Exemplar-Free Video Retrieval

Phani Krishna Uppala

CMU-RI-TR-YY-NN

July 27, 2021

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Deva Ramanan, *Chair*
Aswin Sankaranarayanan, *Chair*
David Held
Achal Dave

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.*

Copyright © 2021 Phani Krishna Uppala. All rights reserved.

Abstract

Video retrieval of activities has a wide range of applications. In the traditional mode of operation, a collection of example videos describing the activities are given and the retrieval technique identifies other samples in a dataset that semantically match the examples provided. However, retrieval using a collection of example videos might not always be feasible, especially in the following two scenarios. The first scenario is when we only have a textual description of a class of videos. The second scenario occurs when the activities under consideration are not temporally localized, making them harder to collect and annotate. For instance, most commonly-used action recognition datasets like Kinetics exploit public sources of videos like youtube for data collection, where all the categories are well localized and can be easily searched and annotated. This strategy does not extend to more complex activities like theft and object abandonment, both of which are not temporally localized, and are hard to annotate.

In this thesis, we describe two video retrieval approaches that work in the absence of visual examples. First, a text based retrieval approach, where a text query allows us to bypass the use of a visual exemplar. Also text embedding models like GPT-2/GPT-3[19] are not trained in a dataset specific manner, ie ... they are trained on all available data on the internet, and contain generalizable knowledge of all the activities in the real world. We will leverage that for developing retrieval models that work in the zeroshot/surprise setup. Since surprise activities are not known during the training time, the activity description/activity name is used during the test time to construct a textual embedding. First proposals are extracted from the video database using the TSM based model. For each proposal a visual embedding is computed. And similarity between video and textual embedding is used for retrieval.

The second approach that we consider is a rule-based unsupervised retrieval framework for categories specific to object transfer. This works by first detecting the objects and persons on a frame by frame basis. Followed by constructing short high-confidence tracklets. These tracklets are further connected in a soft fashion, where each tracklet can be associated with other tracklets such that the cumulative probability of 1. For the soft tracking based method, an annotation pipeline is built that facilities fast annotation of tracks. This works by assuming the high confidence tracklets are readily available, which can be achieved by using a high

association threshold on the existing tracking algorithms. Then the annotation platform only requires user input to map tracklets among each other.

The two approaches discussed successfully avoid using visual exemplars, thereby also avoid all the short comings and restrictions of needing visual exemplars. This demonstrated the plausibility and the effectiveness of exemplar free approaches.

Acknowledgments

I would like to thank my advisors, Deva and Aswin, for their guidance in the past two years.This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of IARPA, DOI/IBC or the U.S. Government.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

Video retrieval, the task of retrieving the desired clip from the database of videos, is becoming increasingly popular. Two factors are fueling this interest. First video form of data is increasingly ubiquitous, led by the wide availability of high resolution cameras capable of recording the videos with ease. Second, deep learning based video understanding approaches becoming more and more accurate. For the first time automated processing of video data to get information is becoming viable. This automated approach towards video processing is essential as the video form of data is often temporally large, making human processing of each video practically infeasible. ie... An hour of video data requires a human annotator to parse through the entire one hour of footage, sometimes frame by frame. Paired with high cost of processing, some video collection applications like security cameras are run all day. Making the requirement for automated processing extremely important.

1.1 Example based retrieval

Most common approach towards video retrieval is exemplar based video retrieval. In this approach an example video is given as input to the retrieval approach, which in turn guides the retrieval of all the videos that semantically match it. This approach is extremely useful and has seen much progress in recent years. However this approach can be limiting in the following cases.

- When there is a requirement for the retrieval of a specific description, that does not fit into a predefined set of categories. For example, consider this real life scenario, when there is a few days worth of security camera footage and a security analyst wants to retrieve all the instances of “Tall man in a white shirt standing next to a red pick up truck”. There are no existing datasets where the above query falls into one of the category labels. Hence it’s not possible to use an existing exemplar for this retrieval. At the same time, it’s difficult to craft a new exemplar through video capture for queries as specific as above.

- When retrieval needs to be performed on the fly, there wouldn’t be enough time to find or craft an example. Especially for large volume applications. Consider YouTube, it is a commonly used web application, where text query is used to retrieve the similar videos. Here the tags attached to the videos are used to match for similarity against the text query. There are two main shortcomings with this approach. First, relying on tags does not have a semantic understanding of the video content. Hence retrieving sub portions in a long video is not feasible. Second, most real life applications like security camera recordings, tags won’t be available. Only method to avoid all the above mentioned shortcomings is to formulate a retrieval approach that works on the fly, without relying on tags.

- When the category of the query is ill posed or there is a large variance in the samples defined by the category. For instance, consider a category object abandonment. It is one of the categories of the MEVA video action recognition dataset[6]. Between different samples of this category there is a large amount of time variance, more specifically between when a person puts down the object and when the person leaves the field of view of the camera. Hence one example is not the best approach to capturing the essence of the category. Also collecting and aggregating multiple examples can be more costly and time consuming. So we must find an effective way to query, that presents the variance of the category. Which in turn helps retrieve all the required samples.

To overcome the above limitations, this thesis formulates two approaches to perform retrieval in an exemplar free manner. First approach uses text as a method for query instead of an exemplar. Whereas the second approach relies on the rules that define the activity, here on to be referred to as the ‘rule based retrieval approach‘.

1.2 Text based retrieval approach

- This allows for input query description to be as specific as required, allowing the algorithm to only retrieve the desired clips.
- Text input can be provided on the fly, allowing for the scalability and usage in high volume applications. As opposed to exemplar based approaches, where either an existing example video clip must be found or newly recorded. Web applications like YouTube take advantage of this to handle the large amount of queries per second.
- By having input as textual modality, allows us to take advantage of the deep learning based textual models. Deep learning based language models have seen ground breaking improvements in recent years. Textual models like GPT-2[19] are trained on multiple datasets, often to the extent of all the available text data on the internet. Leading to a good generalization capability.
- As a simple overview of this approach, textual query is used as input to GPT-2[19] textual embedding generation approach. From the database of videos on which retrieval is performed, a feature extraction is performed for each video. This feature is a representative of the video, which is projected into the same embedding space as that of the textual embedding. Since the textual features and video features in a joint embedding space, the similarity between them can be easily measured through a dot product. And this dot product similarity is used to retrieve the results that are most similar to the query.

