University of Sheffield

Tracking individuals across multiple scene and shots in TV dramas

Team Delta: Wonwoo Soh, Mingyan Zeng, Peng Xu, Yucheng He

Supervisor: Yoshi Gotoh

in the

Department of Computer Science

April 17, 2020

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: _____

Signature: _____

Date: _____

Abstract

This project mainly focus on tracking people in TV dramas. Tracking refers to identifying the people who appear first, and then if the people repeatedly appear, the model can recognize the people. At present, most research on tracking is focused on surveillance view. These studies often achieve excellent performance by tracking the whole body of people appears in surveillance cameras. However, the characteristics of characters appearing in TV dramas views are very different from those of surveillance views, so the tracking methods in surveillance views cannot be fully applied to data in TV drama views. This project analyzes the characteristics of people appearing in TV drama in detail, and removes useless information that will affect the tracking of characters. Because TV drama can contain enough facial information, the project will focus on tracking characters in TV dramas through the identity and recognition of their faces. This project divides tracking into three types of technologies: face detection, face extraction and face comparison, and introduce the structure that can track characters in TV dramas. It shows the demo implemented so far and plans to be implemented in the future.

Contents

List of Figures

Chapter 1

Introduction

1.1 Tracking

If you are fan of sci-fi genre or action-thriller movies, you can easily reminds the specific scenes when you hear about this paper's topic 'tracking individuals across multiple scene and shots in video'. It is recognizing individuals with names by their appearance and track them, sometimes even though they change their appearance, across the moving scene. It sounds unrealistic but the techniques are constantly developed since few decades ago and in some fields, not as much as what in the movie does, but in some degree those techniques are being used.

Before talk about this paper's topic, understanding the major keyword 'tracking' will be helpful to understand the concept of our paper. In European Commission in EUROSUR-2011[4], this documentation describes several concepts in surveillance, and two definitions of them could help us to define what is tracking in this project.

> Identify: to establish the unique identity of the object as a rule without prior knowledge.
> Recognize: to establish that a detected object is a specific predefined unique object.

After think about these definitions carefully, the definition of tracking in TV dramas could be explained. In general, it could be a recognition task, or more specifically, it lies between identification and recognition. For example, the project model takes a video as input. When people appear for the first time in the video, the model will identify the person and establish the unique label to this person, then when this person appears again in the video, the algorithm will recognize this person because the model has identified this person before.

1.2 Motivation and Research Questions

As said before, our topic has been researched by many researchers for many years. While we are researching about the topic, we found out that most works are focus on the surveillance,

the characteristics of Surveillance videos are very obvious. People appearing in surveillance views often include the entire body, and most of them are in a posture of moving or standing still. Most studies have been used to track people in different cameras, this happens in a short period of time (short time tracking), hence the appearance and clothing of the people do not change significantly. Although the details of people in views are unobvious, the resolution will cause the details of the character to be lost (face organ details, such as eyes, nose, etc., or hairstyles, facial expressions, etc.), but through the process of the entire body of the character, people Tracking in surveillance can achieve a good performance.

However, the characters in TV dramas are very different from the surveillance views. For example, the characters in TV dramas do not include the entire body, and the actions of the characters are not limited to standing still or walking. Due to changes in the plot, environmental factors may also affect tracking performance, such as illumination: overexposure in strong sunlight situation and underexposure in dark night situation, or a background that is highly coincident with the color of the character. The appearance of the character will also have obvious changes, such as the character's hairstyle, dress and body. Combining these reasons, we have raised a few questions and carried out research on this project around them.

- How to track individuals on TV drama series?
- How effectively the chosen approach performs on tracking individuals in different types of media, such as TV drama series?
- How to evaluate our works on tracking individuals in TV drama series?

1.3 Aims and objectives

More specifically, the intimate aim of this project is 'tracking individuals in video that containing dynamic scenes'. For more details, we aim to research the topic with few objectives. Throughout the research, we aim to determine how effectively the approach works in dynamic environment. Within that environment, how well the approach could identify the same individual or individuals from different scenes with different characteristics or changed appearances and how many individuals could be identified well with the approach in the scenes with obstructive factors. All these aims can be achieved by objectives of collect the good data set to use that provided from outside source, extract and edit them with specific tool for it, research the appropriate approaches and adapt them with our thoughts and ideas to the data set and finally, make comparison with our own standards. When we adapt the approaches, not only the existing ones, but only we modify them with our own ways to make the difference, which will ideally make the improvement in various measures.

1.4 Structure of the report

This report is divided into several chapters. Chapter 2 presents the initial topics that can be researched about the topic including data analysis. Chapter 3 presents the state of the art of the topic, which shows the literature reviews that are done. Chapter 4 presents the initial studies that our group made with any implementations or demos made. Chapter 5 shows the plan of work with estimation of time consumption and delegations of tasks to each team member. This will be presented by using Gantt chart. Finally, Chapter 6 states the conclusion of this interim paper.

Chapter 2

Initial topic

The initial topic is about the factors that are not directly related to our research topic, but related to it as external factors to consider. In this chapter, it split into two sub-sections, data analysis and tools to use. For data analysis section, we show where we could find the data set, how our data set is chosen with what standards, what the characteristics are and how this data set will be used in our project. For the tools to use section, we show what platforms and tools are going to be used, what are the characteristics and why they are required to use.

2.1 Data analysis

The data set that this project will use is from organization 'TRECVID', conference which sponsored by the National Institute of Standards and Technology and other U.S government agencies, provides sets of data for research purposes. From the conference, they provide different types of videos as data sets.

According to 'TRECVID', for the videos that they provide, there are 3 types. These figures are the example scenes of those type, but not exactly from what 'TRECVID' provides. They are to show those videos are similar to the examples. The first type is the collections of videos from various news magazine, science news, news reports, documentaries, educational programming, and archival videos. The approximate composition of scene for first type is similar to figure 2.1. Second type is the collections of videos from old TV episodes. The approximate composition of scene for second type is similar to figure 2.2. Finally, third type is the collections of videos from airport security cameras and activity detection. The approximate composition of scene for second type is similar to figure 2.3.

