University of Zurich[UZH]

BSc Thesis January 6, 2021

Person Re-Identification in and Across Videos

Michèle Fundneider
of Döttingen AG, Switzerland

Student-ID: 17-726-746
michele.fundneider@uzh.ch

Advisor: **Dr. Luca Rossetto**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

Acknowledgements

I would like to thank Prof. Abraham Bernstein for giving me the opportunity to work on this thesis. A special thank you goes to my advisor Dr. Luca Rossetto for the constant support and encouragement during these six months. He always provided me with useful feedback and suggestions. I am grateful to my family and friends and everyone who has helped me throughout my studies.

Zusammenfassung

Das Ziel von Personenreidentifikation (person re-identification) ist es, alle Instanzen einer bestimmten Person anhand eines Bildes in einer Gallerie von Bildern oder in Videos wiederzuerkennen. Die Forschung konzentriert sich dabei bisher grösstenteils auf die Erkenneung von Fussgängern in Überwachungskameras, Personenreidentifikation ist aber nicht nur in Überwachungsszenarien nützlich, sondern auch für Videoanalysen und Multimedia Retrieval Applikationen, wobei jegliche Art von Videos relevant sind. Um Personen in Videos zu erkennen, muss vor dem Reidentifaktionsschritt ein Personenerkennungsschritt (person detection) ausgeführt werden. Diese beiden Aufgaben verfolgen aber gegensätzliche Ziele, deshalb sind besonders 1-Schritt Methoden, welche diese Aufgaben vereinen, für die Personensuche (person search) geeignet. Wir analysieren hierzu zwei solche 1-Schritt Methoden der Personensuche, Online Instance Matching (OIM) und Norm-Aware Embedding (NAE) und testen wie gut diese auf einem filmbasierten Datenset performen. Um mehrere Personen innerhalb eines Videos zu identifizieren und zu tracken sind Multi-Object Tracking (MOT) Methoden geeignet. Hierbei sind FairMOT und JDE sehr effektiv und schnell, wir testen beide Methoden, um herauszufinden welche uns bessere Reidentifikationsresultate liefert.

Abstract

The goal of person re-identification (re-id) is to recognize all instances of a particular person from an image in a gallery of images or videos. So far, research was mostly focused on the re-id of pedestrians in surveillance cameras. Person re-id is not only useful in surveillance scenarios, but also for video analysis and multimedia retrieval applications, wherein all types of videos are relevant. In order to recognize people in videos, a person detection step must be carried out before the re-id step. However, these two tasks pursue opposing goals, which is why one-step methods that combine these tasks are particularly suitable for person search. We analyze two such one-step methods of person search, Online Instance Matching (OIM) and Norm-Aware Embedding (NAE), and test how well they perform on a movie-based dataset. Multi-Object Tracking (MOT) is another task suitable for identifying and tracking several people within a video. Here, FairMOT and JDE are very effective and fast, we test both methods to find out which one gives us better re-identification results.

Contents

1

Introduction

Person re-identification (re-id) deals with the identification of people across images and videos. The goal of person re-id is to find the queried person amongst a set of images. Person re-id is mainly researched in safety and surveillance applications where the aim is to find a target across different surveillance camera footages. However, video analysis and retrieval applications can make good use of knowing who is visible in a video too and profit from applying a person re-id approach. Person re-id has to tackle different challenges, some of which are illumination changes, occlusions, different viewpoints, low-image resolutions, unconstrained poses and more [Ye et al., 2020]. In recent years and with the advancements in deep learning, re-id methods have achieved remarkable results [McLaughlin et al., 2016], [Ye et al., 2020]. Still, the research scenarios vary from practical applications. Traditionally, person re-identification methods use already cropped images [Ahmed et al., 2015], [McLaughlin et al., 2016]. However, in a real-life scenario a detection step is necessary to find people in the whole scene image. Person search is the concept of bringing these two contradictory tasks together in a single method [Xiao et al., 2017]. Person detection and re-identification have contradictory goals, while person detection aims at finding similarities of all people, person re-id focuses on the differences between them. Multi-Object Tracking (MOT) is another End-To-End Person re-id with the goal of tracking multiple people in a video [Ye et al., 2020].

Videos can bring some advantages as well as disadvantages to the re-id problem. Videos, especially unconstrained videos, can be challenging since they show different backgrounds, camera angles, motion blur and poses. Video-based person re-id has become popular because videos give us more frames/images of a person and therefore contain more useful information.