1.3 Rule based retrieval approach

- We humans can recognize activities without ever having to see an example of that activity, merely by processing the set of rules that describes the activity. Taking the same example of object abandonment from before, we can understand the sign at an airport about not leaving our belongings. Even if we haven't seen a visual example of a person leaving their belongings.
- This approach allows us to bypass dealing with the complexities of variance in the examples in the desired categories. By focusing on the temporal key points that discriminatively define the given category.
- For this approach, existing object detection and tracking algorithms are used for scene understanding. First given video is processed frame by frame to detect all the persons and objects in the video. The persons and objects are formulated as nodes in the graph on a frame by frame basis. Maintaining identity of persons and objects across time is crucial to make activity related inferences. To do this, previously constructed person-object graphs are temporally connected using soft tracking scores. This spatio-temporal graph is later parsed with the rules of the desired category for retrieval.

Chapter 2

Background

The retrieval approaches presented have multiple stages and use different areas of research in computer vision and natural language processing. Following sections discuss the closely related areas in the context of this work.

2.1 Object detection

For any given image, object detection finds instances of all the objects from a predefined set of categories. Object detection is one of the most thoroughly studied tasks in computer vision, especially deep learning based detectors have recently shown rapid improvements. Two main streams of approaches have evolved, first proposal based approaches like faster-rcnn[12]. Second, proposal free approaches like yolo[21].

2.1.1 Proposal based approaches

Proposal based approaches work in a two stage manner. In the first stage object proposals are generated. This backbone network is usually pre-trained on the imagenet dataset[23], which extracts generic features from the input image. Some of the proposal based approaches use multi-scale feature aggregation, this is done by using feature pyramid network(FPN)[16]. This FPN works by taking the features at multiple layers from the backbone network, and combines t. This allows to detect objects of different sizes, which are captured at different layers in the backbone network to be aggregated into a single feature. This single feature is passed on to the next stage. The second stage consists of using detection head and classification heads. The detection head works by taking in the region proposals generated and fine tunes the boundaries of the detection box. Where as the classification head works to predict the class corresponding to each region proposal.

2.1.2 Proposal free approaches

Single shot detectors(SSD) and YOLO fall into the category of proposal free approaches. They have the advantage of a simpler architecture, without having to perform proposal extraction followed by detection and classification. These approaches are tailored for speed, but falls short in terms of AP compared to the proposal based approaches. Recent approaches like RetinaNet[17] have bought renewed interest in the proposal free approaches. RetinaNet deals with the low performance in accuracy by using the improved loss function, ie... focal loss. Focal loss takes into account the large amount of false negatives coming from background in the object detection. These false negatives are easy to classify and form large fraction of the total loss. It effectively mitigates the overwhelming effect of background class by using a reweighting scheme.

In this work, object detection is used as the first stage in the rule based retrieval pipeline. Raw videos are processed to obtain person and object bounding boxes. As the focus of this work is on using the object information for further inferences. Existing state of the art model at the beginning of the work, hybrid task cascade[4], a variant of the mask rcnn is used.

2.2 Object tracking

Object tracking works to maintain the identity of the object/objects across frames. There are two main directions of research in object tracking, single object tracking and multi object tracking.

2.2.1 Single object tracking

Single object tracking works by tracking a single object across the video. Usually single object trackers works directly on the image, and do not expect frame by frame bounding boxes corresponding to the object in the image. Correlation filter based approaches have been dominant in single object tracking for many years. This is a fast and effective algorithm, which works to discriminate between the template and the target. It has inherent transnational in variance makes it works under 2D translations. These correlation filters have been iteratively improved by using deep features [14].

With the advent of Siamese networks[22], a new direction of single object tracking has evolved. Instead of correlation filter based discrimination, this approach uses the feature learning capability of deep networks. Siamese network, which works as similarity function is learned offline, and during the test time discriminates between the template and the target. Usually the bounding box crop of the desired object in the first frame is provided as the template. Following which the Siamese network find the target on a frame by frame basis. Many training techniques have been developed to improve this approach, like taking the template of non first frames as a way to augment training.

Despite of all the improvements, this approach a very visible failure modes, in case of deformable objects and partial occlusions. To overcome this more recently an approach[2] is proposed to leverage the pretrained object detection models like mask-rcnn[12]. Size of the object tracking datasets is usually orders of magnitude smaller than of the object detection datasets. This can be attributed to high cost of annotation of tracking compared to the object detection. Hence using the pretrained backbone from object detection allows to address some of the above mentioned failure modes. Specifically to detect partially occluded objects and maintain similarity between the template and target across mild deformations. Since mask-rcnn is leveraged tracking methods have incorporated tracking along with segmentation. This has also given rise to new field of study known as video object segmentation.

2.2.2 Multi object tracking

In wide range of real life scenarios, it's not often there is a single person/object in the video. Tracking each object separately would incur a huge time complexity, especially for the cases like subway stations or airports that are dense with people. This branch of research address this issue by tracking all the objects and persons in the video feed at the same time. Most approaches for multi object tracking follow the tracking by detection methodology. ie... they except videos to be pre-processed and compute the bounding boxes corresponding to all objects. Once the bounding boxes are computed on a frame by frame basis, they are passed as input to the tracking algorithm. One of the bench marking papers in this area is SORT[3], Simple online realtime tracking. SORT looks at tracking as a dynamic state estimation problem. And uses kalman filters to formulate the state estimation problem. The bounding boxes in each frame, along with the memory accumulated from the previous frame form the states in the kalman filter. Then kalman filter is used to predict the future positions of the bounding boxes. Finally the predictions are associated with bounding boxes available for the next frame. This approach works really well for the simple cases and is extremely fast. However the semantic understanding is lost and abstracted away with the bounding boxes, leading to following failure modes. Object swapping is the most common failure case, this is excepted as the notion of identity is lost. Under occlusions the loss of track is also observed, this can be attributed to loss of identity as well.

To overcome the above limitation Deep SORT [25] is proposed. This work uses a pretrained deep network as a feature extractor, by using localized bounding boxes to extract features and incorporating that into the state of the kalman filter. Many improvements over this have been proposed, different backbones, using mask-rcnn as a backbone. Post processing approaches like merging the tracks[24] to improve the accuracy have also been proposed. This merge happens by using the backbone features and the IoU overlap. Some of the approaches also use epipolar geometry to bring depth consistency, which has been shown to improve accuracy. Recent approaches also attempted to merge the detection step and tracking step, to re utilize the features and reduce the compute. [18].

