Supplementary material for "Personalised CLIP or: how to find your vacation videos"

Bruno Korbar
https://kor.bar
Andrew Zisserman
https://www.robots.ox.ac.uk/~az

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

1 Visual queries for unknown faces

The results show that our model performs well, even for the faces it hasn't necessarily seen during training time. In this section, we clarify the relationships between different training datasets and give examples of succesful (and unsuccessful) retrieved and classified examples for both the Celebrities in Places (CiP) [10] and Celebrities in Action (CiA) dataset.

1.1 CiP

Celebrities in Places contains images for celebrities from the VGGFace dataset [7] (noted as *seen* in Zhong et al. [10], as the backbone CNN is trained on VGGFace), and celebrities from other popular face recognition datasets (unseen). The retrieval network in [10] is also trained on a synthetic dataset for scene retrieval, and *might have* seen the unseen face there, but the face has not been explicitly labeled. A particularly curious aspect of their model is the fact that the network does better in faces-only retrieval on unseen categories than on the seen ones. This has been attributed in [10] to a much more discriminative face descriptor that is obtained during their training.

We similarly train our model on VGGFace, but we omit training on the synthetic dataset, as CLIP [9] has good zero-shot performance on scene retrieval as-is (see table 2 in the main body of the paper). Unlike Zhong et al. [10], who initialise their retrieval model randomly, the CLIP model has seen 400M images, and quite likely some celebrities from the unseen faces too. We show examples of *unseen* faces that our model can retrieve correctly (recall@5), and some it cannot in figure 1

1.2 CiA

Similarly, only around 37% of annotated cast members from the test set are present in VGGFace2 [2], and this increases to 76% when the faces in the CiA training set are included. Thus the network should not have seen around 24% of the test set celebrities during training. We show classification results from table 4 in the main body, broken down by 'seen' (103) and 'unseen' (32) celebrity categories in table 1.

The main takeaway from this table is the fact that unseen faces benefit disproportionately from the addition of the visual query.

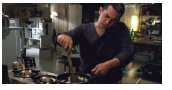
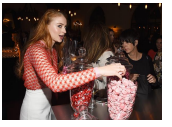
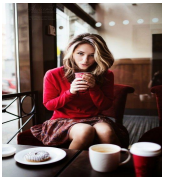
Query	Tgt. Frame Rank	Tgt. Frame
Correct		
Aamir Khan in the kitchen	3	
Adam Driver in the desert	5	
Sophie Turner in the banquet	3	
Incorrect		
Andrea Pirlo in the kitchen	24	
Goran Visnjic in the supermarket	13	
Zoe Levin in the coffee shop	7	

Figure 1: An example of correctly (within top-5) and incorrectly retrieved examples from the CiP dataset [10], where the person was *not* explicitly seen during training. For each example, we provide a rank at which they were retrieved.

2 Real world retrieval example

In order to present a real-world example of personalised retrieval, we apply our model on the entire Season two of the Seinfeld TV show. We want to see how many occurrences of "Michael Richards entering the room" we can correctly retrieve. Specifically, we form our query with the above text and a single randomly selected query image from the top-100 Google face images. See Fig. 1 of the main paper for illustration.

From each episode, we extract two 2 second long video clips each minute (16 frames per clip at 8fps, at : 00 and : 30 timestamp of each minute). For ground truth, the authors watched the Season, noting all occurrences satisfying the query. If an occurrence happens outside of the given time frames, we added additional video clip extracted with the same

Model	Seen	Unseen
random	0.9	3.1
text	31.3	11.1
img	35.8	22.7
text + img	36.7	24.4

Table 1: Performance of our model for person classification on the CiA dataset test set, evaluated on classes that have been seen or unseen during training (not accounting for CLIP [9] pre-training). Numbers are given in % accuracy.

settings to the pool of clips. For 11 episodes, we extracted a total of 483 video clips. By our count, Richards is portrayed entering the room 26 times.

The performance of the model can be viewed in Table 2. Note that Michael Richards (the actor in the role of Kramer) has *not* been seen during training of the model which is not fine-tuned as he is not included in VGGFace or VGGFace2 datasets, however we are still able to localise him in 16 out of top-25 retrieved example.

But what if we had more clips? We try to push this setup by extracting 4 two-second video clips for each minute at the same frame rate as above (at :00, :15, :30 and :45 timestamps of each minute). This yields 994 total clips. Bear in mind that in this scenario, determining positive examples becomes a challenge; for example, authors note that Kramer enters the room at 7:17 in episode 9, however one could argue that he's barley visible in the doors at 7:16. Would that count as a positive example, even if we were not exceptionally precise about it? With that in mind, our results in this scenario still show promise. In the top-50 examples, we retrieve 15 out of 26 positives, with additional 6 clips that could be considered a close-positive.

