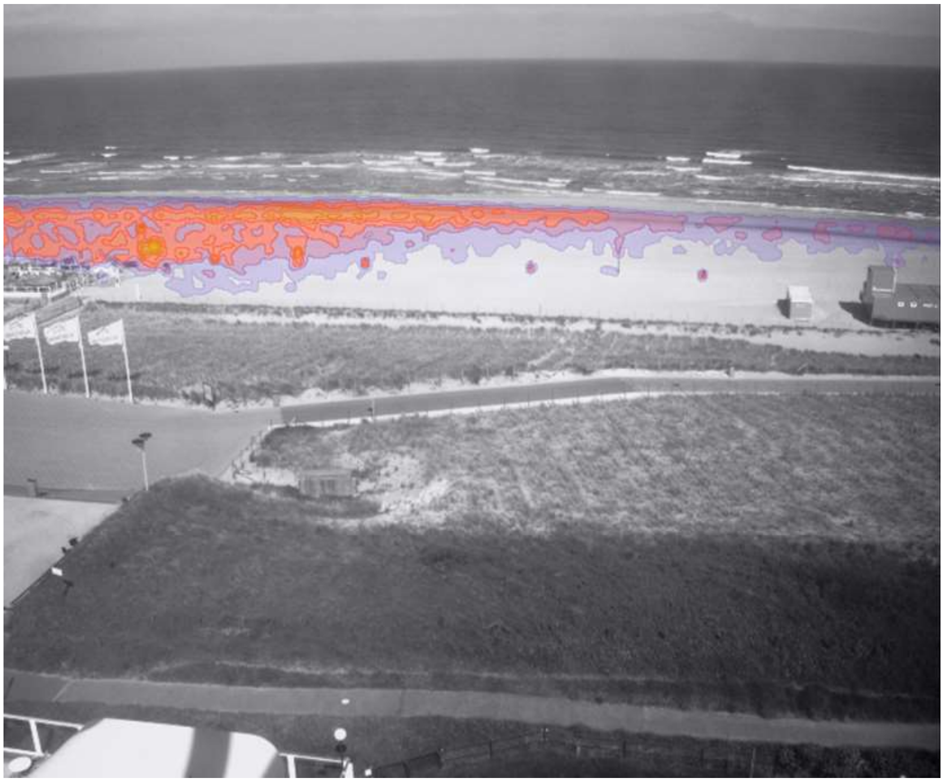
Automatic Quantification of Beach Occupation Using Oversegmentation and Machine Learning



Master Thesis Project

F.J.H. Gulden

December 28, 2017

Project	Master Thesis Project
Type of report	Final report
Author	F.J.H. Gulden
Date	December 28, 2017
Version	Final
Host organisation	Shore Monitoring & Research
Graduation committee	Prof.dr.ir S.G.J. Aarninkhof (*Chairman*) Dr. ir. M.A. de Schipper (*University Supervisor*) Dr. ir. W. Daamen Ir. M. Radermacher Ir. R.C. de Zeeuw (*Company Supervisor*)

Abstract

Decision- and policymakers responsible for the coastal zone aim at combining measures against for instance long-term erosion, with measures that have a positive social, economic impact on the region. Recreational beach usage has a large social economic impact on a region and therefore quantification of the recreational beach usage can provide information on the social, economic situation in a region. In the Netherlands beach usage quantification is mostly performed by manual counting during a limited number of days in the field and this limits the spatial and temporal resolution. The objective of this study is to develop and test a method for accurate, robust and automatic monitoring of the spatial and temporal distribution of the number of beach users on the Dutch coast.

Multiple approaches to quantify the number of persons in an area are reviewed. Comparison of the reviewed approaches showed that a first distinction can be made between methods using a visual light (camera-) sensor and other methods based on the use of Bluetooth, Wifi, GPS-location (all related to phones) and LiDAR. Based on a literature review the visual light sensor is decided upon to best suit beach user quantification. Within the visual light approach a second distinction is made between methods based on the difference in pixel intensity, a method based on variance of pixel intensities over multiple frames and a method using oversegmentation combined with a machine learning framework for classification.

The difference in pixel intensity method is observed most often in literature, but has limitations in conditions that are concerned typical for the Dutch coast (e.g. clouds). The method based on the variance of pixel intensities and the method using oversegmentation are possibilities to overcome the problems described in the studies on the pixel intensity method. Based on preliminary tests the oversegmented machine learning approach is selected because it does not require beach users to move to be detected. Moreover, single snapshots can be evaluated which require a limited data infrastructure in-situ and this is considered advantageous regarding the ease of implementation and cost effectiveness.

The oversegmented machine learning approach divides images into small regions of similar pixels based on pixel gradients. The regions are called superpixels and superpixels can be characterised by significantly more features than the conventional r,g,b relations corresponding to regular pixels evaluated in the differences in pixel intensity method. The availability of an increased number of features provides more options to distinguish between classes during classification and this can be advantageous in difficult (e.g. cloudy) conditions. Classified beach user superpixels can represent multiple beach users due to for instance occlusion and therefore a regression relation between classified beach user superpixels and a manually counted ground truth is determined for conversion of classified beach user superpixels to the number of beach users. Hence, the oversegmented machine learning method for quantification of beach occupation combines an *oversegmented classification model* to classify superpixels into classes (e.g. beach user and sand) and a *regression model* to convert classified beach user superpixels to beach users.

The oversegmented machine learning method has previously been evaluated in the study of Hoonhout et al. (2015) for the classification of coastal images into the classes 'water', 'sand', 'objects', 'vegetation' and 'sky' and this led to the open-source toolbox Flamingo (Hoonhout and Radermacher, 2014b). The current study adapts and develops the Flamingo toolbox for the quantification of beach occupation. The impact of changing parameters of the existing toolbox on the *oversegmented classification model* are evaluated to obtain insight in the parameters that have to be changed to apply the toolbox to the quantification of beach occupation. The influence of the parameters: class aggregation, measures to take account for imbalances in the dataset, regularisation, number of images in the training dataset, image enhancement, addition of artificial channels to enable more (new) features and the required number of features are reviewed. Especially changes in the parameters class aggregation, number of images in the training dataset and artificial channels affect the overall model performance. The effect on the overall model performance of measures to account for imbalances in the dataset is limited. However, these measures can change the relative distribution of precision and recall corresponding to the false-negative and positive rates respectively. The final classification model is trained and validated with a dataset containing 76 manually annotated images, default undersampling to account for the imbalance in the dataset and added artificial channels. A 4-class model with classes beach users, sand water and objects proved to be the best performing class aggregation.

The classified beach user superpixels are converted into a number of beach users with a *regression model* obtained by fitting a second order polynomial regression line to the classified beach user superpixels of the training images and the corresponding manually counted ground truth. Evaluation of the fit shows that the oversegmented machine learning method is a suitable method for quantification of beach occupation indicated by a R^2 of 0.92. The regression model is validated by application of the combined oversegmented classification- and regression models on a new and 'unseen' dataset of 80 images.Validation shows that the regression model is applicable on images that are not used during development of the model (R^2=0.87) and this moreover confirms the suitability of the oversegmented machine learning method. Analysis of the largest errors showed that especially unoccupied beach stretchers and images captured by an unclean lens limit the performance of the oversegmented machine learning method.

The developed oversegmented machine learning method is benchmarked against one of the differences in pixel intensity methods representing the current state-of-the-art. The benchmark shows that the oversegmented machine learning method (R^2=0.87) has a higher performance on the evaluated validation dataset compared to the method representing the current state-of-the-art (R^2=0.76). The difference in performance indicates that the newly developed method is more suitable to the varying conditions associated with the Dutch coast.

Tests of the oversegmented machine learning model on a different camera station than was used for training of the oversegmented classification model, did not lead to satisfactory results. This indicates that the current approach for application on camera stations not used during training is not suitable. Therefore, at this point, the oversegmented machine learning method lacks robustness with respect to the performance on multiple different camera stations. A number of possible causes for the limited performance are treated and provide recommendations for further research.

The presented oversegmented machine learning method, despite its limitations, provides an opportunity to quantify beach occupation with a high temporal and spatial resolution in variable (weather) conditions that are known to limit the performance of the current state-of-the-art methods and are typical for the Dutch coast. The method, therefore, enables the possibility to monitor locations in conditions that with the current state-of-the-art would be difficult to monitor.

Acknowledgements

The past 10 months have been an interesting journey with a lot of new insights and experiences. This journey has been made possible with the help of a number of persons that I would like to acknowledge for their time and input.

First of all, I would like to thank my graduation committee - prof. dr. ir. S.G.J. Aarninkhof, dr. ir. M.A. de Schipper, dr. ir. W. Daamen, ir. M. Radermacher (all TU Delft) and ir. R.C. de Zeeuw (Shore Monitoring and Research) for their insights, support and flexibility during the project.

Next, I would like to thank Bas Hoonhout from Deltares who, together with Max Radermacher, supported me in becoming familiar with the Flamingo toolbox and helped me with the further development of the tool for the quantification of beach occupation. I enjoyed improving my knowledge and skills on programming- and machine learning aspects that I was not familiar with before.

I would like to thank Pieter van 't Hof MSc. of the '*Reddingsbrigade Rockanje*' who helped me with acquiring the webcam dataset at the life guard station on the beach in Rockanje. And, although the images are not used in the final thesis, I would also like to thank Sander Dijkmeijer of DPI Animation House for providing me with a dataset of image acquired at the light house in Scheveningen.

Lastly, I would like to thank all people at Shore for the conversations, laughs and other nice experiences over the past 10 months.