2.3 Tracking Annotation

Rule based retrieval approach proposed in this thesis is evaluated on the MEVA [6] dataset. This dataset released by IARPA has person abandons object category, which fits well for the rule based retrieval. However no annotations are released as part of the dataset. For this reason, this thesis develops a annotation approach specific to annotation of object abandonment. Annotation is at the core of recent advances in computer vision. Many annotations tools have been built to facilitate this. For object detection [7] web interfaces, that works through direct upload have also been developed. These tools [8] have also incorporated semantic and instance segmentation tools. However for videos tools are still not comprehensive, especially to maintain the identity of person/object across the video is still done in a frame by frame basis.

The approach proposed takes advantage of the high confidence tracklets, obtained from the tracking algorithm. Tracking algorithm is tuned to produce short but very high confidence tracks called tracklets. And the human annotation is leveraged only towards joining these high confidence tracklets.

2.4 Language modeling

In the recent approaches language models are often trained on all the available data on the internet, making them best at having generalization capability. This can be evidenced by the GPT-2[19], being used in real world chat bots with free form text input. By leveraging this strong generalization capability, recent works like CLIP [9] are learning a joint embedding space for both textual and visual data. This work uses pretrained models and concepts like joint embedding space from CLIP.

2.4.1 Video text retrieval

Cross modal learning [13] has been the standard approach towards video text retrieval. However with the zero shot learning made viable by better pretrained models has started a new improved direction. Transformer based backbone [1] have been proposed to the multimodal learning required for video text retrieval.

2.5 Video understanding

2.5.1 Action recognition

Action recognition is one of the actively researched areas in computer vision. This has led to collection of large sized datasets like kinetics, something-something , AVA etc. Most of the work is focused on classifying the actions given temporal and spatial localized clips as an input. A different stream of research that works on non localized input is also being explored. This works by first generating proposals from the input videos and later classifying the proposals. Rule based activity retrieval proposed focuses on a different way of exploring video understanding, by looking at video as a continuous path in an image embedding space.

Two stream based approaches [9] that combine appearance and motion are extensively explored. In these approaches appearance information is obtained through the rgb image, and the motion information is captured through optical flow. The features extracted through the two branches are fused in various ways like early fusion, late fusion, cross talk etc...

2.6 Open set

Open set Deep neural networks have made breakthroughs in a wide range of visual understanding tasks. A typical challenge that hinders their real-world applications is that unknown samples may be fed into the system during the inference phase, but traditional deep neural networks or SOTA computer vision methods will wrongly recognize these unknown samples as one of the known classes. Open set recognition (OSR) is a potential solution to overcome this problem, where the open set classifier should have the flexibility to handle unknown samples and meanwhile maintain high classification accuracy in known classes. Consequently, there has been a lot of relevant prior work in this area. Open Set Recognition with Conditional Probabilistic Generative Models [23] proposes an open set classifier which has the flexibility to reject unknown samples posed to the model at test time. In this paper, a novel framework, called Conditional Probabilistic Generative Models (CPGM), for open set recognition is proposed. The core insight of this work is to add discriminative information into the probabilistic generative models, such that the proposed models can not only detect unknown samples but also classify known classes by forcing different latent features to approximate conditional Gaussian distributions.

Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition [9] introduces a probabilistic approach to unify open set recognition with the prevention of catastrophic forgetting in deep continual learning, based on variational Bayesian inference. In order to successfully distinguish unseen unknown data from trained known tasks, the paper proposes to bound the class specific approximate posterior by fitting regions of high density on the basis of correctly classified data points. These bounds are further used to significantly alleviate catastrophic forgetting by avoiding samples from low density areas in generative replay. There is much more prior work which involves rejecting or avoiding the unseen or unknown input samples (that are out of the training data distribution or are from a different output class category) at inference time.

We aim to propose an approach where we perform an Open Set Recognition (OSR) on the unknown or unseen samples but instead of avoiding or rejecting them, we find a way to automatically handle these samples at the inference time. In order to explore this approach more in the open set conditions, we also aim to perform experiments for Video Retrieval from textual input using an image based CLIP model, but on a Video input dataset.

Chapter 3

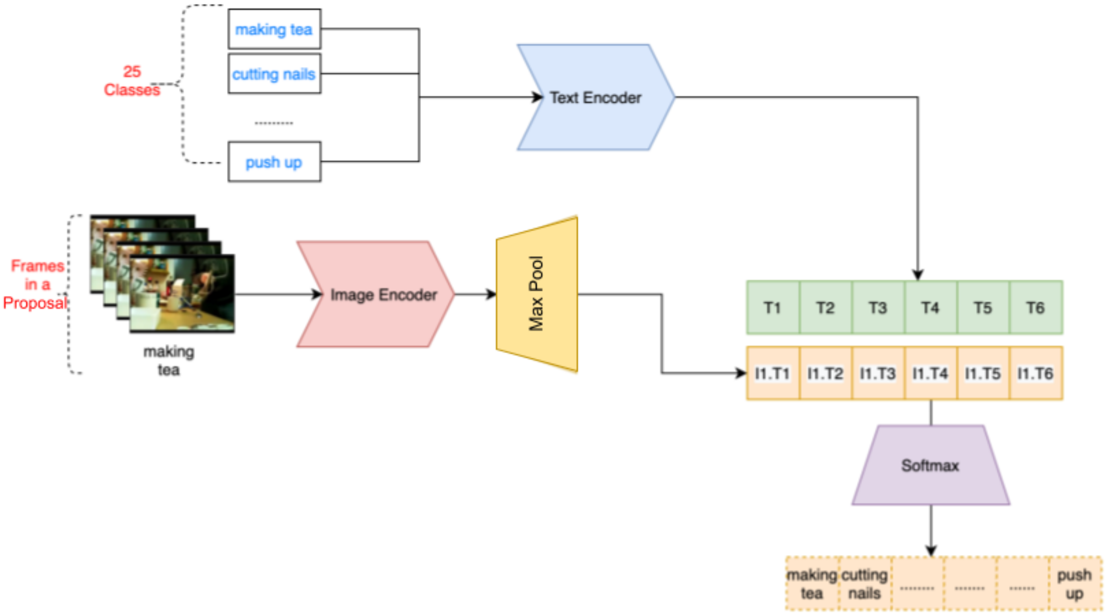
Approach - Text based activity retrieval

Given a set of text queries, $\boldsymbol{T}$, a set of videos $\boldsymbol{V}$, and integer value $\boldsymbol{n}$. We want to learn a mapping that retrieves top $\boldsymbol{n}$ segments of the videos for each of the text query $\boldsymbol{t_j} \in \boldsymbol{T}$.