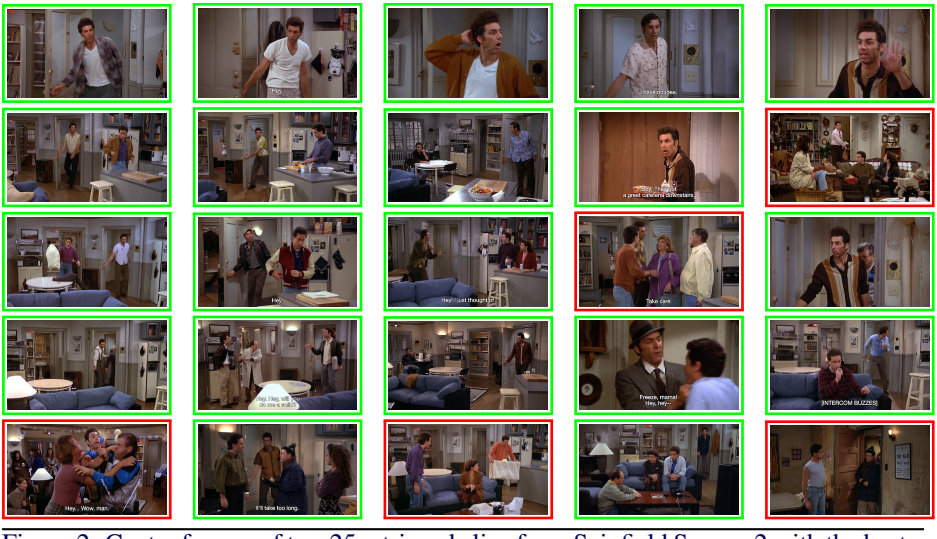


Figure 2: Center frames of top-25 retrieved clips from Seinfield Season 2 with the best model fine-tuned on Seinfeld cast. Frames are sorted from left to right and from top to bottom (top left is rank 1, bottom right is rank 25). Correctly retrieved examples have a green border, whilst incorrectly retrieved examples have a red border. Best viewed in colour.

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

Table 2: Quantified retrieval results from our Seinfeld experiment. Note that the model without fine-tuning has not seen Michael Richards as a character during training for person-awareness.

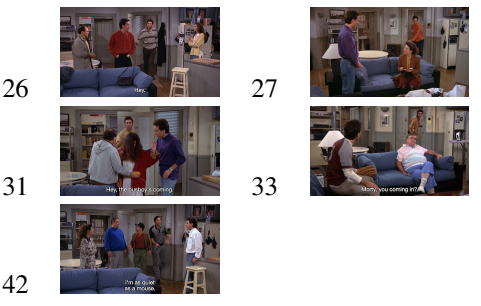


Figure 3: Center frames of examples retrieved outside the top-25 with a corresponding rank to the left of the frame. Note that all of these examples feature Richards' character in the background, totally obscured or potentially out of context.

What if the model sees the characters? Only one of the Seinfeld characters from Season two is present in either of the training sets we fine tune the CiA model on. To improve our chances, we additionally fine tune the model with images containing Seinfeld characters scraped from Google images (500 per character) with a fixed learning rate of $2e-5$ for 3 epochs. This method unsurprisingly yields the best results as seen in table 2. The center frame of the top 25 retrieved results can be seen in figure 2, while the examples missed in the top-25 and their corresponding rank can be seen in figure 3. In the top-25 retrieved examples we count 21 occurrences of Richards entering the room, and all were retrieved in the top 50.

3 Celebrities in Action

In this section, we go more in-depth about our Celebrities in Action (CiA) dataset. We present further data-collection details, show the most common failure cases, and propose further improvements. We show several annotated examples in an external HTML gallery attached.

3.1 Data collection and annotation

To recap, we automatically annotate the video clips from the Hollywood2 [6] and High-Five [8] datasets with the person performing the action using the automatic video annotator by Brown [1].

We scrape 200 images for every cast member automatically found using the IMDB cast list to obtain a common face embedding for each cast member. Faces are then detected and tracked in each clip at 5 frames-per-second, and then an identity is associated with the face track if it is classified as one of the known actors from that film.

