Hand detection, segmentation and tracking from egocentric vision

by

Van-Tien Pham

Submitted to the School of Information Technology and Communication
in partial fulfillment of the requirements for the degree of

Master of Science in Information System and Communication

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

October 2020

Author .....
School of Information Technology and Communication
October 10, 2020

Certified by .....
Thi-Thanh-Hai Tran
Associate Professor
Thesis Supervisor

Accepted by .....
Arthur C. Chairman
Chairman, Department Committee on Graduate Theses

Hand detection, segmentation and tracking from egocentric vision

by

Van-Tien Pham

Submitted to the School of Information Technology and Communication
on October 10, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Information System and Communication

Abstract

Multiple object tracking is the process of assigning unique and consistent identities to objects throughout a video sequence. A de-facto approach to multiple object tracking, and object tracking in general, is to use a method called tracking by detection. Tracking by detection is a two-stage procedure: an object detection or segmentation algorithm fist detects objects in a given frame, these objects are then associated with already tracked objects by a tracking algorithm. Egocentric vision is an emerging field of computer vision that is characterized by the acquisition of images and video from the first-person perspective. In egocentric view, the two human hands are essential in the execution of actions and obtaining their movements and trajectories are the principal cues to define and recognize actions. One of the main concerns of this thesis is to build an automatic tracking by detection software that extract hands positions and identities in consequence frames from egocentric surveillance video supplied by the project âĂIJUnderstanding Human Daily Life Activities from Egocentric Vision Using Advanced Technologies of Deep LearningâĂİ of NAFOSTED. The framework consists of many state-of-the-art detectors from RCNN and YOLO family combined with the SORT or DeepSORT. The thesis then goes on to explore how the stand-alone alone performance of the object detection algorithm correlates with overall performance of a tracking-by-detection system. Finally, the thesis investigates how the use of visual descriptors of DeepSORT in the tracking stage of a tracking-by-detection system effects performance. Results presented in this thesis suggest that the capacity of the object detection algorithm is highly indicative of the overall performance of the tracking-by detection system. Further, this thesis also shows how the use of visual descriptors in the tracking stage can reduce the number of identity switches and thereby increase performance of the whole system. This thesis also presents a new egocentric hand tracking dataset MICAND32 for future researches.

Thesis Supervisor: Thi-Thanh-Hai Tran
Title: Associate Professor

Acknowledgments

First of all, I might want to offer my special thanks to my essential consultant, Prof. Tran Thi Thanh Hai. I am truly appriciated for everything you have guided me all through this.

Many thanks to my colleagues at Viettel High Technology and Industry Copo-ration for supporting me in engineering technique. Also, I might want to express gratitude toward Prof. Vu Hai and the alumnis at MICA Institute for giving me significant suggestions.

Deep inside my heart, I wish to show my gratefulness to my parent and my sister for always inspiring and trusting me in every of my steps.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of object recognition and tracking

1.1.1 Object recognition

In the past decade, modern computer vision technology based on AI and deep learning methods has undergone tremendous development. Today, it is used in applications such as image classification, facial recognition, object recognition in images, video analysis and classification, and image processing in robots and autonomous vehicles. Many computer vision tasks require intelligent segmentation of images to understand the content contained in the image and simplify the analysis of each part. Today's image segmentation technology uses deep learning models for computer vision to understand the real objects represented by each pixel in an image at a level that was unimaginable only ten years ago. Deep learning can learn patterns in visual input in order to predict the categories of objects that make up an image. The main deep learning architectures used for image processing are convolutional neural networks (CNN) or specific CNN frameworks such as AlexNet, VGG, Inception and ResNet. Generally, a deep learning model for computer vision is trained and executed on a dedicated graphics processing unit (GPU) to reduce calculation time.

Image segmentation is a key process in computer vision. It involves dividing the visual input into multiple parts to simplify image analysis. A segment represents an

object or part of an object, and includes a set of pixels or "super pixels". Image segmentation divides pixels into larger parts, eliminating the need to treat a single pixel as an observation unit. Object recognition is divided into three levels:

1. Image Classification: to classify the entire image into categories such as people, vehicles, dogs.

2. Object detection: to detect an object in the image and draw a rectangle around it, such as a pedestrian or sheep.

3. Object Segmentation: to identify the various parts of the image and understand what objects they belong to. Segmentation lays the foundation for performing object detection and classification. There are two type of granularity within the segmentation process itself: semantic segmentation and instance segmentation. Semantic segmentation classifies all pixels of the image into meaningful object categories. These classes are "semantic interpretable" and correspond to the classes in the real world. For example, you can isolate all pixels related to cats and then color them green. This is also called dense prediction because it can predict the meaning of each pixel. Instance segmentation identifies each instance of each object in the image. It differs from semantic segmentation in that it does not classify each pixel. If there are three cars in the image, semantic segmentation will classify all cars into one instance, and instance segmentation will identify each car.