Each video $\boldsymbol{v_i} \in \boldsymbol{V}$ can be arbitrary length, and the activity are embedded in the video. The start and end times of the activity not necessary and seldom coincide with the start and end times of the video $\boldsymbol{v_i}$. This assumption is crucial to as most of videos, like security camera footage are long form and are not trimmed to localize the activity. The above mentioned 'segments of the video' refers to the any sub part of the video $\boldsymbol{v_i}$ that contains the desired activity.

The first step in the retrieval pipeline takes in raw form videos $\boldsymbol{V}$ and extracts the activity proposals from each video $\boldsymbol{v_i}$. The following section 3.2 further elaborates the proposal generation. Following proposal generation, set of text queries $\boldsymbol{T}$ are converted into their embedding representation using the language model. This is further elaborated in section 3.1. Along with text queries, the proposals extracted from the raw form videos are projected into the same embedding space using the video/image model. This step is discussed in section 3.3. Finally for each text query $\boldsymbol{t_i}$ the semantically closer segments of the videos are retrieved using the embeddings generated in the above step, as discussed in 3.4

Figure 3.1: Pipeline for text based activity retrieval, showing text encoder converting activity descriptions into embeddings. Image encoder converting proposals to image embeddings. Followed by similarity measurement to perform retrieval.



3.1 Text embedding generation.

Embedding generation approach takes in a activity description $\boldsymbol{t_j} \in \boldsymbol{T}$ and converts into a vector of size 1 x k. Since we want to capture different semantic aspects of the text query, A sentence embedding model is used. Latest state of the art language model GPT-2 is really good at capturing this information. GPT-2 language model provided through CLIP [20] is used. Prompt engineering is a especially useful technique that is shown by various methods [19, 19] to help generate better embedding representation. Prompt engineering works by formulating various templates, with each of the templates having a empty placeholder that takes in the activity description to form a modified activity description. Set of modified activity descriptions are accumulated. Each of the modified activity descriptions are passed through the language model and averaged to get the prompt engineered embedding representation. The prompt engineered embedding is further used in conjunction with visual embedding as detailed in section 3.4.

3.2 Proposal generation

Since we are trying to retrieve from free form videos, it is essential to get the temporally localized snippets of activities from the videos. Consider security camera footage most of the time there is no activity happening. Unnecessary processing of the empty footage will lead to increase in compute. Hence it makes for a computationally effective approach to first extract proposals. To do this we use the I3D based [10, 11] proposal generation approach. This generates generic proposals our all activities. Apart from making compute efficient proposal generation also helps make visual embedding more noise-free. In other words visual embedding learned using localized proposals compared to large video snippet would capture more semantic information of the activity and less background and camera movements. For each video, proposal generation approach individually extracts the proposals for each of the video. These proposals are further used in section 3.4

3.3 Video embedding generation

The proposal generation approach generates a set of proposal parameters $\boldsymbol{P_i} \in \boldsymbol{P}$ for each of the video $\boldsymbol{v_i}$. The proposal parameters contains $\boldsymbol{v_i}$, $\boldsymbol{t_s}$, $\boldsymbol{t_e}$, $\boldsymbol{\{x_{i1}, x_{i2}, x_{i3}, x_{i4}\}}$ for i $\in$ $\boldsymbol{\{t_s ... t_e\}}$. For each time stamp $\boldsymbol{t_i}$ between $\boldsymbol{t_s}$ to $\boldsymbol{t_e}$. We have a bounding box that when cropped from the corresponding time stamped frame in the video.

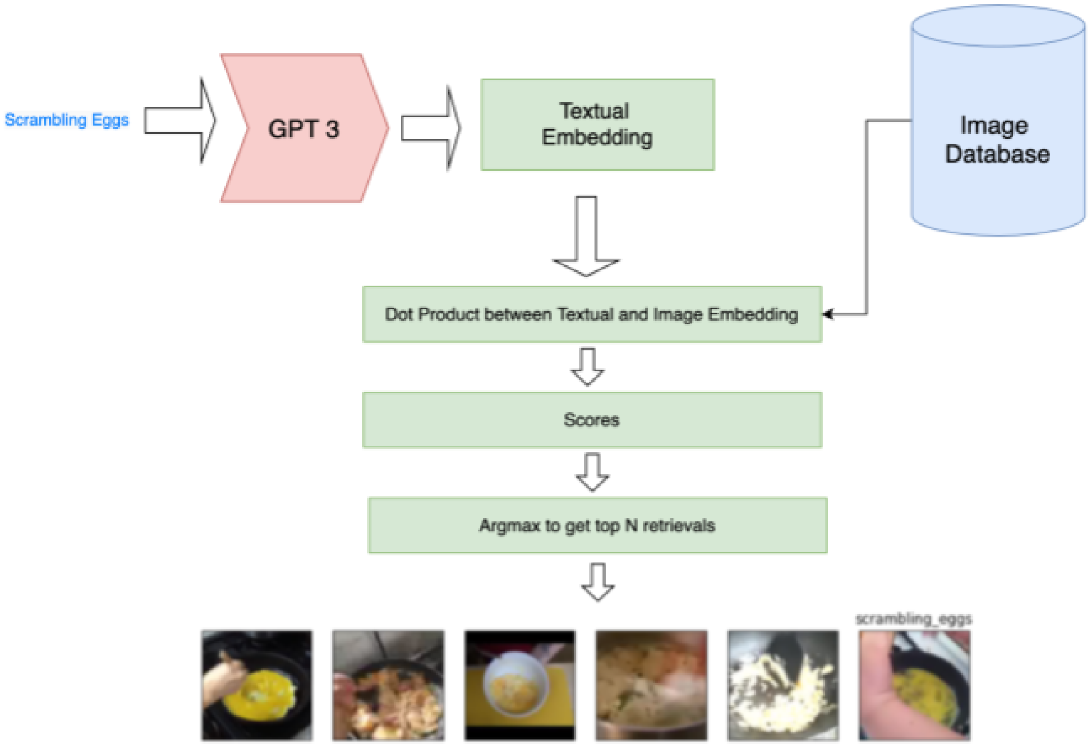
When cropped on a frame by frame basis, we have set of bounding box crops capturing a potential activity. For each set of bounding box, we do the following preprocessing before using a visual model to extract features. Normalization of the image crop, by using subtraction with values (0.48145466, 0.4578275, 0.40821073). Followed by variance normalization using standard deviation as (0.26862954, 0.26130258, 0.27577711). Each normalized image crop is passed through the visual model, RN-50. This pretrained using contrastive learning [20] and is learned to project the visual input to a joint embedding space as that of text. The center 5 with a sub sampling factor are converted into embeddings, followed by max-pooling to get the video embedding representation. This embedding is used in the section 3.4 for retrieval.