Freek Gulden,
December 28, 2017

Contents

1 Introduction

1.1 Context

The intensity of recreational beach usage is an important factor for coastal zone managers as beach users have a large social-, economic impact on a region (Jiménez et al., 2007). The presence of beach users is for instance largely responsible for the positioning of services on the beach and swimmer safety and insight in beach occupation can be used in the planning of these aspects. In order to provide coastal zone managers and other policy makers with reliable data for evidence based decision making, a high spatial- and temporal resolution is desirable because beach occupation is expected to fluctuate across the beach and during the day/year(s).

At this moment it is common practise in the Netherlands to manually count the number of beach users in the field for a few days. Such a manual counting procedure has a limited spatial- and temporal resolution and is labour intensive. This urges the need for a method which is capable of acquiring data with the desired spatial- and temporal detail and is able to process significant amounts of data. To full-fill the last requirement the method should be automated as the amount of data corresponding to the required spatial- and temporal resolution is too large to analyse manually.

The thesis project has been performed at the company Shore Monitoring and Research (Shore). Shore is a small coastal survey, innovation and consultancy company, predominantly operating in the Dutch market but also serving foreign clients. Given the commercial incentive of the company, the proposed method should preferably be easy and cost effective to implement.

1.2 Problem definition

Research projects in the past have led to a number of (slightly) different visual light (camera) methods to map beach occupation, which meet (most of) the above stated requirements. These methods rely on the r,g,b (colour)-intensities of visual images and distinguish people from the beach based on the assumption that beach users form an isolated peak in the pixel intensity histogram.

The intensity histogram presents on the horizontal axis intensity values and on the vertical axis the number of pixels within the image that have a particular intensity value. Most current methods assume that beach and beach users correspond to different intensities in the histogram and that a pixel is either beach or a beach user. If these assumptions hold, the result will be an intensity histogram showing two distinct peaks.

In general the beach is represented by the lighter range in the intensity histogram corresponding to the right peak (Fig. 1.1), whereas people correspond to darker colours in the histogram indicated with the left peak. A threshold can be defined in between these peaks and used to separate the pixels representing persons from pixels representing beach. Note that the area under the peak in the intensity histogram corresponding to the beach is most often larger, because the area under the peak corresponds to a number of pixels and usually more beach than beach user pixels are present.

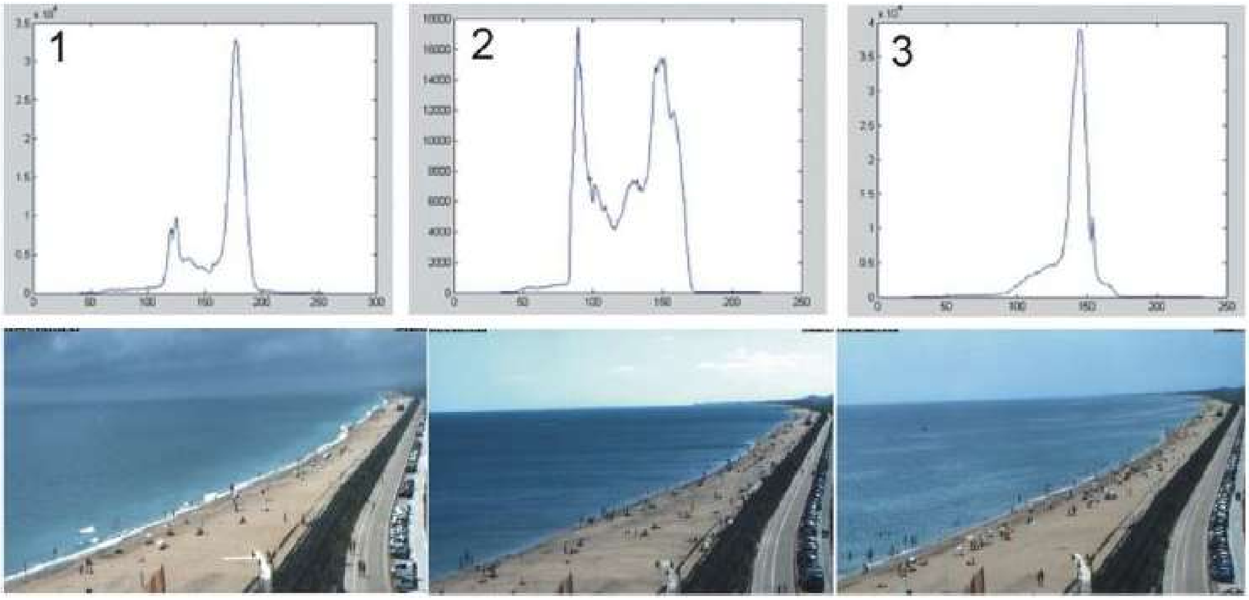
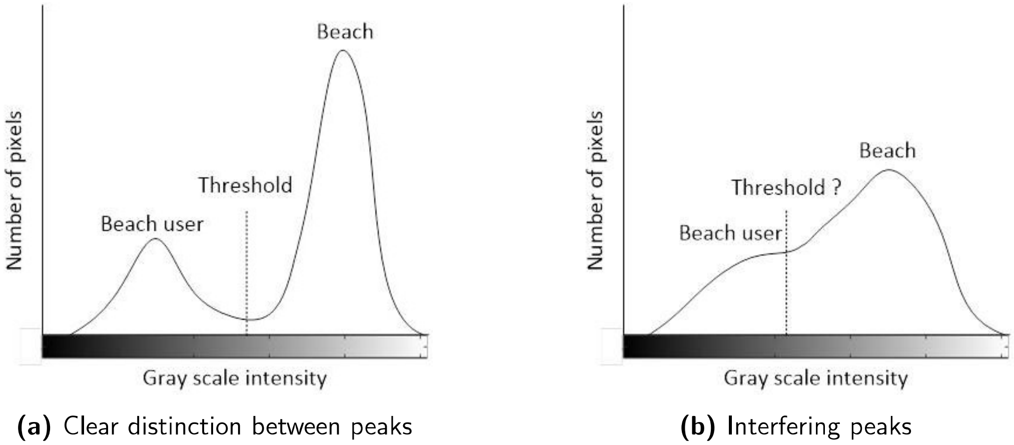


Figure 1.1: Example of changes in intensity histogram for different images: 1) 18/08/2011 at 10h; 2) 09/08/2011 at 15h and 3) 05/08/2011 at 12h at the Lido of Séte beach in France (Balouin et al., 2014). The shape of the histogram is not fixed and changes for different situations. The histogram corresponds to a region of interest defined as the dry beach.

The position of the intensity peaks corresponding to people and the beach may vary within the intensity histogram and moreover the shape of the histogram is not fixed (Fig. 1.1). These changes are caused by deviations in the amount of contrast and the location of the potential threshold dividing the two types of pixels is not fixed (Balouin et al., 2014). Common causes for deviations in the amount of contrast are variation in the height of the sun during the day or variations in weather conditions (e.g. clouds). Some methods have been proposed to reduce the variability of the peaks, but especially cloudy conditions are known to limit the detection accuracy.

Cloudy days are common in the Netherlands and therefore the performance of the existing methods in the Dutch climate is questionable. However, the performance on the Dutch coast has not been tested as most previous studies have been performed around the Mediterranean in the countries: Spain (Osorio et al., 2006; Jiménez et al., 2007; Guillén et al., 2008) and France (Balouin et al., 2014). Only one project with a climate more comparable to the Netherlands (Germany) was found (Kammler and Schernewski, 2004) and this study reported that some images could not be analysed due to bad weather. Unfortunately, no specific details of the detection accuracy of this method are mentioned in this study. The Mediterranean climate differs from the Dutch regarding the amount of sun and clouds and therefore results on performance of projects around the Mediterranean sea cannot be used for the Netherlands.

In addition, the Dutch coast has a rather large inter-tidal area compared to the Mediterranean. The presence of a large inter-tidal area can cause problems as sand in the inter-tidal area is darker than the higher located dry sand. The dark, wet sand will cause the beach peak in the intensity histogram to become wider and shift towards the darker colours which are assumed to correspond to beach users. If this shift is too much, the people- and beach peak can start to interfere (Fig. 1.2), possibly complicating the counting as (parts of) the inter-tidal zone might be marked as beach users.

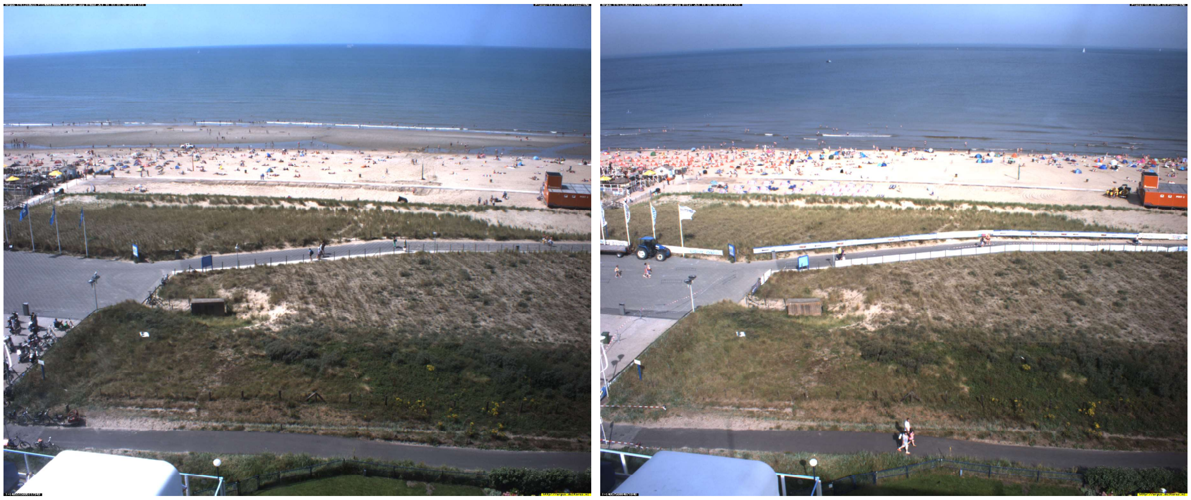


(a) Clear distinction between peaks

(b) Interfering peaks

Figure 1.2: Illustration indicating the possible effect of a large inter-tidal area

A large inter-tidal area, moreover, implicates a significant variation of the beach width during the day and this complicates the definition of a region of interest (ROI). Most of the above mentioned studies define a ROI that contains the permanently dry beach and excludes the sea. On the Mediterranean coast this definition of the ROI is suitable due to the limited tidal variation, but retaining this ROI on the Dutch coast excludes areas that are dry during low tide and can be occupied by beach users (Fig. 1.3). Excluding this area can therefore result in under-predicting the number of beach users and this is undesirable.