Image segmentation helps determine the relationship between objects and the context of objects in the image. Applications include face recognition, license plate recognition and satellite image analysis. Industries such as retail and fashion use image segmentation in image-based searches. Self-driving cars use it to understand their surroundings. Medical imaging-extract clinically relevant information from medical images. For example, radiologists can use machine learning to enhance analysis by segmenting images into different organs, tissue types, or disease symptoms. This reduces the time required to run diagnostic tests.

1.1.2 Object Tracking

Video tracking is a field of computer vision that involves the positioning of moving objects in video. Video tracking has many applications in robotics, motion analysis and video surveillance. These applications usually need to track multiple objects at the same time, which is called multi-object tracking. In video object tracking, the goal is to locate one or more objects of interest or targets in each frame of the video. We usually locate the target by drawing the smallest rectangle or bounding box that contains the target. Video object tracking has a wide range of applications, for example, it can be used for autonomous driving, surveillance, human-computer interaction, motion analysis. There is a close relationship between tracking and detection. Detection includes locating one or more objects in a given image, and the goal of tracking is to locate these objects in the entire video and track which objects along the video frame. In order to track an object, you first need to provide an image of the object to the tracking algorithm, which can be done by a detection algorithm (detection-based tracker) or manually (detection free tracker). Tracking is very crucial because it may help to solve common challenging problems, such as lighting changes, motion blur, zoom ratio changes, occlusion when the target is partially or completely hidden by another object in the video for a period of time, poor image quality. There are two main approaches of tracking: single object tracking (SOT) and multiple object tracking (MOT). Almost all trackers with the best performance are trackers based on Siamese network or Correlation Filter (CF), combined with effective appearance models (CNN function, HOG, color name). In general challenges, most of the highest performance is based on CF trackers, their performance is better than Siam trackers, and for real-time challenges, most of the highest performance is Siam trackers, their performance is better than CF tracking Device. As the name suggests, in multi-object tracking, there are multiple objects to track. It is expected that the tracking algorithm will first determine the number of objects in each frame, and secondly, track the identity of each object from one frame to the next. MOT is a challenging problem: ID switching is difficult to avoid, especially in crowded videos,

and the nature and number of each frame are unknown. Therefore, the MOT algorithm strongly relies on the detection algorithm, and the detection algorithm itself is not perfect. A popular object tracking method is to use a method called tracking by detection. It uses object detection algorithms to detect objects present in the frame. These objects are then tracked by associating the objects in the current frame with the objects in the previous frame using a tracking algorithm. Having a reliable object detection method is crucial, because the tracking algorithm depends on the object detected in each frame. Recently, target detection algorithms based on convolutional neural networks have been able to achieve higher accuracy than traditional target detection methods. The improvement of object detection accuracy promotes the use of one-by-one detection and tracking method for multiple object tracking.

1.2 Context and scope of the thesis

1.2.1 Egocentric vision

Egocentric vision or first-person vision (FPV) is a subfield of computer vision, which requires the analysis of images and videos captured by a wearable camera, which is usually worn on the head or chest, and naturally approximates the camera wearer Vision. Therefore, the visual data captures the part of the scene where the user is focused on performing on-site tasks and provides a valuable perspective to understand the user's activities and their environment in the natural environment.

In recent years, the research community has adopted a self-centered perspective to solve computer vision challenges, such as activity recognition [12] and object detection [11] that are traditionally considered to belong to the field of third-person vision. Since then, self-centered vision has been applied to more complex applications, including video summarization [20] and social interaction analysis [33]. It is worth noting that it has also been extended to the field of healthcare [10], in which static camera systems tend to struggle to a greater degree with regard to privacy issues [28]. Ultimately, self-centered vision is associated with the field of augmented real-

ity to enhance human-centered applications that provide help for specific tasks [21], thereby enhancing human independence; applicable to human capabilities Damaged or reduced condition.

In the medical field, FPV can help build applications that aid people with dementia by recording the sensor carrier's day-to-day activities; In rehabilitation therapy, or motor support in the elderly, physicians are often interested in monitoring the patient's recovery progress through movements and daily activities such as arm lifts. , move the wrist in the grasp object. Currently, such monitoring tools are very limited in hospitals, mainly using the naked eye for observation. Through the automatic analysis and recognition of activities from the series of images captured by the carrier camera (FPV video), the treating doctor can identify and quantify the patient's progression for a therapeutic regimen. value accordingly.

In sports, the use of egocentric cameras is increasingly common. The egocentric systems don't just collect front-view imagery during the journey of speed sports such as cycling and skiing; but also supports analysis of accuracy in sports movements such as golf, basketball. One notable feature is that the FPV image sequence not only contains information about the ego-motion of the object itself, but also the interactive movements between the hand and the subject (shadow, golf club) simultaneously. In sports, the recording of movements at critical moments (ball contact point, hand movement direction) plays an important role in determining the athlete's success or failure.

