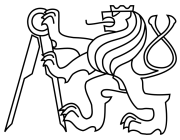
CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY



ASSIGNMENT OF MASTER'S THESIS

Title:	Information extraction about persons in images from a camera
Student:	Bc. Vojtěch Haur
Supervisor:	doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Study Programme:	Informatics
Study Branch:	Knowledge Engineering
Department:	Department of Theoretical Computer Science
Validity:	Until the end of summer semester 2016/17

Instructions

1) Familiarize with the task of extracting characteristics of persons in color and depth camera images from a retail shop and existing approaches, algorithms, and tools that are used for this task.
2) Suggest a procedure that can effectively and efficiently extract information about persons, e.g., their gender, whether they have a shopping cart and if they are customers or staff.
3) Prepare an annotated dataset for subsequent machine learning methods. Implement the proposed procedure for extracting information about individuals.
4) Verify implemented methods on real data and evaluate their accuracy. Suggest further improvements.

References

Will be provided by the supervisor.

L.S.

doc. Ing. Jan Janoušek, Ph.D.
Head of Department

prof. Ing. Pavel Tvrdík, CSc.
Dean

Prague February 18, 2016

Czech Technical University in Prague
Faculty of Information Technology
Department of Theoretical Informatics

Master's thesis

Information extraction about persons in images from a camera

Bc. Vojtěch Haur

Supervisor: doc. RNDr. Ing. Marcel Jiřina, Ph.D.

Tuesday 10^{th} May, 2016

Acknowledgements

I would like to thank Michaela Benešová for doing the demanding work of annotating the supermarket dataset. I would also like to thank Marcel Jiřina for supervising the work, giving valuable advice and encouraging me. I am very grateful to my family for support.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on Tuesday 10$^{\text{th}}$ May, 2016 .....

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Haur, Vojtěch. *0.0.0.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2016.

Abstrakt

Obrazová data zákazníků v supermarketu nesou podstatnou informaci, kterou může využít vedení supermarketu k přijímání obchodních rozhodnutí. Úkolem této práce je zpracovat sebraná obrazová data tohoto typu a prozkoumat možnosti automatizovaného vytěžování informací z nich. Postupem práce je příprava anotace dat, řešerše metod používaných v podobných případech, jejich implementace a vyhodnocení jak jsou vhodná pro tato data. Cílem je vytvoření základů, na kterých mohou stavět další práce.

Klíčová slova Supermarket, anotace datasetu, RGB obrazová data, hloubkové mapy, CNN, selective search, HOG, SVM.

Abstract

Image data of customers in a supermarket contain valuable information, which can be used by the supermarket management for business decisions. The task of this work is to deal with raw dataset of such images and explore the possibilities of automatized information extraction. The approach is to prepare an annotation of the dataset, research the methods used for similar tasks, implement them and evaluate how they can perform on this data. The

aim is to open the topic for future works and provide them with a basis upon which they can expand.

Keywords Supermarket, dataset annotation, RGB images, depth maps, CNN, selective search, HOG, SVM.

Contents

List of Figures

List of Tables

Introduction

Supermarkets can make a good use of information about their customers. Capturing this information can be resource intensive and usage of latest machine learning expertise might decrease the expenses and give access to not easily observable information.

Task description

The ultimate goal is to retrieve as much information about customers from visual data as possible. This work is introductory to this concrete dataset and some of its specific problems. The aim is to provide a starting point for further works which means it needs to cover a wide area of topics and lay the foundations for approaches that will not be carried out because they are out of scope of this work.

The first part of the task is to explore similar works and to get familiar with image processing focused on object detection and human attributes extraction. It incorporates exploration of currently used algorithms for image mining and finding similar datasets and tools that are appropriate for usage on the problem.

The dataset is a raw collection of images with limited preprocessing done on them and another part of the task is to create annotation providing machine learning algorithms with a ground truth inserted by a human annotator. This is crucial because it enables supervised learning which is much more accurate than unsupervised learning.[1]

The next part of the task is to implement and test machine learning methods for extraction of human attributes. This is a very general description of a procedure that consists of several steps such as image preprocessing, object detection, feature extraction and classification.

The last part of the task is to evaluate the proposed and implemented methods and suggest further improvements and alternatives.

Structure

The Introduction chapter opens the task involved with the supermarket dataset and outlines the structure of the work. More exhaustive description of the dataset is in the next chapter called Dataset. The data collection process is clarified, the properties of the dataset are specified and the entities are depicted. Next part of the chapter proposes the dataset annotation logic.

Theoretical part describes research on previous works with the aim to get familiar with existing approaches to similar problems. The chapter also tries to find comparable annotated datasets which would be used to boost training set for classifier training purposes.

The chapter called Experiments describes approaches applied to fulfill the task, namely the annotation tool implementation, object detection and gender recognition. It is followed by the Results chapter which explains the evaluation methodology and evaluates the output of the experiments.

The topics that are ambiguous or are important to note, are examined in the Discussion chapter. Future work describes the approaches that could not fit into the scope of this work, but seem to be important for the upcoming processing of the dataset. The whole work is wrapped up in the Conclusion chapter.

Chapter 1

Dataset

1.1 Dataset description

The supermarket dataset is composed of two sets of images, first are RGB images and second are depth maps of the same width and slightly smaller height. The sets should be paired, for each RGB image there should be depth map with the same name, taken at the same time. These images were captured by a Kinect v1 device placed about 2.5m above ground facing the aisle so that the shelves bellow them were visible (figure 1.5).

There is no accompanying description of the dataset, only the image sets (RGB and depth maps) are provided so the images are identified by their names in the form of mmdd_hhMMss_mss (mss = milliseconds). This work uses the names as identification, to pair RGB and depth images and for reading the time of capture.

There are three groups of different Kinect devices in different placements. The groups are called detergents, coffee and shampoos by the kind of goods that were sold in the aisles where the devices were placed. Furthermore the dataset is split by time of capture. As the data gathering was done in two days, the split is done by days and by hours. This logic is preserved even though it means that groups are unequal in size and object distribution.

The Kinect devices were set to capture images with frequency of 30 frames per second (FPS), which can be perceived as video and more storage space saving algorithms can be used. However, the dataset is stored as png images, the reason is to keep information of each individual frame for easier application of image mining tasks.

There was a computer connected to each Kinect device that was in charge of storing the captured data and also performing simple preprocessing. The

Figure 1.1: Example of RGB image from the supermarket dataset.

primary task of preprocessing was to discard images with no useful information (no moving object in the screen), which formed natural clusters of images by time proximity. This work gained the dataset after the preprocessing and did not have any impact on it.

1.1.1 Entities

There are generally three types of objects that can be seen in the images that are large enough to be considered of some interest. Most important ones are humans, then the shopping carts and shopping baskets. Other objects such as mobile phones, glasses, etc. can be seen held by people but are not frequent enough and are hardly recognizable at the resolution and image quality so they are omitted.

1.1.1.1 Shopping carts and baskets

The difference between a shopping cart (seen in 1.1) and a shopping basket (seen in 1.3) is that the cart is larger, made of metal and only can be pulled on the ground, while the basket is about third of the cart's size, made of plastic and can be either carried or also pulled on the ground.

The fact the baskets can be both carried and pulled is very important. Until

Figure 1.2: Example of a depth map from the supermarket dataset. Corresponds to 1.1.

relatively recent times the baskets in supermarkets could only be carried, but recently they were upgraded with wheels that enable the customer to pull them or carry them as he or she wishes. A basket that is carried is tightly bound to the person or even partially hidden from the camera view and thus should be considered an attribute of the person rather than an independent entity. On the other hand when the basket is pulled, it would make more sense to consider it a stand alone entity.

Both carts and baskets are very welcomed objects for image recognition because of the fact that their variability in the images is very low. That is caused by the fact that the supermarket uses a single type of each and also their load does not seem to affect their appearance, their features such as the ribs are clearly visible whether they are empty or full. Their position also seems to have small variance in the images.

1.1.1.2 Humans

Humans are very variable in appearance both for the reason that they look different, wear various thing and for the reason that their movement incorporates a wide variety of postures. This brings the challenge in a rapidly changing bounding rectangle in both height and width as the person bends, swings arms or stands still.

Figure 1.3: Shopping basket.

As it was stated before, the camera is at the height of 2.5 meter above the ground so the humans are captured at a rather acute angle, which is acuter when the person is standing closer to the shelf beneath the camera. This is one of the biggest differences from similar datasets, which tend to capture human from human level height or less [2]. An example for that is one of the currently most discussed problems which is pedestrian detection for autonomous car driving systems where the camera is inside the car or on its roof. [3]

Many machine learning algorithms are focused on feature recognition from face, especially these trying to classify gender or age [4]. In this dataset there are humans whose faces cannot be seen at all or only in few of the images during their continuous movement through the camera's arc of vision. Even if the face is visible it is far from the resolution and angle that the face based methods use and so they are not applicable to this dataset.

There is a large amount of information that is lost due to truncation when the object is coming into or leaving the scene defined by the camera's arc of vision (figure 1.4). In some of the cases it is impossible to recognize the attributes of a human because there are for example only legs to be seen. For evaluation of some machine learning methods it might be appropriate to remove images with object entering or leaving the scene even at the cost of significant reduction of the data.

Figure 1.4: Information lost due to truncation of object entering or leaving the scene.

1.1.2 Data damage

1.1.2.1 Image distortion

Two of the Kinect devices were corrupt and produced highly damaged images, where large rectangles are shifted, are covering useful image and have strange greenish color. Also the view of the shelf beneath the device is multiplied and sometimes covers more than a half of the image. This damage is unfortunately not repairable and these images need to be discarded.

The distortion does not seem to follow any pattern, it happens to single image in a bulk of valid images and to a big group of consecutive images as well. It can happen to affect as little as one image in hundreds or can also affect half an hour of data capture.

Only the devices in the aisle where detergents were sold were spared of the distortion and so this work primarily focuses on working with this part of the dataset. Whenever it is not stated otherwise, the detergents part of the dataset is the one used for experiments.

The damage of several images also means that a new task of recognizing images damaged this way has arisen in case that the problem with devices persists. However it seems to be a driver problem in Kinect and thus it might

be easier to solve with driver update.

1.1.2.2 Frequency

The ideal frequency of images would be 30 FPS with very little jitter. However, the frequency often dropped and was highly unstable. There are even cases with gaps of 2 seconds, which needs to be kept in mind when using the consecutiveness property of images.

1.2 Dataset annotation

The task of annotation is to provide ground truth for supervised machine learning algorithms and to gain general knowledge about the dataset, because any information gathered by observing only a small fraction of this rather large dataset can be deceiving.

As the annotation process is extremely time consuming it is required to describe as much information as possible to avoid the further need of repeated evaluation by a human annotator. This means that even if not all of the annotation information is used in this work, it might be useful in future works.

1.2.1 Bounding rectangles

First of all it is necessary to annotate the objects in each image by creating their spatial definition, which is called bounding rectangle (also called bounding box in other works). The rectangle needs to be as accurate as possible, because it ough to relieve the subsequent machine learning algorithms from the issue of finding the position of the object in the bounding rectangle.