(a) Low tide

(b) High tide

Figure 1.3: Example of the Kijkduin beach area in the Netherlands during low and high tide. During low tide beach users are present on the inter-tidal area

Because the Dutch climate is expected to deliver unfavourable conditions more often in comparison to the Mediterranean and the larger tidal-range on Dutch beaches, the applicability of existing methods on the Dutch coast is questionable and has to be investigated. Possibly existing methods have to be adapted or new methods have to be developed.

1.3 Research objective

The objective of this master thesis project is to:

Develop/improve and test a method for accurate, robust and automatic monitoring of the spatial- and temporal distribution of beach users on the Dutch coast.

With *accurate* the capability of the methodology to detect humans on the beach correctly (w.r.t. the needs of the end-users) is meant. *Robust* refers to the capability of the methodology to generate sufficiently (w.r.t. the needs of the end-users) accurate numbers in a variety of weather conditions and beach occupation. Consultation of a coastal zone manager of the Province of South Holland and a policymaker of the municipality of The Hague, suggests that an indication of beach occupation (i.e bins of 0-10, 10-100, 100-200 beach users etc.) is valuable for the coastal zone management purposes. Based on this information the final model is considered accurate if it is capable to predict the beach occupation in the correct order of magnitude.

Depending on the application, monitoring can refer to the quantification of the *number* of beach users or the *intensity* of beach users on the monitored coastal stretch. The number of beach users can be useful in studies on for instance the recreational development of a coastal area (are more, the same or less people visiting the beach?) whereas the intensity can be valuable in planning (where do beach users tend to lie on the beach and should litter bins been placed?). A nearly continuous spatial resolution is required to quantify intensity, whereas a temporal distribution in the order of months to years is necessary to quantify trends in the number of beach users on longer time scales.

Although not explicitly included in the objective, the method should preferably be low cost and easy to implement as a result of the commercial incentive of the company.

1.4 Research question

The above defined research objective can be transferred into the following research question:

"What is the most suitable methodology and its most important characteristics regarding accuracy and robustness for automatically monitoring of beach usage on the Dutch coast?

Supporting sub-questions will be formulated in section 3 after the literature review.

1.5 Outline of report

Chapter 2 of this report contains a literature review that elaborates on the available alternatives to count persons observed in literature to finally select the most promising alternative. Chapter 3 defines sub-questions to structure the research in subsequent chapters. Subsequently, chapter 4 presents the methodology that has been used to obtain the results presented in chapter 5. Chapter 6 discusses the results and chapters 7 and 8 provide the conclusions and recommendations, respectively.

2 Literature Review

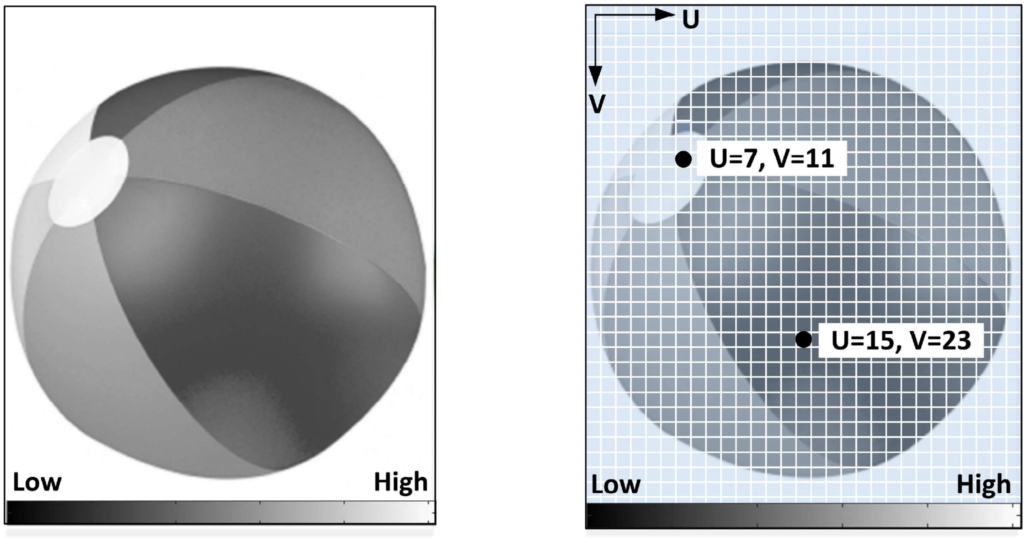
The literature review focuses on automated methods to detect and quantify persons in an area. Because previous work on the quantification of the beach occupation is performed using images, image-analysis alternatives will be examined first. Subsequently an overview of alternative methods to examine beach user quantification and location will be given. This literature review is completed by a concluding section that selects the most promising method with respect to the research objective of this project.

2.1 Digital image-analysis alternatives

Section 2.1 starts with an introductory subsection (2.1.1) on digital image-analysis to provide general background information required to elaborate on the methods in subsequent subsections (2.1.2, 2.1.3 and 2.1.4).

2.1.1 Introduction digital image-analysis

A digital image can be defined as a two-dimensional discretized function $f(u, v)$ (Gonzalez and Woods, 2008) and is captured with a sensor by a sampling process (Young et al., 2007). During sampling the sensor maps a real-world continuous space in x, y-coordinates to a discrete space in u, v-coordinates. Real-world information is assigned to a matrix spanned by u columns and v rows. Every combination of u and v is an entry in the matrix and is called a pixel. Pixels contain the value of the (discretized) amplitude of the function, which is called the intensity (Gonzalez and Woods, 2008).



(a) Continuous image **(b)** Discretized image

Figure 2.1: Illustration on discretization and pixel intensity. Using the grey-scale it can be observed that the pixel corresponding to the dot $(7, 11)$ has a higher intensity than the pixel corresponding to $(15, 23)$, because the latter has a darker grey-level compared to the former.

Dependent on the type of sensor, the intensity of different band-widths (channels) in the electromagnetic spectrum are acquired. A normal, visible light (colour) camera captures for instance the intensity of the Red, Green and Blue (r,g,b) channels, whereas an (thermal) infra-red camera captures the intensity of the thermal channel. Subsequently, (statistical) procedures on pixel intensities can be used to segment, analyse and classify pixels and distinguish classes from each-other. Digital image-sensors used for detecting humans are: visible light, (thermal) infrared or multi-/hyperspectral camera's.

From this group of sensors, the visible light camera is the most common and also the only one observed in previous research on the quantification of beach occupation. Nevertheless the infrared and multi-/hyperspectral sensors have beneficial features: The former is not sensitive to differences in contrast caused by variations in light conditions (Goubet et al., 2006), whereas the latter provides more channels for classification (e.g. visible light *and* infrared (Hwang et al., 2015)). However, visible light sensors are considered more cost effective and it is therefore decided to focus on the visible light sensor in this thesis.

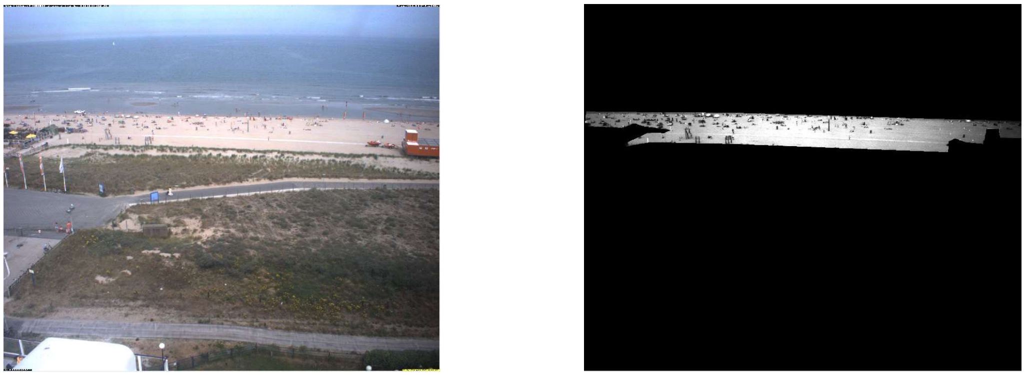
Digital image-analysis procedures are often based on a workflow, which can be separated in five elements (Hendriks, 2014a):

1. **Image acquisition:** Acquisition of the digital image with a sensor.
2. **Pre-processing and restoration:** After acquisition the image might be prepared for further steps or adapted by taking for instance a region of interest out of the original image.
3. **Segmentation:** Distinguish between different parts in the image (e.g. pixels corresponding to persons and pixels corresponding to the beach).
4. **Analysis/ feature extraction:** After segmentation the features of pixels (e.g. colour-intensity) can be analysed in order to find the characteristics of a pixel.
5. **Classification:** Lastly pixels are classified into classes (e.g. beach user and beach) based on the obtained characteristics.