In the case where multiple face tracks with different identities are detected in the scene, we select the one the model is most confident in. Given that the number of training images for person classification is completely balanced, we argue that high confidence for a face track would signify the most dominant (or the clearest) face within the video clip.

If a face-track is not found, we discard the clip. For the test set only, if the face-track is not classified with high confidence (over 0.5), we discard that clip from our consideration. In total, we discard 153 video clips from the dataset.

We manually verify the label correctness on a randomly selected 100 clips from the test set and find the actor annotations to be correct for 97 of them (i.e. annotated actors are visible in the video clip and are performing an action class associated with the video clip).

For clips taken from the High-Five dataset, we additionally annotate them with a high-level place attribute from a ResNet-18 [4] pre-trained on Places365 [11] dataset and manually verify correctness.

We separate the Hollywood2 training set into the training and held-out validation set (for model development) – not according to films but rather according to the clips in the training data. The Hollywood2 test set remains intact, other than clips discarded as noted above. All clips from the High-Five dataset are added to the test set.

Note that not all video clips contain annotations for 'action' and 'scene'. In total, there are 884 clips in our test set annotated with an actor and an action, 576 of them have a scene annotation attached. In total, the dataset contains 135 actor classes, 12 action classes and 10 scene classes. When results are reported, we only report results on the appropriate set of clips: e.g. when reporting results on 'scene' or 'action + scene' retrieval, we retrieve the examples from the appropriate clip-set.

3.2 Annotation failure cases

As our dataset is largely automatically annotated, we naturally observe some label noise. In this section, we discuss the Hollywood2 dataset noise and the three most common failure cases. In the last paragraph, we propose different ways to improve the dataset in the future.

1. Innate scene annotation dataset noise. Parts of Hollywood2 are automatically annotated, which inherently introduces noise into the dataset. The scene settings are potentially

overlapping (would a driveway be considered house-exterior, road, or both?), how to distinguish a hotel room or a bedroom? Would a 12-year-old scene classifier be able to correctly distinguish between them? We find that for the test set, where data has been manually cleaned, these concerns are minimal. If an example is confusing, a scene category is not assigned to it. We do have to acknowledge the noise in the training data, however, in the main body of the paper we show that CLIP-PAD performs well despite the noise.

2. The wrong actor was annotated. On average, there is more than one (annotated) actor in each video clip, however, for a final annotation, we only select the most confident one. This may naturally lead to erroneous cases as the best-seen actor might not necessarily perform an action corresponding to the annotation for that video clip (e.g. **figure from the main paper**). Furthermore, more "famous" actors might also have higher quality images available on Google images, hence leading to better training data for the automatic annotation algorithm. In the example above, Kevin Spacey is both the more famous, and clearer of the two actors in the scene, hence this issue is clearly prominent.

In our preliminary quality control, these issues are rare (2 in 100) due to a strong prior that the actor performing the action is the better seen and/or more famous of the actors in the scene.

3. Ambiguous annotation. Hollywood2 dataset contains multiple actions that require *interaction* [6]: hugging, kissing, handshaking and fighting. For all of these, there can be more than a single correct answer. For example in "Bruce Almighty", Jim Carrey is hugging Jennifer Aniston – whilst both actors are technically correct, only the first one would be accepted as correct.

In our preliminary quality control, we found this to be a common occurrence (16 out of 100),

4. Low recognition confidence and false positives. On average, the named cast of a film contains >200 people, but we only have a limited number of clips and people from each film represented in our data. We go through an effort of manually annotating each video clip with the name of the film it belongs to. In this way, the automatic face annotator only has to choose between the cast of that particular film, but our classification accuracy is still low. In the HTML example gallery, the name is associated with the generated video if the recognition confidence is higher than 0.9, and these clips are relatively rare. We observe that only 195 clips (out of over 1500) are annotated with such high confidence.

This means that whilst being mostly correct, our model is not fully confident in its predictions. In "It's a Wonderful Life" for examples, more than one actor in 3 clips is annotated as 'James Stewart', none with high confidence. In the sample gallery the reader can see that whilst our predictions are correct, confidence is not necessarily high. This is often due to the domain difference between the images sourced from Google Image Search and the character's appearance in the film.

Although we do not find this issue to be concerning (as we only select the most confident annotation which we find correct in 99 out of 100 clips we've looked at manually), improvement in recognition confidence would aid our dataset overall.

Future improvements. We argue that even in the initial release, CiA presents a thorough benchmark for compound retrieval of actions on video. There are future improvements that

could reduce the noise and further address the failure cases above. The optimal way of addressing these issues would be manual annotation which is an expensive and time-consuming solution.