In the field of teaching aids, virtual reality, the use of FPV techniques has brought remarkable results. It can be considered the role of the camera carried on the body as a sixth sense (six-sense). On the one hand, FPV assists in detecting the behavior (e.g. hand gestures) of the subject, on the other hand, the camera carries the observer or the affected person. From there, the system reacts in accordance with the user's requirements. One particular application is the interaction between an observed object and the person carrying a sensor while visiting a museum; The system can on the one hand detect the object that the visitor is interested in, on the other hand identify the behavior (gesture) that the visitor wants to interact with the system (to support

details, or see objects from different perspectives). In educational applications, virtual reality is receiving more and more attention besides medical applications when developing Ego-centric vision systems.

1.2.2 Background project

This thesis falls under the scope of the project âĂIJUnderstanding Human Daily Life Activities from Egocentric Vision Using Advanced Technologies of Deep LearningâĂİ under the sponsorship of National Foundation for Science and Technology Development (NAFOSTED). Egocentric Vision is a carry-on image sensor system in which the person carrying the image sensor plays a central role. The problem of identifying the sensor carrier's activity plays an important role in the development of a number of medical applications such as dementia treatment and functional rehabilitation; in education and sports. The objective of the topic is to identify the sensor carrier's activities through analyzing two important factors: the role of hand gestures in interacting with objects; and the role of the environment in identifying sensor carrier activity. To achieve these goals, the topic will focus on solving the fundamental problems of egocentric-vision such as: condensing video from large data sources (up to millions of images / day), effectively exploiting relationships. system between the sensor carrier and the interactive object / object; and environment / context. Specifically, the project will conduct three main research contents as follows: (1) developing advanced deep learning techniques for the problem of segmentation, identification (context), hand operation under the perspective of sensor to carry; (2) How to effectively combine information sources (such as environmental factors, time, operating status); (3) Based on the proposed techniques, develop some medical applications; to support the elderly and the disabled in daily activities. The direct result of the topic is to form a strong research group on Egocentric-vision at MICA International Research Institute, Hanoi University of Science and Technology at the end. The outstanding feature of egocentric vision is that it provides a first-person perspective by placing a forward-wearing wearable camera on a personâĂŹs chest or head. This provides a unique human-centered view and is set in an optimal way to capture in-

formation that is arguably more relevant to the camera wearer [15]. Naturally, this refers to the surrounding area and its content, usually composed of objects, hands, other people and the background of the scene. Being able to check the angle of the scene and collect all this information clearly, so that higher-level cues can be better inferred, such as quantifying the interaction between hands based on the proximity of the hands [27], and performing objects based on the associated movement [3].

The role of this thesis in the project is to focus on detecting, segmenting and tracking human hand in egocentric videos. Accompanying the theoretical research result is the construction of a pipeline that allows training and integration of detection algorithms with tracking algorithms. The evaluation results on the test dataset of a combination of algorithms will be very useful in detecting and tracking human hands and are a premise for the next research stages.

1.3 Related works

1.3.1 Video object recogniton and tracking related works

Densely Annotated Video Segmentation (DAVIS) challenge [33] is a fairly famous public competition designed for the task of video object segmentation that has been going on for 4 years from 2016 to 2020. It is a benchmark dataset and evaluation methodology for video object segmentation that consists of fifty high quality, full HD video sequences, accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. It provides a comprehensive analysis of several state-of-the-art segmentation approaches using three complementary metrics. These are: semi-supervised challenge, interactive challenge and unsupervised challenge with specific datasets, definitions, rules and evaluation metrics.

The popular Visual Object Tracking (VOT) challenges [17] provides the visual tracking community with a precisely defined and repeatable way to compare short-term trackers, and provides a common platform for discussing evaluation and progress in the field of visual tracking. The goal of the challenge is to build a large database of

benchmarks and organize seminars or similar activities to promote research on visual tracking.

Multiple Object Tracking (MOT) [26] challenge is a well-known benchmark for multi-object tracking that collects a large variety of sequences and provides a framework for the standardized evaluation of multiple object tracking methods. Currently, the benchmark is focused on multiple people tracking, since pedestrians are by far the most studied object in the tracking community. The benchmark includes 2D, 3D and multi-camera challenges. The tracking evaluation is this thesis uses the devkit protocol of the MOT16 which provide several measures, from recall to precision to running time.

1.3.2 Hand recognition related works

One of the first works on the understanding of egocentric activities focused on defining the inner message of the egocentric visual paradigm [12]. They use the extracted visual features to model the relationship between hands, objects, and actions to model activities, and demonstrate the mutual improvement provided by these relationships through bottom-up and top-down models. The project of MICA is also based on the concepts that hands and objects are essential for egocentric action recognition and video understanding.

In [11], the authors have developed a weakly supervised technique that can identify objects by sculpting out objects from large sequences of self-centered activities. In general, this is a difficult task, their algorithm uses domain-specific knowledge from a first-person perspective to make it feasible. The method will automatically segment the active object area, assign some areas to each object, and use semi-supervised learning to disseminate its information. They proved that this method can reliably compare active classes based on the usage patterns of objects.

