Personalised CLIP or: how to find your vacation videos

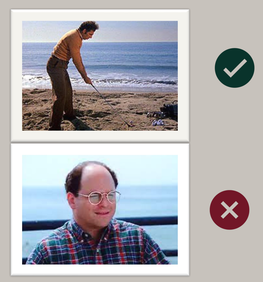
Bruno Korbar, Andrew Zisserman,
Visual Geometry Group

DEPARTMENT OF ENGINEERING SCIENCE
UNIVERSITY OF OXFORD

Problem

What is compound retrieval?

Michael Richards playing golf on the beach.



- ? No model takes multi-modal queries
- ? No benchmark

Data

We present the new "**Celebrities in Action**" dataset for video compound retrieval.

Gathering videos from existing datasets [1,2].

Automatic annotation of the videos [3].

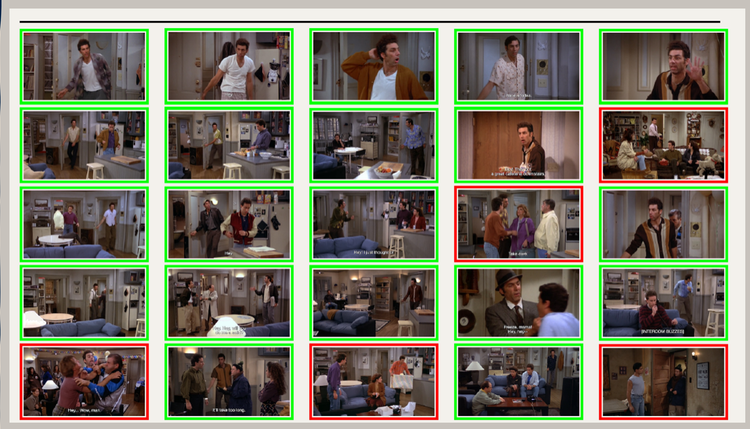
Manual inspection and evaluation.

1. Celebrity doing action.
2. Celebrity in a place.
3. Celebrity in a place doing action.

- ! > 2500 video clips from 69 movies and 5 TV-shows
- ! 3 different retrieval scenarios

"Real world"

Find all clips in season 2 of Seinfeld corresponding to the query: Michael Richards [TOK] entering the room.



And quantitatively:

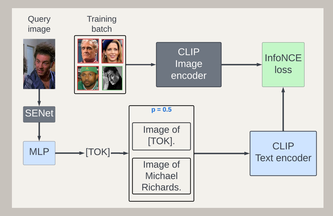
Fine-tuned	2 clips per min			4 clips per min		
	R@25	R@50	last rank	R@25	R@50	last rank
no	0.61	0.92	53	0.46	0.58	113
yes	0.84	1.0	42	0.72	0.88	72

Model

We present a **P**erson **AD**aptive CLIP model: **CLIP-PAD**

- Has knowledge of over 9k celebrities
- Can recognize people from their image or name

A sample application of CLIP-PAD model.



Training the model on VGGFace(2) datasets [4].

Results

- CLIP-PAD outperforms previous retrieval benchmarks and zero-shot CLIP [5].
- Image query helps in cases where person is not known or have been seen in few examples

Method	person query		action		place		action + place	
	text	rgb	R@1	R@5	R@1	R@5	R@1	R@5
CE [21]	✓	*	35.3	65.1	36.7	65.3	32.1	62.5
CLIP [35]	✓		42.4	73.5	58.2	78.1	39.9	72.0
CLIP-PAD	✓		62.1	81.3	67.8	87.0	64.2	81.1
CLIP-PAD		✓	64.5	83.7	68.2	87.3	64.9	81.8
CLIP-PAD	✓	✓	65.0	85.4	71.5	88.6	66.3	82.7

References

[1] **Marszalek *et al.*, "Actions in Context"**, CVPR 2009
[2] **Patron-Perez *et al.*, "High Five: Recognising Human Interactions in TV Shows"**, BMVC 2010
[3] **Brown *et al.*, "Automated Video Labelling: Identifying Faces by Corroborative Evidence"**, MIPR2021
[4] **Q. Cao, *et al.*, "VGGFace2: A dataset for recognising faces across pose and age"**, ICAFGR 2018
[5] **Radford *et al.*, "Learning transferable visual models from natural language supervision"**, ICML2021