Estimating Human pose from depth images using Convolutional Neural Networks

When all you got is a powerful GPU, everything looks like a CNN problem

Bård-Kristian Krohg

Thesis submitted for the degree of
Master in Informatics: Robotics and Intelligent Systems
60 credits

Institute for informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2020

Estimating Human pose from depth images using Convolutional Neural Networks

When all you got is a powerful GPU, everything looks like a CNN problem

Bård-Kristian Krohg

Estimating Human pose from depth images using Convolutional Neural Networks

http://www.duo.uio.no/

Printed: X-press printing house

Estimating Human pose from depth images using Convolutional Neural Networks

Bård-Kristian Krohg

7th January 2020

Abstract

This work is part of a larger project where we explore bringing robotics into geriatric care. The goal of this project is to create a robotic system that can assist in optimizing the use of physical personell, so they are used where they are needed.

This work will focus on capturing information about the user, anonymization of the data, what data is neccessary or ethical to capture, limitations for on-location data processing and what data can be sent for further processing in the cloud, or for human analysis.

We will also implement an ethical data-collection suite for the open-source Robotic Operating System, which can be implemented on a wide variety of robots.

Convolutional Neural Networks has been used for solving object recognition in 2D images with great success. This work aims to use the same techniques to extract 3D human pose from depth images in real time. We will use two multi-staged CNNs, one to encode the location of each joint, and another to encode the association between the joints to do this.

Preface

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Multimodal Elderly Care System

1.1.1 Motivation

As life expectancy increases in Norway, so does the population who needs geriatric care either at home, or in a geriatric facility. Accodring to [3], it is projected that mainly the elderly population in Oslo will increase in the coming years, and that this will lead to increased pressure on the healthcare services.

As part of the effort to let people live independently at home for as long as possible, we propose a Multimodal Elderly Care System (MECS). One of the goals for the MECS is to function as an autonomous safety alarm, a device that lets elderly living at home call the emergency services at the click of a button. However, if the emergency is an accident which renders the user incapacitated, or otherwise unconscious, an alarm that requires interaction will not be of much help. In contrast, the MECS can monitor a user, and warn healthcare personnel in case of an occuring, or predicted, emergency. The MECS will even be able to send contextual information to first-responders, which lets them better prepare for the situation.

We also strive to make the system non-invasive, as this will increase the convenience for the user, in that the MECS will not require any interaction from the user to function. Taking the user out of the operational loop has many other advantages as well. For example, a monitoring system in the form of a smartwatch, will be inconvenient and ineffective the user forgets to put it on.

Further, it is hypothesized that the information gathered by the MECS can help doctors or physical therapists prescribe or recommend health-promoting activities for the user. This could help prevent accidents or lifestyle diseases – which again will help relieve pressure on the communal healthcare services.

1.1.2 Hardware

The MECS is envisioned to be a small, mobile unit as this can be introduced to any home without extensive alterations to the environment, thus lowering the cost of the system and reusability of the units. As the MECS would need a charging station anyway, we propose a master/slave configuration between a stationary and a mobile unit, which should communicate through a secure wireless connection, for example a WLAN. We let the stationary unit take care of the power consuming complex processing, extending the operational time for the mobile unit between battery charges. We will therefore assume the system has access to mid- to high-end personal GPU/processing hardware, when we evaluate the real world practicality/runtime of the algorithm.

To provide as good a service for the user as possible, we believe that gathering many channels of information will be helpful. We wish to learn the users daily activity patterns or vital signs so when unhealthy or risk-filled patterns emerge, preventative actions can be implemented. Human Activity Recognition (HAR), gate/mood recognition, or detection of vital signs all require us to know where the user is in the scene.

This also places some requirements on our system. If the system solely relies on this work to find humans in the scene, we set our lower framerate limit to 8 fps to be able to preform human heart rate aquisition [6][1]. Further, the MECS needs to be able to recognize humans in unstructured environments, in a variety of poses and to diffrentiate between multiple people.

In order to log the users activity patterns, and detect anomalies or deviations in this, we envision the MECS doing HAR.

1.1.3 Privacy

The information gathered by this system should only be available to the users designated doctors/physisians and to the user themselves, and should be treated at the same classification level as any persons medical journal. This means that the dataprocessing from the MECS must happen on-site, or that any cloud processing happens under a user agreement that protects this data from those owning the servers. The same agreement should count for anyone making the hardware/software that is used in the MECS.