The dataset is naturally separated into clusters by the fact that images without any object of interest were discarded by the preprocessing, which means that each cluster starts with the first object entering the scene and ends by the last object leaving the scene. Most of the objects are captured in multiple images in the cluster and especially for humans the images can complement each other with additional information, be it better vision of face, side and front view or just the fact that movement is defined by a sequence of images. During annotation it is crucial to create an ID for each object and keep it for all occurrences of the object in the cluster.

In fact it would be best to keep the same ID for each object, particularly humans, through the whole dataset, however that is almost impossible to do. It would mean that for each human the annotator would need to view all

already annotated humans to find a match or to realize that it is the first occurrence.

1.2.2 Attributes

The biggest benefit to the supermarket is the knowledge gained about its customers. Several properties that can be recognized by looking at a photo of a person were discussed for the annotation of the dataset. These were gender, age, social status, customer/employee disambiguation and the presence of a shopping basket. Gender seems very straightforward with reasonable human level recognition accuracy even though with the camera setup it might not be as good as we would expect in perfect conditions such as people meeting face to face scenario. Age recognition is much more difficult and even human level accuracy is rather poor [4]. Social status is not very well defined and people tend to be very deceptive about it, so it does not seem appropriate for the use in the annotation. Customer/employee is very useful for the interpretation of the findings based on machine learning tasks that would learn from the annotation. Shopping baskets were discussed in the previous section.

It was decided to use these attributes of humans in the dataset: gender, age, customer/employee and the presence of a shopping basket. Gender is binary attribute as expected: MALE and FEMALE values. The age attribute is binned to only three bins: child, young adult and old adult. That is because there is no access to the ground truth and with the quality of images the annotation would be very inaccurate. The binomial customer/employee attribute is highly specific to the dataset and is expected to be crucial for the supermarket business decisions.

The decision to consider shopping baskets to be attributes of humans instead of independent objects is based on the possibility to both carry and pull the basket as it was described in the previous section. Later the decision was found to be rather arguable because shopping baskets pulled on the ground greatly increases the size of human's bounding rectangle which ultimately increase bounding rectangles not only for humans pulling their baskets but for all humans in the dataset. This matter is later described in the Experiments section.

The other recognized object in the dataset is a shopping cart. It is identified by its ID through a cluster and there are no other attributes describing it. This opens the problematic of assigning shopping carts to humans controlling them.

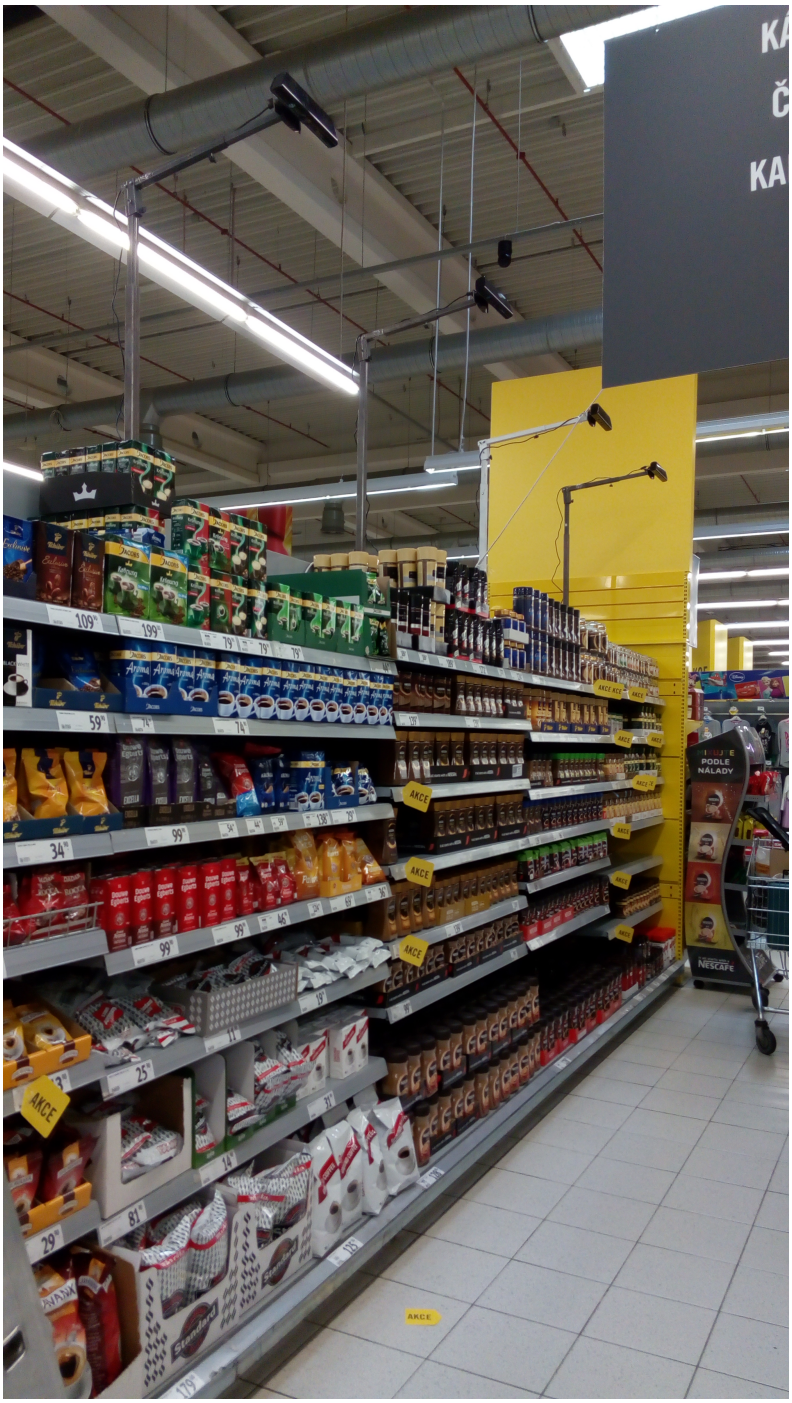


Figure 1.5: The Kinect devices placement in the supermarket

Chapter 2

Theoretical part

This chapter describes research on previous works covering similar topics. It explores the general image classification and object detection approaches and then it describes individual human attributes and how previous works deal with them.

The problematics of supermarket dataset is too complex to be dealt with as whole and need to be broken down to subproblems that have independent solutions. This also applies to the exploration of previous works covering similar topics, because there are many possible projections of what is going on in the images of the dataset. Such projection can be individual images and/or depth maps of humans, human postures, walking patterns, behavioral patterns etc. It is also necessary to consider that this work tries to cover classification of multiple attributes, such as gender or age. Other works tend to cover a single topic to reduce the complexity of the task and keep reasonable scope.

The general idea of an algorithm that should be deployed at the supermarket is a two step procedure. The first step is object detection which needs to pass the object locations and types to the next step that should recognize the attributes of these objects and any additional information about them for the supermarket to work with.

2.1 Object detection

The task of object detection corresponds to the identification of objects in the image and output their spatial representation as bounding rectangles. One part of object detection is classification which identifies the contents of

a detected bounding rectangle. Usages of classification relate to each task of this work in some extend and so need to be covered properly.

2.1.1 Image classification

The classification is one of the ground problems in machine vision and for this reason there is a high amount of competition which brings very advanced state-of-the-art methods. There are multiple famous challenges [5] such as ILSVRC [6], CIFAR-10 [7], STL-10 [8]. These challenges are defined by their respective datasets, which might be useful for boosting the training set for machine learning tasks on the supermarket dataset.

2.1.1.1 Traditional approaches

Traditional machine vision approaches split the classification process into two parts [9], feature extraction with algorithms such as scale invariant feature transform [10], histogram of oriented gradients (HOG)[11], bag-of-visual-words (BOW)[12] followed by classification with a standard classifier such as very popular support vector machines (SVM). [12]

There is no exact step between the era of convolutional neural networks (CNNs) and traditional approaches, in fact these two take turns in popularity. In the 1990s CNNs were more popular, in the 2000s it changed to HOG and its derivates combined with various classifiers, particularly SVM and the popularity returned back to CNN in recent years. [12]

2.1.1.2 CNN

The results show that deep convolutional neural networks achieve the highest accuracy rates in the challenges described in [5]. CNN differ from the classical approaches in the manner that here is generally one large neural network which does both feature extraction in its lower layers which form in fact various filters and classification which is performed in upper fully connected layers that are similar to layers of a multilayer perceptron. [13]

Major issue in using neural networks for image classification is that the fully connected layers used in multilayer perceptron have poor scalability for this kind of problem [14]. There is quadratic growth of state space relative to the size of the image in the means of memory, which is essential for training using GPU.

There are also increasing demands for the neurons to learn that nearby pixels generally describe more important patterns than distant ones. Providing exhaustive information of the whole image to each neuron in the input layer leads to overfitting[14].

The key concept in convolutional neural network, also called CNN or ConvNet are the specialized layers which serve various purposes, but when joined together they create feature extraction for the fully connected layers or even other manners of classification. [15] Most used layers are:

- **CONV** convolutional layer is perhaps the most interesting, it forms filters for convolution
- **POOL** pooling layer decreases the spatial size by applying max operation to a region, which downsamples the representation
- **RELU** performs elementwise operation such as thresholding and serves to apply non-linearity
- **INPUT** first layer which reads input image without modifying it
- **NORM** normalization layer, particularly useful for nets with unbound activations [16]
- **FC** this is classical fully-connected layer known from multilayer perceptron

It is important to note that all the layers are differentiable and that the filters formed by CONV layers alone or in combination with other layers are learnable.

Convolutional layer is a neural representation of filters. It is applied to the activations of the previous layer in the manner of convolution, however it retains the differentiable property of a layer of neurons. CONV is defined by many hyperparameters that define its behavior:

- **W**: input volume size, gives amount of pixels in each spatial dimension (usually x and y axes)
- **F**: receptive field size, specifies how many previous activations can a single neuron see
- **S**: stride, tells the move distance when the convolutional layer is shifted to next position of the previous layer activations
- **P**: padding, expands the border parts of input volume with zeros in order to enable convolution even in positions at borders where convolution would no longer be possible. Without padding the border areas would be quickly omitted by the network, losing information in the process.