The steps in this workflow can be performed in multiple ways and especially the combination of choices for the steps 3-5 result in different alternatives. In total three fundamentally different approaches have been observed. The first approach corresponds to alternatives assuming the existence of two peaks in the intensity histogram, one representing beach users and one representing the beach. It was already mentioned in the problem definition (section 1.2) that these methods show limited detection accuracy in cloudy conditions or situations with a varying background colour (e.g. inter-tidal area). The second and third approach aim at overcoming these issues by respectively using the variance of pixel intensities over multiple frames or applying oversegmentation combined with a machine learning framework for classification. The following three subsections elaborate on the working principle of the three alternatives separately.

2.1.2 Method I: Use of differences in pixel intensity

Previous research on (automatic) quantification of beach users has especially been based on differences in (colour-) *intensity* of pixels. The method based on differences in pixel intensity uses single r,g,b-images from the visible light band. During pre-processing a Region of Interest (ROI) is defined and taken out of the image (Fig. 2.2). The exact definition of the ROI differs slightly between the studies, but is in general focused on the *dry* beach. Besides the extraction of an ROI, some researchers convert the image to a 8-bit grey-scale image (Kammler and Schernewski, 2004; Balouin et al., 2014) or analyse only the red colour band (Guillén et al., 2008).



(a) Original image **(b)** Pre-processed image

Figure 2.2: Example of pre-processing

The differences in pixel intensity method uses the intensity of pixels to segment beach users from the beach via a thresholding process. A pixel corresponding to the beach, shows a peak at a different location in the intensity histogram than one corresponding to objects (i.e. beach users). The division between these two peaks is set as a threshold and after analysis of the pixel intensities, used to classify the pixels.

The definition of the threshold differs between the studies, but most authors state that this threshold is not a fixed value and will change for different light/contrast conditions. Some define the threshold where three neighbouring grey tones contain more than 20 pixels each (Kammler and Schernewski, 2004), whereas others have determined three spectra classes with corresponding threshold and use an automated process to find the class corresponding to a certain image (Balouin et al., 2014). Moreover, the thresholding itself differs slightly. Pixels can be classified pixel by pixel (e.g. Kammler and Schernewski, 2004), whereas pixels can also be compared with the intensity of its neighbours in a kernel and classified based on statistics of the kernel (e.g. Jiménez et al., 2007).

Subsequently, the classified beach user pixels have to be converted into the number of beach users and two approaches have been observed. The first approach derives a regression relation between the classified pixels and a counted ground truth (e.g. Guillén et al., 2008), whereas the second approach determines a relation between the number of pixels and the area used per person (e.g. Balouin et al., 2014). Because persons are identified in the $u, v-$ coordinates of the image, the image first has to be rectified to $x, y-$ coordinates that represent true distances in the real world.

Although there are slight differences in the methods to detect beach users based on intensity values, all have been reported to have limited skill in low visibility conditions (e.g. rainy- or cloudy days) or with variations in background (sand) colour.

2.1.3 Method II: Use of variance of pixel intensities over multiple frames

An alternative to the method using differences in pixel intensity of single snapshots, is to use the pixel variance of multiple images. Trygonis et al. (2015) describe a method which uses multiple r,g,b-images to determine the *variance (σ^2)* of the pixel intensity of the same pixel over multiple frames instead of the intensity of a pixel in a single frame. During prepossessing an ROI is defined including the swash zone and no image enhancement steps have been mentioned.

To determine the variance of a pixel, the changes in intensity of that particular pixel are compared between multiple frames and quantified with the standard deviation ($= \sqrt{\sigma^2}$). The standard deviation can be calculated for every pixel in the image and assigned to a matrix with the same size as the original image. The result of this procedure is the variance image (Fig. 2.3). A pixel with a low (or zero) variance value indicates that the intensity in that pixel did not differ (much) between the analysed images, whereas a high variance value indicates a lot of variation in pixel intensity.

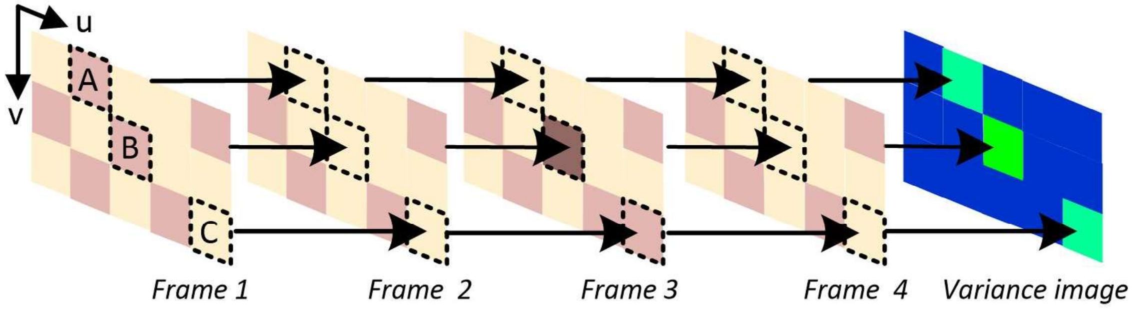


Figure 2.3: Illustration of four frames with 15 pixels and the corresponding variance image that can be created from the differences in pixel intensity between the frames. Pixels A, B and C have a changing colour (and thus intensity) between the snapshots and therefore show up in the variance image.

The phenomenon that induces variation and will be visible in the variance image is dependent on the timescale and amount of analysed images. This can be explained with the definition of the standard deviation (Dekking et al., 2005):

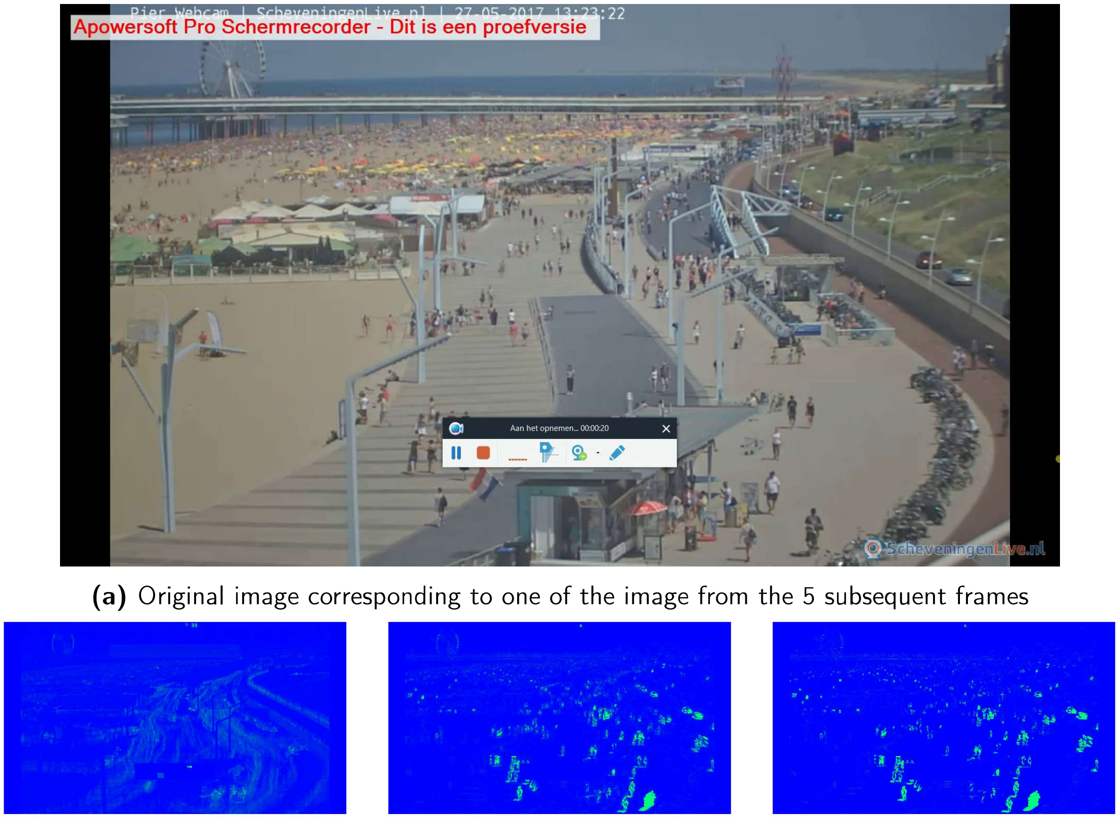
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} \tag{2.1}$$

In which:

x_i = intensity of evaluated pixel in particular frame
$\overline{x}$ = mean intensity of evaluated pixel over N frames
N = number of frames

During analysis on long timescale, instantaneous variations in the intensity (e.g. a person on the beach) will have a limited impact on the numerator while the large amount of frames significantly increases the denominator. As a result of this instantaneous variations will be averaged out, while persistent variations (e.g. a waving flag in windy conditions) will be visible. If only a few images from a short period are analysed, the impact of an instantaneous variation on the nominator will be relatively large while the denominator is rather small. The result of this is that the instantaneous variations will be visible as well.

