Analysing various methods for object extraction and the classification of kitchens

Lars Essenstam, s1868179*
University of Twente, Faculty of EEMCS, Enschede, The Netherlands

(Dated: June 28, 2019)

In the last few years there have been great successes in the application of deep and machine learning for the use of both object detection and classification. However, when there is a limited amount of data available for many different classes, accuracy is low and decent results can often not be obtained. This research aims to show various case-specific methods to analyse the data and to extract important features to improve classification, such as the Hough transform and mean shift segmentation. A convolution neural network, Alexnet has been trained using both the raw data and the extracted features. When training and validating the network using the raw data an accuracy of 28% has been obtained. When applying extracted features, the handles of the kitchen, to the same network accuracy improved from a 28% to an accuracy of 41%. This increase of thirteen percentage points shows that significant improvement is possible when extracting features before training a network.

Keywords: Deep learning, object detection, hough lines

I. INTRODUCTION

Currently, many image classification methods make essential use of machine or deep learning techniques. Throughout the last years several mayor steps have been taken in improving both the accuracy and training time of these techniques, such as by Krizhevsk et al. [1] with their convolutional neural network Alexnet. This work has been aided by large image datasets such as ImageNet, consisting of over fourteen million images that can be used to pre-train these networks. Currently, Alexnet has been applied successfully for many applications such as object detection [2] and segmentation [3]. These achievements spurred more interest in creating better performing networks, which means a more accurate classification, better computational times or both. Much research has been devoted to these, resulting in networks such as Googlenet [4] and SqueezeNet[5]. Googlenet focused on creating better computation times by decreasing complexity, and focused on increasing accuracy, while Squeezenet focused on severely decreasing computation times. While these methods give promising results, many of them require a large amount of training data. While many deep and machine learning techniques have proven to be effective, all of them need an extensive amount of training data per class. This paper describes some methods to achieve results even though many classes are present, using a convolutional neural network in combination with other methods to segment and improve results relative to simply applying a deep learning network on the raw data. Many of these methods will be determined by carefully examining the available data and see what can be used for a more accurate classification. The specific case that these images will be applied to is that of kitchens, which may seem an odd subject.h owever it can be quite relevant for many people and companies. People living in social housing can request to have their kitchen replaced to the company providing the housing. Whether this request will be granted depends on the age of the kitchens, which makes it useful to be able to extract his automatically.

II. PROBLEM ANALYSIS

When analysing the problem and data, it has to be noted that not every image is the same in forms of quality, orientation and size. For this specific problem, two types of data will be available. Firstly, a more ideal type of data, which usually represents a clean and neat kitchen without objects that could disturb. A good example of this can be seen in figure 1. Usually, this image is taken from the front of the kitchen in a way that most objects are visible. This allows for a more structural analysis of the images, which will be described in section III. The other type of data is the more practical, unstructured data. In this other, unrelated objects are visible and the photo can be taken from various angles. Two examples of this can be seen in figure 2. Due to this, a more structural approach will be less viable to this data. Therefore, the problem can be split into two parts. On one hand the structural approach can be used to extract features that can be used to classify the image. On the other hand there is the classification part of the problem, building something that can automatically estimate the age of the kitchen. Part of the unstructured data will be used in this stage. The more structured data can also be used for this, however it is mostly unlabeled and can therefore not all be