Within those three video sets, we chose to use the TV episodes as our data set to use after compare them. As mentioned in aim and objective section, our aim requires the scenes to be very dynamic, which means many things happen within the scene. As we can we in figure 2.1, the news and documentaries are very static. Those static scenes will be good for image recognition as very little changes made during the scene, but that is not what our project

Figure 2.1: Example of News magazine, science news, news reports, documentaries, educational programming, and archival video[1]

aims to. image recognition is only part of our project. For the airport CCTV, as we can see from figure 2.3, the scene is dynamic in some sense since people continuously move around and their behaviours are not identical. Another advantage is the CCTVs are widely spread in the airport, it's easy to follow and track certain individual by linking the scenes from different videos. However, many studies and researches already exists and some of techniques are actually implemented already in some fields. Furthermore, each CCTVs are fixed to view same place, the degree of angles of scenes are same and there are no changes of environment, so that the environments are not dynamic at all. Another consideration is the resolution of the videos. as it can be seen from figure 2.3, the resolution of scenes are not great that we cannot clearly view the face of each individuals. With considering all these factors, TV episodes were the best type to fit with our project as they have very dynamic scenes compare to other two. Within one episode, the environment changes by moving camera and viewpoint changes by switching it to other cameras in same scene. Objects move when interacted with the characters and sometimes they become obstacles to recognizing images. Furthermore, in whole series, the same character shows up in different episodes. Sometimes the appearances or outfits of the character change through out the series, which gives a challenge whether the tracking techniques can identify the character successfully or not.

The data set we got was used in 'TRECVID' in 2013. The title of TV series is 'EastEnders', a British soap opera which has beeen broadcast in BBC One since 1985. There are 244 video

Figure 2.2: TV Episodes[2]

files, which is about 300GB and 464 hours running time in total. Each video is a week;s worth and the format is MPEG-4/H.264. The truth table is provided in 'TRECVID' website. Since not all nearly 300GB of videos can be viewed and used for our project, we have decided to collect short scenes from the videos so that we can have approximately an one hour video with differently categorized situations. Collecting those scene are done manually by using video editing tool and also will be manually annotated in future to compare the output with the output produced by applying our project. The master shot reference table is also provided. Master shot means the shot from start to end of the scene before any transition of camera view is made and the characters shown in the scene change. This possibly can be used when annotating the scene manually or when analysing the data since each master shots always contains the same characters within the scene.

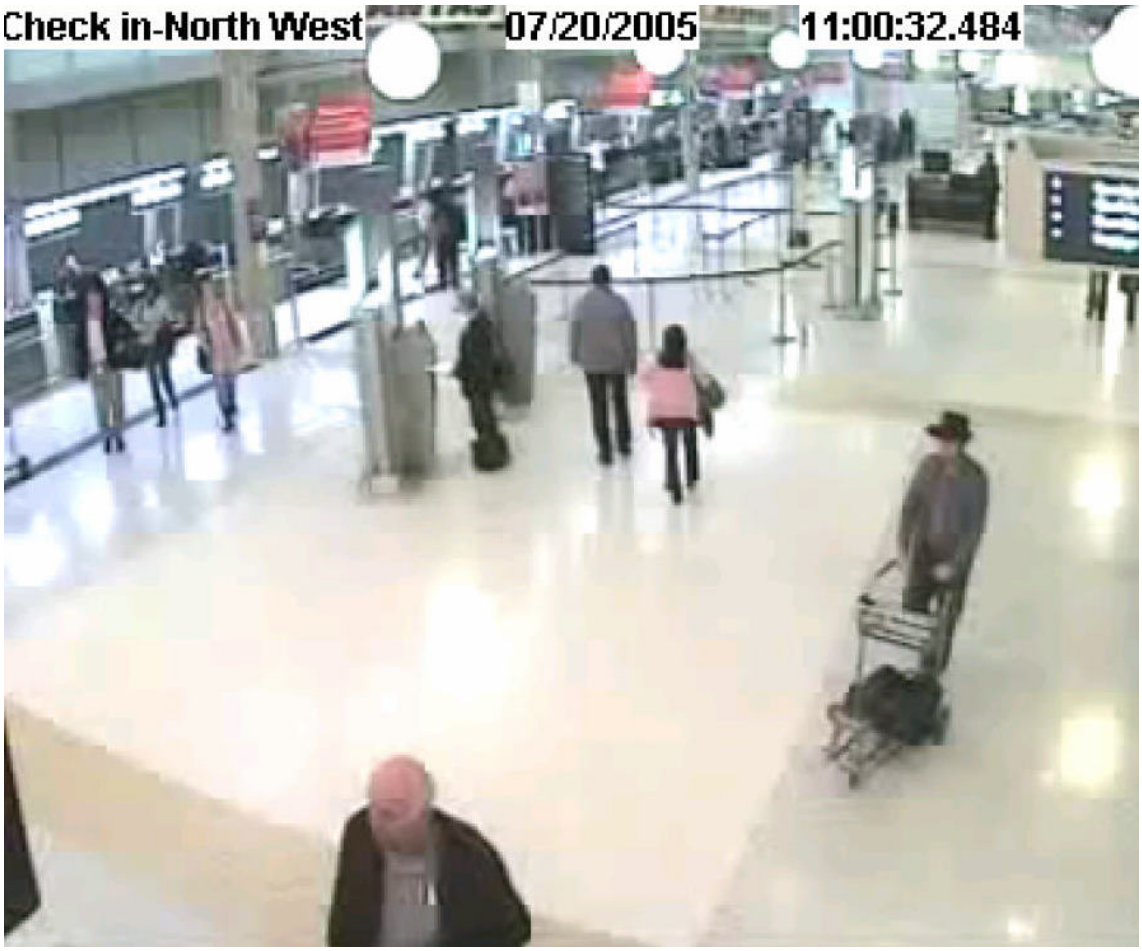

Figure 2.3: Example of Airport Security Cameras Activity Detection[3]

2.2 Tools for implementation

This project uses python 3.6[5] as the main language. Python3 provides a rich open source library of Computer Vision and Deep Learning, which makes implementing related methods and research easier. Other extension mainly technologies in python version involved are opencv[6], keras[7], tensorflow[8], dlib[9],etc.

