FAST OBJECT INSTANCE SEARCH IN VIDEOS FROM ONE EXAMPLE

Jingjing Meng, Junsong Yuan, Yap-Peng Tan, Gang Wang[*]

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 637819

ABSTRACT

We present an efficient approach to search for and locate all occurrences of a specific object in large video volumes, given a single query example. Locations of object occurrences are returned as spatio-temporal trajectories in the 3D video volume. Despite much work on object instance search in image datasets, these methods locate the object independently in each image, therefore do not preserve the spatio-temporal consistency in consecutive video frames. This results in sub-optimal performance if directly applied to videos, as will be shown in our experiments. We propose to locate the object jointly across video frames using spatio-temporal search. The efficiency and effectiveness of the proposed approach is demonstrated on a consumer video dataset consisting of crawled YouTube videos and mobile captured consumer clips. Our method significantly improves the localized search accuracy over the baseline, which treats each frame independently. Moreover, it is able to find the top 100 object trajectories in the 5.5-hour dataset within 30 seconds.

Index Terms— spatio-temporal trajectory, object instance search in videos

1. INTRODUCTION

Object instance search in videos aims to search for and localize the spatio-temporal trajectories of a particular object in the 3D video volume. This problem arises from the need of fine-grain analysis on big data, such as object-instance-level annotation on YouTube videos for contextual ads. It is closely related to the rich body of work on object instance search in images [1] [2] [3] [4] [5] [6] [7] [8][9]. Similarly to them, we focus on instance search from a single example.

Although great progress has been made in object instance search on large-scale image datasets in the past decade [1] [10] [3] [4], little has been done on videos, which concerns about locating the query object as spatio-temporal trajectories in the video space [11]. Although named Video Google, [4] treats each frame independently and returns a ranked list of key frames or shots of a video that contain the object. The temporal consistency between the video frames is totally ignored except for when rejecting unstable regions across frames in a shot. Similarly, the TRECVID object instance search challenge searches for shots that contains a topic without concerning about localization or how many frames in the shot contains it [12] [13] [14] [15]. Therefore, in essence these approaches are still tailored for images (i.e. frames). Of course, a naive approach to achieve spatio-temporal localization is to use image-based approaches to search for and localize the object on every video frame. In other words, spatial localization is independently performed on each frame regardless of temporal constraints. Therefore, this approach neither guarantees spatio-temporal smoothness of the trajectories nor global optimality.

We explore how to effectively incorporate the spatio-temporal context into an efficient search framework for large-scale videos. Specifically, we adopt our previous Randomized Visual Phrase (RVP) [10] to efficiently generate frame-wise confidence maps. The key difference from [10] is that we do not rely on individual confidence maps to determine the location of the object using a heuristic threshold, rather, we jointly consider all confidence maps across the entire video volume to obtain the globally optimal spatio-temporal localization of the object using Max-Path search [16]. Although Max-Path search [16] has been used for video event detection and pedestrian detection. We are the first to innovatively adapt it to the problem of object instance search in videos. Our previous work [11] is closest to this paper in that it combines Hough Voting and Max-Path search to locate object centers in a video sequence. However, it produces the trajectories of object center rather than the object itself. Moreover, it is not efficient for large scale videos.

We evaluate our system on a 5.5-hour consumer video dataset, using 10 external objects as queries. Compared with the baseline RVP that searches and locates the object independently in each video frame, our approach significantly improves the search performance by considering the spatio-temporal consistency in consecutive video frames.

Our contributions are three-fold. First, by innovatively exploring temporal information in videos, our approach significantly improves object instance search and localization accuracy over the baseline method that evaluates each frame

[*]This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office. It is also partially supported by the Nanyang Assistant Professorship (SUG M58040015) to Junsong Yuan.