[*] Correspondence email address: l.essenstam-1@student.utwente.nl

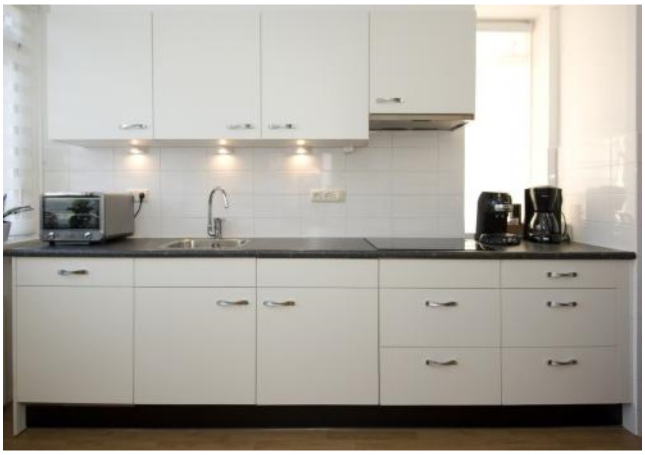


Figure 1. One example of the structured data

used. The structured data can be used to analyse various types of pre-processing. One more difficulty of the problem is the size of the available. For the structured data there is little data available, which makes it hard to either train a deep learning network or to recognise features that are available in more years. For the practical dataset there are more images available. A total of 28 classes with 270 images. While these are some images, it is on average around 10 images per class which is very little to effectively train a classifier. Therefore, it has been decided to merge the data to 4 different classes to increase the amount of images per class. As all images are of kitchen, it is highly likely that kitchen from years that are close to each other have common features. There can be different trends in the shape of the cabinets to different handles. Moreover, the materials used can be quite different, as well as the type of object that exist. According to advise given by experts who manually date the kitchen, they can be dated by looking at the hood, oven and the handles of the kitchen. These are therefore things to look at.

III. STRUCTURAL ANALYSIS

In order to properly analyse the image several characteristics and features specific to the data can be found. Firstly, as can be seen in figure 1 there are usually lots of straight lines in an image. The cabinets form rectangles using these straight lines, however many objects such as the oven are also rectangles. This could be used to segmentate the kitchen into smaller parts where features could be extracted better. One method that can easily be used to detect these straight lines is the Hough transform, which will therefore be explored further. Secondly, the cabinets are always the same width. There exists a certain standardised width of either 40 cm, 50cm or 60cm. This can be used to scale the image or can be used to estimate the size of different objects. Thirdly, in a kitchen many different objects and features are always oriented or located in a related way. The stove will most often be underneath the cooker hood. The oven will usually be right below the stove, possibly in the same rectangle. Moreover, the cabinets at the top are often straight above the cabinets at the bottom, resulting in one straight line across the entire image. This information can be used to help classify in a kitchen as the location relative to other objects can also be used to aid or validate the classification.Moreover, as the company ordering the assignment has advised that the age of the kitchen can be seen clearly by the handles, cooker hood and by the oven, these will be things to investigate. For this it is useful to segmentate the image so these features can be analysed separately, as different objects can have different features could have different ages if one of them has been replaced. However, as changing one object does not change the label and age of the kitchen this must be accounted for.

IV. METHODS

A. Mean Shift Segmentation

An easy method to segmentate the image is mean shift segmentation. Mean shift segmentation uses a grayscale image and uses the brightness to cluster pixels. First, a random point will be picked with a circle around it. The pixels in this circle will be scanned and the mean will be calculated. Next iteration, the center of the circle will be put as at this mean. This will continue until the mean moves less than a pre-determined threshold. Then a cluster of pixels will be created of the circle. When a cluster has been formed, it is checked if there are different clusters created previously that have a similar mean. If this is the case, they will be merged to one cluster. This process will be continued until all pixels have been put into a cluster. Both the size of the circle and the threshold influence how many segments the image will be segmented.

B. Edge detection and Hough transform

The goal of the edge detection and Hough transform is to extract straight lines from the images in order to segmentate them in relevant parts. Because the Hough transform is implemented easier and works better when an edge detection is applied beforehand, an edge detection will be done first. For the edge detection it has been chosen to use the Sobel's method. Usually, the Sobel operator will estimate the gradient in eight directions.