Movement-based egocentric action recognition is described in [23]. According to the motion and color-based features and trajectories extracted from the video frames, the interaction points of the hand and the object, the object, the head and the self-movement are declared as actions, but the position of the modeled hand is not paid

special attention. [25] describes another multi-modal approach to egocentric activity recognition. Here, the hand segmentation network, the object localization network and the network trained by the motion flow are combined to predict the action.

In [4], an architecture based on two-stream visual segmentation was used to predict the interaction area between hands and objects in a video stream and model them as actions. In [8], the concept of hand-object interaction is further explored, and the object shape-related grasping related to modeling actions is detected. The end-to-end approach also includes [2], where in order to recognize actions, the network is trained on paired frames and optimized for action recognition, object segmentation and inter-frame object interaction, as well as their training targets associated with recurrent networks. This thesis also intuitionistic that hands and objects are the basis of self-centered actions, but it emphasizes the need to clearly detect the areas and positions of hands and objects to recognize actions.

In the field of egocentric action recognition, a lot of work has been done in the explicit exploration of opponents and objects and their temporal relationships. Hand detection, segmentation and recognition techniques were developed [22] [5] [6], and the results were used to model behaviors or activities. In this thesis, I rely on a single-frame object detector to detect hands. In [1], hand-based activity recognition from an egocentric perspective is discussed. Before inferring the activity, the EgoHands dataset and the egocentric hand detection and segmentation pipeline were developed. This is one of the manual works, showing the difference between relying on hand detection or segmentation and using manual label for activity classification. In the literature [27], activity recognition based on egocentric hands is considered, in which the distance between the detected hands or the distance between the detected hands and the objects marked as activities is regarded as the feature of activity classification.

In [16], the egocentric human action recognition is solved by explicitly using the existence and location of the region of interest in the scene without further use of visual features. Their understanding that the human hand is very important in performing actions, and focused on its actions as the main clues to define actions is similar to this thesis, but I focus on comparison the affection of the detector (RCNN

Table 1.1: Problem formulation.

Input	Sequence of frames from an egocentric surveillance video
Output	Trajectories of the tracked hands, includes: • The number of human hands in the current frame • The location of each hand, represented by a bounding box for the detection algorithm, or a set of pixels for the segmentation algorithm • Identity of each hand across the video

and YOLO family) and tracker (SORT [7] or DeepSORT [31]) on locating the hands and capturing its movement. Finally, in all the works above, the suggestion on the choice of a suitable methods for online tracking applications is unavailable.

1.4 Problem formulation and assumptions

Table 1.1 define the input and output of the problem in this thesis. The goal of this thesis is to solve the problem of detecting, segmenting and tracking human hand objects in the video from the first perspective. Along with researching, confirming the theory, and proposing a hand tracking by detection system, I built a software to detect and track human hands in videos from the first perspective. The tracking by detection approach uses the combination of a detector and a tracker as following:

- Detector: Faster RCNN, Mask RCNN, MaskRCNN with region based, YOLOv3, YOLOv4 or Ground-truth
- Tracker: SORT or DeepSORT

The inference results on test dataset are used to analyze the effect of phase detection algorithm selection on hand tracking phase, and also the efficiency of the DeepSORT algorithm compared to the SORT algorithm in the tracking phase.
The main contributions of this thesis are three-folds:

- First, a hand detection, segmentation and tracking from egocentric vision pipeline is built. This platform contains many state-of-the-art detection algorithms in RCNN and YOLO families, and also 2 tracking algorithm SORT and Deep-SORT. Other algorithms can be simply integrated into this framework.

- Second, a comparative evaluation of combination of a detector and a tracker is conducted and analyzed on both term of accuracy and performance. Consequence, the recommendation for choosing the method on egocentric hand tracking applications is given.

- Third, MICAND32, a new egocentric hand detection, segmentation and tracking datasets is built during this thesis, accompanied with labeling, visualizing and evaluation tools. Part of this thesis is inherited from a part of my previous work [MAPR2020]. Code in this thesis is open-source and will made available at .

Chapter 2

Methodology and Datasets

2.1 Tracking by detection approach

Tracking multiple objects is the task of assigning unique and consistent identifiers to multiple objects in a video series. This article investigates an object tracking technique called "tracking by detection". Tracking through detection is a two-stage process: the object detection algorithm first detects the objects present in the frame; then the detection is done. These objects are then linked to those that have been tracked through a tracking algorithm. Usually, the object detection algorithm and tracking algorithm are completely separate from each other, so they can be analyzed separately. Object detection is the process of detecting specific categories of objects in an image, examples of these categories are things such as pedestrian or car. The purpose of the object detection algorithm is to locate and classify objects belonging to any popular category. Therefore, for each detected object, the object detection algorithm produces an estimate of the object's location, size, and category. The position and size of the detected object are usually represented by a bounding box, which is a rectangular box surrounding the object. The range of the detected object can also be defined by a segmentation mask, which is a pixel-level mask of the object. Due to recent advances in the field of image classification, target detection has made considerable progress. This progress is attributed to a breakthrough in how to use CNN for image classification [18]. The target detection algorithm considered in this