In this work, we will introduce and compare different methods for the task of person re-id. We explain the applicability of these methods on unconstrained videos. We will focus on two different tasks, person search and MOT. For person search we mainly compare two methods, Norm Aware Embedding (NAE) [Chen et al., 2020] and Online Instance Matching (OIM) [Xiao et al., 2017]. We select a subset of the MovieNet [Huang et al., 2020] dataset to compare how efficiently they can re-identify people in keyframe images within a movie and across different movies. The NAE method of [Chen et al., 2020] outperforms the OIM method of [Xiao et al., 2017]. [Zhang et al., 2020] and [Wang et al., 2019] introduce accurate and fast MOT methods. We compare the two by testing it on different short video clips to see which can identify and track people

more accurately. The newer [Zhang et al., 2020] is slightly less accurate at identifying people in this scenario. Finally, we reflect on our results and discuss some problems and propose some changes and additions to improve the applicability of person re-id on unconstrained videos.

2

Related Work

We can address the person re-identification problem in videos through different approaches. The general person re-identification methods use cropped images to match identities from query images to gallery images [Zheng et al., 2016]. Person search refers to the approach to use non-cropped gallery images and combining the detection and re-id step into one step [Xiao et al., 2017]. Multi-Object Tracking (MOT) is used for tracking multiple people in videos, pose estimation and tracking can be used for the similar purpose of tracking people in videos by their pose.

2.1 Person Re-Identifcation

While image-based person re-id has already been researched broadly, video-based re-id has gained more attention in recent years and has become increasingly popular [McLaughlin et al., 2016], [Ye et al., 2020]. This is mainly due to the advantage of having multiple frames making it is easier to identify people [Pathak et al., 2020]. There are various advantages of using video data, such as the usage of temporal information, the larger sample size of a person and the more natural usage of person re-id. With video-based person re-id there are also different challenges, these include video sequences which have different length or frame-rates, occlusions, and inaccurate tracking [McLaughlin et al., 2016].

Today the majority of methods make use of deep learning techniques. [Li et al., 2014] and [Ahmed et al., 2015] propose Convolutional Neural Networks (CNN) for person re-id, where the input consists of a pair of cropped pedestrian images and they use binary verification loss functions to train their parameters. [McLaughlin et al., 2016] also use a deep learning approach for the video-based re-identification problem. They propose a recurrent deep neural network architecture which joins recurrency and temporal-pooling of appearance data with representation learning. They use a Siamese network architecture, in which for each person's video sequence it learns an invariant representation. The data from all time-steps is combined into a single feature vector for the whole input sequence due to the introduction of temporal pooling and recurrent layers in their network leading to an improved performance.

However, person re-id methods rely on cropped images. For our use case we need whole scene images, this is where object detection is needed to detect people in an image.

2.2 Person Search

In a real-world scenario we have to find a person from a gallery of non-cropped images. Person search is the task where you find a person in a gallery of full images by a cropped image of that person [Xiao et al., 2017]. Thus, person detection and re-id are needed. [Xiao et al., 2017] propose a CNN where pedestrian detection and person re-id are jointly handled. Their CNN uses a pedestrian proposal net to produce the bounding boxes and an identification net to extract features for comparison with the target person. They adapt with each other during joint optimization. They further propose an Online Instance Matching (OIM) loss function to learn identification features more effectively for better scalability. [Munjal et al., 2019] use this OIM and extend it, they propose a query-guided end-to-end person search network (QEEPS). First, they add a query-guided Siamese squeeze-and-excitation network (QSSE-Net) which leverages global context from query and gallery images, second, a query-guided region proposal network (QRPN), which produces proposals relevant for the query, third, a query-guided similarity subnetwork (QSimNet), which learns a query-guided re-id score. [Chen et al., 2020] present a Norm-Aware Embedding method (NAE), which splits the person embedding into norm for detection and angle for re-ID. Furthermore, they add a pixel-wise extension (NAE+) to reduce the misalignment. Compared to the above mentioned one-step methods, NAE outperforms them on the two common datasets of person search, namely CUHK-SYSU [Xiao et al., 2017] and PRW [Zheng et al., 2017]. Person search is useful to our use case since we also need to find people from whole scene images.