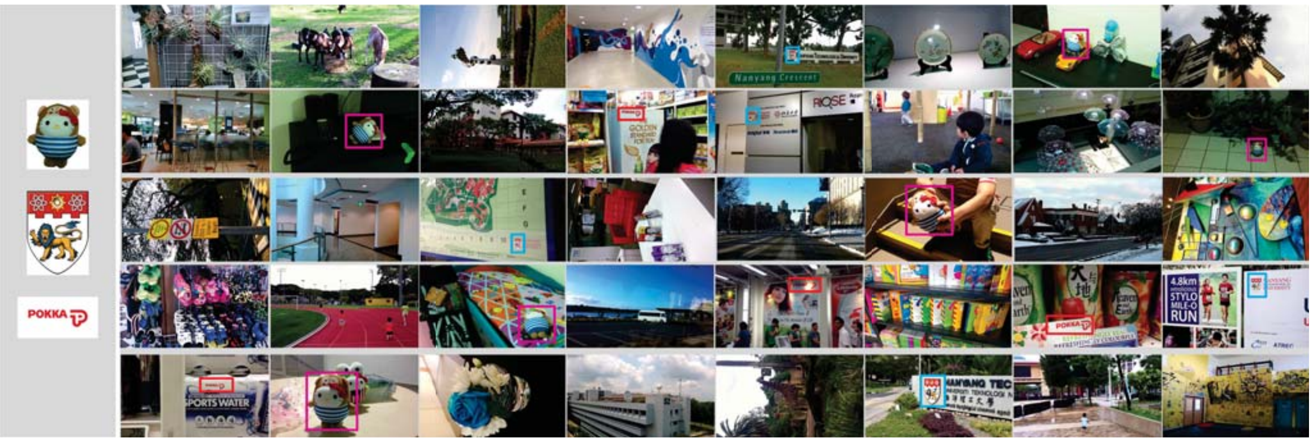


Fig. 1: Example videos from our dataset with annotated object locations (only one key frame is shown for each video). Three example queries are given on the left: KittyB, NTU and Pokka (Table 1). Their ground truth locations are annotated in magenta, blue and red bounding boxes, respectively. This dataset consists of consumer clips crawled from YouTube and mobile captured personal videos. The scenes are diverse and mostly cluttered, while the target objects only occupy a small region in the video frame. These make object instance search and localization very challenging on this dataset. Best viewed in color and enlarged.

independently. Second, our approach is efficient and scalable. It can return the top 100 trajectories of an object instance in 5.5-hour videos in less than 30 seconds. Third, we make a benchmark video dataset available to the research community, which provides per-frame bounding box annotations of object instances (Object-Instance-Videos (OIV): https://sites.google.com/site/jingjingmengsite/research/data).

2. ALGORITHM

If we treat the entire video database as a long video sequence $\mathcal{V} = \{\mathcal{F}_1, \mathcal{F}_2 ... \mathcal{F}_n\}$, where $\mathcal{F}_i \in \mathcal{V}$ is a $w \times h$ sized frame with temporal index i, then given a query object Q, we want to find in $\mathcal{V}$ all spatio-temporal trajectories of instances of Q.

To simplify the problem, let us assume that the trajectories are non-overlapping. Therefore, we can divide $\mathcal{V}$ into m short video chunks, i.e. $\mathcal{V} = \{V_1, V_2 ... V_m\}$. The problem can now be recast to finding in each video chunk $V_i \in \mathcal{V}$ *one* trajectory has the highest likelihood of object occurrence.

We denote the collection of all spatio-temporal trajectories in V_i as $\mathcal{T}_i = \{T_{i1}, T_{i2} ... T_{il}\}$, where l is the total number of object trajectories in V_i. Each $T_{ij} \in \mathcal{T}_i$ can be further represented as a sequence of bounding boxes, i.e. $T_{ij} = \{\mathcal{B}_{ij_1}, \mathcal{B}_{ij_2} ... \mathcal{B}_{ij_k}\}$, where k is the total number of frames in trajectory T_{ij} .Our objective is to find for each $V_i \in \mathcal{V}$:

$$T_i^* = \underset{T_{ij} \in \mathcal{T}_i}{\arg\max}\, s(T_{ij}) \tag{1}$$

where $s(T_{ij})$ is the confidence score of trajectory T_{ij}, which measures the likelihood of object occurrence. A high score indicates a high likelihood that the object appears along the trajectory, and a low score indicates otherwise. The trajectory confidence score $s(T_{ij})$ is calculated as the sum of the confidence score of bounding boxes along T_{ij}, i.e.

$$s(T_{ij}) = \sum_{\mathcal{B} \in T_{ij}} s(B) \tag{2}$$

Once we have the best trajectory T_i^* for each video chunk, we can then return the ranked results of all trajectories.

2.1. Spatio-temporal Localization using Max-Path Search

To solve (1), we propose to run Max-Path search [16] on $V_i \in \mathcal{V}$. Max-Path can obtain the global optimal trajectory with proven lowest search complexity, which is linear to the video volume size ($O(whn)$). Similar to [17], multiple paths can be located by repeating the search after removing the current best path from the confidence map sequence.