paper usually consists of a CNN designed for image classification, and then has an algorithm-specific additional structure around the CNN. CNN is called the backbone of the algorithm, and the algorithm-specific structure is called the meta-architecture. By convention, this paper will identify object detection algorithm through its meta-architecture. CNN-based object detection algorithm can be divided into two different groups: single-stage and two-stage detectors [9]. The two-stage detector first is possible bounding boxes by segmenting the image into regions of interest, and then CNN classifies these regions in the second stage. The single-stage detector is bounding box and class estimates in a single forward pass of the image through the CNN. Traditionally, two-stage detectors have achieved higher accuracy at the expense of speed compared to single-stage detectors. However, the recently introduced loss function Focal loss [24] makes the accuracy of a single-stage detector close to that of a two-stage detector. The trade-off between speed and accuracy is the main design choice, as studied in the paper [13]. The tracking algorithm in the "tracking by detection" framework is responsible for assigning unique identifiers to the tracked objects and establishing object associations between frames. This article focuses on target detection algorithm, and only considers two different tracking algorithm: SORT and DeepSORT SORT stands for simple online and real-time tracking. It is a deliberately simple tracking algorithm that uses a Kalman filter [14] to estimate the future position of an object, and uses the Hungarian method [19] for frame-to-frame correlation. Deep SORT is an extension of SORT that incorporates appearance information when performing object association between frames.

2.2 Object detection and segmentation algorithms

2.2.1 RCNN model family

2.2.2 YOLO model family

2.3 Object tracking algorithms

2.3.1 SORT

2.3.2 DeepSORT

2.4 Egocentric vision datasets

2.4.1 GTEA family datatsets

2.4.2 EgoHands dataset

2.4.3 Micand32 dataset

something

Chapter 3

Proposed Framework

3.1 Proposed framework: tracking by detection

The framework proposed in this thesis consists of 4 stages: 1. data preparation stage, 2. training stage, 3. inference stage, 4. evaluation stage. Initializing with data preparation stage, the GTEA family and EgoHands datasets is collected and is pre-processed in order to keep only suitable and related ground-truth samples. These 2 datasets are used to construct EHTA model zoo by training hand detection and segmentation models from the first-person view. The annotators use semi-automatic EHTA tool to create the new dataset Micand32 descripted in chapter 2, sub-section 2.4.3. Next is the models training stage. Three datasets including Micand32S, GTEA family and EgoHands datasets is used as fuel to train the hand detection and segmentation from egocentric vision, family of RCNN models and family of YOLO models. At the same time, a deep appearance descriptor is trained on distinct hand groups ground-truth in Micand32S. While the original DeepSORT's descriptor is made in people tracking context, the descriptor trained in the thesis is made well suited for deep metric learning in this egocentric hand context. Detail training techniques are reported in section 3.2. Following the training stage is the inference stage in which all the videos in Micand32 datasets is tested. For each sequence of n frames, each frame in turn is fed respectively into the detection and segmentation phase to get the prediction which locate the handâĂŹs bounding boxes, masks and their confidence scores. This phase