2.3 Multi-Person Tracking

We will separate Multi-Person Tracking tasks into two categories: Multi-Object Tracking (MOT) and Object Tracking based on Pose Estimation. MOT is the task of estimating the paths of multiple objects of interest in videos. MOT was mainly addressed with two separate steps: first, a detection step, where targets in single video frames are found; second, an association step, where detected targets are assigned and connected to existing trajectories [Wang et al., 2019]. The system needs at least a detector and an embedding (re-id) model. Most MOT methods use two-step methods, where object detection and re-id are handled separately [Mahmoudi et al., 2018], [Wojke et al., 2017]. In both categories remarkable progress has been made and for each task the two-step methods can select the most applicable method. Thus, the accuracy and performance are very good. Nonetheless, they tend to be very slow since the two tasks do not share computations and are unable to perform inference at video rate. Therefore, One-Shot MOT recently gained popularity [Wang et al., 2019]. In One-Shot MOT object detection and identity embedding are accomplished in one network. Yet, tracking accuracy is lower compared to two-step methods. [Wang et al., 2019] integrate the two steps into a single network, avoiding re-computation. They propose a Joint Detection and Embedding Model (JDE) which simultaneously outputs detection results and the corresponding appearance embeddings of the detected boxes. [Zhang et al., 2020] identify

three factors critical to accuracy: (1) anchors are not suited for re-id, (2) multi-layer feature aggregation, (3) lower-dimensional features are better for MOT. They propose a simple baseline which jointly considers these factors: an anchor-free object detection approach to estimate object centers, parallel branch for estimating the pixel-wise re-id features which are used to predict the objects' identities, Deep Layer Aggregation operator in the backbone network. Compared to [Wang et al., 2019], the FairMOT model [Zhang et al., 2020] performs better on the MOT Challenge datasets [Milan et al., 2016] due to anchor-free object detection.

When looking at methods for the task of Object Tracking by Pose, [Girdhar et al., 2018] follow a two-stage approach, where the tracking stage computations are significantly less expensive and are not limited by the number of instances per frame. [Xiao et al., 2018] propose a simple baseline methods for human pose estimation and tracking. The goal is to reduce the complexity of the existing methods while maintaining their effectiveness. For their pose estimation they use the backbone network ResNet [He et al., 2016] and add deconvolutional layers on it. Their pose tracking is based on the greedy matching method as in [Girdhar et al., 2018], only modifying the use of optical flow based pose propagation and similarity measurement.

We will look further into the two MOT methods described above, since MOT can be used in our use case for detecting and tracking people within a video and is the only approach that already works with actual videos instead of images.

3

Selected Methods

In this section, we will further describe the selected methods for person search: OIM [Xiao et al., 2017] and NAE [Chen et al., 2020] and for MOT: JDE [Wang et al., 2019] and FairMOT [Zhang et al., 2020]. We will propose a possible pipeline for combining these two tasks.

3.1 Person Search

We compare two person search methods NAE [Chen et al., 2020] and OIM [Xiao et al., 2017], both test their methods on the CUHK-SYSU [Xiao et al., 2017] and PRW [Zheng et al., 2017] datasets. CUHK-SYSU is a person search dataset containing street snaps in an urban city and movie snaps with people, in total the dataset consists of 18,184 images with 96,143 bounding boxes and 8,432 labeled identities. People with less than half body appearances or abnormal poses are not annotated and people who change their clothes are not included. PRW is another person search dataset using images from six security cameras containing 11,816 frames with 932 different identities. On CUHK-SYSU OIM achieves Mean Average Precision (mAP) of 75.5 and top-1 Cumulative Matching Characteristics (CMC) of 78.7, while NAE achieves mAP of 91.5 and top-1 of 92.4, on PRW OIM achieves 21.3 mAP and 49.9 top-1, NAE achieves 43.3 and 80.9 respectively. [Chen et al., 2020] build their method based on [Xiao et al., 2017] and display better results on these two datasets.

3.1.1 Joint Detection and Identification Feature Learning for Person Search

[Xiao et al., 2017] propose a joint deep learning framework in a single convolutional neural network (CNN). The framework accepts a whole scene image as input and uses a Stem CNN for converting the raw pixels of the image to convolutional feature maps. They construct a pedestrian proposal net that will predict the bounding boxes of people on top of the feature maps. Next, the bounding boxes are led to the identification net with RoI-Pooling [Girshick, 2015]. This results in L2-normalized 256-d features for each bounding box. For inference the gallery people are ranked by their feature distances

to the target person. In training, they propose an Online Instance Matching (OIM) loss function to be placed on top of the feature vectors to track the identification net, along with several other loss functions for multi-task training of the proposal net. In the dataset there exist labeled identities, unlabeled identities, and background clutter. They maintain a lookup table to store the labeled identity proposals and a circular queue for the unlabeled identity proposals.

The goal of OIM is to maximze the expected log-likelihood. Their OIM loss compares the mini-batch sample to all identities, leading to similarity between the underlying feature vector and the target one, and pushing it away from others.