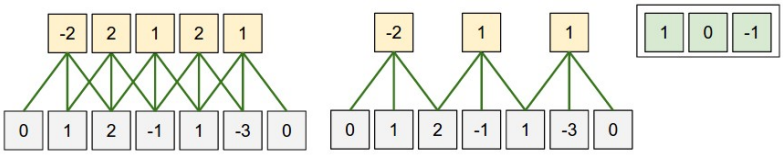


Figure 2.1: Illustration of spatial arrangement. In this example there is only one spatial dimension (x-axis), one neuron with a receptive field size of F = 3, the input size is W = 5, and there is zero padding of P = 1. **Left:** The neuron strided across the input in stride of S = 1, giving output of size (5 - 3 + 2)/1+1 = 5. **Right:** The neuron uses stride of S = 2, giving output of size (5 - 3 + 2)/2+1 = 3. Notice that stride S = 3 could not be used since it wouldn't fit neatly across the volume. In terms of the equation, this can be determined since (5 - 3 + 2) = 4 is not divisible by 3. The neuron weights are in this example [1,0,-1] (shown on very right), and its bias is zero. These weights are shared across all yellow neurons (see parameter sharing below). Image and caption borrowed from [14]

The illustration of the hyperparameters' meaning is in Figure 2.1

Pooling layer has important role in reducing the spatial size, enabling further layers to work on significant information already recognized by the convolutional layers. Typical size of the region is 2x2 with stride of two, efficiently dropping 3/4 of the information in favor for the most substantial 1/4 in the case *max* function is used (which is the most common). There are other functions such as *average*, which was popular in the past, but it has been found out that *max* outperforms all of them. In order to properly function in backpropagation algorithm, the layer should remember which of the previous activations were selected to pass the gradient to the right neuron. [14]

By definition the CNNs are deep neural networks, which means the negative aspects apply to them, especially the problems with gradient vanishing or exploding in stochastic gradient descent training. Various approaches to fight this issue are deployed, such as static initialization instead of random, partitioning the network into smaller networks with added auxiliary classifiers training feedback and discarding the auxiliary additions later [17], etc. However training is still very computationally intensive.

2.1.2 Localization

The other part of object detection is localization, finding the spatial arrangement of objects in the image. There are generally two approaches, region proposal and regression, however the former appeared to be superior to the latter [15] [14].

Regions with CNN features (R-CNN) described in [15] is a three step approach for object detection. First step is object location proposal performed by selective search algorithm, second is feature extraction by specific CNN which only has the lower layers and third is classification with SVM classifier.

2.1.2.1 Selective search

Selective search [18] was proposed in an environment where state-of-the-art method was an exhaustive search in combination with weak classifiers and methods that are computationally cheap. The exhaustive search approach has several specifics such as using class dependent width to height ratio of proposed rectangles and using weak classifiers to preselect interesting proposals for a cascade of classifiers.

Selective search uses segmentation in the image to produce object location proposals. The required properties of the algorithm are firstly **high recall**, making sure all objects in the image are proposed at the cost of also proposing uninteresting rectangles and secondly **fast computation** so that the method is not a bottleneck for the whole object detection algorithm.

The concept of selective search presented in [18] is to find segments of the image with a method prone to oversegmentation. The goal is to obtain small regions that do not spread over multiple objects. After the initial segmentation the regions are greedily joined together to form larger segments until a single one covers the whole image. Object location proposals are formed from the segments created during the joining mechanism, including the initial state.

2.2 Gender recognition

Gender recognition is a highly popular problem in computer vision for the reason that it can bring very valuable information from abundant resources that are images of people. Most of the applications however focus on recognition based on face, which is not useful for supermarket dataset as was described in the Dataset chapter.

2.2.1 Body

In the paper [9] the problematics of gender recognition from body are introduced. The authors of the paper needed a dataset for evaluating the methods they were discussing and state that at the time of the research there were no available databases of human bodies with gender annotation, which caused them to use MIT pedestrian database 2.2 [9] [19] which they manually labeled.

Figure 2.2: An image of human from the MIT pedestrian database [19].

The properties of the MIT pedestrian database are crucial for the comparison with the supermarket dataset. In the dataset used in the paper the people are either captured from the front (47%) or from the rear (53%), with standard size of 128x64 pixels and centering of the person in the image. Notable is also the fact, that there were 600 males and 288 females and 31 images were not labeled because the annotators were not able to recognize the gender.

The proposed methods in [9] correspond to the time the paper was published, which was January 2008. The authors first experimented with Canny edge detector for feature extraction, which they later replaced by histogram of oriented gradients. For classification the used algorithms were random forests and Adaboost using decision stumps. In the paper there is also introduced new part-based gender recognition (PBGR) algorithm which partitions the image by grid layout to describe parts of human body which can give various information such as skirt being usually worm by women. Each of the parts has a weak learner and the output is given by an ensemble of the learners.

Reported results of [9] are generally around 70% for most of the method

combinations (Canny edge detector, HOG, raw image + random forests, adaboost, PBGR) with best accuracy $76 \pm 1.2\%$ for frontal view and $74 \pm 3.4\%$ for rear view.

The possibility to fully reproduce the methods and results of the paper is hampered by the differences in the MIT pedestrian database and the supermarket dataset. The later has much more difficult setup with humans being captured from any angle, mostly from the side and they are generally not well centered in their bounding rectangles because of shopping baskets being considered attribute to humans. Another similar issue is that there is variation in bounding rectangle sizes while in MIT dataset all humans are in bounding rectangles of the same size.

More recent work [2] is very similar to the supermarket dataset because it focuses on gender recognition from RGB and depth images captured by a Kinect device, which means it deals with similar amount of information. The only missing component of information is the video part of supermarket data which would allow the classifier to skip some of the worse images and use the better ones for prediction.

The paper [2] describes the creation of a dataset captured for the purpose of evaluating this method. The dataset has been collected using Kinect v2 and ASUS Xtion Pro Live (internally similar to Kinect v1) mounted at about 1.5m above ground. There were 118 participants, 64 females and 54 males, with ages ranging from 4 to 66 with the mean being 27 years. There were various standing and walking patterns performed by the participants that should cover all of the possible angles at which the camera can capture humans as well as poses.

There are several methods described in the paper as previous work on the topic, most notable are HOG and Convolutional-recursive neural networks. These methods are then challenged by new proposed tessellation generation approach which splits the 3D bounding volume of human into subvolumes called voxels.

The results presented are highly promising for the achievable accuracy in this problematics. HOG performs in a similar manner as in [9] with about 70% accuracy in most experiments with an exception in easiest case that is standing pose, where it has about 84%. The other methods are more consistent, the neural network has 83-86% accuracy and tessellation method has 85-91%.

Unfortunately, the results are mostly based on data collected by the more advanced Kinect v2 device, which seems to offer far better depth resolution than is available in the supermarket dataset and because of that the accuracy

obtained particularly by tessellation method might not be achievable in this setup. Also, the difference in Kinect placement can be an issue.

2.2.2 Gait

One of the projections of the dataset is gait which refers to human walking patterns. The manner of walking can be used not only for gender recognition, which is one of the tasks of this work, but also for human identification [20]. It is a highly interesting field for this work as it brings an independent view of the data as opposed to image classification which generally processes single images.

Gait has recently seen increased interest in machine learning researches [21] because the data collection is simpler than data collection focused on other biometric features such as face, which rely on good resolution [20]. A camera for video capture can be placed further away from the human and be less disturbing. However even with lower resolution needed there are higher requirements for hardware because of the amount of images needed to capture gait.

The research of the topic is still in its early stages and therefore there are several approaches to tackle the problem. As with other image processing methods the main concern is feature extraction to give the appropriate interpretation of the images to subsequent machine learning algorithms, especially classification. Two of the many methods there are silhouette-based and model-based approaches [20].

Silhouette-based approach focuses on background subtraction trying to leave only the human in the image and then construct the feature vector from the edge of the human. Model-based approach tries to recognize body parts of the person, mainly the torso and legs and further work with them in manner of calculating angles, estimating stride and others. [20]

The CASIA B dataset [22] is interesting for this work from the point it contains gender annotations. Another notable fact is that the public part of the dataset is silhouettes only, which is very close to what depth maps look like. The gradient in depth maps can be converted to binary image that represents foreground and background with easy thresholding and separation of humans in the image in the case there are multiple of them.

The experiments run on CASIA B dataset presented in [21] show key concepts of utilizing gait in works on supermarket dataset. The reported results

of correct classification rate (CCR) are very high, the paper states 97.7% average CCR for the ideal scenario where the human is captured at the same angle for both the training set and the testing set. The additional results cover the scenario where the human is given a bag to carry for the capture of testing set only. This lead to average CCR of 67.8%. Another scenario for testing set was a clothing change which was represented by the human wearing a coat, which lead to a drop of average CCR to 28.9%. In the case there was difference in angle of capture for training set and testing set there is great variability reported, but 18 degrees difference lead to a decrease of CCR of more than 75%. The exhaustive report on the experiments is available in [21].

The accuracy of experiments on CASIA B dataset give us a warning that the applied methods are lacking in robustness and a lot of work would be necessary to apply these methods on the supermarket dataset, where customers can wander to shelves and their silhouettes are specific because they are pulling shopping baskets or pushing shopping carts.

The idea of using information retrieved from the CASIA B dataset needs to be abandoned for another reason. That is because the cameras used to capture the dataset were set in the manner that side view was obtained whereas in the supermarket dataset cameras captured humans in variable view from top to side according to which part of the aisle the human was walking in.

Any algorithm for classification based on gait needs to consider that there are numerous periods when the human in the supermarket dataset is standing still because he or she is watching the shelf, talking to a cell phone, etc. Proper preprocessing needs to eliminate such periods in order to present the next part of the algorithm with images of human walking, especially if another dataset is to be used to expand the training set.

2.3 Age

Correct estimation of age offers access to wide variety of external knowledge from surveys and researches and can prove to be invaluable for the supermarket. The annotation of the dataset however uses only very coarse classes of age, because the estimation is very difficult even at human level given the quality of images. That means the possible external knowledge would be unavaible even if the classification was extraordinarily good.

Human age prediction is a very popular field in machine learning. There are several datasets such as MORPH [23] available and there are many pub-

lications covering the topic [24] [25] [26] [27] [28]. Main distinction from the supermarket dataset is that these data and methods are focused on age prediction from faces and the datasets operate with real ground truth, not human estimation, which is rather poor in this domain [4].

Age is generally a cardinal variable and so both classification and regression approaches are viable if the granularity is years as it is natural for human age. However the classes used in the supermarket dataset (child, young adult and old adult) are no longer cardinal and so only classification can be used.

There are many methods for age estimation from faces, such as wrinkle pattern matching [28] , biologically-inspired features [26], active appearance models [27] and so on. The criterion of correctness is MAE (mean absolute errors) calculating how far is the estimation from the ground truth on average.

However, with the setup far different from all the previous works, the age estimation for the supermarket dataset will need to search for approaches developed for other image classification tasks. Methods performing well for gender recognition can perhaps be utilized for age estimation as well if they are not too specialized.

2.4 Basket and customer/employee

The two last human attributes in the supermarket dataset that were not discussed yet are the presence of shopping basket and the disambiguation between customers and employees. It would be naïve to expect these very specific aspects to be researched in some previous works.

For shopping baskets the task seems very similar to shopping cart recognition and methods that perform well for shopping carts can be expected to work with baskets. The main challenge should be to merge these procedures with human detection to follow the course set by the dataset annotation with baskets being human attributes and expand their bounding rectangles.

The task of disambiguation between customer and employee appears to be one of the most specific challenges in the dataset, because there seems to be a clear difference only in clothing. There is a hypothetical possibility to find disambiguation in behavior, however that is a greatly complex problem.