However, because of the assumption that many straight lines exist in the vertical (north/south) and horizontal (east/west) direction it is opted to only use the filters in these directions. Moreover, the weights have been adjusted a little to allow for a stronger detection. The filters to be applied are as follows:

$$h_1 = \begin{matrix} 1 & 4 & 1 \\ 0 & 0 & 0 \\ -1 & -4 & 1 \end{matrix} \text{ and } h_2 = \begin{matrix} 1 & 0 & -1 \\ 4 & 0 & -4 \\ 1 & 0 & -1 \end{matrix}$$

For better accuracy, the filters can be rotated to have all perspectives. The filters will be applied separately using convolution, and can later simply be added together to create a total combined image. After the edge-detection has been applied, the image will be converted a logical image where pixels are either a 0 or 1. From this point, the Hough transform can be used to detect straight lines, which can be useful for the segmentation of image. A straight line can easily be described by $y = ax + b$. However, when a line is exactly vertical this description cannot be used as this would mean that a would approach infinity which gives computational errors in many cases. Instead the Hesse normal form can be used: $r = xsin(\theta) \cdot ycos(\theta)$ where r is the distance from the origin to a certain point on a pixel and θ is the angle between the x-axis and the line connecting the point the the origin. The Hough transform can be implemented rather easily.

As every line can be described by r and θ it can simply be started by creating a plot containing both these values, where r ranges from 0 until the diagonal length of the image and where θ ranges from -90 till 90 degrees. Next, for every pixel of the image, loop over the angles θ and determine using the neighbours of the pixel if there could be a straight line there. If this is the case, it should increment the value in the correct location of the r θ plot. If this is done for every pixel in the image that is a logical one it will increment the value in the r θ plot more often if there are pixels on the same line, therefore the highest values in the r θ plot will likely be straight lines in the original image.

When the Hough transform has been created, the peaks in the image can be found. As both the range from the origin and the angle to the x-axis are known a line can be derived and plotted over the original image of the kitchen. These lines can be used to segmentate the image into relevant parts.

C. Deep learning network

When analysing the two types of data, the structured data will not be used for the deep learning network as there simply is not enough data available, only 15 images. For the unstructured data, there exists little data for each class as well. Moreover, many features, like the hood and oven, are not visible in the images.

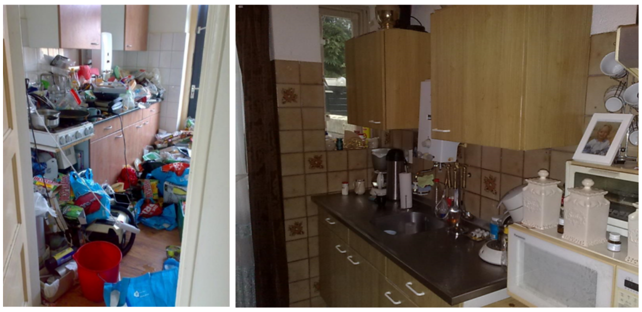


Figure 2. Example images of a kitchen from 2012 (left) and 1990 (right)

The amount of classes ranged from the year 1984 until 2017, of which 28 contained at least 1 image. Because it is not viable to train on just a handful of images the classes will be merged according to the decade the kitchen was constructed or replaced. This leaves four classes, the eighties, nineties, zeros, and tens. This reduces the amount of classes and makes training easier and more realistic. For these classes there is more data available which should give more realistic results. To setup a baseline, a convolutional neural network is used to train and classify the images. For this it has been opted to use Alexnet[1] as it is relatively easy to implement and has achieved good results in the past. All implementation of this will be done using matlab 2019a as it is easy to use for deep learning application with its many toolboxes.

As can be seen in two example images in figure 2. From these and the other images it becomes clear that the cooker hood is often not present in the kitchen or not visible in the image. Therefore this cannot be used to classify the images correctly as only very few images have a visible cooker hood. Moreover, a clear oven cannot be found in many of the images. Many kitchens do not have an oven, these kitchens usually have a small self-bought oven or microwave from which the age is unknown. In other cases the oven is not visible in the image, which makes it impossible to extract it. For these reasons the oven cannot be used to classify the images. Secondly, there is not a lot of data available. The total dataset is not very large, it consist of 270 images, which is not a lot to train and validate a convolutional neural network. However, many images are either duplicates, or the same kitchen is given two different labels, meaning it is present in two different years which of course can never be possible. Out of the features to recognize a kitchen, only one remains; the handles. Even though some images do not contain a handle a big portion of them do and therefore they can be used to classify the images.