3.1.2 Norm-Aware Embedding for Efficient Person Search

As previously mentioned [Chen et al., 2020] propose an efficient person search model. Compared to OIM [Xiao et al., 2017], NAE [Chen et al., 2020] uses the embedding norm as the confidence of person/background by removing the original region classification branch. The Norm-Aware Embedding is implemented as follows: global average pooling and a fully connected layer is applied just as in OIM, where we receive the feature vector x, then this vector is, unlike in OIM, decomposed into a norm r and a 256-dimensional vector called angle Θ. The magnitude of norm r is squeezed to the range [0,1] to interpret it as the detection confidence. The NAE results from this modified norm and the angle Θ.

3.2 Multi-Object Tracking

In this section, we will compare two MOT methods, JDE [Wang et al., 2019] and FairMOT [Zhang et al., 2020]. The methods are evaluated on the public datasets of the MOT challenge [Milan et al., 2016] with FairMOT reporting better results on these datasets.

3.2.1 Towards Real-Time Multi-Object Tracking

The Joint Learning of Detection and Embedding (JDE) [Wang et al., 2019] outputs the location and appearance embeddings of target people simultaneously in a single forward pass. The model detects targets in the first step. In the appearance embedding the distance between two identities in successive frames should be smaller than those with different identities. The method receives a video frame as input and passes it through a backbone network to receive feature maps. The smallest feature map, which has the strongest features, is up-sampled and joined with the second smallest feature map by skip connection. Lastly, the prediction heads are added at all scales. The prediction heads contain several convolutional layers and produce dense prediction maps. These predicition maps consist of three tasks: 1) the box classification results, 2) the box regression coefficients, and 3) the dense embedding map.

[Wang et al., 2019] modify the standard RPN [Ren et al., 2015] in two aspects. They redesign the anchors to fit to the targets, in this case pedestrians. The values for the dual

thresholds used for foreground/background are selected, they determine boxes with an Intersection over Union (IoU)>0.5 as foreground and an IoU<0.4 as background. The advantage of these thresholds is that false alarms are suppressed, which is especially important for occlusions.

To achieve an embedding space where results of the same identity are close to each other, [Wang et al., 2019] use triplet loss. Two challenges arise with this formula of triplet loss. The first one being the big sampling space in the training dataset. The other challenge is that training can be unstable and converge slowly. [Xiao et al., 2018] optimize over a smooth upper bound of the triplet loss to solve this problem. The upper bound loss is similar to the cross-entropy loss. They conclude that the cross-entropy loss outperforms the other two losses, triplet loss and upper bound loss.

3.2.2 FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking

[Zhang et al., 2020] build their framework based on JDE [Xiao et al., 2017] but lay the focus on treating the detection and re-id tasks equally, instead of prioritizing the detection and performing detection and re-id in a cascade style, therefore calling it FairMOT. FairMOT uses two homogenous branches, one for the detection of objects and the other for the re-id feature extraction. Thus, FairMOT removes the unfair disadvantage of the re-id task and obtains better re-id features leading to satisfactory MOT results. The anchor-free detection branch predicts objects centers and sizes as position-aware measurement maps, it is built on top of CenterNet [Zhou et al., 2019]. A modified version of Deep Layer Aggregation (DLA) [Zhou et al., 2019] is applied on the ResNet-34 backbone, resulting in a model named DLA-34. Heatmaps, object center offsets and bounding box sizes are estimated by three heads on the DLA-34. The heatmap head estimates the object center locations, the offset head accurately localizes objects, and the box size head estimates the box width and height. The re-id branch predicts a pixel-wise re-id feature to define the object centered at the indicated pixel. They built a convolutional layer with 128 kernels onto the backbone features to extract the re-id features for every location.

3.3 Pipeline

In order to bring the person search and MOT task together in a single framework, we analyze the possibility of a potential pipeline. As a first step, it is necessary to segment the shots in a video since MOT only works on one-shot videos without sudden camera view changes. The shot keyframes will be given into the person search framework returning the bounding boxes and IDs. The one-shot videos are fed into the MOT framework respectively. The MOT bounding boxes of the first frame must match the bounding boxes and IDs of the corresponding shot keyframe.

4

Experiments

In this section, we evaluate the selected methods for person search and MOT. As a first step we describe the datasets used. We will then evaluate the results generated on these datasets to detect how efficient and accurate the methods are and which method has the best performance. All methods are tested on one NVIDIA GeForce RTX 2080 Ti.

4.1 Datasets

Person search and MOT methods are applied to different use cases and under different settings, therefore we need two different datasets and settings for them. For the person search methods we will use a selected subset of the MovieNet [Huang et al., 2020] dataset and test if a person can be found in the keyframes of a movie. For the MOT methods we will use short one-shot video clips and test the effectiveness of identifying and effectively tracking all targets.