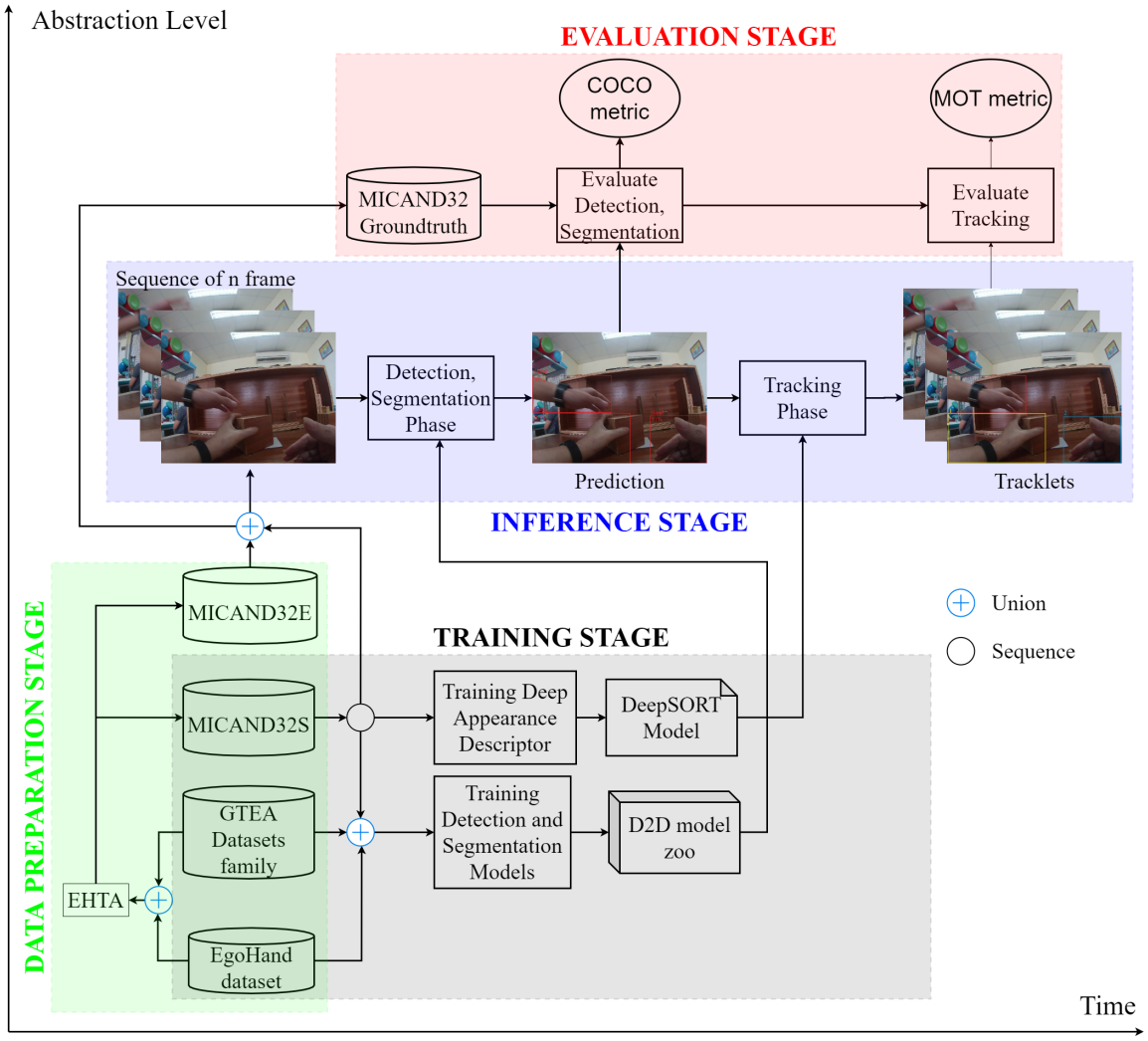

Figure 3-1: Overview of proposed framework: D2D. The x-axis represents the time flow of the 4 stages. The y-axis regards the increasing degree of abstraction level of the stages.

requires user to select detection or segmentation algorithm from pre-trained D2D model zoo. Frames with detections are then fed into tracking phase. This phase also requires user to choose tracking algorithm SORT or DeepSORT. The tracking phase indicates the identifications of egocentric hands with their trajectories in the whole video. Section 3.4 explains in detail the inference procedure. Finally, the evaluation stage encounters. Detection bounding boxes and segmentation masks is be evaluated by comparing with Micand32âĂŹs ground-truth. This result is measured in COCO format. The handâĂŹs tracklets is evaluated and reported in MOT Challenge format. Detail evaluation criteria is described in section 4.1.

3.2 Training stage

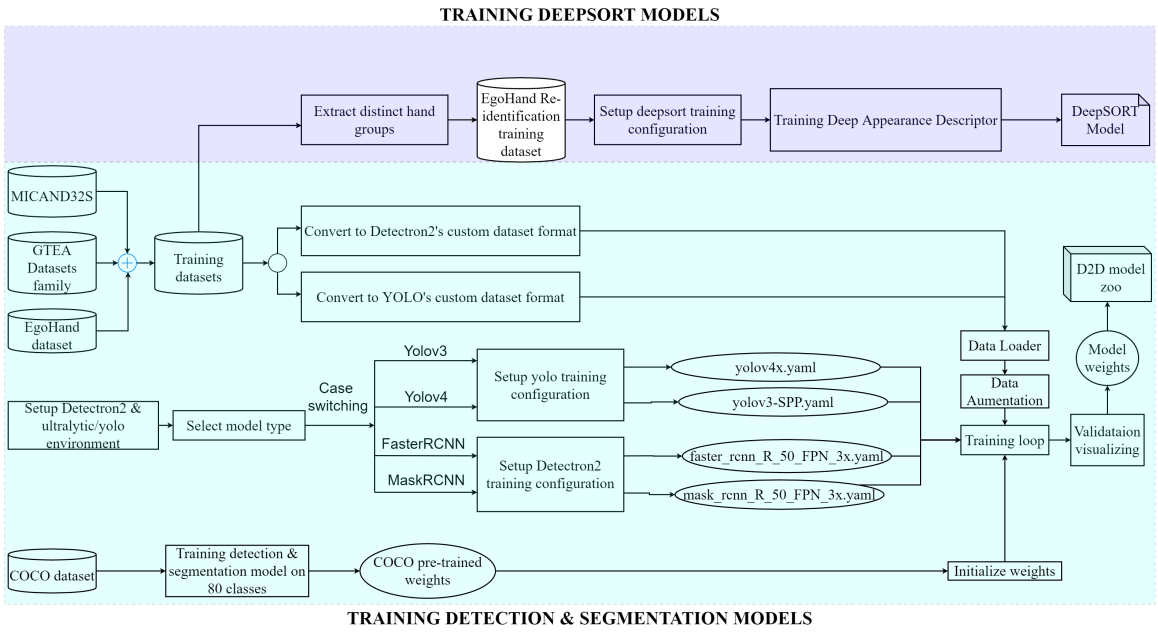

Figure 3-2: Workflow of the training stage.