Chapter 3

State of the art

This chapter mainly discuss relevant topics in the field of object detection and re-identification that is useful in this project. First, a evaluation of R-CNNs in the field of object detection are introduced. Second, a brief introduction of application of R-CNNs in face detection. Third, relevant researches on people re-indentification is discussed.

3.1 R-CNNs

Convolutional neural network(CNN) was heavily used for image classification, since Krizhevsky et al.[10] show a deeply trained CNN with higher image classification accuracy on the 2012 ImageNet large Scale Visual Recognition Challenge (ILSVRC). However, the application of CNN on field of object detection is empty during the time. SIFT[11] and HOG[12] is the solution for canonical visual recognition task, PASCAL VOC object detection, but the progress has been stalled during 2010-2012. Girshick et al. [13] proposed a region-based(R-CNN) which has higher object detection performance on PASCAL VOC challenge, compare to HOG-like approaches. R-CNN consists of 2 stages. In the first stage, use of selective search method to generate a set of region proposals[14] for the input image, which have been approved on object detection[15]. In the second state, the set of region proposals was transformed to bounding box and fed to a refined AlexNet[10] and then map to an SVM. The input image is classified by the AlexNet + SVM model and the bounding box of classified images is regressed to reduce the localization errors.

The contribution of the R-CNN is that it transforms CNN on classification tasks to the field of object detection with significant performance improvement on PASCAL VOC challenge. However, the efficiency of the R-CNN is unsatisfied due to (1) every proposals(approximately 2000 for each input image) are forward pass through the CNN in order to extract features, (2) CNN, SVM bounding and box regression is trained separately. To overcome this problem, Girshick [16] proposes the Fast R-CNN based on his previous research [13]. The approach run CNN exactly once for each input image. The fixed length feature map is extract from feature map of last convolutional layer and then feeding to region of interest (RoI) pooling

layer. The fast R-CNN jointly train the CNN, SVM and bounding box regression. Softmax is used instead of SVM as the classification layer of the CNN. A linear regression layer is added parallelly to the softmax layer to tighten bounding boxes. The performance of Fast-CNN on PASCAL VOC is increased from 53.7%[13] to 66%[16], and the speed is 9 times faster.

The first step of the object detection is to generate a set of potential bounding boxes or regions of interest. The proposals were generated by using a fairly slow hand-crafted model such as selective search [17]. The approach consists of two networks, the first, called Region proposal Network (RPN) for generating proposal that share convolutional layer with the second network which can be any other state-of-art object detection network[10][16] for refining proposals and classifying objects. By sharing the convolutional feature map, only one CNN needs to be trained, the computing cost for generating proposals is nearly cost-free(e.g., 1ms per image).

The RPN is kind of fully convolutional network (FCN)[18] and is designed for predict

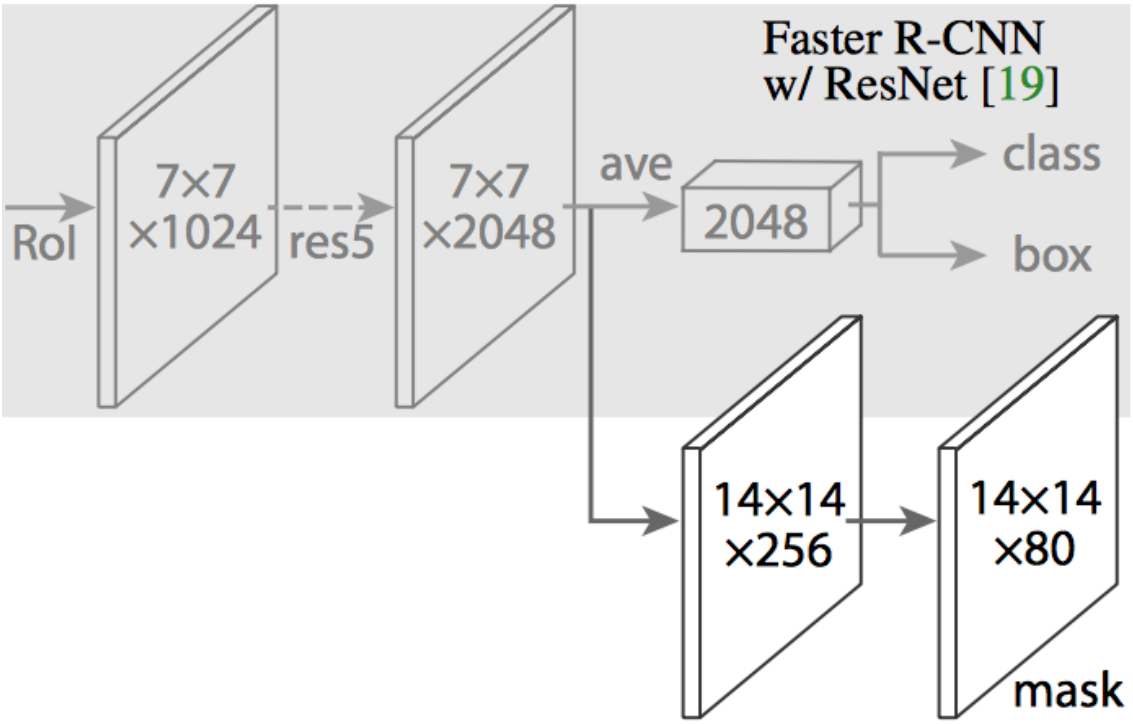

Figure 3.1: An FCN is added on the top of Faster-RCNN

proposals with a wide range of scales and aspect ratios. To deal with, they introduce anchor in RPN. An anchor is at center of each sliding window location on the shared convolutional feature maps and is scored for predicting how good is the anchor. Following the default setting of [9]. RPN generate 9 anchors with preset scales and aspect ratios. The 9 anchors contain tree scales(128 * 128, 256 * 256, 512*512), and each scale contains 3 aspect ratios(1:1, 1:2,

2:1). Therefore, the number of anchors is 9*W*H, if the size of shared feature map is W*H. For the generated anchor, the RPN, first determine whether the anchor is the foreground or the background which means to determine whether has covered the object. The second is to refine the coordinate of anchor which belong to foreground.