The MECS is also capable of gathering additional data that is not relevant for the healthcare personell. For example, by using depth sensors we are able to create 3d maps of the users home or environment. this is helpful for the MECS, however it

[1]If the maximum human heart rate is 220 bpm, and we want to measure it accuratley using video sequences produced by the MECS, our sampling frequency needs to be higher than $7.\overline{3}$ Hz in order to satisfy the Nyquist rate.

is not relevant for the healthcare personel. (with the exception of first responders, which could get the information either through a descriptive message – "the patient is in the bathroom on the second floor, no elevator, steep stairs" or via the actual internal map the MECS has created for its own internal navigation.)

1.2 Human Pose Estimation

For The MECS to function as intended, it is imperative that it can detect humans reliably. And for HAR we need accurate pose estimation.

Chapter 2

Background

2.1 Convolutional Neural Networks

2.2 Pose Estimation

In this work we are heavily inspired by [**cao2017realtime**] that uses two convolutional neural networks to find human pose. One of the networks finds the probability for the 2D location of a set of joints, we get N confidence maps for the locations of each N number of joints. The other creates M PAF the probability maps for M number of limbs[1]. We then use the results from this PAF to find out which joints from the first result should be connected.

Using the method described in [**cao2017realtime**] we however don't get any information if some pairs of joints are missing, leading to an incomplete skeleton.

A lot of research[2] has been going into extracting human pose either from RGB images, or using depth images. Although many methods exists such as *Histogram of Gradients (HoG)* classifiers,

A lot of research has been done in estimating human pose in two dimensions, as this is where we have quite large datasets, such as the MPII, or the Human 3.6M datasets [1, 2].

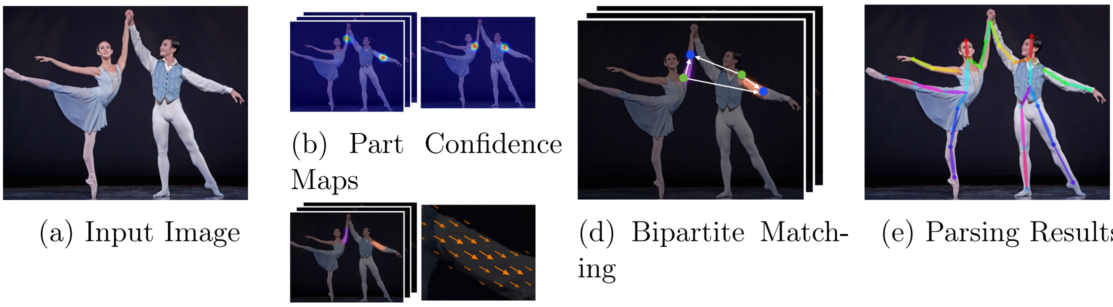
Background extraction

HoG classifiers

Motivation for 3D pose To train any network using supervised learning, we need large amounts of training data. One of the goals for the MECS project is to do *Human Activity Recognition (HAR)*, so one can track the user from day to day and look for patterns that could lead to worsening living conditions. We also want

[1] In this work we'll stick to the convention of using the term limb to describe any connection between any pair of body landmarks, which we call joints.

[2] TODO: Cite Research



(a) Input Image

(b) Part Confidence Maps

(c) Part Affinity Fields

(d) Bipartite Matching

(e) Parsing Results

Figure 2.1: The pipeline described in [**cao2017realtime**]. The input image ?? is fed into the two networks which produce joint detections in confidence maps ?? and PAFs ??. They then preform bipartite matching in ?? to determine which detected joints should be connected by a limb. ?? shows the finished results.

to be able to recognize the activity from any viewpoint, and this is where a 2D approach will lack robustness. This is because any HAR model trained solely on 2D data, will only be able to recognize the activity from the views it has seen the activity being preformed. A 3D approach will provide us with robustness in respect to view-independentness.

Chapter 3

Human 3D pose from depth images

We implemented an algorithm for extraction of human pose in 3D. Applying methods used on 3-channel (RGB) images to depth images, we show that the same methods can be used to extract objects in 2d images, can be used to extract objects in depth images as well, when it comes to human pose.

3.1 Training data preparation

Skeleton models

3.2 2D detection transfer

(First ideas.)

Torso placement and tree structure for placing.

Fit to human standard model (rules for symmetry and lengths)

Occlusion problem, and interpolated points, visual hull constraints

We train the network on both depth images and a kinematic model of each 3D ground truth location.

3.3 Pose from CNN over depth maps

We create a separate 'side-view' detection map for each 'frontal' detection map. This reduces the convolution operations, since we don't have to convolve over the whole 3D space.

Chapter 4

Experiments

Chapter 5

Future Work

Heart/respiration rate monitoring using frequency search in changing rgb and depth-pixel values for automatically selected RoIs.

Mood detection on facial expressions.

Human activity recognition using 3D pose provided by the method proposed in this paper.

Train the network over a larger dataset in unstructured environments and with multiple people present.