3.4 Retrieval using the embeddings.

This is the last step in our pipeline for retrieval. We use the texual and visual embeddings generated in the last stages to perform retrieval. For our given set of videos $\boldsymbol{V}$, we extract the proposals, compute the embedding for each proposal. These set of embeddings form the database, which only needs to be computed once. For any number of text queries, we can reuse the database at our disposal again and again. This also allows us to add more videos by just computing the embedding for new videos and adding them to the existing database. This would not effect the old data.

Since textual and visual embeddings are in a joint embedding space. The similarity can be computed by a simple dot product. So for a given text query the embedding $\boldsymbol{e_t}$ is computed. Then dot product between $\boldsymbol{e_t * e_{pi}}$ for all proposals $\boldsymbol{p_i}$ is computed. And embeddings with top k scores, their proposals and the corresponding videos are retrieved.

Figure 3.2: For a given text query, ”scrambling eggs”, all pre computed embeddings from database are taken to compute dot product similarity. The argmax over resultant scores is used for retrieval.


Scrambling Eggs
GPT 3
Textual Embedding
Image Database
Dot Product between Textual and Image Embedding
Scores
Argmax to get top N retrievals
scrambling_eggs

Chapter 4

Approach - Rule based activity retrieval

Given a set of activities, $\boldsymbol{A}$, and for each activity $\boldsymbol{a_i}$ there are a set of rules $\boldsymbol{R_i}$. There rules describe the activity as we would a novel activity to a human. For example, person abandons object would have the following set of rules.

- A person entering the camera field of view by possessing an object.
- The person disassociates himself from the object.
- Person does not comeback into possession of the object with in next 5 minutes.

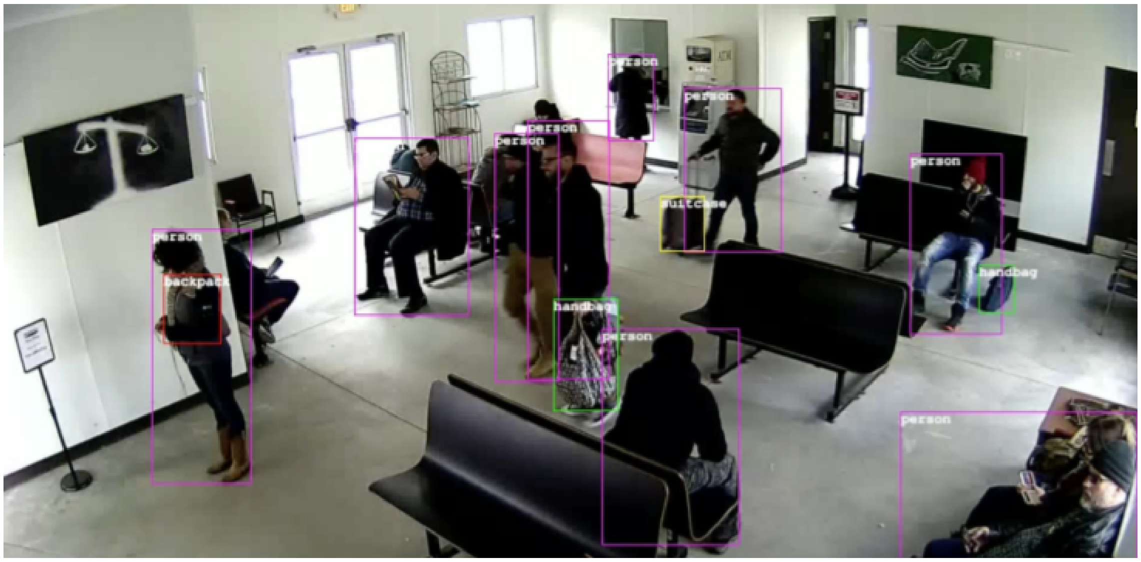
Along with the activities and rules describing them, we also have a set of videos $\boldsymbol{V}$, and integer value $\boldsymbol{n}$. We want to learn a mapping that retrieves top $\boldsymbol{n}$ segments of the videos for each of activity.

Similar to the approach in chapter 3, each video $\boldsymbol{v_i} \in \boldsymbol{V}$ can be arbitrary length, and the activity are embedded in the video. The start and end times of the activity not necessary and seldom coincide with the start and end times of the video $\boldsymbol{v_i}$.

4.1 Object detection.

Similar to the case in 3 the raw form videos are long, and mostly consists of no activity. In other words, desired activities are temporally sparse. Prepossessing these raw form videos to extract desired information, makes our algorithm more compute efficient. In this step we use the object detection method, hybrid task cascade [4, 5]. It is a variant of mask-rcnn model, one of the most robust object detection and segmentation model. The raw form videos are processed on a frame by frame basis to extract the object and person bounding boxes. These relevant categories from COCO [15], person, backpack, suitcase, handbag are selected for further processing. Backpack, suitcase and handbag categories are grouped into object category, while person category is separately processed. The bounding box detection's from hybrid task cascade are first denoised using a non maximal suppression step. Followed by thresholding person detections by 0.7, and object detections by 0.3. Figure 4.1 shows different object categories, before grouping in different colors. It also labels the object category in the form of overlay text. The following person and object detection bounding boxes along with their confidence scores are stored for further processing in the next stages.

Figure 4.1: Image from MEVA dataset, with bounding boxes showing the persons and object. Bounding boxes are color coded pre object grouping, person in marron, backpack in light green, suitcase in yellow.