The initial observation of differences in the way customers and employees dress is that while customers obviously can wear virtually anything, employees are

required to wear an uniform, which in the case of this particular supermarket is a grey T-shirt with rarely observable small logo of the supermarket written in white letters. That means that all the discussed feature extraction methods that focus mostly on structure in the images, such edge detectors, SIFT, HOG and to some extent most of the convolutional filters in the CONV layers of CNNs are of little use for this task, because the spatial features are same to any ordinary T-shirt a customer can wear. Some color based feature extraction methods need to be utilized for the task.

Chapter 3

Experiments

The chapter Experiments describes the process of implementing both the methods researched in the previous chapter and the author's own algorithms to explore the problematics and propose solutions to the tasks. At the beginning the annotating tool is outlined, then a highly explorative experiment is described that tries the power of CNN on the supermarket dataset. The chapter continues with section object detection, which is generally split into detection for humans and for shopping carts. The last part describes gender recognition as the experiment for human attributes classification.

The general algorithm that is proposed by this work is composed of two parts, object detection and attribute classification. These are the cornerstones that need to be done, however this work does not intend to outline the complete algorithm in details as it is introductory work for supermarket dataset and the experiments are largely explorative. This means that multiple approaches to solve a single problem might be carried out and there are probably methods that will be discarded because they are not performing as well as they are expected to.

3.1 Annotation tool

A specialized tool there was created for the purpose of annotation. It was built using Surmon platform [29], loads the RGB images from the dataset, displays them and enables the user to draw rectangles into them. The user assigns either human or shopping cart type to the drawn rectangle and if human is selected, it also queries for information forming the dataset: gender, age bin, customer or employee and presence of shopping basket attributes.

One of the main requirements for the tool was to speed up the annotation

process and because of that the annotation tool displays four images at the same time. When user draws a new rectangle, the tool searches the previous images for rectangles of the same type (human or cart) and assigns the new one to the match, if the user does not explicitly state that it is a new independent rectangle. If the assignment is successful, the new rectangle retains the ID of its former state, all the attributes and there are rectangles interpolated into images that are in-between the images containing the old and new rectangles.

Interpolation offers up to three times speedup when the tool displays 4 images concurrently (figure 3.1). That is because the tool displays the images by their time sequentiality and when the user is done with the images currently shown to him or her, the tool displays the last image as the first and loads three new images. When everything is perfect, the user only annotates the last image of the newly displayed ones and the interpolated rectangles fill the two middle images.

The tool uses fully automatic persistency into json format, which is popular nowadays and most programming languages offer neat tools for loading from the json file and constructing objects from the loaded data. Each image is represented by a single object and the amount of small json files needed to cover the whole dataset would demand large amount of disk space because of filesystem block size. The file format used for output is special format called jsons, which stores single json as a string line. This enables having a rich text file and decreases both the storage space requirements and the number of disk I/O operations when loading the dataset annotation.

3.2 Cart classification

One of the first experiments with the supermarket dataset was to train convolutional neural networks to evaluate their classification capabilities. Because it was early in the annotation process when only a few hours were annotated, it was decided to try out the easy task of image classification — which means presence of an object of concrete class in the whole image. Although each of the hours used for this experiment had more than a thousand of images, there were in fact only a couple of unique humans with one of the employees reappearing multiple times. There was a considerable chance that the network would only learn attributes specific to these people such as the color of their clothing and thus overfit instead of learning general knowledge. For this reason shopping carts were selected for the recognition, because they look very similar to each other and it is possible to consider images of a single cart

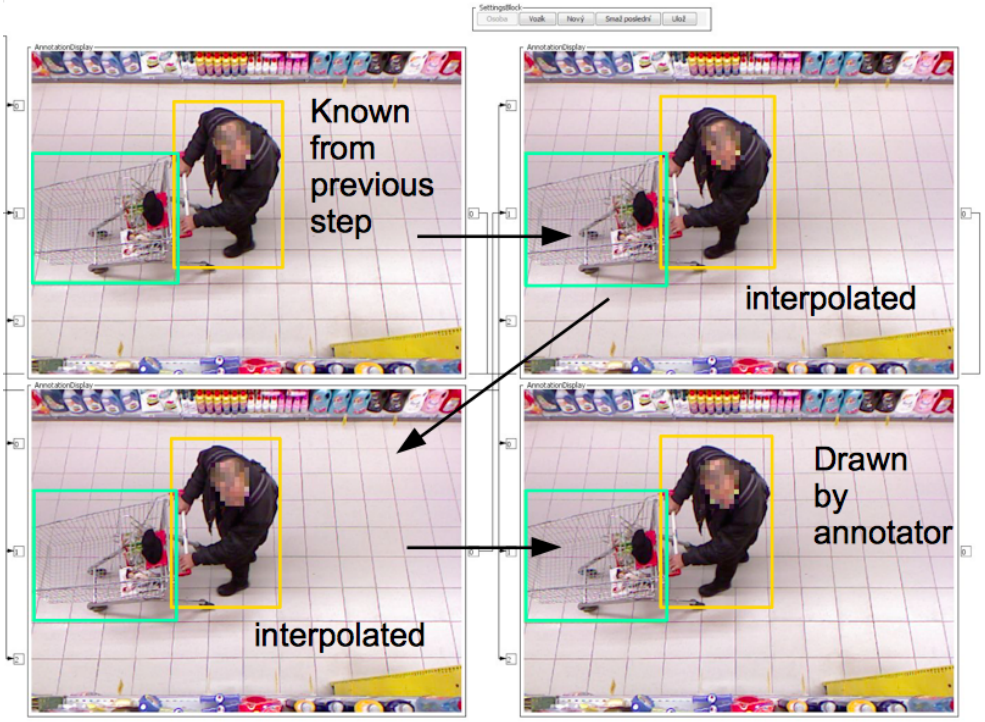


Figure 3.1: Annotation tool implemented for the annotation of supermarket dataset. The image displays are interactive and a human annotator draws rectangles into them. Note that the displays are original size of the images in the dataset to preserve information during annotation.

moving in the camera to be independent.

The framework caffe [30] was used with two networks implementations, CaffeNet and GoogleNet which have been pre-trained on ILSVRC 2012 [6]. This means that the lengthy process of training deep neural network, especially the lower layers has been done and it was only needed to fine tune the upper fully connected layers to recognize shopping carts.

Training of the networks was done using GPU to speed up the process. Because the GPU used has only 2GB of dedicated memory it was necessary to decrease the batch size to 16 and to downsize the images by half to 320x240 pixels. Resizing of images is a common practise with CNN, e.g. because it is favorable to have each dimension of the image divisible by 2 multiple times to avoid truncation at pooling layers [14].

3.3 Annotation automatization

Annotation of the dataset is one of the goals of this work and also a necessary step for classification of the customers' attributes as it provides ground truth for the classifier to learn from. The amount of objects in the image varies from most common case where there is a single human (sometimes pushing a shopping cart) to seven objects and that can be even exceeded with the theoretical limit being much higher. This means that the human annotator needs to draw a bounding rectangle around each annotated object to assign the annotation to it. However, this is by far the most time consuming step and needs to be eliminated in order to speed up the annotation process.

Automatization of annotation means that machine learning algorithms are deployed to recognize the objects that are recognized in the annotation process and draw the bounding rectangles around them in order to eliminate the most time consuming part of annotator's work and allow him or her to only insert the most valuable piece of information, which are the attributes of humans as described in the earlier part of this work.

The automatization is also very similar to a part of the final algorithm which should be deployed in the supermarket, object detection. Object detection would be responsible for recognizing and capturing individual humans and other desired objects in the frame of the camera and passing them to the next part of attributes extraction pipeline of the algorithm.

3.4 Object Detection

Object detection means finding the bounding rectangles for objects in the image. This task is significantly harder than image classification [15], because it contains image classification as a subproblem and is furthermore required to give precise object locations as well. Various methods are applied, however it seems that the sliding window approach is the most precise. [12] [15]

For both applications — the speeding up of annotation and as part of the deployed algorithm — the accuracy of object detection is crucial for the subsequent machine learning algorithms. These are very sensitive to having the object very well cut out of the original image to on one hand provide all the available information and not to truncate any part of the object but on the other hand not to include too much of the surrounding which is a noise and will deteriorate the following classification procedures.

Accurate cutting out of the desired object from the original image is the field where including shopping baskets as human attribute and not individual

objects like shopping carts has proven to be the biggest problem. And that is for two reasons, first of which is that a classifier needs to learn that shopping baskets are part of customers even though only a portion of the customers are equipped with them. The other is that if the customer pulls the basket, the bounding rectangle size increases more than twice. On the annotated part of the dataset the average width of a customer is 165 pixels, but maximum width is 340 pixels.

The increased width of a customer means that about half of the bounding rectangle's volume is noise and the person is not centered in the rectangle so any classifier needs to take that into consideration and work with spatially independent features or would need drastically higher amount of data to learn.

Different approaches were used for the object detection of the two recognized objects (humans and shopping carts). The reason is that the original dataset offers a different amount of information covering these two objects because depth mapping does not capture shopping carts, especially when they are empty. Either the depth sensor on the Kinect device has trouble picking on the ribs of the cart or the preprocessing step of the data collection filters the output of Kinect for carts. Either way it is obvious from the dataset that most of the times the shopping carts are not present in the depth map, even though the customer's selected goods are.

The Fast RCNN [31] is an out-of-box solution to this problem and an extensive effort was made to train this kind of network, but unfortunately without success. An instance of the network trained on the PASCAL VOC dataset which contains humans was available for download and was tested for the purpose of finding bounding rectangles for humans. From few testing images it became apparent that the network underperforms on supermarket dataset as it had problem finding humans even in the easiest cases and was giving false positives in the background. This means that different approach was proposed to solve this problem.

3.4.1 Humans

Humans are clearly distinguishable in the depth map part of the dataset (see 3.2) which offers an interesting application of clustering approach based on the depth images. The two main issues that need to be solved with this approach is the occasional presence of the items in the shopping cart in the depth map and multiple humans so close to each other that their images in the depth map connect which eliminates some of the simple clustering methods.

Figure 3.2: Example of depth map with two humans. Note one of them is pulling a shopping basket.

The applied clustering algorithm is mean shift analysis [32] due to its ability to operate without prior knowledge of the cluster count unlike other algorithms, e.g. popular k-means. Both the algorithm itself and function estimate_bandwidth were provided by the python library sklearn.cluster [33].

The first step of the implemented algorithm finds the brightest pixel in the whole image which represents the highest point of objects the Kinect device captured in the depth map. This requires scanning the whole image. It is followed by another full image scan that thresholds pixels by two values, first is static given as a parameter of the algorithm which serves the purpose of eliminating non human objects (keeping mainly items in cart in mind), second is dynamic calculated from the found maximum lowered by a constant given as a parameter. The pixels that pass this thresholding are collected to a temporary dataset consisting of their x and y values.

With the temporary dataset collected, the bandwidth is calculated by the method estimate_bandwidth to give it as mandatory parameter to mean shift. The next step is running mean shift analysis on the dataset in order to identify the humans in the image by the centers of clusters. This step is the most computation intensive and is sensitive to the parameter for calculating dynamic threshold in the first part of the algorithm, because it controls the amount of points that are given to mean shift.