So far, we have reviewed use of CNN features to effectively locate different objects in an image with bounding boxes. [13][16][17]. Kaiming et al.[19] has extended the R-CNN-like approaches to go a step further by carrying out pixel level segmentation rather than just bounding boxes. Mask R-CNN extends Faster R-CNN by adding FCN [10] branch to output a binary mask for each RoI, which parallel to the existing object detection approaches(e.g., Faster R-CNN). The RoI Pooling layer extracts a small feature map from each RoI, and quantizes the floating-number RoI. The quantization misaligns between RoI and the extracted features which has negative impact on pixel level mask, while may not have effects the performance of classification. A refined RoI pooling layer is proposed to make the mask work as expected, called: RoIAlign. Instead of quantization, bilinear interpolation is used on RoIAlign to avoid misalignment.

Recent study on face detection has demonstrated impressive results by suing R-CNN-likes

Figure 3.2: The RHS represent the pixel level mask of the car(LHS), the quantization leads a misalignment

models. Huaizu and Erik[20] report state-of-the-art result on face detection by directly training Faster-RCNN on the large scale WIDER face dataset[21]. Xudong et al.[22] take the idea a step further by combining feature concatenation, hard negative mining, proper calibration of key parameters to Faster-RCNN. As the consequence, the approach was ranked as one of the best approach on the FDDB benchmark[23]. Coincidentally, Cakirouglu et al.[24] has applied Mask-RCNN to field of face detection. The Mask-RCNN is pre-trained on the face examples collected from PASCAL-VOC, and has been test on the WIDER face dataset.

3.2 Face Re-identification

The people re-identification research has grown rapidly over the past two decades, and most of these methods are based on camera setting, sample set, appearance-base, non appearance-based, and body model. The figure 3.3 shows the category of people re-identification research approaches. Most studies are based on the surveillance view. The characteristic of the surveillance view is that the people appearing mostly contain the whole body. Although it is impossible to distinguish some small features, such as the use of eyes, ears, nose and other features. People re-identification in surveillance view can still use body characteristics, such as body shape, face shape, hairstyle, and other macro characteristics for re-identification and even make 2d or 3d models to complete people re-identification objective. However, due to the limitations of TV drama itself, shapes of the characters' bodies are often inconsistent, even only heads of characters appear. Therefore, researching the face re-identification area can better achieve the requirements of this project.

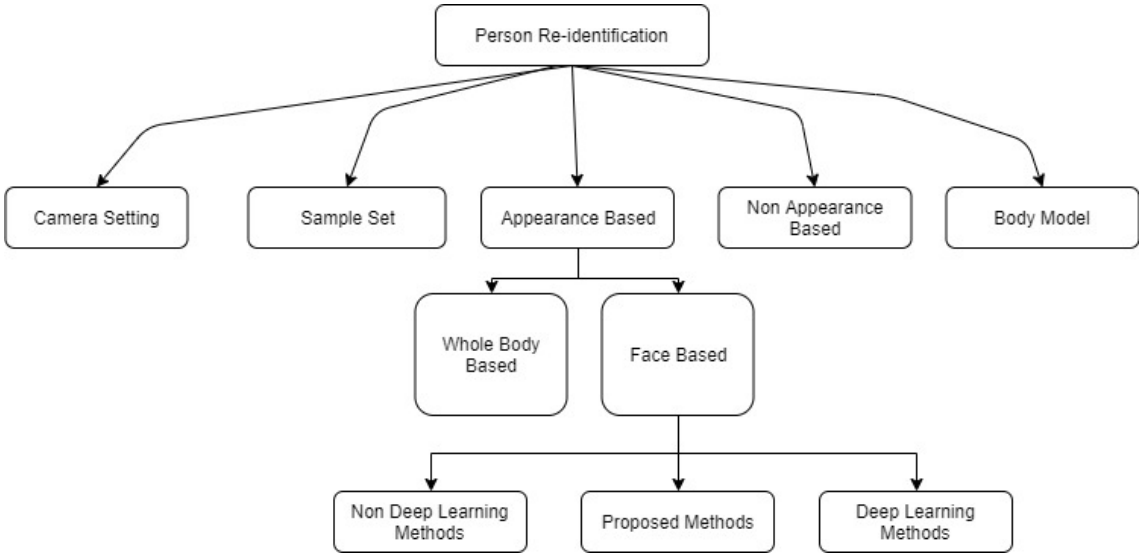

Figure 3.3: People Re-identification

Facial landmarks is an approach for locating facial structures, capable of presenting the eyes, eyebrows, nose, mouth, and jaw of a human face. This technique has been successfully applied in face re-identification with an outstanding performance. Facial landmarks is a subset of Shape prediction problem. This technique requires the face ROI image as an input (such as the bounding box of the face) to locate all key points (face structure) of the corresponding face shape. There are currently two methods of facial landmarks problem that are widely used. One is the Millisecond Face Alignment with an Ensemble of Regression Trees facial landmark detector proposed by Kazemi and Sullivan[25]. This method uses labeled training data, where the label includes specific 2d coordinates for each face structure. By using these training data, a regression face landmark model will be trained, and this model has excellent

performance in predicting facial landmarks. The other method is the supervised descent method proposed by Xiong X and De la Torre F. [26] to obtain a facial landmarks detector. This is a widely used high performance regression approach. However, this method has two main disadvantages.Jianwen Lou et al.[27] pointed out these disadvantages. One is that this method is highly dependent on the local optimal algorithm. In some cases, this algorithm cannot find the best local optimal value. Another disadvantage is that during the learning process of the algorithm, it may learn in a contradictory descent direction, which will also cause the algorithm to fail to obtain the best performance. Therefore, there are many Facial landmarks detectors derived from SDM.