The proposed method by Trygonis et al. (2015) uses the above described behaviour to segment (moving) beach users from the background. The background containing the persistent variations is created by randomly sub-sampled frames from a 10 minute period (Fig. 2.4b). Moreover, a variance image based on 5 subsequent frames sampled at 5Hz is created to indicate instantaneous variation (Fig. 2.4c). The variance image based on 5 subsequent frames also contains background variation and therefore the background image is subtracted from the instantaneous image. The result is a variance image only containing instantaneous pixel variations (Fig. 2.4d). Because humans are the predominant source of variation on the beach, the obtained image with instantaneous pixel variations is assumed to indicate persons on the beach.


Pier Webcam | ScheveningenLive.nl | 27-05-2017 13:23:22
Apowersoft Pro Schermrecorder - Dit is een proefversie
Aan het opnemen... 00:00:20
ScheveningenLive.nl

(a) Original image corresponding to one of the image from the 5 subsequent frames

(b) Background variance image **(c)** Instantaneous variance image **(d)** Substracted variance image

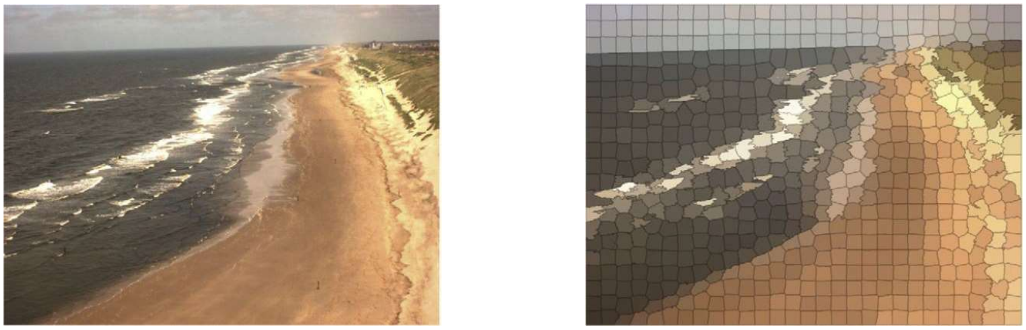
Figure 2.4: Example of procedure with variance images

After obtaining the subtracted variance image and analysing the standard deviations, again a threshold can be used to distinguish pixels with a high standard deviation from pixels with a low value. Subsequently, standard image operators can be used to ultimately count the number of beach users. Note that this method will only count *moving* beach users because not moving beach users do not induce variation in pixel intensity, which is necessary to be captured by this method.

2.1.4 Method III: Oversegmentation and trained machine learning framework for classification

Another option to possibly overcome the issues of the pixel intensity difference method is the classification framework presented in Hoonhout et al. (2015). This method analyzes single r,g,b- snapshots and no image enhancement or pre-processing steps are mentioned. Although the presented results are focused on the segmentation of coastal images into major classes like object, sand, sky, vegetation and water, the algorithm is said to be a general classification framework for coastal images.

The essential difference between this method and the method using differences in pixel intensity, is the use of an oversegmentation procedure. Oversegmentation is a process wherein superpixels are formed based on pixel gradients (Achanta et al., 2012). Superpixels contain multiple pixels with similar colour characteristics that are located close to each other within the image (Fig. 2.5) and this method classifies these superpixels instead of single pixels.



(a) Original image

(b) Oversegmented image

Figure 2.5: Example of oversegmentation (Hoonhout et al. (2015)). In the oversegmented image superpixels are indicated with a grid.

The added value of superpixel formation is an increase in the available features for classification purposes. Regular pixels provide information on a limited number of features as for instance the r,g,b- colour intensities, while superpixels have a lot more characteristics that can be expressed in features (e.g. Fig. 2.6). The result is an increase in available features from a few related to colour intensities of single pixels to more than 1000 features related to superpixels used in Hoonhout et al. (2015).

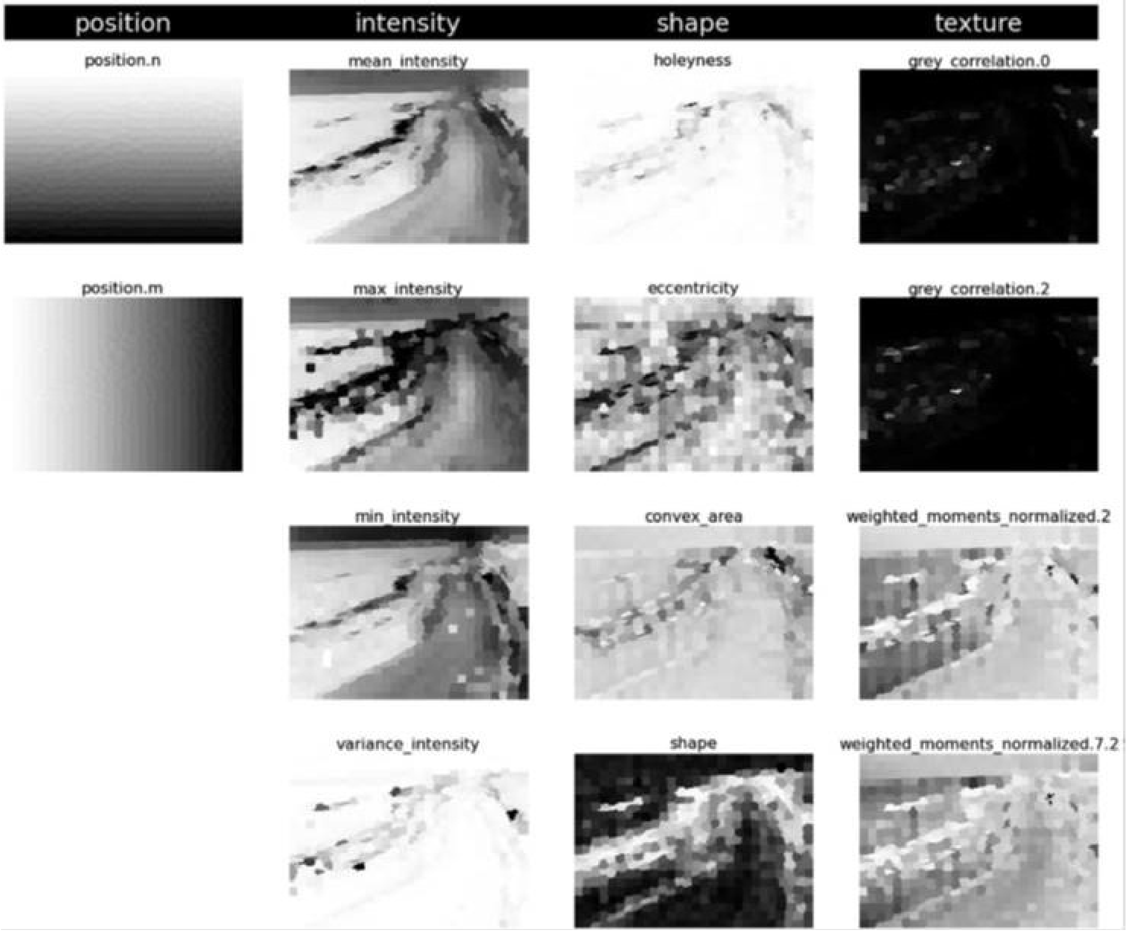


Figure 2.6: Example of features that can be extracted from superpixels (Hoonhout et al., 2015)

For situations where the regular pixel features (e.g. colour intensity) differ clearly between the proposed classes, these can be sufficient (note: this is the working principle of Method I using differences in pixel intensity), but if they start to interfere due to for instance the earlier mentioned inter-tidal zone, it becomes difficult to classify the pixels correctly. The availability of much more features potentially provides additional options to distinguish between pixels in difficult to classify images, which makes this method an interesting alternative.

The relation between features and classes is not known upfront and manually deriving them would be a complex and time-consuming process. Therefore, use is made of a machine learning framework that learns a classification model the correspondence between features and classes. In general, two approaches to the learning task can be adopted (Fig. 2.7).

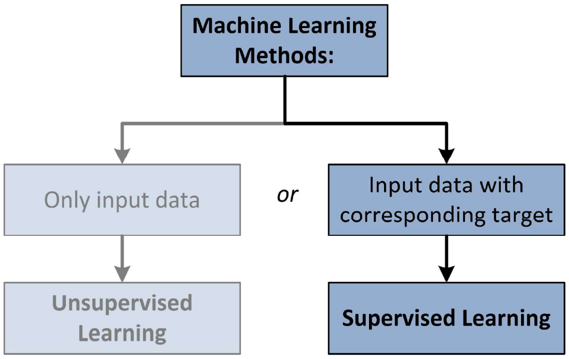

Machine Learning Methods:
Only input data
or
Input data with corresponding target
Unsupervised Learning
Supervised Learning

Figure 2.7: Illustration on approaches to learning task. The oversegmented machine learning method described in Hoonhout et al. (2015) adopts supervised learning.

The classification of visual images by features of superpixels is a learning problem with input data that can be assigned by targets (class-labels) after manually annotating a sufficient amount of images. This is an example of *supervised* learning, because the computer is learned what the class-label corresponding to the input-data is. The opposite of supervised learning is unsupervised learning in which the machine will get input data and searches for relations that lead to (on beforehand unknown) output labels. The study of Hoonhout et al. (2015) is based on supervised learning.