The other pixels need to be assigned to their respective clusters. Another thresholding is done in order to eliminate possible noise and to allow the algorithm to output bounding rectangles of different sizes because the clusters consist of more or less pixels. All of the pixels passing thresholding in this step are assigned to clusters and for each cluster the bounding rectangle is defined by leftmost, topmost, rightmost and bottommost pixel.

3.4.2 Carts

As it was stated earlier, shopping carts are not captured in the depth map and detection of them needs to be done in the RGB map. With very well performing deep neural networks recognizing the presence of a shopping cart in the image the proposed algorithm tests various rectangular parts of the image for the presence of a shopping cart and tries to select the best rectangle with positive response from the cart recognition CNN.

The most straightforward method to find rectangles to test for presence of a shopping cart is exhaustive search [18], which guarantees to iterate over the best rectangle, however the quadratic complexity makes it too computationally demanding to use.

3.4.2.1 Primitive cart detection method

Primitive cart detection method is based on the idea of slicing off parts of the image for as long as the image contains a cart — which is information provided by the CNN. This basic method should only work when there is exactly one cart in the image.

The first part of the algorithm tests if there is a cart present in the image at all to make sure it is working on a valid image. Then the algorithm cuts off slices of the image from left, right, top and bottom. The slices size is given as a parameter to the algorithm relative to the image, e.g. one fifth. Slicing from each direction continues as long as the CNN recognizes that there is a cart in the remaining image and also stops cutting from a direction and returns the last slice if if was recognized to contain a cart.

The method is appropriate only for images containing a single shopping cart and requires an extension to be properly deployed for the task of finding bounding rectangles of carts. This extension would be based on splitting the

image into several parts, e.g. finding ten vertical and ten horizontal strips and testing those strips (single ones and neighbor pairs) for the presence of a cart. After finding the strips containing shopping carts the image would be split to multiple images that would be passed to the implemented primitive cart detection method which would find the bounding rectangles.

Because the main part of the primitive cart detection method did not perform well, as is further described in the Results section of this work, the extension enabling to recognize multiple carts in a single image was not implemented.

3.4.2.2 Candidate object locations method

Part of the Fast R-CNN training is the selective search algorithm [31] which proposes object locations for the network to tell what is the distribution of classes of the objects in the rectangels. The method described in this section is based on approximating the R-CNN run by combining trained CNN and selective search, but letting the network do the recognition instead of using another layer of SVM classifiers, how it was done in [15].

Selective search is a state-of-the-art method for object location proposal and is described in detail in the Theoretical part section. The implementation from DLIB python library[34] named find_candidate_object_locations was used. It produces thousands of object location proposals for an image of the supermarket dataset (figure 3.3). However the neural network was trained on downsized images and it has best accuracy using images of the same resolution. So the images for this application were downsized to half (320x240 pixels) and proposed object location count dropped by about ten times to the matter of hundreds rectangles.

3.4.2.2.1 Single selected rectangle The first approach was to select a single rectangle from the proposed set, which would ideally have the best coverage of the actual cart. It was observed that the proposal method offers even very large rectangles covering more than half of the image and also one of the few parameters of the find_candidate_object_locations function is the minimal size of regions from which the rectangles are constructed, effectively eliminating the smaller rectangles from the set. Both these findings lead to the statement that the smallest rectangle containing a cart according to the CNN is the best one. Because testing the rectangles with the CNN is computationally intensive, very simple optimization in the form of sorting the set of rectangles in ascending order by area is used.

Single rectangle approach suffers from two problems, one of which is crucial. The lesser in importance is the realization that the neural network confirms the presence of a cart even for smaller rectangles that only partially contain

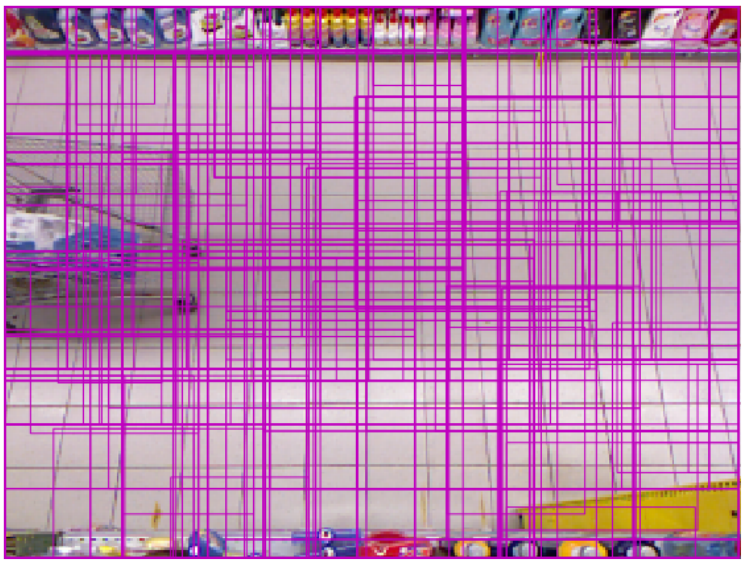


Figure 3.3: All object location proposals in RGB image from the supermarket dataset.

a cart (FIGURE C.3) or they are too small to cover the whole cart (figure C.4). The crucial problem is that the single rectangle setup does not recognize multiple carts in the image, which is one of the mandatory requirements for the algorithm.

3.4.2.2.2 Multiple selected rectangles The final version of the algorithm selects **n** smallest rectangles containing a cart where **n** is a parameter given by the user. Selected rectangles are then tested for intersections and are separated into groups. This means that rectangles in a group are overlapping among themselves and have no overlap with rectangles in other groups. Optionally very small groups (relatively to the **n** parameter) are discarded because they can be false positives from the neural network.

Each group is represented by an interpolated rectangle which is created by the mean value of all the rectangles in the group. The aim in doing so is to mitigate for described problems with too small rectangles and rectangles that only partially cover the cart.

The parameter **n** of selected rectangle count affects the probability that the algorithm finds rectangles for all of the carts in the image. However if it is

too big, the larger rectangles are also selected and that not only decreases the similarity with ground truth bounding rectangle, but also increases the chance that there will be an overlap joining the groups of two different carts together.

3.4.2.2.3 Enhancing Selective search During the implementation of the described algorithm it became apparent that often there are no rectangles that would be very similar to the ground truth or it is hard to pick them from the large set of proposed rectangles. While the selection of multiple rectangles and their interpolation tries to fight the problem, another approach is to give the find_candidate_object_locations function more information which is available in the dataset. One such information is that there is a very stable background in the supermarket dataset which the function does not know about as it works on individual images. This knowledge can be passed to the algorithm as background subtraction.

3.4.2.2.3.1 Background subtraction The background images were calculated for each hour as a mean of all RGB images in the hour. At this scale there were ghosts, a weak images of people or other objects standing too long in a certain spot. An example of a ghost can be seen in image 3.4.

The ghosts should not have a major impact on object location proposal by the means that the proposal method increases the number of rectangles trying to capture the ghost in addition to real objects. The method already favors recall and thus produces an abundant amount of proposals. A huge problem can occur when the ghost is strong and consists of a shopping cart that is recognized by the CNN, then the whole part of the supermarket dataset that shares this flawed background would have false positive detection of the ghost cart.

When the background is subtracted from an image that would be passed to object location proposal the shopping cart is clearly visible as well as the ghosts, which are however slightly weaker (can be seen in 3.4). The rest is very dark which means that the background subtraction worked well, however there is noise of individual bright pixels.

3.4.2.2.3.2 Gaussian smoothing Gaussian smoothing from the scipy [35] library was deployed as noise reduction method. The smoothing removed the noise and blurred the image. The difference in appearances of the image before and after smoothing suggested that the smoothing should not remove

Figure 3.4: Example of background image with a ghost caused by a human standing for a long time at one place.

any valuable information from the image. Nevertheless the method blurred sharp edges which could decrease the quality of segmentation step in selective search.

The find_candidate_object_locations was applied to the smoothed image and performed significantly worse than on the unmodified image. It proposed about half the amount of rectangles and the similarity to ground truth of the best rectangles in the set was much worse. The interpretation of this phenomenon can be that the function segments the image based on mostly color and background subtraction combined with gaussian smoothing which removes most of the color from the image.

The algorithm ceased using background subtraction and smoothing as it proved to have bad impact on find_candidate_object_locations function.

3.5 Attribute classification

The attribute classification is highly dependent on the distribution of the classes in the supermarket dataset. As it can be seen in the Results chapter, Annotation section, the annotated dataset is rich in the matter of number of instances, however there are few unique humans.

Figure 3.5: An image after background subtraction.

The classification of gender seems feasible, because there are 57.25% females and 42.75% males, which means reasonable amount of instances for both classes. On the other hand, the customer/employee attribute has 85.75% customers and only 14.13% employees. Even worse is the situation with age distribution. There are no captured children in the dataset and only 10.94% humans labeled as old. The rest are young adults with 89.06%. It was decided not to carry out classification experiments on these attributes with bad distribution.

The gender classification was performed and the results are evaluated in the Results chapter. The shopping basket classification would be also feasible, however this work rather focused on shopping cart detection, which incorporates classification and it seems logical to consider the problems very similar and apply the machine learning tasks for cart classification for the shopping baskets in similar fashion.

3.5.1 Gender classification

Images of humans are needed for the task of gender classification. They are cropped from the RGB dataset images by their bounding rectangles taken from the annotation. The annotation is used instead of using the object de-

tection algorithms described earlier in order to work with the best quality data and avoid discussion of how much the performance of object detection affects the estimation of humans' attributes.

Most of the classifiers work on objects represented by feature vectors of the same length. To ensure that the construction of feature vectors by any method will result in the same length for each image it is logical to resize the images to same sizes. Alternative approaches such as transformation of vectors after they have been extracted from the images with various sizes would be too complex.

To achieve the same sizes of images there would be a simple method of resizing the images after cropping by the exact values of bounding boxes. That would lead to a very high amount of skew of the shape applied to virtually all the images.

The approach used to obtain the images was to find out the average size of human bounding boxes: width=165 height=199 and maximum size: width=340 height=279. The first step to obtain the dataset was selecting the standard size of human image. Then the bounding boxes which were smaller in any dimension than the standard were expanded in the dimension equally to both directions if possible, or as close to equal expansion as possible. The problem that could prevent equal expansion was that the bounds were to close to the border of the whole image. Last step was the hard resize of the image to match the standard size. Only images larger than the standard were affected in this step. The standard size was set to: width=280 height=220, which is biased to the maximal size, so that a minimum number of humans would be skewed.

The negative impact of expanding the bounding rectangles is that there is added background. Also if the rectangle was not expanded with equal addition to both directions in the dimension (e.g. if the height is 40 pixels smaller than standard, both the top and the bottom of the rectangle should be expanded by 20 pixels), the human would be no longer centered in the image, increasing the demands for the classification step to be spatial invariant.

Gender recognition was performed by histogram of oriented gradients combined with random forests and support vector machines. The aim of these two approaches was to reproduce the results presented in [9] because the setup was very similar. GoogleNet was also trained and evaluated.

