5483	0.5075	0.4563	0.4805	0.4541	0.4128	0.4325	0.3952	0.3553	0.3742
SM+SE+SW	0.8300	0.5916	0.6908	0.8151	0.5525	0.6586	0.7580	0.5014	0.6036	0.6485	0.4091	0.5017
SM+SE+SW+MM	**0.8575**	**0.6883**	**0.7636**	**0.8212**	**0.6556**	**0.7291**	**0.7738**	**0.5434**	**0.6384**	**0.7076**	**0.4281**	**0.5335**

VSAL: Experiments - Visualization

Comparison of spatial similarity matrices
with or without sequence encoder

w/o

w/

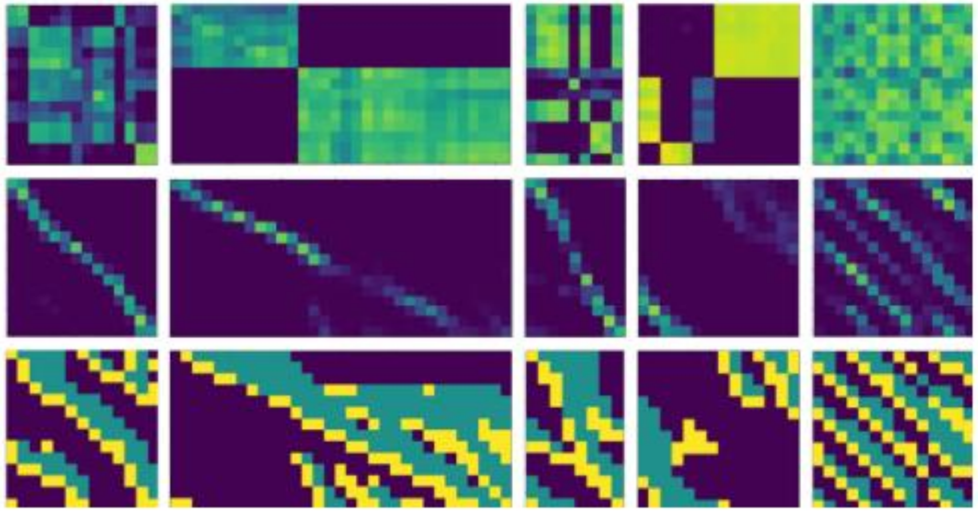


Comparison of spatial similarity matrices
with or without sequence encoder

S

MM

SM



QnA