The training stage shown in 3-2 consist of parts: 1. training detection and segmentation models, 2. training DeepSORT model. Input of both parts are the combination of 3 datasets: Miand32S, GTEA family and EgoHands dataset. The first part will generate D2D model zoo which consists 4 types of models, 2 from yolo family and 2 from RCNN family. The second part trains a CNN for deep appearance descriptor

for cosine metric learning [30] in DeepSORT. Detail implementation is explained as follow.

3.2.1 Training detection and segmentation models

Initially I setup programming environment which consists 2 frameworks: Detectron2 [32] for RCNN model family and Ultralytics [29] for YOLO model family.

Detectron2 is a complete rewrite of the previous version Detectron, and it originates from maskrcnn-benchmark. The platform is now implemented in and powered by the PyTorch deep learning framework. Through a new modular design, Detronron2 is flexible and scalable, and can provide fast training on a single or multiple GPU servers. Detectron2 includes high-quality implementations of state-of-the-art object detection algorithms, including DensePose, panoptic feature pyramid networks, and numerous variants of the pioneering Mask R-CNN model family also developed by FAIR. Its scalable design makes it easy to implement cutting-edge research projects without having to spend the entire code base. The requirements of installing Detectron2 includes: Linux with Python 3.6+, Pytorch 1.4+ and torchvision that matches the PyTorch installation, OpenCV need by demo and visualization. In this thesis, I build Detectron2 from source. The compilers gcc and g++ version 5+ are required, and ninja is recommended for faster build. After having these prerequisites, I clone the Detectron2 repositories and install it via pip from the local clone.

The Ultralytics open-source research into future object detection methods is represented via this repository [29]. The requirements of install Ultralytics is Python 3.8 or later with all dependencies includes Cython, matplotlib, numpy, OpenCV, pillow, PyYAML, scipy, tensorboard, tqdm, pycocotools, scikit-learn, seaborn, coremltools and onnx. I build Ultralytics from source by cloning their github repository and install via pip.

After installing programming environments, users have to select the model type to train. There are 4 main models integrated in this thesis's framework, the RCNN family consists FasterRCNN and MaskRCNN, while the YOLO family consists Yolov3 and Yolov4. Depending on case of model selection, D2D switches to appropriate branch of

setting up training configurations. For the RCNN family, Detectron2 supports different backbone network architectures such as ResNET {50, 101, 152 }|, FPN, VGG16, etc. In this thesis, I choose the standard configs, FasterRCNN_R_50_FPN_3x and MaskRCNN_R_50_FPN_3x with backbone Resnet 50 layers, feature pyramid network 3x architecture. The config file is saved in a ".h" format file.

3.3 Training Deep Appearance Descriptor for DeepSORT

3.4 Inference stage

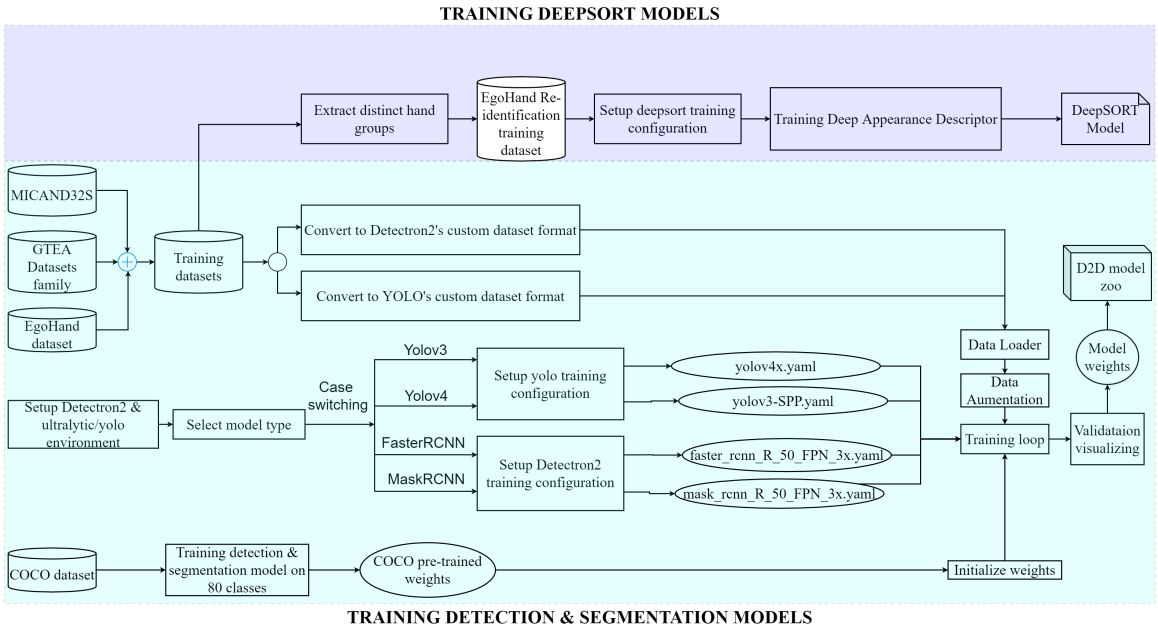

Figure 3-3: Workflow of inference stage.