Chapter 4

Results

The results chapter describes the outcomes of the experiments presented in the previous chapter. First of all, the annotation is evaluated and the statistics that have been learned by annotating the data are described. After that the image classification is evaluated. The object detection section composes of three parts, methodology describing how the object detection is generally evaluated in this work and other two parts specifying this evaluation with human and cart detection. The last section focuses on attribute classification.

4.1 Annotation

There are 72119 humans in 59947 images. In the supermarket dataset the humans are assigned ID independently in each cluster, the annotation does not solve the reappearing humans in multiple clusters and so they are considered different individuals. This signifies, that the number of 411 assigned ID does not mean that there are 411 unique humans. It is not possible to find that out without huge effort to match all of the humans manually.

The human attributes give valuable information, which should be provided to the supermarket management. It is necessary to look at the distributions of the attributes to have better understanding how to perform machine learning on them. The statistics presented here are based on the number of human bounding rectangles, not individuals from the dataset point of view.

- Gender:
  - female: 41289 (57.25%)
  - male: 30830 (42.75%)

- Customer/employee:
  - customer: 61930 (14.13%)
  - male: 10189 (85.87%)
- Age:
  - child: 0 (0%)
  - young_adult: 64230 (89.06%)
  - old_adult: 7889 (89.06%)
- Basket:
  - with basket: 12573 (17.43%)
  - old_adult: 59546 (82.57%)

There are 16043 shopping carts in 15469 images.

4.2 Cart classification

At the time of this experiment there were only two annotated hours, both from the detergents aisle, day 17. Their sizes are 1073 images for hour 07 and 1603 images for hour 08. The labels for such experiment were based on the presence of a cart in the image and so there were only two classes. For hour 07 there are 290 images with a cart in them and for hour 08 there are 133.

It was decided to use hour 08 as a training set and hour 07 as a testing set by their absolute sizes and not by the class distributions, which could turn out to be wrong. The evaluation of this experiment also needs to take into consideration that the used hours were smaller than average hour, so the results might not be representative. However the subsequent usage of the GoogleNet network shows that if has over 90% accuracy even on other annotated hours.

The two CNNs used for the experiment were GoogleNet and CaffeNet, both distributed with caffe framework [30] and for both there is pre-trained model available from the Model ZOO [36]. For both networks the setup of training was as close as possible to the default values.

In order to train using GPU, which was NVIDIA GeForce 680MX, which has 2GB of internal memory, the images were downsized to half of the size (320x240 pixels). It was also necessary to decrease the batch size for GoogleNet.

The table 4.1 shows the progress of training CaffeNet, which proved to be

Table 4.1: Accuracy of CaffeNet during training

iterations	**accuracy**
2000	79.96%
3000	91.91%
4000	87.74%
5000	82.56%

the inferior of the two CNNs used. It is obvious that during training the net overfitted to the training set and its performance for the test set started to deteriorate. The training was re-run and stopped at about 3000 iterations to get the best configuration available.

GoogleNet was configured to perform the testing after 4000 iterations, after which its accuracy was about 97%. The result is not exact due to the fact that the net has three loss layers (figure C.2) for output of various depth and the result was obtained by averaging. In further iteration the accuracy dropped again.

4.3 Object detection

4.3.1 Evaluation methodology

There are two parts of evaluation for object detection, the similarity of the bounding rectangle location and size to ground truth and the accuracy of classification of what is captured by the bounding rectangle. Determining the accuracy of classification is fairly straightforward only with some nuances. Such nuance can be that the classifier returns a vector of classes with corresponding confidences instead of the most probable class. On the other hands evaluation of the bounding rectangles accuracy can have several approaches and thus the most appropriate one needs to be picked.

Accuracy of classification is not considered in this section, because the object detection presented is independent for each class of the object — humans and shopping carts. So it only propagates into first evaluation metric that is the number of bounding rectangles of the given class detected should be the same as in the ground truth. This is the simple measure telling the portion of correct number of matches to all images the algorithm has processed. In this section this metric will be called matching accuracy.

Bounding rectangle similarity is evaluated only for the cases where the correct number of rectangles has already been proposed. Each of the ground truth rectangles is assigned the highest similarity among the proposed rectangles by

the equation:

$$similarity(A, B) = \frac{intersection(A, B)}{union(A, B)}$$

The similarity metric is mean value of all best similarities to the ground truth rectangles.

This approach has been presented in [3], which also states that rectangles with similarity over 0.5 can be considered a good match. This work does not filter good and bad matches of rectangles, it rather presents mean of the similarities.

The goal values of matching accuracy and similarity that should be satisfactory for subsequent machine learning methods are higher than 90% for matching similarity to offer tolerance for difficult cases (such as object entering and leaving the frame) and over 50% similarity to follow the metric presented in [3] .

4.3.2 Humans

Human detection has been performed on the hour 09 of day 18, because the algorithm is time demanding and optimization of the algorithm would take inadequate amounts of time. The hour is composed of 3053 images, out of which 2406 contain a human. The final results will need to be reevaluated on a different part of the supermarket dataset to confirm the outcome is consistent. As it was discovered later, important note on this hour is that there are no shopping carts captured in it.

Table 4.2: Grid search results. First number is matching accuracy and beneath is similarity

Upper threshold (right) Lower threshold (down)	0.3	0.4	0.5	0.6	0.7
0	0.5777 0.6916	0.5912 0.6978	0.5538 0.7124	0.4219 0.7578	0.2239 0.7794
0.1	0.5777 0.6632	0.5912 0.6701	0.5538 0.6869	0.4219 0.7314	0.2239 0.7486
0.2	0.5777 0.5385	0.5912 0.5472	0.5538 0.5866	0.4219 0.6269	0.2239 0.6366
0.3	0.5777 0.4805	0.5912 0.4907	0.5538 0.5329	0.4219 0.5776	0.2239 0.5954
0.4	0.5777 0.3714	0.5912 0.3810	0.5538 0.4217	0.4219 0.4839	0.2239 0.5046
0.5	0.5777 0.2268	0.5912 0.2329	0.5538 0.2619	0.4219 0.3488	0.2239 0.3920

The proposed algorithm is sensitive to the threshold parameters. A necessary step in evaluation and optimization of the algorithm for usage was to search for the best configuration of these parameters. With a rather limited spectrum and the danger of overfitting, the simple grid search method seemed appropriate for the task.

The table 4.2 shows the results of the grid search. Upper threshold influences the number of found bounding rectangles, thus affecting mainly matching accuracy and indirectly similarity as well because similarity is only measured when there is a matching number of rectangles. Lower threshold only influences the sizes of the rectangles and only affects the similarities, which can be seen in the table as there are same values of matching accuracy in the columns.

The results of the grid search have given promising directions to parameters 0.4 for upper threshold and 0.0 for lower threshold, which the algorithm continues to use. The priority for choosing these values is the highest matching accuracy. The similarity is passing the goal value in about half of the cases and so the matching accuracy is the most important information for the choice of parameters.

Finding the cluster centers was the part of the algorithm not performing on the level that needed to be met. To enhance the performance of mean shift analysis it was necessary to make sure there was an accurate value of the most important parameter, the bandwidth. It was estimated with function estimate_bandwidth based on the statistical properties of the dataset, which changed from image to image and this was identified as the most unstable part of the algorithm.

With the given setup it appeared quite natural to set the bandwidth as a constant. Two values, 35 and 50 were tested based on the estimation of bandwidth from the estimate_bandwidth method run on an image with two people in the figure 4.1, which gave the result of approximately 34. With the value of 50 the matching accuracy rose to over 91%, which appeared satisfactory and it seemed not necessary to optimize this part of the algorithm.

Concurrently to improving the algorithm by setting constant value of bandwidth another approach was implemented. The value of height was omitted in the construction of the temporal dataset of points, which the mean shift would use to find the cluster centers. This was for the reason, that straight up adding it to the dataset as another dimension of the feature vector made little sense. It would increase the distance of temporal dataset points that were on different height levels and because of that the clustering algorithm would not converge to the center of the cluster, which is often the head and has a far different height level than most of the points.

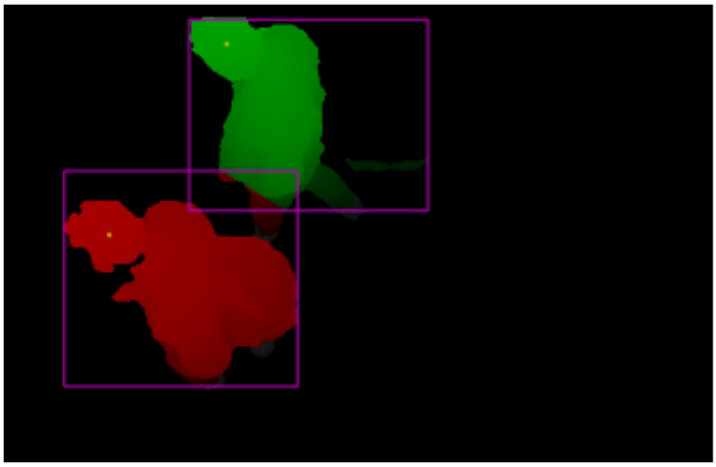


Figure 4.1: Vizualization of mean shift. Yellow dots are cluster centers, color denotes assignent of depth points to clusters and magenta are method proposed bounding rectangles.

The height of the points is however too valuable information to be omitted. Although adding more points to the temporal dataset is not favorable due to increased computational time of the already resource hungry clustering algorithm, it has been discovered that adding points with height at least 70% of the highest point twice and points with 90% height triple times increases the matching accuracy to 94.57%.

The table 4.3 shows that on hours (hours 08, 09) that have few shopping carts in them the algorithm is performing as expected. However the matching accuracy drops for the cases (hours 10, 11, 12) when there are shopping carts, especially ones filled with goods. As it was described, the problem is that the shopping carts are occasionally captured in the depth map.

4.3.3 Shopping carts

4.3.3.1 Primitive cart detection method

The method did not meet the expected output. There is no assessment of performance on sufficient number of images because the method was unable to detect the shopping carts even for the images it was being optimized for.

Table 4.3: Performance of human detection on various hours (day 18)

hour (down)	matching accuracy	similarity
08	0.9283	0.7366
09	0.9457	0.7194
10	0.8119	0.6459
11	0.6829	0.6498
12	0.7605	0.6458

Figure 4.2: An example of primitive cart detection method's output for simple instance.

The most important parameter for the method is the number of slices the image is divided into. This parameter was exhaustively examined and the best values of 5 or 6 slices were giving unsatisfactory outputs. A result can be seen in figure 4.2.

4.3.3.2 Candidate object location method

The candidate object location method was evaluated on the hour 10 of the day 18, which has 1476 shopping cart which the method needs to detect. The results are again measured in the two metrics described in methodology — matching accuracy and similarity.

The obtained results were 89.05% matching accuracy and 58.29% similarity. That can be considered close to satisfactory as it corresponds to the expected values.

4.4 Gender classification

The methods examined in the Theoretical part were applied for gender classification. These are histogram of oriented gradients combined with random forest [9] and support vector machines [12]. The convolutional neural networks were also tested.