Supervised learning can be applied in multiple subsequent steps (Hoonhout et al., 2015):

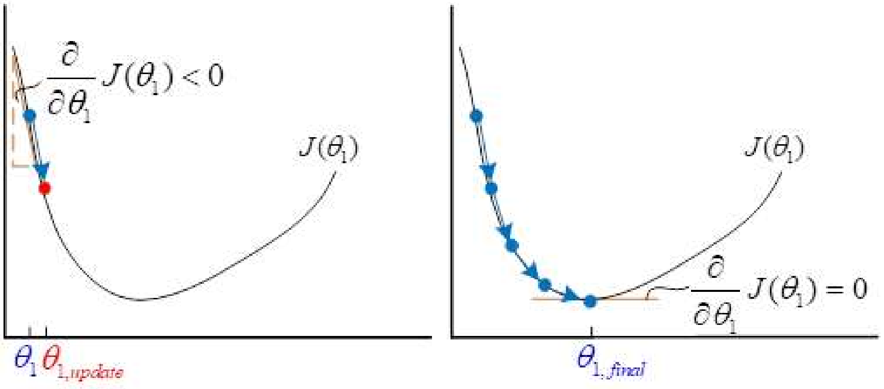
1. Manually annotate the class of superpixels for a representative amount of images in order to create a dataset for training.
2. Extract and normalise the features from the annotated images.
3. Split the training dataset in a part for the actual training and remain a part for testing/validation of the model.
4. Train the model by minimising a cost function via iterating training data in order to end up with a parameter vector describing the relation between features and the classes.
5. Test/validate the performance of the classification model and redo training if performance is not satisfying.

In general the objective of classification is to create a model that best divides the data in classes (e.g. beach users and beach). During training a hypothesis $h_\theta(x)$ is selected and fitted to the data by varying parameters (*unary potentials*) $\theta_0, \theta_1 ..., \theta_n$ related to features x_0, x_1 ... , x_n (A. Ng, 2012d).

The optimal values for the unary potentials are obtained by minimising a cost function $J(\theta)$ with *gradient descent*. The cost function penalises incorrect classifications and therefore minimising the cost function results in a classification model with the lowest number of incorrect classifications. Gradient descent is a procedure that evaluates the gradient of the cost function at the location corresponding to the values of $\theta_0, \theta_1..., \theta_n$. Subsequently, the fastest direction to arrive at the minimum of the cost function is determined based on the gradient and the unary potentials $\theta_0, \theta_1..., \theta_n$ are simultaneously updated (A. Ng, 2012c):

$$\theta_{j,update} = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_j) \tag{2.2}$$

The update of the unary potentials represents a small step towards the minimum of the cost function and the step size is controlled by the learning parameter α in eq. 2.2. After multiple iterations the minimum of the cost function is obtained corresponding to a zero gradient. Therefore, the unary potentials are not updated anymore and the final (optimal) values are obtained (Fig. 2.8).

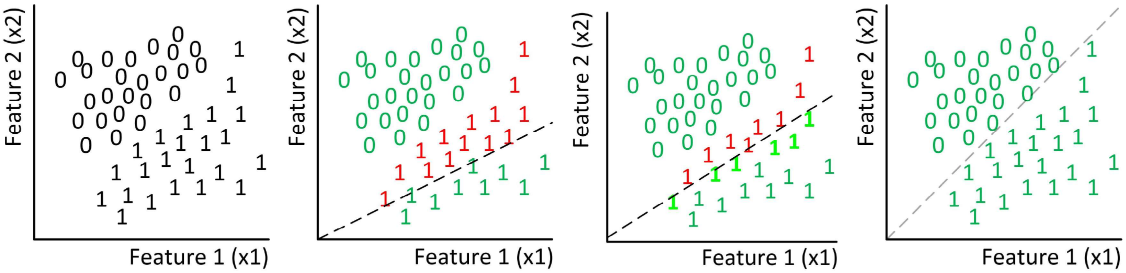

$\frac{\partial}{\partial \theta_1} J(\theta_1) < 0$
$J(\theta_1)$
θ_1 $\theta_{1,update}$
$J(\theta_1)$
$\frac{\partial}{\partial \theta_1} J(\theta_1) = 0$
$\theta_{1,final}$

(a) First iteration

(b) Final iteration

Figure 2.8: Illustration of principle of minimising cost function J_{θ_1} with gradient descent for one unary potential. The gradient at the evaluated point on the cost function is analysed and used to update the value of θ_1 with eq. 2.2. This procedure is repeated until the gradient becomes zero which corresponds to the minimum of the cost function.

The foregoing can be illustrated with a (simplified) two-class example in a two dimensional feature space (Fig. 2.9). In this example a division has to be made between classes 0 and 1. The data points correspond to training instances (i.e. superpixels) with particular feature1 and feature2 characteristics. Because the training dataset is annotated, the class labels corresponding to the data points are known and it is possible to plot the data points together with their class label (Fig. 2.9a). The division of the two classes in this example can be made with a simple linear model in the feature1-,feature2- plane (Fig. 2.9d). A good hypothesis for this linear model would be the definition of a straight line: $x2 = \theta_0 + \theta_1 x1$ (output = x2 intercept + slope) in which θ_0 and θ_1 can be changed to find the best division.



(a) Annotated dataset **(b)** First iteration **(c)** Second iteration **(d)** Final iteration

Figure 2.9: Example of effect gradient descend. The number of incorrectly classified data points reduces with every iteration as a result of the increasing slope. The increase in slope is caused by the updates of θ_1 as a result of gradient descent.

For this simple example it can readily be seen that the line $x2 = x1$ (and thus $\theta_0 = 0$, $\theta_1 = 1$) divides the data best (Fig. 2.9d). However, the computer does not know this upfront and will for instance start with $\theta_0 = 0$ and $\theta_1 = 0.5$ (Fig. 2.9b). During subsequent iterations the values of θ_0 and θ_1 are updated resulting in changes in the intercept and gradient of the line (e.g. Fig. 2.9c). Eventually the cost function is minimised and the optimal values of θ_0 and θ_1 are obtained together with the corresponding optimal division (Fig. 2.9d). This learns the model what combinations of feature 1 and feature 2 correspond to the classes. A superpixel with for instance high feature 1 value and a limited feature 2 values is very likely to be of the 1-class, whereas a superpixel with limited feature 1 and high feature 2 values will most likely correspond to the 0-class. The unary potentials are stored in a parameter-vector and can subsequently be used for the classification of new data points.

The above example is a very simplistic representation of the learning principle of a machine learning algorithm. In reality the algorithm will have more than 2 dimensions due to the availability of numerous features as a result of oversegmentation. Moreover, the line does not have to be straight and can get curved by taking into account higher, non-linear feature relations, which require more advanced model types. In the example both features are valuable to the classification of the classes (a high feature 1 value can also lead to a 0-class data point if the feature 2 value is relatively large and therefore information on the feature 2 value is required), whereas in reality it is possible that particular features have no distinctive value among the different classes.

After training, the superpixels of unseen images are classified by a prediction function that classifies the superpixels such that the prediction function is maximised (Hoonhout et al., 2015).The existing algorithm is capable of distinguishing between the classes: sky, water, sand, vegetation and objects. To detect people the object class should be sub-divided further in order to be able to distinguish for instance persons from buildings.

2.2 Analysis of alternatives

Besides monitoring with images, alternatives using other signals can be used to detect persons. For the completeness of this study, these alternatives are also briefly explored.

2.2.1 Bluetooth

The first alternative method detects persons with the use of Bluetooth-signals emitted by portable devices like a mobile phone (O'Neill et al., 2006). When a person enters the range of the receiver and has Bluetooth switched on, the person will be counted. Application was for instance found in the train station of Utrecht Central in the Netherlands (Voskamp, 2012). A disadvantage about this method is that Bluetooth should be switched on, such that only 10% of the passing pedestrians was detected (Voskamp, 2012). Moreover, new Iphones and Android cell phones are almost undetectable because of a built-in function which switches off Bluetooth if it is not used. The obtained spatial information of this method is limited as the range of the receiver is limited.

2.2.2 Wifi and GPS-location

Following the use of Bluetooth emitted by mobile phones to monitor pedestrians, more methods hav been found that make use of mobile phones. The use of mobile phones for data acquisition, mobile sensing, is becoming increasingly popular (Lane et al., 2010). Other signals that might by interesting are Wifi and GPS. Nevertheless, Wifi and GPS have the same disadvantages as Bluetooth since they also need to be switched on by the user. For the latter is it moreover necessary that people download an app which they allow to collect their location.

2.2.3 LiDAR

Lastly, approaches using multiple single-row laser-range(LiDAR) scanners were found. The method of Zhao and Shibasaki (2005) uses for instance laser scanners with a view of 270°on approximately 20cm above ground level to detect pedestrians. The assumptions is that movement is caused by the feet of walking people. Beneficial about the use of laser technology is that it is not sensitive to weather and light conditions. However, the range is limited which is unfavourable for the spatial resolution and also the amount of additional acquired spatial information (position water line etc.) is limited.

Although the above described alternatives show limitations, they are capable to detect humans. Therefore, the alternatives will be included in the considerations regarding the correspondence of the alternatives characteristics and the research objective in the next section.

2.3 Conclusion

The preceding subsections focused on automated methods to detect and quantify the number of persons in an area, but did not treat the suitability of the methods with respect to the research objective. This subsection elaborates the correspondence between the methods and the research objective in order to identify the most promising method.

The reviewed alternatives are combinations of a sensor to acquire the data and an algorithm to process the data and ultimately quantify the number of persons in an area. Both the choice for the sensor and the choice for an algorithm influence the suitability of an alternative with respect to the research objective. This section will first focus on the suitability of different acquisition sensors (subsection 2.3.1). Subsequently, the choice for the most suitable method will be substantiated (subsection 2.3.2).