3.5 Evaluation stage

Detail evaluation criteria and measurement metrics is reported in chapter 4, section 4.1. The evaluation stage includes 2 parts: (1) evaluate detection and segmentation

results; (2) evaluate tracking results. For detection and segmentation result, this thesis adopts the evaluation API from Detectron2 repository and the py-cocotools package. For tracking result, performance is measured according to the framework presented in [26]. The authors provide evaluation scripts for official development kit of MOT Challenge. MOT16 was chosen since it is a compilation of other many metrics developed in an attempt to standardize evaluation of multiple object tracking. This devkit requires Python 3.7+, Matlab R2020a, matlab python engine, pandas and pytz. Tracking result will be saved in simple comma-separated value (CSV) files.

Chapter 4

Experiments

4.1 Evaluation criteria

4.1.1 Object detection metrics

4.1.2 Object tracking metrics

4.2 Experimental results

4.2.1 Egocentric hand detection and segmentation result

4.2.2 Egocentric hand tracking result

4.3 Discussions

4.3.1 Object detection: tradeoff between accuracy and speed

4.3.2 The superiority of DeepSORT over SORT

4.3.3 Impact of detection method over tracking result

4.3.4 Complexsity of 4 types of patients's actions

4.3.5 Challenging cases

Chapter 5

Conclusion

5.1 Conclusions

5.1.1 Accomplishment

5.1.2 Drawback

5.2 Future works

Appendix A

Tables

content...

Table A.1: Armadillos

Armadillos	are
our	friends

Appendix B

Figures

Figure B-1: Armadillo slaying lawyer.

Figure B-2: Armadillo eradicating national debt.

Bibliography

[1] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957, 2015.

[2] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. *CoRR*, abs/1806.06157, 2018.

[3] Gedas Bertasius, Hyun Park, Stella Yu, and Jianbo Shi. First-person action-object detection with egonet. 07 2017.

[4] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. First person action-object detection with egonet. *CoRR*, abs/1603.04908, 2016.

[5] A. Betancourt, M. M. Lopez, C. S. Regazzoni, and M. Rauterberg. A sequential classifier for hand detection in the framework of egocentric vision. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 600–605, 2014.

[6] Alejandro Betancourt, Pietro Morerio, Emilia Barakova, Lucio Marcenaro, Matthias Rauterberg, and Carlo Regazzoni. Left/right hand segmentation in egocentric videos. *Comput. Vis. Image Underst.*, 154(C):73âĂŞ81, January 2017.

[7] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *CoRR*, abs/1602.00763, 2016.

[8] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016.

[9] Karanbir Singh Chahal and Kuntal Dey. A survey of modern object detection literature using deep learning. *CoRR*, abs/1808.07256, 2018.

[10] Aiden Doherty, Steve Hodges, Abby King, Alan Smeaton, Emma Berry, Chris Moulin, SiÃćn Lindley, Paul Kelly, and Charles Foster. Wearable cameras in health: The state of the art and future possibilities. *American journal of preventive medicine*, 44:320–3, 03 2013.

[11] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288, 2011.

[12] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, page 407âĂŞ414, USA, 2011. IEEE Computer Society.

[13] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[14] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[15] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[16] G. Kapidis, R. Poppe, E. Van Dam, L. Noldus, and R. Veltkamp. Egocentric hand track and object-based human action recognition. In *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 922–929, 2019.

[17] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84âĂŞ90, May 2017.

[19] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1âĂŘ2):83–97, 1955.

[20] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012.

[21] Teesid Leelasawassuk, Dima Damen, and Walterio Mayol-Cuevas. Automated capture and delivery of assistive task guidance with an eyewear computer: The glaciar system. In *Proceedings of the 8th Augmented Human International Conference*, AH '17, New York, NY, USA, 2017. Association for Computing Machinery.

[22] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2013.

[23] Y. Li, Zhefan Ye, and J. M. Rehg. Delving into egocentric actions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015.

[24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[25] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016.

[26] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[27] Thi Nguyen, Jean-Christophe Nebel, and Francisco FlÃşrez-Revuelta. *Recognition of Activities of Daily Living from Egocentric Videos Using Hands Detected by a Deep Convolutional Network*, pages 390–398. 06 2018.

[28] D. Townsend, F. Knoefel, and R. Goubran. Privacy versus autonomy: A tradeoff model for smart home monitoring technologies. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4749–4752, 2011.

[29] Ultralytics. Ultralytics. https://www.ultralytics.com., 2019.

[30] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. *CoRR*, abs/1812.00442, 2018.

[31] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017.

[32] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[33] R. Yonetani, K. M. Kitani, and Y. Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2629–2638, 2016.