Deep face feature is another main approach to achieve people re-identification objective. With AlexNet[28] winning first place on ImageNet in 2012, face detection and face recognition objectives uses deep learning approaches have exploded in recent years. Deep face feature approach uses a hierarchical structure method to analyze the shape of input face, illumination conditions, and expression of the face at different levels, which greatly improves the effects of face detection and recognition. A large number of methods explore better effects by combining different network architectures and loss functions. According to the comparison and analysis of Deep face feature methods by Mei Wang and Weihong Deng[29] in 2019, The method proposed by Jingtuo Liu et al.[30] in 2015, combined with CNN-9 neural network architecture and triplet loss function for learning, achieved a performance of 99.77%. Rajeev Ranjan et al[31] published the L2-performance method uses ResNet-101 neural network architecture and L2-Softmax loss function and achieved a score of 99.78%. In 2018, Jiankang Deng et al.[32] made an Arcface architecture used the ResNet-100 neural network architecture and arcface loss function to achieve 99.83%.

Chapter 4

Initial studies

For this project, the characteristics of the characters in TV dramas people re-identification are different from the characteristics of the people in surveillance people re-identification. When characters appear in TV dramas, most of them appear only with faces. Even if part of the body appears, the part of the body that appears is often not fixed: half, three-quarters of the body or whole body; The actions of people appearing are often not fixed: reaching out and raising hands, even more complex daily movements involve their hands, legs and head; The cloth of the characters in the TV drama will also change frequently according to the change of the plot. All of these points will make it difficult to achieve the re-identification target. But TV dramas provide a lot of close-ups of people's faces with detailed faces features, which makes the face re-identification a more appropriate approach. This project is aiming to track the characters in the TV dramas. The areas of Machine Learning and Deep Learning are developing rapidly, this technology contributes to a lot of areas recently, but it requires a lot of data and corresponding labels. TV dramas contain a sufficient amount of data, but the data has no label to train the neural network. Therefore, the objective of this project can only be achieved by combining deep learning, face landmark and face comparison approaches.

4.1 Main Structure

Because the TV drama data set for this project does not contain labels, it is impossible to simply use the deep neural network to train the character label in TV drama as a training data set, and then use the remaining TV series as a test set to find all the characters. The essence of people re-identification objective in this project is to find several methods to achieve face detection, face extraction and face recognition. The figure 4.1 shows the structure approach of this project. The first target is to find the faces of the characters in the TV drama data set. According to the above section described, the project will choose a pre-trained facial mask R-CNN[24] to find the ROI (bounding box) of all faces. After uses face detection to find the faces of all people, the next step needs to apply a face extraction method to get the face landmarks of all faces. The method proposed by Kazemi and Sullivan[25] will be used as the start of the initial study, because the method is included in the dlib library [9], which

makes the initial implementation and analysis of the project easy. After obtaining landmarks through this method, a global similarity transformation approach will align all landmarks. Finally, the joint Bayesian approach is used to compare and classify the similarity between pictures.

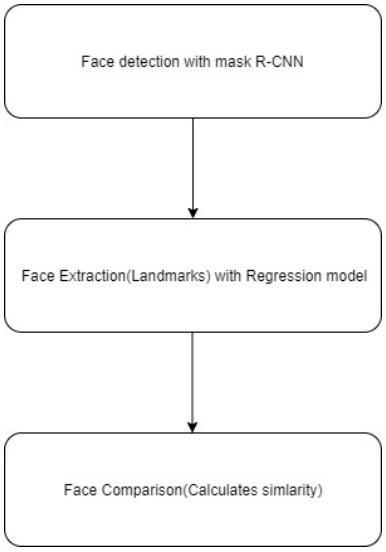

Figure 4.1: Structure of the System

All the characters appearing in the input videos have no labels, so the project needs a logical structure to describe how to compare and classify the detected characters. We introduce a class set. The class set mainly contains all the classified characters. The set is empty at the beginning. We assume that when a frame as input import to model and a face is detected, if the class set is empty, the detected face will be added to the class set as a new class and represented as person_1. Can deduce the rest from this, the input video is divided into pictures of individual frames, and then faces of characters will be detected one by one. For another situation, we assume that there are person_1, person_2 ... person_n in the class set. If a frame as input imported and one or more faces are detected, these faces will be successively compared with the classified faces in the class set. By finding the class with the highest score, the detected face will be considered to belong to this class, and the face picture will be added to the person class. If the similarity scores of all person classes are very low, then it is considered that the detected face has not appeared before, the class set will add a person_n+1 and then put the detected face into it. Figure 4.3 shows how the detected faces are classified with the faces in the class set.

