Learning human motion features and trajectory predictions in large changing environments

Robin Eberhard, Stefan Dörr

***Abstract*—This study present a new approach to detecting humans in service environments. Making use of the latest advances in machine learning, fast scalable applications can be modeled without having knowledge about the inner workings. We present a leg detection tool, that operates on laser scan data, classifying each point as human or non-human. The results are further used to learn to identify people, allowing for precise operation of an autonomous robot in the service industry. This also includes prediction of trajectories, which will be used for movement of in crowded scenes.**

***Keywords*—IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

Autonomous robots are making their way from the production industry to the service industry. This change introduces a new set of problems, that deal with interactions of people and robots. Some of these problems can be simplified by first detecting persons, which then allows for example, to follow people, approach or evade them while navigating through an environment.

All of these problems can be solved with model-based solutions, but since neural networks are becoming more and more viable, we introduce a new approach, that does not require knowledge about people or their movements. Instead, the network will find characteristics on its own and place unique identifiers on persons. This is achieved by clustering the laser scan, but using additional information we find from the trained network.

Having the information and history of positions of people, we can learn trajectories, behaviour and intentions when the robot is serving customers. We provide a way to model the trajectories using Long short-term memory (LSTM) cells in combination with a Mixture density layer, which outputs a set of normal distributions, similar to *paper from graves*.

We will go through the models that are used to train the program, as well as some benchmarking and comparisons with similar projects.

II. Model architecture

The detection and identification of legs and therefore people based on laser scans is separated into several steps. First, the laser scans are convoluted in order to classify each point as either human or non-human. The classified scans are typically clustered based on the euclidian distance to each other. This approach does not work well, when people are close together or even when they are moving, as the program would then either cluster too many legs or see two legs as separate people respectively.

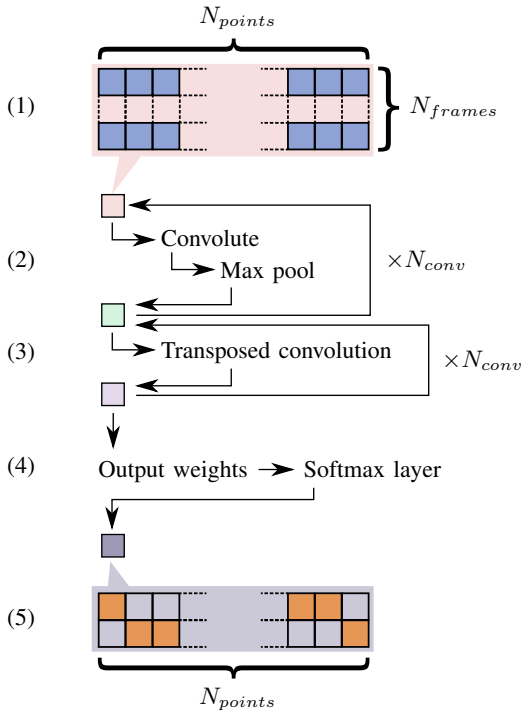

Fig. 1. **Laser scan classification.** (1) The input is a 2D array of $N_{points} \times N_{frames}$ containing distances from a laser scan. It is fed into (2), where features are extracted using convolutions and downscaled using a pooling layer. After N_{conv} iterations of (2), the now one-dimensional output is upscaled in (3) into the shape N_{points}. (4) multiplies the output and applies the softmax function, resulting in probabilities of each point belonging to class human or non-human in (5).

This is the base for the second step, where we use information from the convolutional layers as additional parameters to the clustering algorithms. The clustered points are used to identify a pair of legs as one person. The positions can then be fed into the third and last step of the setup, which aims to predict trajectories of individual people. The network takes a set of positions, that are run through a LSTM cell, which implicitly works as a memory for the program and holds the information necessary to build future trajectories.

A. Classification of the laser scan

In order to classify a laser scan, we take a snapshot of the current scan in equal timesteps. In each snapshot, we find one dimensional arrays of length N_{points} containing distances r_i to the closest intersections from the laser beams.

The network will find characteristics in the objective shapes of the laser scans. To also account for movement, N_{frames}

snapshots are composed together. The outputs are probabilities for each point in the most current snapshot to be in either class human or non-human.

Therefore, the mathematical description of the model can be described by an input of size $N_{frames} \times N_{points}$ and an output of size $2 \times N_{points}$.

The output is modelled from the input as outlined in Figure 1 and described in more detail in the following. When convolutional layers are used for classification, a network will train the weights such that they represent characteristics for a class. In this way, the characteristics are first of geometrical nature and become more abstract the more layers there are.

The input is therefore fed into a convolutional layer and then downsampled using a max pooling layer. This is repeated N_{conv} times, after which the output is a one dimensional array of length smaller than N_{points}.

To gain a classification for each point in the most current laser scan, the output is upscaled using transpose convolutions, again extracting features. Multiplication by output weights and applying the softmax function results in the desired shape $2 \times N_{points}$, where each entry is the probability of the point belonging to class human and non-human..

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.