4.1.1 MovieNet

[Huang et al., 2020] present MovieNet, a dataset for movie research. The dataset consists of 1,100 movies with a variety of annotations relevant for movie analysis, one of which is the person re-id annotation, where they provide the respective bounding boxes and person IDs. These IDs contain only the top 10 cast according to IMDb, while the other identifications are labeled as "others". In sum, they present 1.1M person re-id annotations with 3,087 labeled cast and 364k "others". They provide shot keyframes in 240 pixels resolution.

For our experiments we selected a subset of the MovieNet [Huang et al., 2020] dataset. In a first part, we select a randomly chosen query image of an actor with the respective person ID and bounding box and set the remaining annotated images of that movie as gallery images. The objective is to search though all gallery images and compare the identified bounding boxes and people to the query image. Tab. 4.1 shows the different subsets we selected for the single movie person search. In the first column we name the subset, the second and third column display the movie and actor with their respective IMDb IDs, the fourth column shows the selected query image name and for reference we display the number of gallery images in the last column. In addition, we search

for the selected query image across two different movies to test whether or not the search accuracy decreases compared to searching within one movie. In Tab. 4.2 the selected subsets for the multiple movies person search are collected, they are made up of a combination of the subsets of the single movies. For the query images from Tab. 4.1 we search through the respective movie plus another movie with the same actor in Tab. 4.2.

Set	Movie	Actor	Query Image	Gallery Size
single1	tt1375666	nm0000138	shot_1429_img_0.jpg	5242
single2	tt0407887	nm0000138	shot_0513_img_1.jpg	2355
single3	tt0315327	nm0000120	shot_0554_img_0.jpg	1195
single4	tt0338013	nm0000120	shot_1082_img_0.jpg	3350

Table 4.1: MovieNet Subset Selection for Single Movie

Set	Movies	Actor	Query Image
multiple1	tt040788, tt1375666	nm0000138	shot_1429_img_0.jpg
multiple2	tt040788, tt1375666	nm0000138	shot_0513_img_1.jpg
multiple3	tt0315327, tt0338013	nm0000120	shot_0554_img_0.jpg
multiple4	tt0315327, tt0338013	nm0000120	shot_1082_img_0.jpg

Table 4.2: MovieNet Subset Selection for Multiple Movies

4.1.2 Video Clips for MOT

In order to test the MOT methods on how effective they are in tracking IDs in unconstrained videos (i.e. movies), we select three short one scene clips and manually annotate the first and last frame. In a next step we test these clips on the MOT methods JDE [Wang et al., 2019] and FairMOT [Zhang et al., 2020]. Their methods accept as input image sequences or videos and provide as output bounding box annotations with person IDs. In Tab. 4.3 we show the selected clips, their frame length and the number of matching IDs from the first to last frame.

Clip	Length (# Frames)	# Matching IDs
1	113 frames	11
2	98 frames	5
3	112 frames	2