Note that with a single example, we do not train a discriminative classifier, but directly match the query against the dataset. Hence the resulting confidence maps are not discriminative for Max-Path search [17]. To introduce negative values, we add a negative threshold to each confidence map, similar to [17]. To accommodate object appearance variations in different video chunks, instead of a fixed threshold [17], we set the threshold adaptively to be proportional to the average pixel-wise confidence score of each video chunk (excluding zero confidence maps). Denote the total number of non-zero confidence maps in video chunk V_i as $\mathcal{N}_{NZ}$, the threshold is:

$$MP_{thres} = \beta \frac{\sum_{\mathcal{F} \in V_i} \overline{s_{\mathcal{F}}}}{\mathcal{N}_{NZ}} \tag{3}$$

$\overline{s_{\mathcal{F}}} = \frac{\sum_{p \in \mathcal{F}} s(p)}{|\mathcal{F}|}$ is the average pixel-wise confidence score of frame $\mathcal{F}$. Here $|\mathcal{F}|$ is the total number of non-zero pixels in frame $\mathcal{F}$, and $s(p)$ indicates the confidence score of pixel $p \in \mathcal{F}$.

2.2. Efficient Frame-wise Confidence Map Generation via Randomized Visual Phrases

Before Max-Path search, we need to first obtain the confidence map for each frame $\mathcal{F} \in V_i$. To this end, we utilize the Randomized Visual Phrases (RVP) approach for image datasets [10]. By averaging the matching scores over multiple randomized visual phrases (spatial context), RVP has shown to be able to better handle appearance variations than other state-of-the-art object search methods on large image datasets [5] [6] [7] [8] with a higher efficiency. However, solely relying on RVP to obtain the object location in each frame has two drawbacks. First, RVP depends on a heuristic segmentation coefficient α to locate target objects in an image [10]. Second, its performance drops with insufficient rounds of partition, as the confidence map would not be salient. However, by using Max-Path search to jointly evaluate confidence maps across each video chunk (2.1), our approach removes the dependency of localization on the segmentation coefficient α, and can afford faster evaluation on each frame (with fewer rounds of partition), then leverage the spatio-temporal consistency to boost localization accuracy.

2.3. Coarse-to-Fine Search

To further improve efficiency, we build two sets of inverted file indexes: one on coarsely sampled frames, and the other on the entire dataset frames. We use the coarse index to fast filter low-confidence video chunks before we proceed to per-frame search, which is more computationally expensive. Given a query, the baseline RVP is first run on the coarse index to obtain confidence scores of sampled frames. Then an initial ranking of all video chunks is produced based on the average confidence score of sampled frames. Only for those top ranked video chunks, per-frame confidence maps are generated using the fine index and we run Max-Path for re-ranking.

3. EXPERIMENTS

3.1. Dataset

We conduct our experiments on a video dataset of 5 hours 38 minutes with a spatial resolution of 800×450. It consists of consumer clips crawled from YouTube (around 2 hours) and mobile captured personal videos (the remaining). As can be seen in Fig.1, the dataset covers diverse topics and most scenes are cluttered, which make object instance search and localization very challenging on this dataset.

The videos are uniformly cut into 150-frame chunks (preserving clip boundaries), resulting in 4,347 video chunks. We manually annotate each frame that contains any of the query objects (Table 1) to obtain the ground truth object locations. In total 237 video chunks with a total length of approximately 20 minutes are annotated.

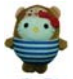
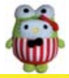
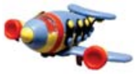
Hessian-Affine detectors [18] are used to extract interest points, which are then described as 128D RootSIFT descriptors [19]. We use FLANN [20] to build a vocabulary of 250K words by sampling 1 key frame per video chunk. As in [4], a stop list is used to remove the top 5% and bottom 10% frequent words in the dataset.

The coarse inverted index file of 230MB is built by sampling every 30^{th} frame and indexing over 20M interest points. The fine inverted index takes 7GB, which indexes the entire dataset of $600,375$ frames and over 600M interest points.

3.2. Queries

Table 1 summarizes the statistics of the 10 query objects used in our experiments. Among them, KittyB, KittyG and Plane are 3D objects, while the remaining 7 are 2D objects. Note that to make a fair evaluation, we do not extract objects from the testing video frames as queries, but use external query images instead. Specifically, for the 3D objects, we take one picture from a single view for each as the query. For the 2D objects, we use Google search with the text object name to obtain the query image. The query images are shown in Table 1 as well. In all our experiments, only a single query image is used to search an object.