One option would be the classification of scenes in automatically annotated and not annotated video clips using a modern scene classifier. This would potentially introduce additional categories to the data, but it would increase the variety of the dataset. It would also require additional rounds of manual annotations to keep the test set completely noise-free.

We could also optimise the video auto-annotator [1] for our films. The results can be improved by manually cleaning the automatically-scraped Google images, or by running multiple iterative rounds of automated annotations. Specifically, we could discard all cast members that we know are not in the selected clips, gather additional data on those that are, and re-annotate the lot until the annotations (and their confidence) change no more. This would hopefully reduce the number of false positives and low-confidence classifications.

Even without the potential improvements, we believe our dataset is already a formidable benchmark for various compound retrieval scenarios.

4 Additional experiments

Bellow, we expand on the existing experiments.

4.1 Further benchmarks on CiA

Method	Text	Image	R@1	R@5
CE [5]	✓		32.1	62.5
DE [3]	✓		35.1	66.2
CLIP [9] (0-shot)	✓		39.9	72.0
CE [5]*	✓	✓	39.1	71.5
DE [3]*	✓	✓	39.9	71.8
CLIP-PAD*	✓	✓	57.8	78.1
CLIP-PAD	✓	✓	66.3	82.7

Table 3: Additional results on our CiA dataset. [3] and [5] have been finetuned using default parameters. [3] does not have a dedicated face embedding module which might impact its performance negatively. Models using only text effectively perform text-to-video retrieval, while models using text and image combine a visual query (an image of the actor) with the text query – see last paragraph in sec. 4 of the main paper for details on query formation). Note that models marked with "*" use the two-stage querying process as described in text.

Despite the existing benchmarks, compound image retrieval is a fairly unexplored task. Similarly, our method is the first one disambiguate between the more "traditional" text-to-video retrieval and our proposed task of compound query video retrieval.

Furthermore, most video-retrieval methods that we are aware of encode video and text in separate streams or do not allow querying with a multi-modal query. However, we apply two modern retrieval methods [3, 5] for which code is available, and compare it further to our model. Specifically, we compare the performance for the 'action + place' metric on the CiA

Method	High five			Hollywood 2			CiA
	high-fiving (40)	hugging (48)	kissing (43)	kissing (103)	get out of car (57)	drive car (102)	all classes
CLIP	49.1	47.5	54.9	55.1	81.8	85.0	73.1
CLIP-PAD	57.3	55.6	59.4	59.9	83.0	85.3	75.7

Table 4: Breakdown of action classification performance per-class for select classes. For evaluation, we follow the same protocol as in Table 4 of the main body of the paper, using only text as a query.

dataset using uni-modal and multi-modal queries as outline below. To retrieve the correct clip based on a text query only, we feed in the textual query in a form "*person* doing *action*" (where *person* and *action* are defined in the data) to each method's respective text encoder and use that to find the most similar video clip. As these models are not capable of multi-modal retrieval, when using the combination of text and image, methods marked with '*' in Table 3 are using a two-stage process. In the first stage, we rank the clips according to the similarity to the textual query and the similarity to the visual query (image of the target actor, processed by the visual encoder provided by the respective methods). In the second stage, we take the intersection of the two ranked lists by considering the top n elements of each list, where $n >= k$, until k common elements are found. In order to make the comparison fairer (as our model is capable of using multi-modal queries), we also apply our model in the same fashion and denote the result as 'CLIP-PAD*'. Note that *(a)* there is a clear benefit of a compound query compared to the intersection of ranked lists, and *(b)* our multi-modal method outperforms all others still.

4.2 Looking closer at action classification

We notice that some action classes can be recognised with significantly less accuracy than others. As we can see in table 4, examples coming from the High-five dataset tend to score lower on that particular benchmark. We note that the number of examples is not as indicative of a performance nor is there a major distribution shift between the datasets (for example, 'kissing' has equal performance for examples coming from High-five and examples coming from Hollywood2).

References

[1] Andrew Brown, Ernesto Coto, and Andrew Zisserman. Automated video labelling: Identifying faces by corroborative evidence. In *Multimedia Information Processing and Retrieval (MIPR)*, 2021.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018.

[3] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[5] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019.

[6] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[7] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[8] Alonso Patron-Perez, M. Marszałek, Andrew Zisserman, and Ian D. Reid. High five: Recognising human interactions in TV shows. In *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, 2010.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[10] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Faces in places: compound query retrieval. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[11] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.