Combining the two systems described above, we can describe the overall operation of the project system. Figure 4.3 shows how the system works overall. First of all, take a TV drama video as input, the system will divide the video into pictures according to the number of frames. Face detection technology will detect the face in each picture and output the

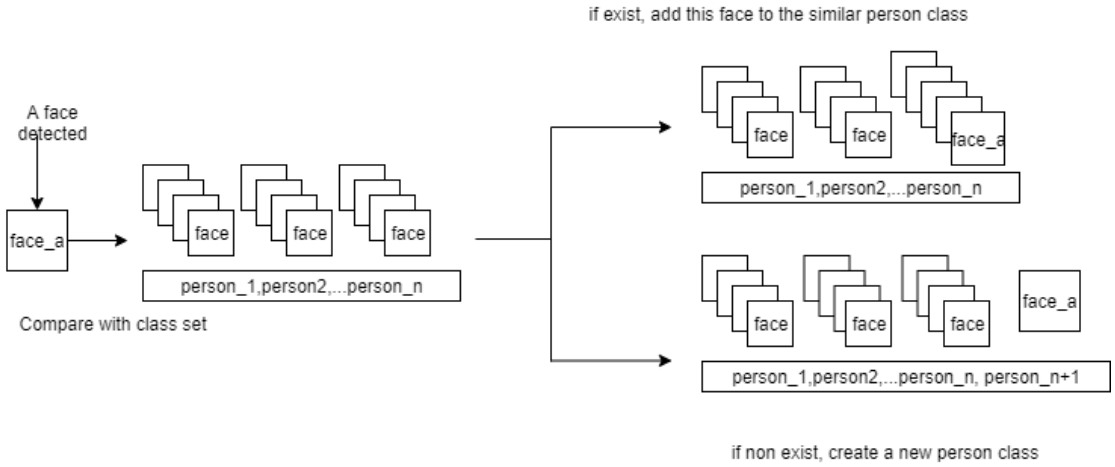

Figure 4.2: Comparison structure

location information (roi information) of the bounding box. By using the face extraction method to extract comparable face information, these faces will be compared with the faces in the class set, and the detected faces will be classified into the class with the highest similarity score. If the class set is empty or the similarity scores of the detected face in the class set are low, then a new person class is added and the detected face is placed in it. Then the system will detect the next picture until all the pictures in the input video have been detected.

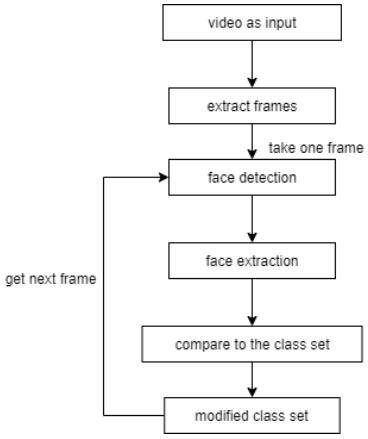

Figure 4.3: Overview structure

The structural method involves the model that needs to be trained, they are not models with explicit requirements. These models are all used to complete a specific broad objectives (face detection, get landmarks from box-bounded faces). These approaches avoid the problem of training models for the project but dataset has no labels, because the models used in this project can be pre-trained from other labeled datasets and then applied to the project.

This project is trying to find a combination of the best performance in a TV drama series people re-identification task among different face detection and face extraction methods. The face detection and face extraction methods in the above structure system description can be replaced. This project will conduct more exploration and analysis in the next semester.

4.2 Evaluation Metrics

In this subsection, the evaluation metrics used for evaluating the performance of people tracking and re-identification will be introduced in detail.

According to the literature, evaluation metrics that are designed for evaluating multi-object trackers have been introduced [33] [34]. However, there are two reasons determine that those metrics are not suitable for this work. Firstly, those metrics are frame-based essentially. Compare to the measurement for frame-based performance, in video-based information retrieval applications, if a person or an object in a shot is completely detected is more important. If it is not, to retrieve information from a shot is impossible. Whereas in fact, although a person or an object is tracked only in a snap, it is still possible to retrieve that shot. However, in frame-based metrics, this kind of information is lost. Secondly, those metrics require the tracking system to associate the components of a fractional track. In this project, the tracking system is not doing that, it is done in the process of retrieval implicitly instead. Therefore, a more suitable evaluation metric which is introduced by Mika Fischer et al. [35] in article *'Person re-identification in TV series using robust face recognition and user feedback'* will be used in this project instead of using any one of the metrics mentioned above.

4.2.1 Metric for overall tracking performance

The first evaluation metric designed in this project is to calculate the percentage of correctly detected occurrence of persons in the video, no matter how short the time for a person is showing up. It gives an idea of the number of detection that is completely false as well. The design principle of it is described as follows. Firstly, the number of persons that are detected by the tracking system in the video needs to be figured out. In order to do so, how the detected tracks are associated with the ground truth labels that are established in the initial data analysis state should be defined. To be specific, in each frame, a ground truth label will be assigned to the track which is closest to it, but only if the ground truth label is covered by the track. A simple majority vote will be used on the ground truth labels that are already associated with the frames in a track to assign the whole track a ground truth label. As a result, each track will either has an associated ground-truth label, or not, which can happen for the case that none of the face contours in the track are close to a ground truth label, which caused by false detections or the detection of background characters. Combine these together, the *track precision* and *recall* are defined below:

$$track\ precision = \frac{tracks\ associated\ with\ a\ label}{tracks\ found} \tag{4.1}$$

$$track\ recall = \frac{tracks\ associated\ with\ a\ label}{labeled\ tracks} \tag{4.2}$$

Last but not the least, duplicated detected tracks for the same person in a shot will be counted as once, otherwise, the results of the track recall would be too overestimate to the actual performance.

4.2.2 Metric for tracking quality

The first metric does not concern about the tracking quality of the tracking system, so a second metric needs to be built. For this metric, all that we need to concern is the number of found face contours that actually cover the correct face and the number of faces in the track that is actual labeled are covered by the detected tracks. For this sake, the things described above need to be computed for each labeled track, then sum those results altogether and average the sum. The metric named *average frame precision* and *recall*, AFP and AFR in abbreviation, comes out below:

$$AFP = \frac{1}{N}\sum_{t=1}^{N}\frac{correct\ face\ contours\ associated\ with\ track\ t}{face\ contours\ associated\ with\ track\ t} \tag{4.3}$$

$$AFR = \frac{1}{N}\sum_{t=1}^{N}\frac{correct\ face\ contours\ associated\ with\ track\ t}{face\ contours\ in\ track\ t} \tag{4.4}$$

where N is the number of successfully detected labeled tracks by the tracking system. This metric tells the tracking quality. If the majority of the occurring faces in the shot are covered, AFR will be very close to 1. If fewer false detections are contained in the tracks, i.e tracks that switch from person A to person B and tracks lose the tracked face, then AFP will be very close to 1 as well. Obviously, the tracks that will only be concerned with this metric are those could establish a correspondence to a person. And the frame precision and recall would mean nothing for all the other tracks.