Table 1: Annotated Object Instances in the Video Dataset

No.	Name	Illustration	No. of Ground Truth Trajectories
1	100Plus		24
2	Ferrari		16
3	KittyB		34
4	KittyG		18
5	Maggi		32
6	NTU		29
7	PKU		15
8	Plane		25
9	Pokka		13
10	Starbucks		31

3.3. Evaluation Metric

We evaluate the search performance using trajectory mean Average Precision (mAP) of the 10 objects. The Average Precision (AP) of each object is calculated on the top 100 returned trajectories. Since the precision depends on whether a returned trajectory is *relevant* to the ground truth, we adopt

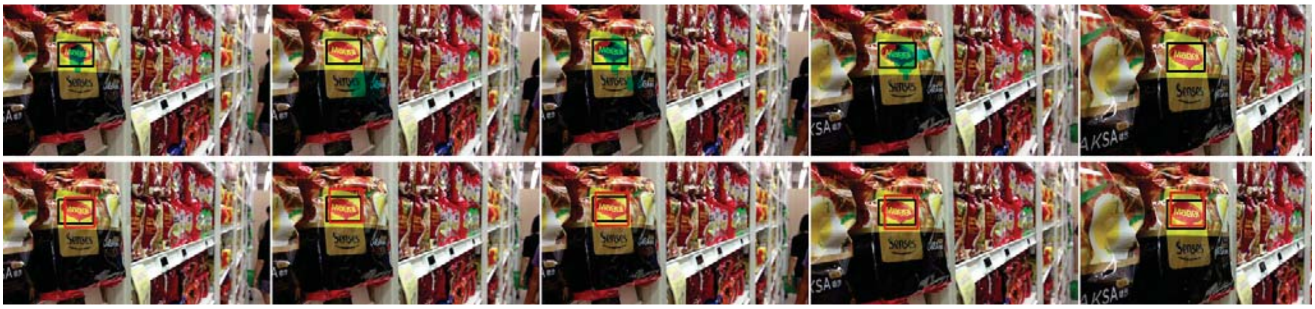


Fig. 2: Example localized search results of ”Maggi” . The segmented locations from the baseline [10] are highlighted in cyan in the top rows, our search results are marked with red boxes in the bottom rows, and ground truth are marked with black boxes.

Sequence Frame Detection Accuracy (SFDA) [21] to measure this relevance. Only when SFDA is above a threshold, a returned trajectory is considered *relevant*. We skip the definition of SFDA due to space limit (refer to [21]).

3.4. Results

We followed the settings in [10] to use $\mathcal{K} = 200$ rounds of partition with a segmentation coefficient $\alpha = 3.0$ to generate the initial ranking in the coarse round. In the fine round, we fix $\mathcal{K} = 50$ and $\alpha = 3.0$ for both the baseline RVP and our approach (RVP-MP). For all experiments, we use the Max-Path algorithm with multiple scale extension at a fixed aspect ratio of $1 : 1$ and a local neighborhoods of 3×3. The spacial step is set to 10 pixels and the temporal step is set to 1 frame. The negative coeffient β is set to -2.0 for all queries.

3.4.1. Efficiency

All experiments are conducted on a quad-core dual-processor machine with 2.30GHz CPU and 32GB RAM, without GPU. We parallelized both RVP and Max-Path search in 8 threads and implemented the algorithms in C++. For coarse ranking and filtering, the average time of the 10 objects is 0.833 seconds. And it takes another 28.738 seconds on average to obtain the top 100 trajectories for each query (3.883 seconds to generating the frame-wise voting maps and 24.855 seconds for Max-Path search). Therefore, the average search time excluding I/O is 29.57 seconds on the 5.5-hour video dataset.

3.4.2. Accuracy

Fig. 2 shows a qualitative comparison of example trajectories that are returned by our proposed RVP-MP and the baseline RVP. We can see that by exploring the spatio-temporal consistency across consecutive video frames, RVP-MP is able to filter false alarms and reduce missing detections, therefore produce more accurate localization of object instances.