2.3.1 Most suitable sensor

Previous studies on beach usage quantification proved that r,g,b- images acquired with a digital camera can be used to automatically monitor a beach area with reasonable spatial- and temporal resolution. Nevertheless, accuracy and robustness showed limitations in certain conditions for images processed with the algorithm based on pixel intensities (subsection 2.1.2). Promising alternative algorithms based on variance images (subsection 2.1.3) and oversegmented images (subsection 2.1.4) indicate that the limited accuracy and robustness are not necessarily related to the sensor, but can be the result of the type of algorithm that is used.

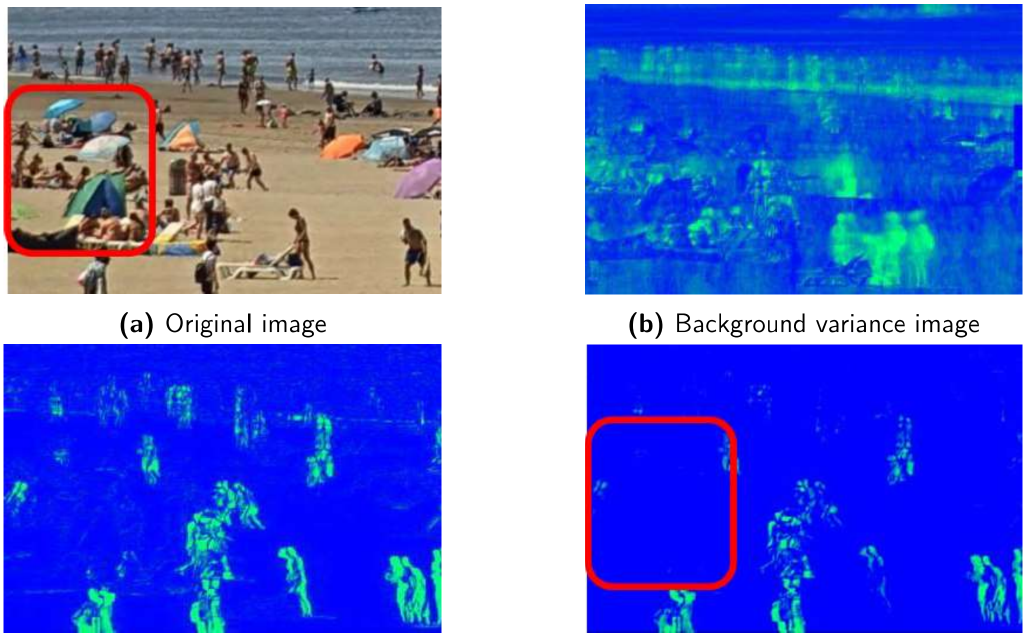
The Bluetooth, Wifi and LiDAR sensors have a limited range, which conflicts with the spatial distribution requirement of the research objective. The spatial detail could be increased by deploying more sensors, but this limits the cost effectiveness and ease of implementation. GPS-location of mobile phones can deliver the requested spatial and temporal distribution, but has a questionable robustness caused by its dependence on the willingness of people to turn on the sensor and allow third parties to use their location.

Based on the above it is concluded that the digital r,g,b-sensor (camera) is the most suitable sensor regarding the objective of this research project. The choice for the digital r,g,b-sensor is supported by the existence of algorithms that are capable to extract other coastal state indicators based on r,g,b-images (Davidson et al., 2007). The possibility to use a sensor that captures multifunctional coastal data is considered to be a big advantage when serving coastal zone managers and other policymakers involved with the coast.

2.3.2 Most suitable method

The algorithms based on variance and oversegmented images potentially solve the limitations of the pixel intensity algorithm for conditions with low contrast, resulting in an increased robustness. However, the variance-algorithm uses variations in pixel-intensity to detect beach users, implicitly assuming that beach users move because otherwise no variation in pixel intensity will be observed.

Preliminary tests showed that non-moving persons are a realistic scenario for sunny days with a lot of sun-bathing people lying/sitting on the beach. To this end a webcam-stream (scheveningenlive.nl) was recorded and cut to frames. Subsequently, variance images (background and instantaneous) have been created and subtracted. The results show that moving persons show up well in variance images, but lying/sitting people merge into the background and this limits the applicability and therefore robustness (Fig. 2.10).



(a) Original image

(b) Background variance image

(c) Instantaneous variance image

(d) Instantaneous - background

Figure 2.10: Example of not/limited moving people (inside the red square) that do not show up in the variance image.

The algorithm using oversegmented images does not rely on movement and therefore is not restricted to moving persons. Another advantage of the oversegmented approach over the variance image approach is the use of single snapshots, because single snapshots correspond to a limited amount of data that can be transferred on internet connections with a relatively low capacity. Variance images on the other hand use snapshots from 10 minutes at 5 fps resulting in a significant amount of data requiring either a high capacity internet infrastructure or a local server to process the snapshots in-situ and transfer only the created variance image. Both options hamper the ease of implementation and moreover result in a less cost-effective solution.

From the above it is concluded that the use of the recently proposed method of oversegmented images combined with a machine learning framework for classification is the most suitable alternative regarding the research objective of this master thesis project. This research project therefore focuses on the development of the oversegmented machine learning method.

3 Supporting research sub-questions

In the last section of the literature review it is concluded that the oversegmented machine learning method is the most suitable method and this provides a theoretical answer to the first part of the main research question: *"What is the most suitable method?"*. Supporting research (sub-)questions are required to verify that the oversegmented machine learning method is indeed a suitable method for beach user quantification. Moreover, supporting research sub-questions are necessary to determine the most important characteristics regarding accuracy and robustness corresponding to the second part of the main research question.

In this section four supporting sub-questions are formulated to structure the research in subsequent chapters.

3.1 Sub-question 1

The oversegmented machine learning method is not used to classify beach users before and to quantify the performance and accuracy, it is first necessary to adapt and develop the method for this application. This process is supported by the first sub-question:

Q1. Which parameters have to be adapted or added to make the oversegmented machine learning method capable to classify beach users?

3.2 Sub-question 2

The oversegmented machine learning method classifies superpixels, but the relation between classified beach user superpixels and the actual number of beach users is not known. Ideally a superpixel contains one beach user, but for instance occlusion can cause a different relation. The second sub-question investigates the relation that should be used to convert classified beach user superpixels to a number of beach users:

Q2. What is the most suitable approach to convert classified beach user superpixels to the number of beach users?

3.3 Sub-question 3

The oversegmented machine learning method is selected because oversegmentation enables additional features that are expected to increase the accuracy for conditions that proved to be difficult to classify with the current state-of-the-art differences in pixel intensity method. The third sub-question verifies the performance of the current state-of-the-art in the variable conditions associated with the Dutch coast. Moreover, the performance of the current state-of-the-art method is compared with the performance of the oversegmented machine learning method to obtain insight in the added value of the oversegmented machine learning method:

Q3. How does the oversegmented machine learning method perform in comparison to the current state-of-the-art in the variable conditions associated with the Dutch coast?

3.4 Sub-question 4

Robustness was introduced as the capability to quantify beach occupation in variable weather and beach occupation conditions. The oversegmented machine learning method includes a training procedure on a specific number of images and this introduces another aspect of robustness: robustness of the method on images captured by camera stations not present in the training data. The image size (resolution) and camera height can for instance differ between stations and this might limit the performance of the model on images from stations that are not included in the training data. The fourth sub-question focuses on the applicability of the oversegmented machine learning method on images from a camera station (e.g. a webcam) that was not included in the training process:

Q4. What is the performance of the oversegmented machine learning method on images from a camera station that was not included in the training procedure?

The next chapter (chapter 4) elaborates on the methodology that is used to answer the above formulated sub-questions. Subsequently, chapter 5 presents the results of the performed research on the sub-questions, whereas chapter 6 discusses these results and tries to generalise them. Chapter 7 uses the results on the sub-questions to formulate conclusions and answer the main research question. Finally, chapter 8 treats the most important recommendations for further research on the quantification of beach user occupation with the oversegmented machine learning method.

4 Methodology

The most promising method obtained in section 2.3, combines oversegmentation with a machine learning framework. The combination of oversegmentation and machine learning for classification of coastal images has previously been explored (Hoonhout et al., 2015) and resulted in the open-source Flamingo toolbox (Hoonhout and Radermacher, 2014b). This study develops and applies the Flamingo toolbox on quantification of beach user occupation by adapting and adding steps to the original workflow presented in Hoonhout et al. (2015).

This chapter starts with an introduction to the existing toolbox. Subsequently, the workflow presented in Hoonhout et al. (2015) is reviewed to indicate the steps/parameters that can be changed or added to make the Flamingo toolbox applicable for quantification of beach occupation. Moreover, the relation between the possible changes/added steps and the research sub-questions is explained in the introduction. Subsequent sections explain the steps in the workflow and elaborate possible changes/additions. The methodology is finished by a section that summarises the treated steps to possibly change/adapt the original workflow. Moreover, the final subsection explains the order in which the results of changing/adjusting steps are treated in the subsequent chapter on the results.

4.1 Introduction Flamingo toolbox

The Flamingo toolbox has been built in the open-source programming language *Python* (Python Software Foundation, 2017) around the packages *scikit-image* (Van der Walt et al., 2014), *scikit-learn* (Pedregosa et al., 2011), *OpenCV* (OpenCV.org, 2015) and *Pystruct* (Müller and Behnke, 2014). The toolbox contains documentation, the source-code and modules for image (over)segmentation, classification and rectification. The toolbox uses a general workflow that is adapted for the quantification of beach user occupation (Fig. 4.1).