The images will first be divided in the respective classes. They will then be applied directly to the

AlexNet network. As all inputs images must be of the same size, 227x277, they must first be resized to this size. The network consists of a total of 14 layers including five convolutional layers, three maxpooling layers, two fullyconnected layers with one ReLu layer in between. The last two layers consist of a softmax layer and of an output layer. The convolutional layers contain filters that will be adapted each epoch if it improves accuracy. These convolution network contain the most learnables or parameters, however the fully connected layers also contain learnables when connecting the layers.

Seventy percent of the data will be used to train the network while 30 percent will be used to validate and test the trained network. This division has been chosen to keep some data for validating and testing the network that has not been used during training, so the trained network can be properly tested.

The next step is looking at the handles, which are visible in most images and are given as a tip as something that can be use to recognize the kitchen. Due to there being many different types of the handles and because many handles are orientated differently they have been cut out of the kitchen images manually. At the same time they have been rotated and flipped so they are all seen from the same perspective. This resulted in a total of 170 images of handles, which have been divided in the categories. Seventy percent is fed into the AlexNet network as training data and thirty percent is used as validation. It has been opted to use exactly the same network as when setting the baseline so a fair comparison can be made between the two.

V. RESULTS

A. Mean Shift Segmentation

Mean Shift Segmentation has been applied to some images, an example of the results can be seen in figure 3. When looking at the results, it can be concluded that the segmentation is not very useful to extract features. Although some features, such as the handles and the oven can be extracted there is a lot of effect from unrelated objects located in the kitchen, such as decoration or painting, as well as effects from light and shadows. This creates segments that cannot be used to extract different features. Therefore this will not be used for further analysis.

B. Edge Detection and Hough lines

The edge detection delivered good results as seen in 4. Both the filters for the north and south direction, described horizontal and vertical in the image, worked well and the combination showed the edges well enough to apply them to the Hough transform. When looking at the peaks in figure 5, indicated by green squares it can be seen that all of them are centered around an angle of either -90 or 0 degrees. This indicates that there are many straight line in the image that are centered around these degrees. When using these maxima to plot the lines as seen in figure 6 it can be concluded that the Hough transform works well enough to extract the lines of the cabinets and of other kitchen devices. This can be used to further segmentate the image.

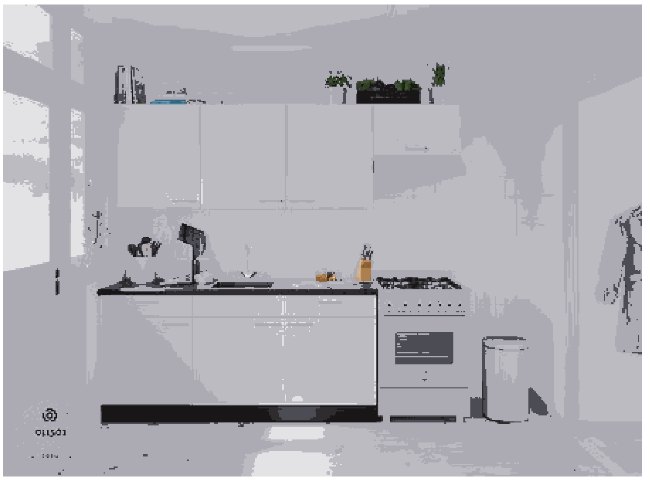


Figure 3. Means shift segmentation of a kitchen

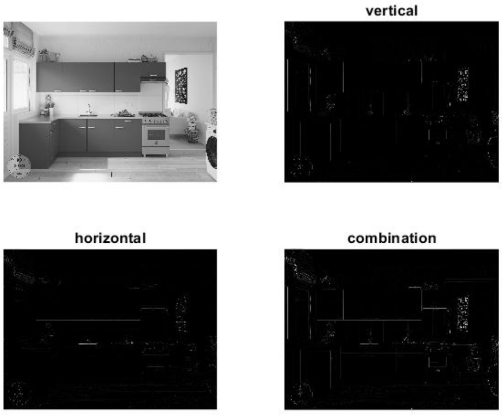


Figure 4. Sobel Edge detection applied to a kitchen

C. Deep learning

First, the baseline is set with the Alexnet network applied directly to the data. The resulting confusion