4.2 High Confidence tracklet generation.

Most of the activity movements in videos are smooth, ie... the variation across frames is quite small. This translates directly to the bounding boxes from the previous stage. Except when there is a spatial crossover between two entities(persons or objects), the bounding boxes are isolated from each other and move in a tractable fashion. We take advantage of this predictability and group the bounding boxes of the same object that are temporally localized into one 'tracklet'. Each tracklet corresponds to a series of bounding boxes that belongs one object and whose identity is easy to maintain. SORT [25] a multi object tracking algorithm is used for this. SORT takes in set of bounding boxes for each frame and uses kalman filter based state prediction and association to form tracks. A very high association threshold is used, this ensures no false associations are made. This comes with a trade of some true associations are not being made. This is expected and will be taken care of in the next stages of the pipeline. Using the above mentioned steps a set of tracklets are generated for each video, these are further associated to construct identity graph in the next stage.

4.3 Soft tracking - constructing the identity graph

Multi object tracking approaches work by making associations between bounding boxes in different frames. These associations are take binary value, either 0 or 1. This hard association strategy followed by tracking approaches gives tracks. We formulate the approach where the association does not have to be binary. Intuition behind this approach, is to allow the tracking approach to capture the uncertainty in it's own association. This works by constructing a graph, termed as identity graph. The tracklets form the previous step form the nodes of this graph. The edges for this graph are populated to capture the instance identity. In other words, the tracklets that are likely to be of same identity will assigned a higher score. For each tracklet, it's end time stamp is taken and all the new tracklets that begin with in a time window are selected. Edges of the graph from the old tracklet to all the new tracklets are populated. Any given edge score is calculated using the dot product similarity between the visual embeddings of both tracklets. A visual embedding corresponding to the tracklet is computed as the average of visual embeddings for all the frames in the tracklet. This process is repeated for all the tracklets, in all the videos. Each identity graph generated per video is stored for retrieval as discussed in 4.4

4.4 Performing retrieval

The identity graph constructed in the section 4.3 is used in this stage for retrieval. An additional activity specific requirements are calculated. For example person abandons object activity requires person identity, object identity and person-object association. A simple spatio-temporal closeness is used as a metric to compute person-object association. Each path traversal in the identity graph represents a potential track of an person/object. And the score obtained by multiplying all the edge scores in this path form the liklihood that path is valid. For any given set of rules all the paths are ranked based on the score for that set of rules. The highest scoring paths are retrieved as the desired activity.

Chapter 5

Results

MEVA dataset is used to experimentally evaluate the rule based and text based retrieval approaches. For fair comparison metrics used for exemplar based approach are reported. [11] are used.

For the text based retrieval approach mean-nAUDC@0.2tfa of 0.8651 and mean-pmiss@0.04tfa of 0.8929 are achieved. For reference, exemplar based retrieval [11] achieved mean-nAUDC@0.2tfa of 0.76 and mean-pmiss@0.04tfa 0.65 respectively. Qualitative reuslts are shown in the figures 5.1 5.2

Method	mean-nAUDC@0.2tfa	mean-pmiss@0.04tfa
Text based	0.86	0.89
Exemplar based(I3D)	0.76	0.65

A total of 52 abandoned objects are annotated subset of the MEVA dataset, which contains 3 ground truth abandoned objects. Out of 52 retrievals 3 are true positives and 49 false positives. The method has high false alarm rate, this is expected as the supervision is very limited in the form of rules. However the approach was to all 3 of the true positives accurately.

Method	Retrievals	TP	FP	FN
Hard tracks	2140	0	2140	3
Soft tracks	52	3	49	0
GT tracks	3	3	0	0

Figure 5.1: Top 10 retrieved results for the categories specified on the right.

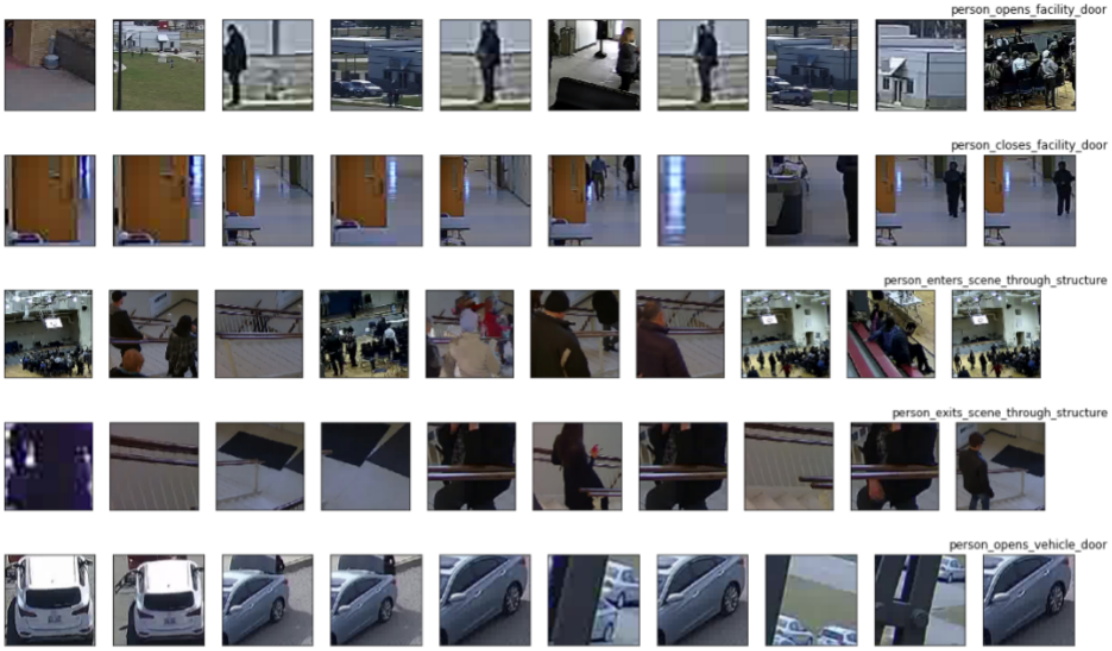
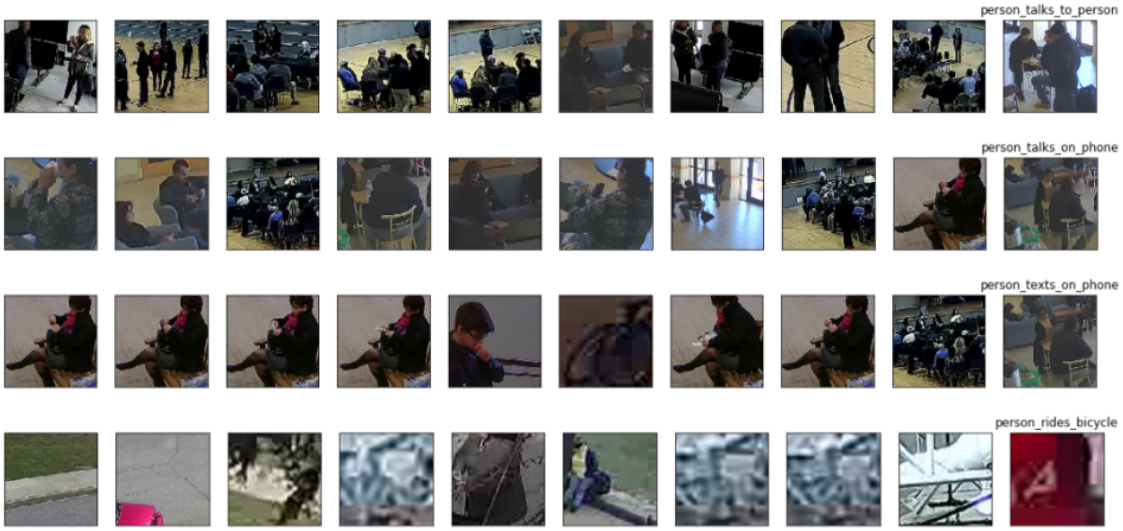