Table 4.3: Short Youtube Video Clips for MOT

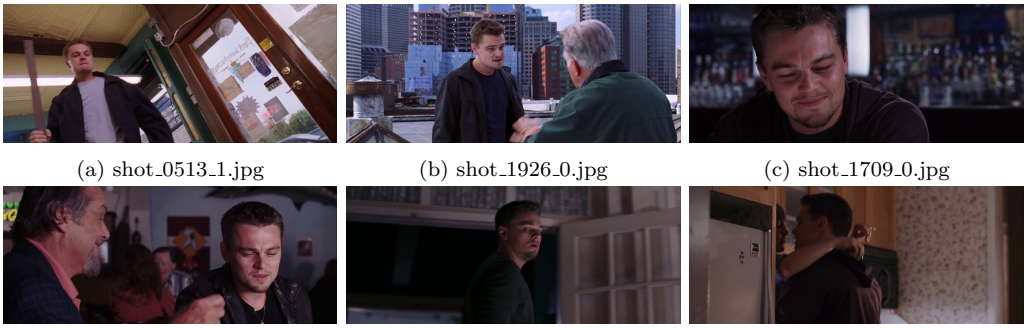
4.2 Results and Comparisons

NAE vs. OIM

We measure our results on the same performance metrics as in [Xiao et al., 2017] and [Chen et al., 2020], Recall and Average Precision (AP) for person detection and mAP and CMC top-1, top-5, and top-10 for person re-id. A person bounding box is only ranked if the IoU to the ground truth is larger than 0.5. The evaluation results for the single movie datasets are summarized in Tab. 4.4 which conclude the following points. The detection results are recorded in the second and third column of Tab. 4.4, which show us that both Recall and AP are higher with NAE than OIM for all tested datasets. Therefore, we conclude that NAE's detection performance is higher. For the re-id results we compare the results from the fourth, fifth and sixth column of Tab. 4.4, which show us the mAP, top-1, top-5, and top-10 accuracies. The performance varies greatly between the movies, while single1 only reaches a recall of 25.52 and a mAP of 1.52, single3 achieves 66.69 recall and 18.50 mAP. [Xiao et al., 2017] state that with increasing gallery size the search performance decreases. The top-k predicitons are either 0.00, if there is no correct prediction in the top-k images, or 100.00, if there is at least one correct prediction, since we only test it on one query image. On these metrics NAE outperforms OIM too. This concludes that the detection quality and the re-id accuracy of NAE is more advanced and the overall person search method of NAE outperforms OIM. When comparing the results for the single movie datasets from Tab. 4.4 with the results for the multiple movie datasets from Tab. 4.5 we discover that the search across movies does not decrease the search accuracy, it even increases it from 4.38 to 5.08. The increase in accuracy (using NAE) contradicts the statement from before, where performance decreases with increasing gallery size.

Method	Set	Recall	AP	mAP	top-1	top-5	top-10	Time (it/s)
OIM	single1	12.95	3.24	0.18	0.00	0.00	0.00	6.16
NAE	single1	**25.52**	**6.99**	**1.52**	0.00	0.00	**100.00**	**12.26**
OIM	single2	18.00	4.59	1.19	0.00	0.00	100.00	6.26
NAE	single2	**34.65**	**9.77**	**7.07**	0.00	**100.00**	100.00	**12.34**
OIM	single3	38.49	16.80	6.66	0.00	100.00	100.00	6.94
NAE	single3	**66.69**	**26.62**	**18.50**	**100.00**	**100.00**	100.00	**10.92**
OIM	single4	18.69	6.37	0.58	0.00	0.00	0.00	7.04
NAE	single4	**44.12**	**17.59**	**2.06**	0.00	**100.00**	**100.00**	**8.76**
OIM	Total	18.04	5.71	1.12	0.00	9.68	28.76	-
NAE	Total	**36.60**	**7.22**	**4.38**	**9.68**	**57.53**	**100.00**	-

Table 4.4: Comparison of OIM and NAE on MovieNet Single Movie Subset



(a) shot_0513_1.jpg (b) shot_1926_0.jpg (c) shot_1709_0.jpg

(d) shot_0783_1.jpg (e) shot_2162_0.jpg (f) shot_1492_2.jpg

Figure 4.1: Sample Query Images from MovieNet [Huang et al., 2020] for Movie tt040788 and Actor nm0000138

Set	mAP	top-1	top-5	top-10
multiple1	2.34	0.00	0.00	100.00
multiple2	4.52	0.00	100.00	100.00
multiple3	11.20	100.00	100.00	100.00
multiple4	4.47	0.00	0.00	0.00
Total	5.08	18.72	50.00	81.28

Table 4.5: NAE on MovieNet Mutliple Movies Subset

Tab. 4.6 summarizes the result for using different query images of the same person in the same movie. The respective images are displayed in Fig. 4.1 to display the differences in angle and body/face view. As can be seen in Tab. 4.6, the mAP varies from 3.33 to 7.07, and only two images have their top-1 prediction being correct, but everyone has their highest agreement in the top-5. Essentially this insight shows us that picking a fitting query image can have an impact on the overall performance.

Query Image	mAP	top-1	top-5	top-10
shot_0513_1.jpg	7.07	0.00	100.00	100.00
shot_1926_0.jpg	4.19	0.00	100.00	100.00
shot_1709_0.jpg	5.40	0.00	100.00	100.00
shot_0783_1.jpg	6.04	0.00	100.00	100.00
shot_2162_0.jpg	5.31	100.00	100.00	100.00
shot_1492_2.jpg	3.33	0.00	100.00	100.00
shot_0726_1.jpg	4.32	100.00	100.00	100.00
Total	5.94	33.33	100.00	100.00

Table 4.6: Comparison of Different Query Images and their Performance on MovieNet Subset single2 on NAE Method