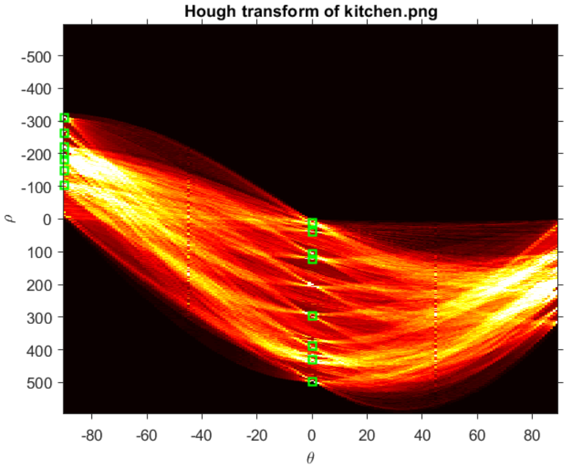


Figure 5. Hough transform including peaks of a kitchen

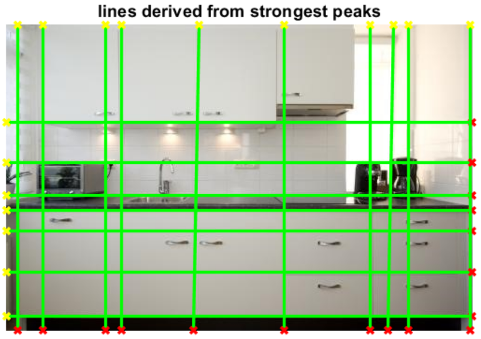
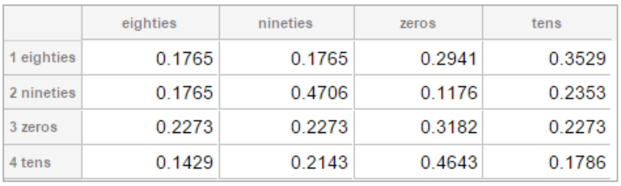


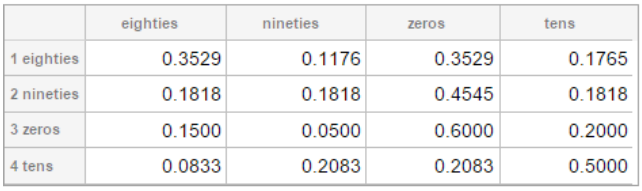
Figure 6. Lines derived from hough transform of a kitchen

matrix can be seen in figure 7. The total accuracy of this image is 28%. The expected accuracy if there would have been random picks would have been 25% when using four classes. Although there is a tiny difference, it is only 3% which makes the results and trained network not very worthwile as this could also be the network being trained on other aspects that the images randomly have in common. Moreover, it can be seen that for example in the class labeled eighties most images have been classified as tens, which is the largest difference there can be between two categorie. Although these results do not bode well for further experiments, they can be used as a baseline to compare further experiments to. Next, the handles have been extracted manually from the images to test whether extracting specific features, like the handles, can give an improved accuracy. The total accuracy is 41% which is an improvement compared to the 28%. It can also be noted that the class nineties has not improved significantly, instead only 18% gets classified as nineties. This could be caused by the size of the training data. For the class zeros there were more images available, which means more features can be extracted causing more images to be classified as zeros. This can be confirmed by the fact that also from the eighties and tens a significant proportion is classified as zeros. Another reason for this is that if there exists more training data from one class, more images will be classified as that class because there is a bigger chance it is correct.