To quantitatively compare our approach with the baseline and evaluate how the number of partition rounds affects trajectory mAP, we fix the segmentation coefficient α at 1.0, the SFDA threshold at 0.3, and increase partition rounds $\mathcal{K}$ from 1 to 100. As can be seen in Fig. 3, our proposed method consistently outperforms the baseline RVP with sufficient rounds of partitions (above 10 rounds). In general, more rounds of partitions will improve the mAP for RVP-MP. However, different from [10], we observe a much faster convergence around 50 rounds of partitions for both methods. For the baseline RVP, this is likely because we rank the trajectories based on the average score of all segmented regions in a video chunk. Therefore the final ranking is less sensitive to false alarms in individual frames. For RVP-MP, it can be attributed to the spatio-temporal consistency across video frames that Max-Path utilizes to boost localization accuracy on less salient confidence maps (resulting from fewer partition rounds). With 50 rounds of partition, RVP-MP achieves a trajectory mAP of 25.68%, which improves the baseline of 19.60% by 31%.

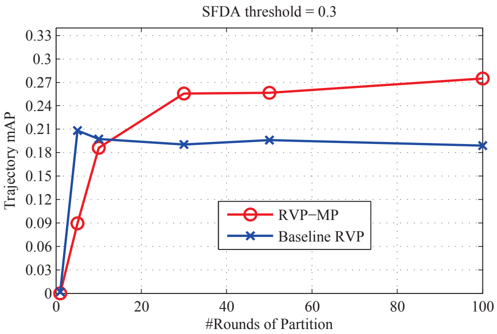


Fig. 3: The impact of partition rounds on search accuracy. Our approach (RVP-MP) consistently outperforms the baseline when the number of partition rounds is above 10.

4. CONCLUSION

We present an efficient approach to search and location object instance in large video volumes from a single example. Rather than locating the object instance independently on each frame, as is done by existing methods for image datasets, we innovatively formulate the problem as spatio-temporal search of the optimal object trajectories in videos. Our proposed approach significantly improves the state-of-the-art object instance search method [10] in trajectory mAP on a challenging consumer video dataset.

5. REFERENCES

[1] Ran Tao, E. Gavves, C.G.M. Snoek, and A.W.M. Smeulders, "Locality in generic instance search from one example," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 2099–2106.

[2] Yuning Jiang, Jingjing Meng, Junsong Yuan, and Jiebo Luo, "Randomized spatial context for object search," *Image Processing, IEEE Transactions on*, p. to appear, 2015.

[3] Xiaohui Shen, Zhe Lin, Jon Brandt, Shai Avidan, and Ying Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn reranking," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[4] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2003.

[5] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[6] Christoph H. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2009.

[7] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[8] Josef Sivic and Andrew Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 591–606, 2009.

[9] Jingjing Meng, Junsong Yuan, Yuning Jiang, Nitya Narasimhan, Venu Vasudevan, and Ying Wu, "Interactive visual object search through mutual information maximization," in *Proc. ACM Multimedia*, 2010.

[10] Yuning Jiang, Jingjing Meng, and Junsong Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[11] Jingjing Meng, Junsong Yuan, Gang Wang, Yap-Peng Tan, and Jianbo Xu, "Object instance search in videos," in *Proc. International Conf. on Information, Communications and Signal Processing*, 2013.

[12] Takahito Kawanishi, Akisato Kimura, Kunio Kashino, Shin'ichi Satoh, Duy-Dinh Le, Xiaomeng Wu, and Sbastien Poullot, "Ntt communication science laboratories and nii in trecvid 2010 instance search task," in *TRECVID'10*, 2010, pp. –1–1.

[13] Duy-Dinh Le, Cai-Zhi Zhu, Sebastien Poullot, Vu Q Lam, Vu H Nguyen, Nhan C Duong, Thanh D Ngo, Duc A Duong, and Shinichi Satoh, "National institute of informatics, japan at trecvid 2012," .

[14] Cai-Zhi Zhu and Shin'ichi Satoh, "Large vocabulary quantization for searching instances from videos," in *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, New York, NY, USA, 2012, ICMR '12, pp. 52:1–52:8, ACM.

[15] Yan Yang and Shinichi Satoh, "Efficient instance search from large video database via sparse filters in subspaces," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, 2013, pp. 3972–3976.

[16] Du Tran, Junsong Yuan, and David Forsyth, "Video event detection: From subvolume localization to spatiotemporal path search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 404–416, 2014.

[17] Du Tran and Junsong Yuan, "Optimal spatio-temporal path discovery for video event detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[18] Michal Perd'och, Ondrej Chum, and Jiri Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 9–16.

[19] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Computer Vision and Pattern Recognition*, 2012.

[20] Marius Muja and David G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISSAPP'09)*. 2009, pp. 331–340, INSTICC Press.

[21] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and Jing Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 319–336, Feb 2009.