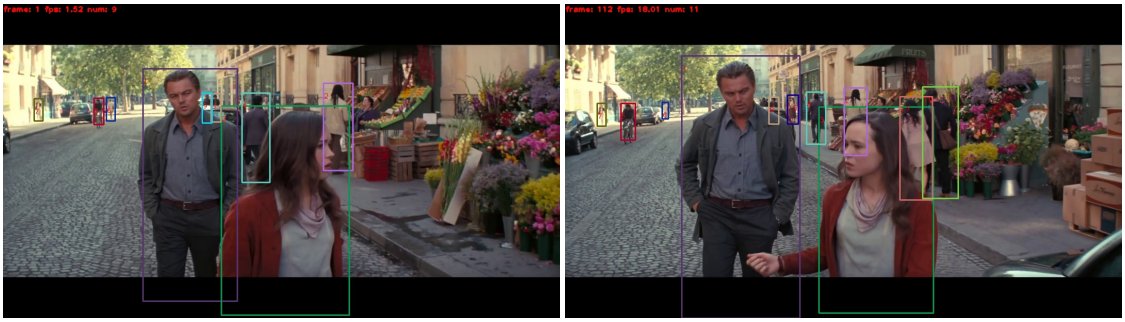
JDE vs. FairMOT

We measure the detection results for the two MOT methods with Recall and Precision and the re-id task with how many matching identities in the first and last frame are correctly identified. We measure and compare the computation time and the number of frames processed per second (it/s). The results are displayed in Tab. 4.7, which provides us with the following insights. Since we only have a small dataset with single videos there are only a small amount of identities to be detected, hence the precision is very high with sometimes 1.0 meaning all boxes were predicted. However a perfect precision of 1.0 is rather unrealistic in a larger dataset. Since our main focus lays in tracking how many identities have been correctly tracked from the first to the last frame, the correct IDs proportion is the most relevant. We observe that on all tested videos the correct IDs proportion of JDE is higher than FairMOT. Which leads us to the conclusion that JDE has a better tracking accuracy than FairMOT. The duration and iterations per second are also lower for JDE, but since we only have three very short clips, this information is meaningless. Taking into consideration these aspects, JDE outperforms FairMOT on the given video clips.

MOT Comparison					
Method	Set	Precision	Recall	Correct IDs	Time (s) / (it/s)
JDE	Clip1	0.85	**0.71**	**6/11**	**15:54 / 7.11**
FairMOT	Clip1	**0.89**	0.67	5/11	22:02 / 5.13
JDE	Clip2	1.0	**0.93**	**4/5**	**13:21 / 7.34**
FairMOT	Clip2	1.0	0.87	3/5	21:08 / 4.64
JDE	Clip3	1.0	**1.0**	**2/2**	**14:22 / 6.82**
FairMOT	Clip3	1.0	0.5	0/2	20:22 / 5.50
JDE	Total	-	-	**12/18**	∅7.09it/s
FairMOT	Total	-	-	8/18	∅5.09it/s

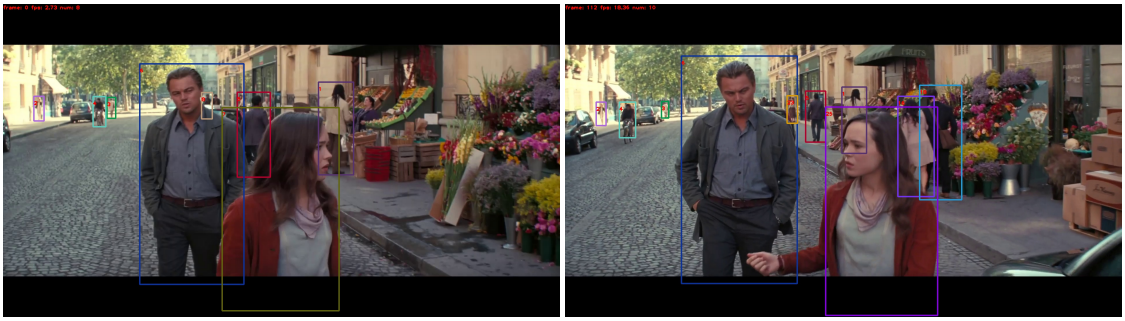
Table 4.7: Comparison of JDE and FairMOT on Our Selected Video Clips

Fig. 4.2 visualizes the result of running Clip1 through the JDE demo algorithm [Wang et al., 2019]. On the left we can see the first frame of the clip displaying all detected bounding boxes and their assigned person IDs, almost all people are detected and those who are not, are in the background and occluded. Compared to the picture on the right, which shows the last frame of the clip, we are able to detect that all except one person in the background were successfully re-identified. Fig. 4.3 shows us the results of the FairMOT method [Zhang et al., 2020] when run under the same settings as before. Compared to Fig. 4.2 it shows that FairMOT detects only 8 people in the first frame compared to 9 in JDE and the re-id of them is not as accurate.



(a) JDE First Frame (b) JDE Last Frame

Figure 4.2: JDE Results Clip 1



(a) FairMOT First Frame (b) FairMOT Last Frame

Figure 4.3: FairMOT Results Clip 1

5

Discussion

In this section, we will critically discuss the results and propose improvements to achieve our goal.