Train an accompanying network that takes a sequence of estimated limb positions and their probability as input, and trying to refine the estimation based on earlier detection. This could also be done through a Kalman filter.

This should all accumulate in an LSTM network for predicting diseases. – requires dataset aquired over possibly years, dispersed over many users, and their daily activities, as possible. Other factors that should be taken into consideration is environmental factors such as humidity, temprature and weather. (As they may be risk factors for certain conditions such as heatstroke or depression.) With such a diverse dataset we could possibly do PCA to determine certain risk factors for different diseases.

Train and test network on the Human 3.6M dataset using TOF data

Chapter 6

Conclusions

Appendices

.1 Depth sensors

Since depth sensors are widely used in different robotics applications for tasks such as SLAM, odometry and object detection, we selected this as our main source of information for monitoring the user. There are mainly three different technologies to choose from: *Structured light*, *Time-of-Flight (ToF)* and *Stereo vision*.

.1.1 Stereo vision

uses two cameras that are observing parts of the same scene. In commercial packages the cameras are usually calibrated, so we have measurements to put into the camera matrix as well as the rotation and translation between the two camera matrices.

However, to get a 3D structure, we need to find common feature points between the two cameras. To do this, we can use various feature descriptors such as ORB, SWIFT and SURF. When good matches has been found between the images, we measure the disparity between the points, and triangulate the distance. The depth measurements for the rest of the image are then calculated by matching pixels close to the found featurepoints.

Since this is an optically based technology, it will work well in well-lit scenes that contain many unique featurepoints. If we operate in an homogenous environment with few, or similar textures it will be difficult to find featurepoints to map the environment. An example of this could be on the seabed or inside buildings with limited light conditions, for example during a blackout.

.1.2 Structured light

uses a projected pattern of light points onto the scene which is registered by a calibrated camera. Usually, the projected light pattern and camera operate in the infrared part of the electromagnetic spectrum[1]. This means that in locations where one can expect a lot of IR radiation, this technology will not work very well. Since the IR radiation from the sun usually is much stronger than the one emitted from the projector on the sensor, this technology will not work well outside in well-lit conditions. It will however work inside and in conditions where no external light source are provided.

In addition, since the light is structured and the sensor is calibrated, we can skip the step where we find common featurepoints to triangulate the distance which we have to do in the stereo vision case.

[1]The Microsoft Kinect V2 sensor uses a wavelength around ~827-850nm stated by a developer in this forum. TODO : fix source

.1.3 Time-of-Flight

cameras uses the known constant of c to calculate distances in the image, by measuring the time a light-pulse emitted from the camera uses to be reflected onto the camera sensor. For example, Microsofts Kinect v2 uses a specialized ToF-pixel array in conjunction with a timing generator and modulated laser diodes to obtain per-pixel depth images [5].

As with structured light sensors, this is susceptible to interferrence from external light sources, or specular surfaces, and has limited range because of light fall-off. However, since the distance calculations are timing based, we can obtain framerates up to 30 fps in the Kinect v2 sensor [4].

.2 Robotic Operating System

In order to make the system easier to use and available to as many platforms as possible, it was decided to create it for the Robotic Operating System (ROS). ROS is a collection of libraries and a runtime environment making communication with different modules and programs on the robot possible.

.3 Classifiers

write a bit about what classifiers are, and how we use them to find the different keypoints in the image.

Bibliography

[1] Mykhaylo Andriluka et al. '2D Human Pose Estimation: New Benchmark and State of the Art Analysis'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[2] Catalin Ionescu et al. 'Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.

[3] Helseetaten Oslo Kommune. *Oslohelsa – Kortversjonen, Oversikt over helsetilstand og påvirkningsfaktorer*. June 2016. URL: https://www.oslo.kommune.no/getfile.php/13139280/Innhold/Politikk%20og%20administrasjon/Statistikk/Oslohelsa_kortversjon.pdf.

[4] Elise Lachat et al. 'Assessment and Calibration of a RGB-D Camera (Kinect v2 Sensor) Towards a Potential Use for Close-Range 3D Modeling'. In: *Remote Sensing* 7.10 (Oct. 2015), pp. 13070–13097. ISSN: 2072-4292. DOI: 10.3390/rs71013070. URL: http://dx.doi.org/10.3390/rs71013070.

[5] John Sell and Pat O'Connor. 'XBOX One Silicon'. Hot Chips 25. Aug. 2013. URL: http://www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.10-SoC1-epub/HC25.26.121-fixed-%20XB1%2020130826gnn.pdf.

[6] Hao-Yu Wu et al. 'Eulerian Video Magnification for Revealing Subtle Changes in the World'. In: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31.4 (2012).