	eighties	nineties	zeros	tens
1 eighties	0.1765	0.1765	0.2941	0.3529
2 nineties	0.1765	0.4706	0.1176	0.2353
3 zeros	0.2273	0.2273	0.3182	0.2273
4 tens	0.1429	0.2143	0.4643	0.1786

Figure 7. Confusion matrix of an Alexnet implementation on the raw data



	eighties	nineties	zeros	tens
1 eighties	0.3529	0.1176	0.3529	0.1765
2 nineties	0.1818	0.1818	0.4545	0.1818
3 zeros	0.1500	0.0500	0.6000	0.2000
4 tens	0.0833	0.2083	0.2083	0.5000

Figure 8. Confusion matrix of the Alexnet implementation on the augmented handles

VI. DISCUSSION

From the results from the hough transform can be decided that it works quite well for the structured data. It can easily recognize many straight lines and therefore can extract cabinets with handles and other features of kitchen. However, the hough transform cannot be applied to the unstructured data either because there are many unrelated objects in the image and there are not many straight lines completely visible. For a final implementation, it could be requested that there are no unrelated objects in the kitchen and that the picture is taken from the frontside of the kitchen. Taking pictures from the frontside of the kitchen could make sure that all features are visible. Essentially, to make better use of the Hough transform the unstructured data should become more structured. Moreover, the result given by the Hough transform as seen in 6 is not ideal. There are some objects, such as windows in the background that contain straight lines which are also automatically derived from the image. This is on one hand a logical effect, however as the window is not a feature that can be used to date the kitchen it is not very useful and makes segmentation of other points harder. The deep learning method applied to the unstructured data worked decently. When applying to the raw data, results close to a random choice could be found. This implies that there are not many big features that the kitchen in one class have in common as the network could not achieve a high accuracy, even when given more time to train. When training and validating using the extracted handles accuracy improved significantly, however it is still not in the range where it becomes useful. The amount of classes has been reduced to four in order to achieve an accuracy of 41%, which is still not impressive. On the other hand, the handles are only one of the suggested features. When a combination of the handles, cooker hood and cabinets can be used accuracy will probably improve. This would require a different dataset where more objects and features of the kitchens are visible.

VII. CONCLUSION

During this paper two types of data, a more structured one and an unstructured one, were analysed by using for example the Hough transform. To classify the data correctly, a deep learning network has been set up. Not enough data was available for the structured data therefore the unstructured dataset was used to train and validate. A baseline was set resulting in a near-random accuracy of 28%. When using extracted handles accuracy improved to 41%, which is not usable on its own. However, it does show that extracting features and training and validating them seperatly increases results, which is promising for further research. From the results, several notable conclusions can be made. Firstly, although the dataset was not of high quality and a basic method using Alexnet only resulted in an accuracy of 28%, there was a significant improvement when using the handles. From this the conclusion can be drawn that from the kitchen itself there exists very little features that the network can train on, however from the handles more information can be found. Moreover, it can be concluded from figure 6 that applying the hough transform and using it to create lines and segmentate the image could work for a higher quality dataset, however it will not work for a small or lower quality dataset that was available in the end. In the end, the result of 41% accuracy shows some possibilities, however better data and more research is required to come to a usable conclusion.

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks* (University of Toronto, Toronto, 2012).

[2] R. Girshick, J. Donahue, T. Darell, J. Mailik, *Rich feature hierarchies for accurate object detection and semantic segmentation* (2014).

[3] E. Shelhamer, J. Logt, T. Darell *Fully Convolutional Networksfor Semantic Segmentation* (2016).

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Aguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. *Gooing Deeper with Convolutions* (Google Inc. University of North Carolina, Chapel Hill, University of Michigan, Ann Arbor, 2015

[5] F. N. Iandola, S. Han, M. W. Moskewiez, K. Ashraf. W. J. Dally, K. Keutzer *SquuezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size* (2017).

[6] Milan Sonka, Vaclav Hlavac, Roger Boyle *Image Processing, Analysis, and Machine Vision*