To protect the evaluation of the experiment from errors caused by object detection the annotation bounding rectangles were used to crop the human images from the supermarket dataset RGB images. This should offer clear view on how the classifiers can perform in the task.

The classical methods of histogram of oriented gradients combined with either random forests or support vector machines used hours 07-13 of the day 18 as a training set. The training set therefore composes of most of the data. For testing purposes the hour 14 of the same day was used. It has 10286 images containing human, with 11422 human bounding rectangles, out of which 6713 were females and 4709 males. Distribution of classes is approximately 69% to 41%.

HOG combined with random forests showed the accuracy 54.34% and HOG combined with SVM 51.49%. This means that the algorithms were unable to learn how to classify gender in his setup.

The setup for CNN was a slightly different, it used all of the annotated humans, meaning 72119 images. However it had larger testing set of 20000 humans. The testing set was picked by listing all of the images by the time of their capture and truncating the last 20000 images. This is the best effort to avoid splitting images from one cluster to both the training set and the testing set, which would allow the neural network to see the particular human in training set and evaluation would be done on overfitted network.

The distribution of the classes is better than in the previous setup, there are 11515 females and 8485 males in the testing set. That is about 57.58% females. The classifier needs to surpass this value to show it has learned how to distinguish gender.

The table 4.4 shows that the CNN was able to learn some information from the dataset and performs better than random guessing. However the behavior of the network is very unusual, it would be expected to improve the accuracy on the testing set to the point of overfitting to the training set at which point the accuracy should start dropping for the testing set. The repeated drops in accuracy suggest that the network is not stable in the learning the fundamental knowledge and picks various features. It would not be same to assume

Table 4.4: The evaluation of GoogleNet during training. The values are classification accuracy on the testing set. The columns represent the three different output branches of GoogleNet.

output branch (right) iteration (down)	1	2	3
1000	0.402375	0.4315	0.410375
2000	0.65625	0.6345	0.655125
3000	0.607125	0.6195	0.64675
4000	0.595125	0.597625	0.626875
5000	0.616875	0.62025	0.63425
6000	0.679375	0.66175	0.698875
7000	0.66	0.6165	0.572875
8000	0.609125	0.611125	0.63125
9000	0.614875	0.620375	0.629125
10000	0.6015	0.61525	0.610875
11000	0.583875	0.609875	0.589625
12000	0.59375	0.57775	0.57525
13000	0.610375	0.61975	0.618375
14000	0.66625	0.678125	0.675375
15000	0.6095	0.612125	0.636625
16000	0.613625	0.600125	0.599
17000	0.64725	0.666625	0.65625
18000	0.653375	0.601625	0.594
19000	0.658625	0.651875	0.64625
20000	0.653125	0.640375	0.646

the reported accuracies on another data.

Chapter 5

Discussion

The Discussion chapter debates some of the ambiguous topics and topics that this work did not touch, but it seems important that they are not omitted completely.

5.1 Human detection results

The logic of the algorithm for human detection is based on the fact that there are just two recognized object types in the dataset (humans and shopping carts). Most of the times the shopping carts are not captured by the depth sensor and the depth images have very little noise, which allows the clustering approach to perform well even though it is unsupervised.

The evaluation on hours 10, 11, 12 of the day 18 showed two flaws in the logic. Firstly, the shopping carts are captured by the depth sensor when they contain goods the customer put into them. Secondly, the depth sensor does not capture humans standing far from the Kinect device. These phenomena can be seen in figures C.5 C.6.

The solution to the first flaw would be to subtract the height points representing the shopping cart by the knowledge of shopping carts locations from the cart detection. The cart detection algorithm is supervised, so it should be able to distinguish between a cart and a human, while the human detection would only work on the leftover depth points clusters, which it would assume to be humans.

The solution for the second flaw is much harder even to propose. The reason for that is that there is no information about the humans in the depth map and the whole algorithm of human detection cannot solve the problem. A

whole new approach based on the RGB image would need to be proposed.

One of possible approaches is clustering based on RGB images after background extraction and gaussian smoothing, which bear a resemblance to the depth map, but is a lot more noisy. It is interesting from the point that it offers the opportunity to reuse existing methods.

5.2 Shopping baskets

It was shown in the Results section that the consideration of shopping baskets as properties of humans is causing an issue for attribute classification. The main problem is that it makes centering of humans in their bounding rectangles almost impossible.

As it was discussed in the section describing dataset annotation, the other option of having the baskets as stand alone objects is also not favorable, because it would require more work for the human annotator as he or she would need to draw more rectangles. Also there would be the necessity of assigning the baskets to the humans controlling them for the big picture which the supermarket management would want to see from the system.

5.3 Computational demands

Algorithms presented for object detection are very computationally demanding and the processing of a single hour can take several hours (depending mainly on the size of the hour) of computation. While the efficiency of the algorithms is not a primary requirement specified in the assignment of this work, being able to process the data on the fly would decrease the demands for data space that is consumed rapidly by image data that is buffered to be processed.

From the dataset it is apparent that the amount of images that pass the preprocessing step and thus need to be evaluated by the algorithms described in this work is about 15% even in the hours with high traffic. This means that the algorithm can be expected to only process about 5 frames per second for consistent 30 frames per second data stream with only a single hour buffering of images.

However, even with this tolerance the algorithms as they are presented are slower. There are several approaches to the speed-up, easiest of all is parallelization. As none of the algorithms presented works with the sequentiality of the images, this can be done trivially only at the cost of higher power con-

sumption and better hardware.

More sophisticated optimization approaches can also be deployed. It is common in image processing to downsize the image to drastically decrease computational time. The information lost by the process varies from algorithm to algorithm.

For the method used for human detection the most time consuming part is finding the cluster centers by mean shift analysis, which is dependent on the temporary dataset size. Therefore it makes sense to downsize the image only for this part of the algorithm. The shift of the cluster center that would be caused by downsizing does not directly affect the performance of the algorithm as long as the center generalizes the cluster well and so there is good hope that there would be little to no losses in metrics evaluating the algorithm.

The shopping cart detection, on the other hand, does not seem appropriate for the downsizing approach. The fine trained convolutional neural network is sensitive to scale of the image and performs best on images of the same size as the ones that it saw during training. That means having special CNN for the amount of downsizing applied and there is no guarantee that the CNN trained smaller scale would perform as well as the one presented in the Results section. Also, downsizing in this algorithm would directly affect the bounding rectangles, which means cumulative error of a pixel in each of the four dimensions for each downsizing (in the worst case).

Chapter 6

Future Work

6.1 Centering humans in their bounding rectangles

The problem that humans are not well centered in their bounding rectangles has been discussed in previous chapter. Better performance of all human attributes classification can be generally expected if the centering problem is solved and thus it should be one of the directions the future works should follow.

The problem is caused by considering shopping baskets as attributes to humans. Recognizing if the human is equipped with a basket is the first step in this future approach proposal. With high accuracy classifier which would recognize shopping baskets the algorithm could tell apart which bounding rectangles need further cropping because they contain a shopping basket and which should not be cropped, because there is only the human in them.

The aim of the cropping algorithm would be to remove the surrounding background as well as the shopping basket. This would lead to the loss of visual information about the basket, but with the knowledge that the person is equipped with one, it shouldn't be a problem to consider insignificant information lost. The cropping could be based on the depth map, which gives good information about humans as it was shown in human detection experiment.

6.2 Video

One of the main directions that future works should examine the video logic of supermarket dataset. That is using the subsequent property of the images. It can help eliminate the problem of n object entering or leaving the scene and other cases such as the human facing away from the camera of being covered

by another individual, which cause the loss of information. This became apparent during the manual annotation, where especially for difficult cases the annotator needed to go through the whole cluster capturing a human to recognize the attributes needed for the dataset.

One of the approaches that should not be forgotten is using the classifier confidence in its output. This information is often lost when the classifier chooses the class by selecting the highest score among possible classes. Working with the confidence, for example by discarding images with very low winning class score or emphasizing the ones with big difference between winning class score and the score of second class, can improve the classifier performance.

6.3 Shopping carts assignment

The supermarket dataset annotation does not carry information about connection between a shopping cart and its respective owner. These are covered as independent objects in space. However, the knowledge if the shopping cart belongs to the concrete human is valuable and it is worthwhile to extract it by assigning the cart to a human in the same cluster.

Most appealing approach to the assignment is finding the correlation of the objects' movement. There would arise some minor problems like that the representations of the objects, bounding rectangles, change shape throughout the cluster. That can be solved by calculating the centers of the rectangles or by calculating the distance of the rectangles.

However, the main problem is that without the ground truth the approach cannot be easily evaluated. It would be good to consider the assignment of a cart to a human only an attribute of human (similar to shopping basket) and evaluation would be done by comparing classification of other human attribute (such as customer/employee) with and without the added cart attribute.

Conclusion

The introduction to the supermarket dataset problematics has been made and the tasks stated in the assignment fulfilled. A two step procedure for extracting information about humans has been proposed and it has been experimented with algorithms executing the independent parts of the procedure.

Previous works on the comparable problematics and analogous datasets have been explored. Most of the proposed approaches were based on these works. Even though there were examples of setups very similar to the supermarket dataset, the comparison of the results was rather approximate.

One of the tasks to carry out on the supermarket dataset was to prepare an annotated dataset. The annotation was proposed and discussed. According to the proposition an annotation tool was implemented and then given to the human annotator. A considerable amount of data was annotated by the annotator.

The first part of the procedure proposed for extracting information about humans is object detection. Extensive work has been done on this topic, because it is crucial for successive machine learning algorithms, with human attribute recognition in particular. The proposed and implemented algorithms are an adequate basis for future work with reasonable performance.

The second part of the procedure is human attribute recognition. There has been a sizable effort put into research on this topic and some of the researched methods were tested for selected attributes. The research and the obtained results suggest that the issue is rather complex and a vast amount of future work is needed for satisfactory results.

The tasks described in this work's assignment were successfully completed.

Bibliography

[1] Kordík, P.; Šlapák, M. Computational Intelligence Methods. 2015, course MI-MVI Lecture 2. Available from: https://edux.fit.cvut.cz/courses/MI-MVI/_media/lectures/02/lecture02.pdf

[2] Linder, T.; Wehner, S.; Arras, K. O. Real-time full-body human gender recognition in (RGB)-D data. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, IEEE, 2015, pp. 3039–3045.

[3] Dollar, P.; Wojek, C.; Schiele, B.; et al. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 34, no. 4, April 2012: pp. 743–761, ISSN 0162-8828, doi:10.1109/TPAMI.2011.155.

[4] Han, H.; Otto, C.; Jain, A. K. Age estimation from face images: Human vs. machine performance. In *2013 International Conference on Biometrics (ICB)*, June 2013, ISSN 2376-4201, pp. 1–8, doi:10.1109/ICB.2013.6613022.

[5] Benenson, R. What is the class of this image ? 2013–2016, [Online; accessed 4-May-2016]. Available from: http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html

[6] Russakovsky, O.; Deng, J.; Su, H.; et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, volume 115, no. 3, 2015: pp. 211–252, doi:10.1007/s11263-015-0816-y.