person_opens_facility_door
person_closes_facility_door
person_enters_scene_through_structure
person_exits_scene_through_structure
person_opens_vehicle_door

Figure 5.2: Top 10 retrieved results for the categories specified on the right.



Chapter 6

Conclusions

Exemplar free retrieval approaches discussed showed really promising results. This can be mainly attributed to well formulated text embeddings that have the capability to capture diverse aspects. The concept of joint embedding space is also crucial as it made possible to connect text and visual data in a seem less manner. Further improvements for this approach can come from exploring ways to formulate more effected joint embedding space. The rule based approach is effective at capturing desired activities from minimal supervision. However suffers from large amount of false positives, and this can be the place for large improvements.

Appendix A

Appendix

List of few figures that are used for visualization of the above approaches.

Figure A.1: Visualizing the networks attention map, obtained through backprop. (a) and (e) correspond to the rgb image input to the network. (b) and (f) are 7*7 attention heatmaps computed for words person and chair respectively. (c) and (g) are smoothed versions of (b) and (d). (d) and (h) overlay heatmaps on top of rgb image to show that person and chair heatmaps activate at spatially relevant locations.

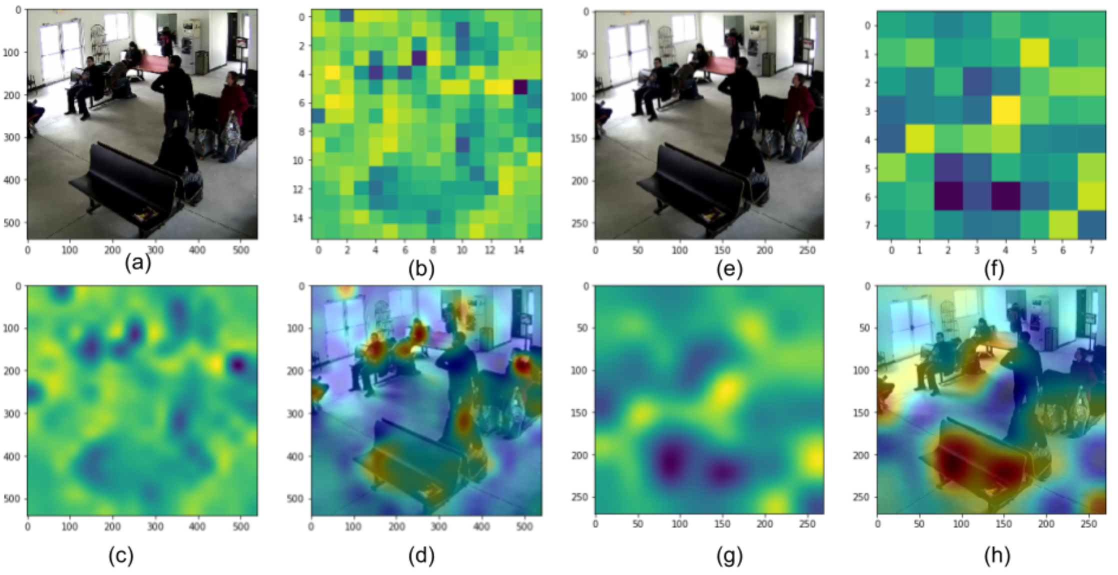


Figure A.2: Dot product similarity across 37 activities in MEVA dataset. We can observe that the similar activities have similar text embeddings. White color represents higher similarity.