4.2.3 Metric for track switch rate

In addition, a metric for checking if track switches happened, i.e if faces of two different persons are covered in one track. It is built as follows:

$$track\ switch\ rate = \frac{tracks\ covering\ more\ than\ one\ person}{tracks} \tag{4.5}$$

4.2.4 Metric for person re-identification

The precision and recall will be used here again to evaluate the person re-identification component. The tracks output by the tracking system are treated as the base set. The metric is built as follows:

$$precision = \frac{retrieved\,tracks\,that\,are\,associated\,with\,the\,correct\,person}{retrieved\,tracks} \tag{4.6}$$

$$recall = \frac{retrieved\,tracks\,that\,are\,associated\,with\,the\,correct\,person}{tracks\,that\,are\,associated\,with\,the\,correct\,person} \tag{4.7}$$

To make the results more representative, every track generated by the tracking system will be used as the initial query set, which means evaluation will be done on every possible re-identification, and the results will be averaged.

4.3 Initial Works

At the beginning, we tried to implement face detection in the first step. But because the pre-trained model for face detection could not be found, and because of the time limitation, the goal of face detection could not be achieved currently. We found a pre-trained model made by Christian Clauss et al.[36] with using mask RCNN for objection detection and applied it to our TV-dramas dataset. The figure 4.4 show the bounding box of mask R-CNN where various objects can be found. Some papers point out that using the mask R-CNN architecture and labeled face as training data, the trained model can achieve the face detection objective. This verifies the feasibility of our use of mask R-CNN for face detection. Hence in the following stages, we will implement the structure mentioned in Section 4.1 step by step, and compare the performance of different model combinations.

Figure 4.4: Mask R-CNN merge detection

Chapter 5

Plan of work

Brief plan of work until the end of this project can be viewed from figure 5.1. So far, most of the initial researches are done and further works are implementation of the approaches to the data set and compare the results. In order to do it, editing the data should be finished before approaches are ready to analyse the data. For second half of the year, it's split into four phases and first three are mostly for further and actual research for implementation and testing of approaches. Last phase will be for writing the final report and preparing the final presentation and journal article. Again, figure 5.1 is the brief plan at this stage, it will be updated constantly through out the project. It is not shown in the figure 5.1 yet but it is also considered to work during the Easter holiday if the project got delayed work or not much progresses are made.

GANTT CHART

PROJECT TITLE	Tracking individuals across multiple scene and shots in video	Team	Team Delta
PROJECT MANAGER	Yoshi Gotoh	DATE (L	24 Nov 2019

WBS NUMBER	TASK TITLE	ASSIGNED TO	START DATE	DUE DATE	DURATION (DAYS)	% OF TASK COMPLETE
1	**Initial project progress**					
1.1	Interim report	All team	25/11/19	18/12/19	24	0%
1.1.1	Initial topic	Yucheng	25/11/19	18/12/19	15	0%
1.1.2	State of art	Mingyan	25/11/19	18/12/19	15	0%
1.1.3	Proposed advancements	Peng, YC	02/12/19	18/12/19	11	0%
1.1.4	Demo, initial experiements etc	Peng	09/12/19	18/12/19	10	0%
1.1.5	Plan of work	All team	12/12/19	18/12/19	7	0%
1.2	Data selection	Wonwoo	25/09/19	18/12/19	24	0%
2	**Project Implementation / Testing**					
2.1	Data analysis	Wonwoo	10/02/20	10/04/20	63	0%
2.1.1	Data edition	Wonwoo	10/02/20	23/02/20	14	0%
2.1.2	Data annotation	Wonwoo	24/02/20	08/03/20	14	0%
2.1.3	Test of data	Wonwoo	09/03/20	10/04/20	35	0%
2.2	Tracking (RCNN)	Mingyan	02/10/20	10/04/20	63	0%
2.2.1	Select and analysis Approaches	Mingyan	10/02/20	08/03/20	28	0%
2.2.2	Impliment approaches	Mingyan	24/02/20	10/04/20	49	0%
2.3	Comparison(Feature Extraction)	Peng	02/10/20	10/04/20	63	0%
2.3.1	Select and analysis Approaches	Peng	10/02/20	08/03/20	28	0%
2.3.2	Impliment approaches	Peng	24/02/20	10/04/20	49	0%
2.4	Test and Evaluation	YC	02/10/20	10/04/20	63	0%
2.4.1	Approaches analysis	YC	10/02/20	23/02/20	14	0%
2.4.2	Approaches Implementation	YC	24/02/20	09/03/20	21	0%
2.4.3	Evaluate model performances	YC	09/03/20	10/04/20	35	0%
3	**Assessment preparation**					
3.1	Final report	All team				0%
3.1.1	Investigate topic		23/03/20	14/05/20	51	0%
3.1.2	State of art		23/03/20	14/05/20	51	0%
3.1.3	Proposed advancements		23/03/20	14/05/20	51	0%
3.1.4	Any testing conducted		23/03/20	14/05/20	51	0%
3.1.5	Clear statement of innovation		23/03/20	14/05/20	51	0%
3.2	Final presentation preparation		06/05/20	20/05/20	14	0%
3.3	Journal article preparation		06/05/20	20/05/20	14	0%
3	**Module Assessment**					
3.1	Interim report deadline		19/12/19	19/12/19	1	0%
3.2	Final report deadline		14/05/20	14/05/20	1	0%
3.3	Final team presentation		21/05/20	21/05/20	1	0%
3.4	Journal article deadline		21/05/20	21/05/20	1	0%