5.1 Problems

The results from using the MovieNet dataset [Huang et al., 2020] on the person search methods to find the requested person are rather poor. When tested against the CUHK-SYSU [Xiao et al., 2017] and PRW [Zheng et al., 2017] dataset as presented in their papers the results are much better and the accuracy is significantly higher. Possible reasons for the weak performance are as following. When looking at the MovieNet dataset, it contains a large variety of images, such as crowds, face close ups, side portraits and people from the back, this makes it inherently difficult to detect all people. The methods were not trained on the MovieNet dataset, although the CUHK-SYSU dataset contains some images from TV shows and movies, it is not as diverse and has only a selection of easily recognizable people. Furthermore, we only use 240 pixels images which lead to some restrictions in recognition. We chose the query images randomly, as already stated in Tab. 4.6, different results can be achieved by selecting a different query image of the target person. Selecting a different query image does affect the search performance slightly but does not affect the person detection results. Another factor aggravating the re-id of a target person in a movie is the appearance changes, the character may wear different clothes or hairstyles during the movie. Since person re-id is mostly based on full-body appearances relying on factors such as clothing colors, it can be challenging to identify when they change.

5.2 Improvements and Applicability

The existing methods are mostly designed for surveillance use cases where a target person is searched among pedestrians. For video retrieval and analysis however other types of videos are even more relevant. In a movie setting the scenes change a lot from different angles and viewpoints. Since the shots often show only the face without the body, face recognition methods could help fill that gap. Datasets for the task of person re-id in

videos as per our definition are difficult to find. Most re-id datasets use already cropped images or images scraped from videos with few frames and in similar pedestrian settings. To our knowledge there does not exist a dataset that satisfies the exact criteria needed. A dataset similar to the MovieNet dataset would be useful. The main contributions to be added to that dataset would be a better image or video resolution, annotated videos, and additionally face annotations. Training on this dataset could also improve the person search performance.

6

Conclusions

In this work, we analyse different person re-id methods and their applicability to unconstrained videos. We select two End-to-End person re-id approaches, Person Search and MOT, with two different methods each. We test the person search methods on a subset of the MovieNet [Huang et al., 2020] dataset to test their performance on movie images. For this task NAE [Chen et al., 2020] gives us the best results. However the actual search performance is rather poor on this dataset for various possible reasons such as low image resolution and appearance changes. For MOT we compare the methods on one shot video clips. In this setting, efficient multi person tracking can be achieved with JDE [Wang et al., 2019], performing in near realtime on one NVIDIA GeForce RTX 2080 Ti. A proposition is to combine the two in a single framework, where we perform shot segmentation to run person search on the shot keyframes and MOT in between the keyframes. Further studies should be invested in a possible integration of face recognition to increase the search performance accuracy and in creating a diverse dataset for person search in unconstrained videos.

References

[Ahmed et al., 2015] Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916.

[Chen et al., 2020] Chen, D., Zhang, S., Yang, J., and Schiele, B. (2020). Norm-aware embedding for efficient person search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12612–12621.

[Girdhar et al., 2018] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., and Tran, D. (2018). Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Huang et al., 2020] Huang, Q., Xiong, Y., Rao, A., Wang, J., and Lin, D. (2020). Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*.

[Li et al., 2014] Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

[Mahmoudi et al., 2018] Mahmoudi, N., Ahadi, S. M., and Rahmati, M. (2018). Multi-target tracking using cnn-based features: Cnnmtt. *Multimedia Tools and Applications*, 78:7077–7096.

[McLaughlin et al., 2016] McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334.

[Milan et al., 2016] Milan, A., Leal-Taixe, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking.

[Munjal et al., 2019] Munjal, B., Amin, S., Tombari, F., and Galasso, F. (2019). Query-guided end-to-end person search. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Pathak et al., 2020] Pathak, P., Eshratifar, A. E., and Gormish, M. (2020). Video person re-id: Fantastic techniques and where to find them (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13893â13894.

[Ren et al., 2015] Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

[Wang et al., 2019] Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2019). Towards real-time multi-object tracking.

[Wojke et al., 2017] Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric.

[Xiao et al., 2018] Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking.

[Xiao et al., 2017] Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ye et al., 2020] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. H. (2020). Deep learning for person re-identification: A survey and outlook.

[Zhang et al., 2020] Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2020). Fairmot: On the fairness of detection and re-identification in multiple object tracking.

[Zheng et al., 2016] Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future.

[Zheng et al., 2017] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., and Tian, Q. (2017). Person re-identification in the wild.

[Zhou et al., 2019] Zhou, X., Wang, D., and KrÃhenbÃ$\frac{1}{4}$hl, P. (2019). Objects as points.

List of Figures

List of Tables