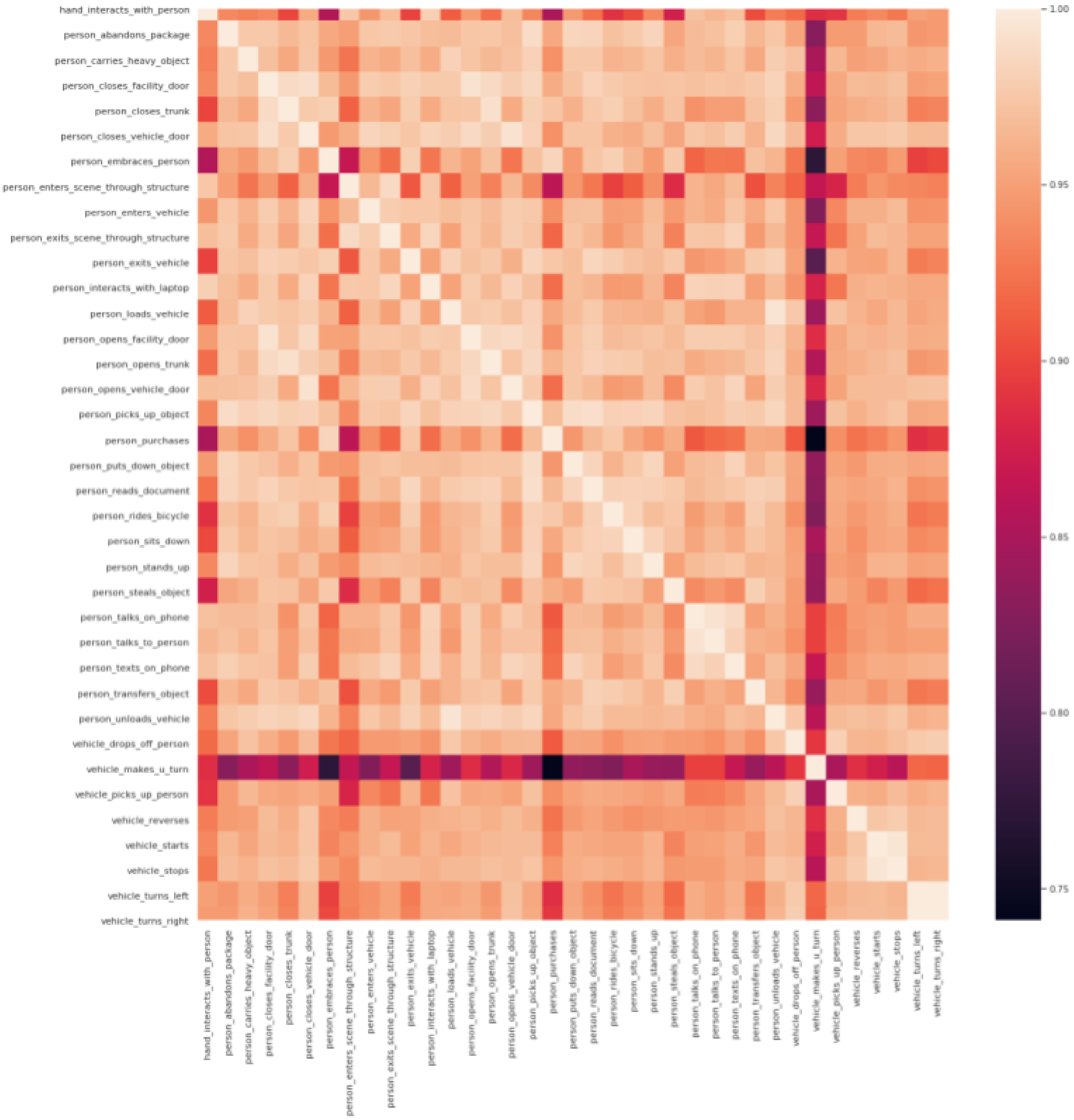
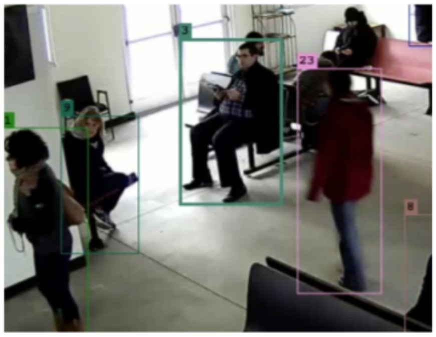
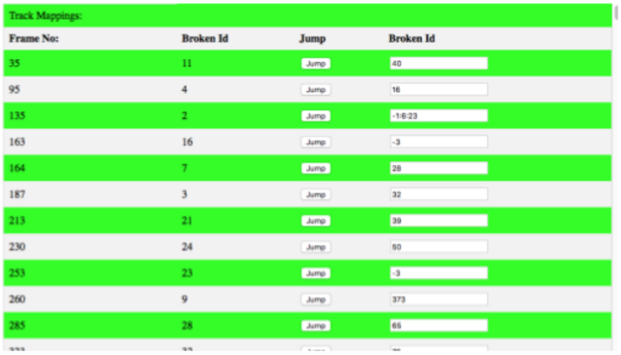


Figure A.3: Annotation pipeline devised for connecting tracks. Visual interface is design automatically jumps to the tracklet end temporal location. This alleviates the need for temporal parsing by human annotator.





Track Mappings:

Frame No:	Broken Id	Jump	Broken Id
35	11	Jump	40
95	4	Jump	16
135	2	Jump	-1:6:23
163	16	Jump	-3
164	7	Jump	28
187	3	Jump	32
213	21	Jump	39
230	24	Jump	60
253	23	Jump	-3
260	9	Jump	373
285	28	Jump	65

Bibliography

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2021. 2.4.1

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00103. URL http://dx.doi.org/10.1109/ICCV.2019.00103. 2.2.1

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, Sep 2016. doi: 10.1109/icip.2016.7533003. URL http://dx.doi.org/10.1109/ICIP.2016.7533003. 2.2.2

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. *CoRR*, abs/1901.07518, 2019. URL http://arxiv.org/abs/1901.07518. 2.1.2, 4.1

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4.1

[6] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021. 1.1, 2.3

[7] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). 2016. Version: X.Y.Z, Accessed: INSERT$_{D}ATE_{H}ERE.$2.3

[8] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-

6/19/10. doi: 10.1145/3343031.3350535. URL https://doi.org/10.1145/3343031.3350535. 2.3

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. URL http://arxiv.org/abs/1812.03982. 2.4, 2.5.1, 2.6

[10] Joshua Gleason, Rajeev Ranjan, Steven Schwarcz, Carlos Castillo, Jun-Cheng Chen, and Rama Chellappa. A proposal-based solution to spatio-temporal action detection in untrimmed videos. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 3.2

[11] Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Real-time detection of activities in untrimmed videos. In *The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2020. 3.2, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.322. URL http://dx.doi.org/10.1109/ICCV.2017.322. 2.1, 2.2.1

[13] Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. Temporal tessellation: A unified approach for video analysis, 2017. 2.4.1

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. 2.2.1

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4.1

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 2.1.1

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 2.1.2

[18] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking, 2020. 2.2.2

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. (document), 1.2, 2.4, 3.1

[20] 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, 2021. 3.1, 3.3

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 2.1

[22] Soumava Roy, Mehrtash Harandi, Richard Nock, and Richard Hartley. Siamese networks: The tale of two manifolds. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3046–3055, 2019. doi: 10.1109/ICCV.2019.00314. 2.2.1

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 2.1.1, 2.6

[24] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet, 2018. 2.2.2

[25] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. 2.2.2, 4.2