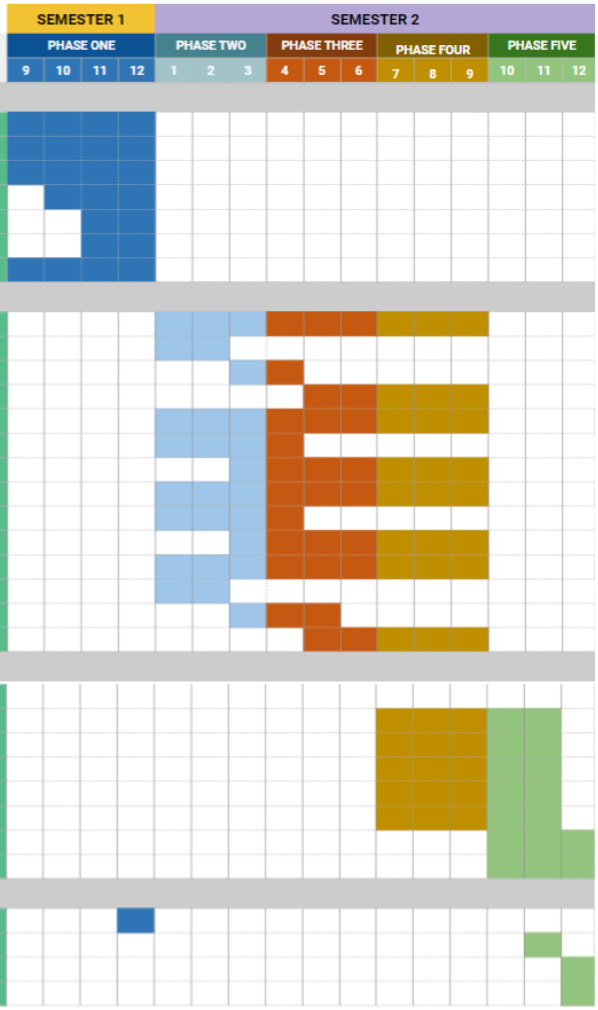

Figure 5.1: Gantt chart for plan of work

Chapter 6

Conclusion

This project analyzes tracking and divides the task into two parts, identity and recognition. Through the study of current tracking research, we found that most of these research focus on solving the tracking problem in the Surveillance View by analyzing and processing the body of people. However, the characters in the TV drama camera setting do not have a complete body shape, and more of the faces of characters appear in the lens. Therefore, the TV drama views tracking method will be different from the tracking body of people in surveillance, and the face processing will get better performance. Since TV drama data set has no label, this becomes the biggest problem in constructing algorithms and verifying results in this project. Due to lack of labels, the project cannot use deep learning approach to directly perform face tracking. After analysis, the project divides the tracking tasks into three main steps: face detection, face extraction and face comparison. By introducing the concept of class set, characters can be automatically identified and recognized by algorithm. Due to time constraints, this project has only verified the feasibility of using Mask RCNN to complete face detection. For future work, this project will imply the entire system structure and try multiple face detection, face extraction and face comparison methods to find the best combination.

Bibliography

[1] Alex. (2018) Boston man confesses to killing pedestrain and fleeing scene during tv interview, gets arrested. [Online]. Available: https://grabien.com/file.php?id=400105

[2] (1967) Star trek (the original series) season 1, episode 28, "the city on the edge of forever". [Online]. Available: https://www.flickr.com/photos/29069717@N02/25767008757/in/photostream/

[3] (2005) Brisbane airport: The missing cctv footage. [Online]. Available: https://www.expendable.tv/2012/08/brisbane-airport-missing-cctv-footage.html

[4] E. COMMISSION, "Regulation of the european parliament and of the council; establishing the european border surveillance system (eurosur)," 2011.

[5] P. S. Foundation, "Python," https://www.python.org/, December 2019, access on 19/12/2019.

[6] Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015.

[7] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.

[8] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[9] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, 2012.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, p. 91–110, 2004.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik, "Recognition using regions," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[15] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[16] R. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NIPS*, pp. 91–99, 2015.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[20] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 650–657.

[21] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[22] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[23] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[24] O. Cakiroglu, C. Ozer, and B. Gunsel, "Design of a deep face detector by mask r-cnn," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2019, pp. 1–4.

[25] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," 06 2014.

[26] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," 06 2013, pp. 532–539.

[27] C. X. W. Y. e. a. Lou, J., "Multi-subspace supervised descent method for robust face alignment," *Multimedia Tools and Applications*, vol. 78, December 2019.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386

[29] M. Wang and W. Deng., "Deep face recognition: A survey." https://arxiv.org/abs/1804.06655, April 2018.

[30] T. B. Z. W. C. H. Jingtuo Liu, Yafeng Deng, "Targeting ultimate accuracy: Face recognition via deep embedding," https://arxiv.org/abs/1506.07310, June 2015.

[31] R. C. Rajeev Ranjan, Carlos D. Castillo, "L2-constrained softmax loss for discriminative face verification," https://arxiv.org/abs/1703.09507, March 2017.

[32] N. X. S. Z. Jiankang Deng, Jia Guo, "Arcface: Additive angular margin loss for deep face recognition," https://arxiv.org/abs/1703.09507, Jenuary 2018.

[33] S. R. Bernardin K, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing 2008:10*, 2008.

[34] O. J. B. S. Smith K, Gatica-Perez D, "Evaluating multi-object tracking," *In: Proc. of the CVPR workshop on empirical evaluation methods in computer vision*, no. 2, p. 36, 2005.

[35] M. Fischer, H. K. Ekenel, and R. Stiefelhagen, "Person re-identification in tv series using robust face recognition and user feedback," *Multimedia Tools and Applications*, vol. 55, no. 1, p. 83–104, 2010.

[36] C. Clauss and W. abdula, "Mask r-cnn for object detection and segmentation," https://github.com/matterport/Mask_RCNN, 2017.