[7] Krizhevsky, A. The CIFAR-10 dataset. 2009, [Online; accessed 4-May-2016]. Available from: http://www.cs.toronto.edu/~kriz/cifar.html

[8] Coates, A. STL-10 dataset. 2011, [Online; accessed 4-May-2016]. Available from: http://cs.stanford.edu/~acoates/stl10

[9] Cao, L.; Dikmen, M.; Fu, Y.; et al. Gender recognition from body. In *Proceedings of the 16th ACM international conference on Multimedia*, ACM, 2008, pp. 725–728.

[10] Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, Ieee, 1999, pp. 1150–1157.

[11] Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, IEEE, 2005, pp. 886–893.

[12] Everingham, M.; Van Gool, L.; Williams, C. K.; et al. The pascal visual object classes (voc) challenge. *International journal of computer vision*, volume 88, no. 2, 2010: pp. 303–338.

[13] BVLC. bvlc_reference_caffenet. 2015, [Online; accessed 5-May-2016]. Available from: `https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet`

[14] Stanford. Convolutional Neural Networks (CNNs / ConvNets). 2016, [Online; accessed 28-April-2016]. Available from: `http://cs231n.github.io/convolutional-networks/`

[15] Girshick, R.; Donahue, J.; Darrell, T.; et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] Krizhevsky, A. cuda-convnet. 2012, [Online; accessed 4-May-2016]. Available from: `https://code.google.com/p/cuda-convnet/wiki/LayerParams`

[17] Szegedy, C.; Liu, W.; Jia, Y.; et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] Van de Sande, K. E.; Uijlings, J. R.; Gevers, T.; et al. Segmentation as selective search for object recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2011, pp. 1879–1886.

[19] Oren, M.; Papageorgiou, C.; Sinha, P.; et al. Pedestrian Detection Using Wavelet Templates. 1997, pp. 193–99.

[20] Nixon, M. S.; Carter, J. N. Automatic Recognition by Gait. *Proceedings of the IEEE*, volume 94, no. 11, Nov 2006: pp. 2013–2024, ISSN 0018-9219, doi:10.1109/JPROC.2006.886018.

[21] Yu, S.; Tan, D.; Tan, T. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, 2006, ISSN 1051-4651, pp. 441–444, doi:10.1109/ICPR.2006.67.

[22] Zheng, S. Gait Recognition. 2010, [Online; accessed 4-May-2016]. Available from: `http://kylezheng.org/gait-recognition/`

[23] Ricanek, K.; Tesafaye, T. MORPH: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, April 2006, pp. 341–345, doi:10.1109/FGR.2006.78.

[24] Guo, G.; Mu, G. Simultaneous Dimensionality Reduction and Human Age Estimation via Kernel Partial Least Squares Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 657–664.

[25] Guo, G.; Fu, Y.; Dyer, C. R.; et al. A probabilistic fusion approach to human age prediction. In *Proc. CVPR-SLAM Workshop*, 2008, pp. 1–6.

[26] Mu, G.; Guo, G.; Fu, Y.; et al. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, ISSN 1063-6919, pp. 112–119, doi:10.1109/CVPR.2009.5206681.

[27] Cootes, T. F.; Edwards, G. J.; Taylor, C. J. Active Appearance Models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Springer, 1998, pp. 484–498.

[28] Kwon, Y. H.; Lobo, N. D. V. Age Classification from Facial Images. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 762–767.

[29] Surmon, l.t.d. Surmon. 2012, [Online; accessed 3-May-2016]. Available from: `http://surmon.org`

[30] Jia, Y.; Shelhamer, E.; Donahue, J.; et al. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[31] Girshick, R. B. Fast R-CNN. *CoRR*, volume abs/1504.08083, 2015. Available from: `http://arxiv.org/abs/1504.08083`

[32] Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 24, no. 5, 2002: pp. 603–619.

[33] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, volume 12, 2011: pp. 2825–2830.

[34] King, D. E. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, volume 10, 2009: pp. 1755–1758.

[35] Jones, E.; Oliphant, T.; Peterson, P.; et al. SciPy: Open source scientific tools for Python. 2001–, [Online; accessed 2016-05-08]. Available from: `http://www.scipy.org/`

[36] BVLC. Model Zoo. 2016, [Online; accessed 5-May-2016]. Available from: `https://github.com/BVLC/caffe/wiki/Model-Zoo`

Appendix A

Acronyms

CNN convolutional neural network, also called deep neural network

RCNN Regions with CNN features

hour It is main working dataset subset, group of images specific to one device and taken in a concrete hour. If not otherwise stated, the images were taken by the device placed in detergents aisle.

FPS Frames per second, frequency with which the camera captures images

json JavaScript object notation

CCR Correct classification rate

Appendix B

Contents of enclosed medium

- `README.txt` ..... the file with CD contents description
- `annotation_tool` ..... the directory with the annotation tool
- `experiments` ..... the directory python source files to experiments
- `text` ..... the thesis text directory
  - `DP_Haur_Vojtech_2016.pdf` ..... the thesis text in PDF format
  - `DP_Haur_Vojtech_2016_assignment.pdf` . the assignment of the thesis in PDF format
- `source_thesis` ..... the directory of LaTeX source codes of the thesis

Appendix C

Image appendix

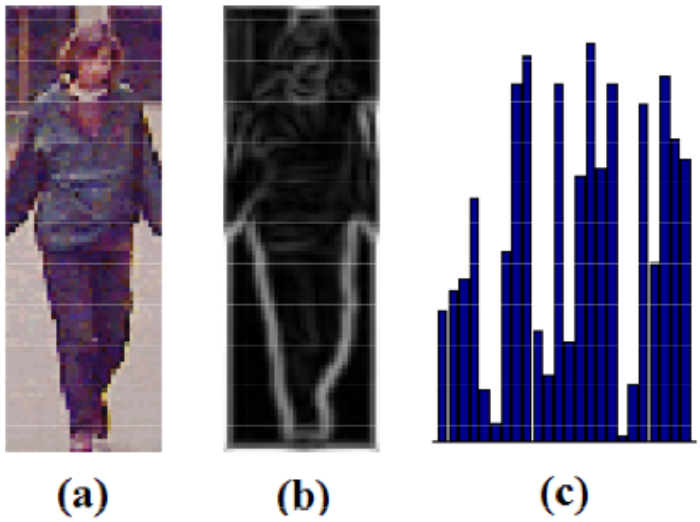


Figure C.1: Histogram of Oriented Gradients [9]

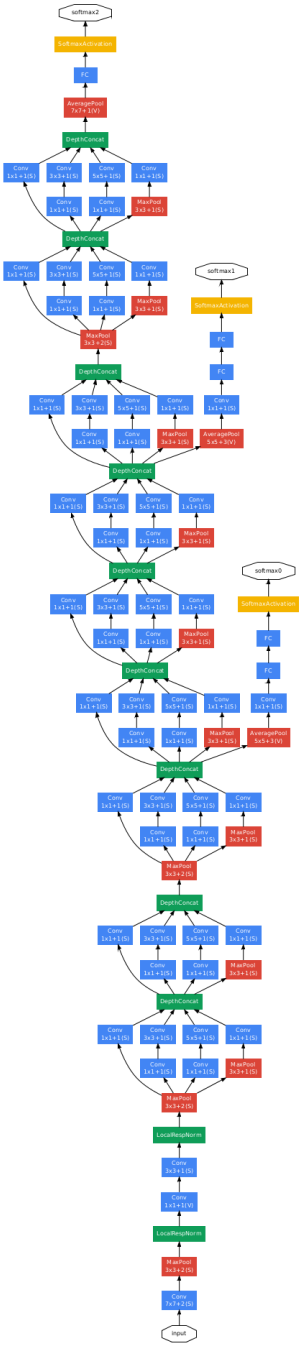


Figure C.2: GoogleNet structure. There are three output branches. [17]

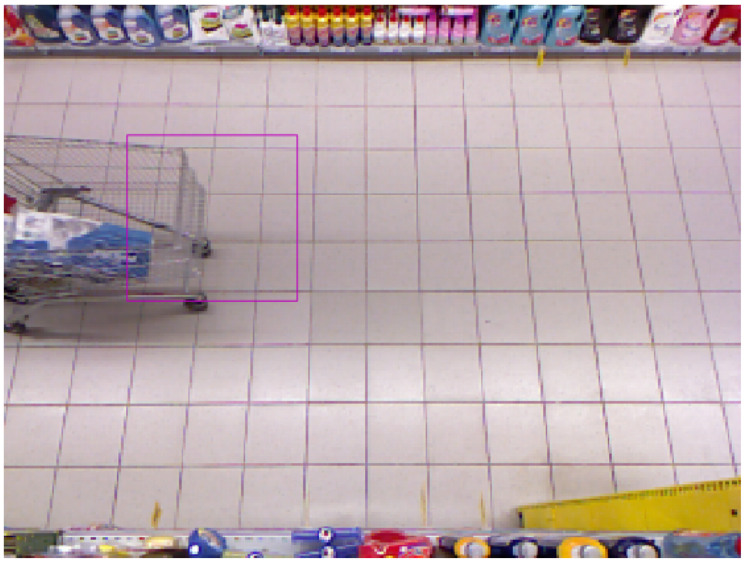


Figure C.3: Example of a partially matching bounding rectangle which the CNN recognizes as containing a cart.

Figure C.4: Example of too small bounding rectangle which the CNN recognizes as containing a cart.

Figure C.5: A problematic image for human detection, RGB counterpart to C.6.

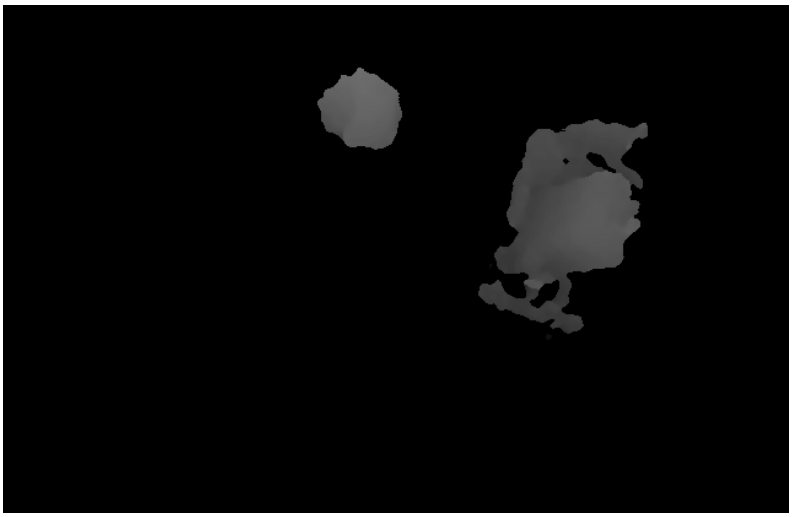


Figure C.6: A problematic image for human detection, depth counterpart to C.5. Note that there is large cluster representing a cart and that the right human is not visible in the depth map at all.