The workflow shows that the *oversegmented machine learning method* applied to beach user quantification, combines an *oversegmented classification model* and a *regression model*. The oversegmented classification model classifies images in beach user superpixels and superpixels corresponding to other classes (e.g. sand and water), whereas the regression model converts the classified beach user superpixels to beach users. The steps in the oversegmented classification model correspond to the original workflow presented in Hoonhout et al. (2015) with the addition of an image enhancement step. Changes in these steps have to make the oversegmented machine learning method capable to classify beach user superpixels, which corresponds to the first research sub-question.

The steps in the regression model are required to convert the classified beach user superpixels to the actual number of beach users and this corresponds to the second research question. The validation step in the regression model only validates the *conversion* from beach user superpixels to beach users because the oversegmented classification model has already been validated. The oversegmented classification model puts 25% of the training data aside *before* model training and therefore these images are not seen by the model during training. The trained classification model is used to classify the unseen images and comparison of the model predictions with a manually annotated ground truth returns the validated (classification) model performance.

After development and validation of both the oversegmented classification and regression model, the oversegmented machine learning method can be used to quantify the beach occupation. To obtain information on the performance of the developed oversegmented machine learning method with respect to the current state-of-the-art, it is benchmarked against one of the differences in pixel intensity methods as introduced in section 2.1.2 to answer the third research sub-question.

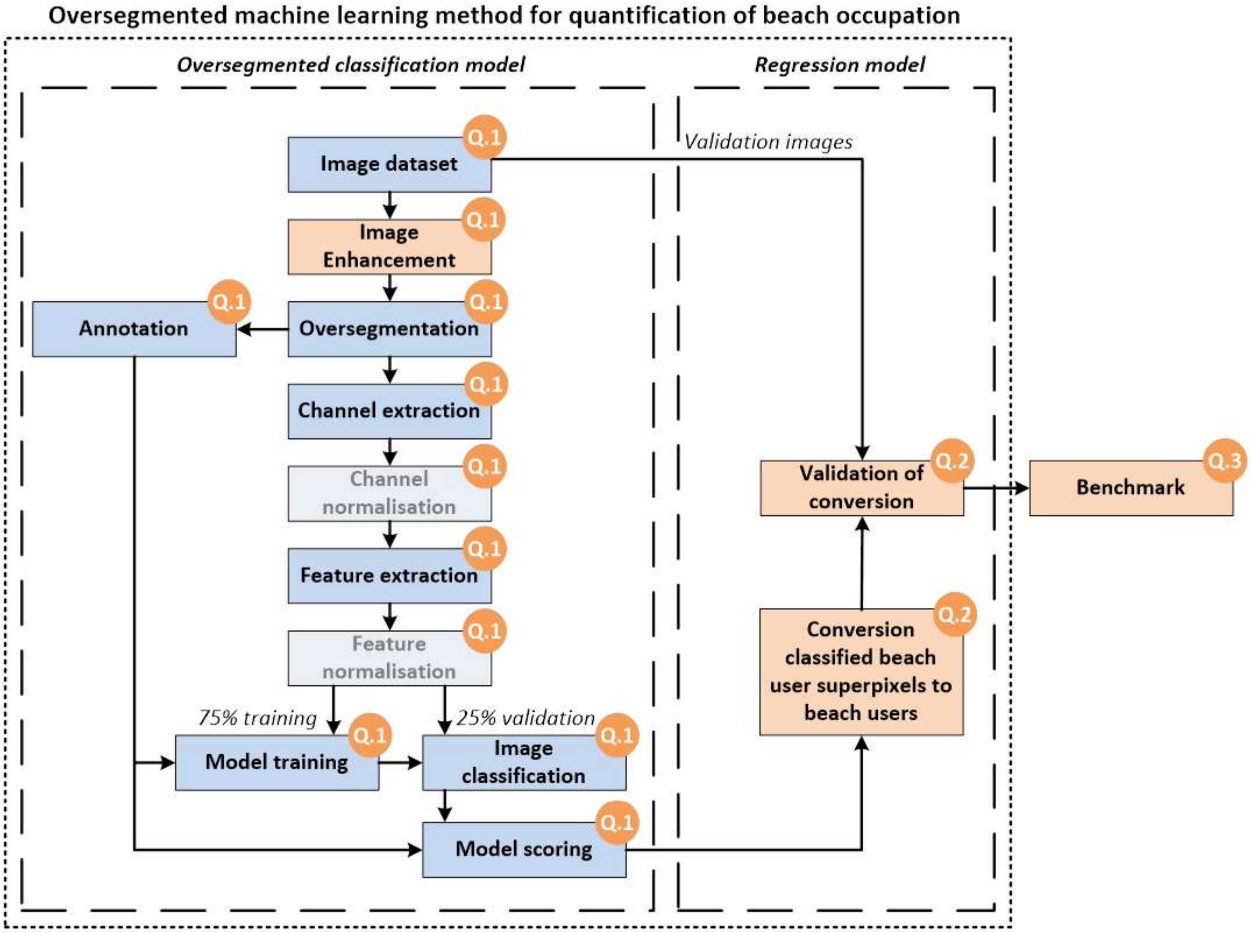


Figure 4.1: Adapted workflow of classification algorithm (after Hoonhout et al. (2015)). The lighter boxes indicate steps that are fixed and the orange boxes represent added steps. The other steps involve parameters that might be changed. The orange circles indicate the sub-question that a step corresponds to.

Investigation of the fourth sub-question is done by revisiting the oversegmented classification and regression models with images from a camera station not used during training. The detailed approach to the investigation of this sub-question with respect to the flowchart is elaborated in section 4.11.

The coming sections elaborate the contribution of steps to the oversegmented machine learning method and indicate possibilities to adapt steps and make the method applicable on beach user quantification. Moreover, the approaches to test the impact of the proposed changes are treated. First the image dataset and annotation (4.2) are treated and subsequently sections on oversegmentation (4.4), channel and feature extraction and normalisation (4.5), model training (4.6), image classification and model scoring (4.7), the conversion of beach user superpixels in beach users (4.8), the validation of the conversion (4.9), the benchmark (4.10) and finally the applicability on images from a camera station not used during training (4.11) follow. This chapter is finalised with a summary in section 4.12.

4.2 Image dataset and annotation

To examine robustness in different weather conditions, beach occupation and camera stations a dataset containing data from two different camera stations at two different locations on the Dutch coast is used (Fig. 4.2).

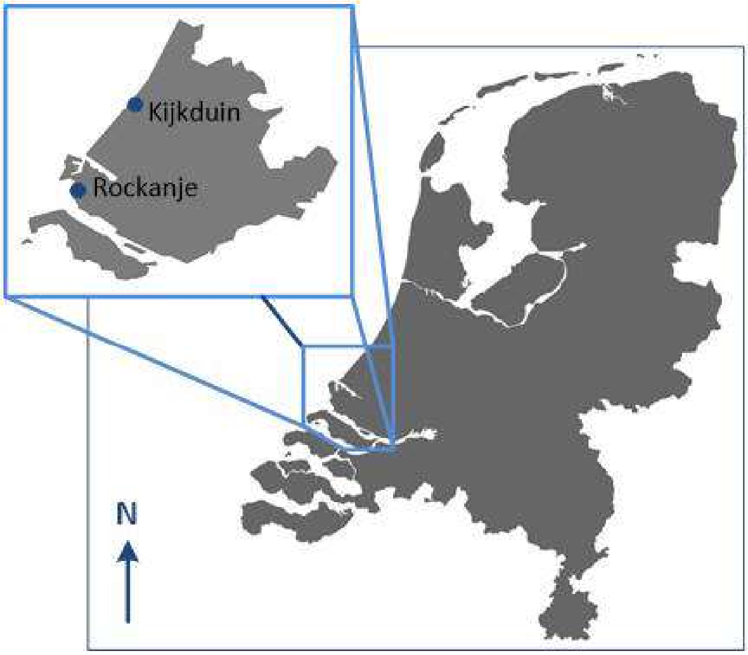


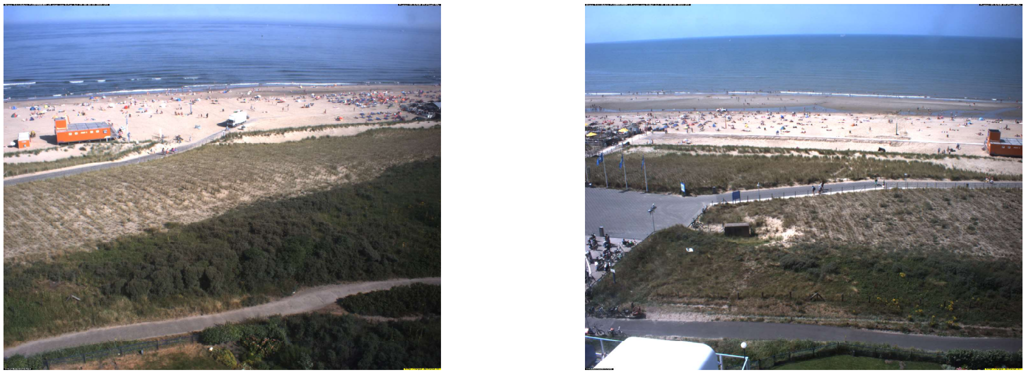
Figure 4.2: Used camera stations along the Dutch coast

This subsection first elaborates the specifications of the camera stations. Subsequently, the data-subsets used for training of the oversegmented classification model and development of the regression model is treated. Next, the dataset for validation of the regression model is elaborated and this section is finalised by a subsection on the dataset used for testing the applicability of the oversegemented machine learning method on images from a camera station not used during training.

4.2.1 Camera stations

Argus station Kijkduin

The first data source is the Argus-station in Kijkduin. This station consists of multiple camera's mounted on the roof of a hotel near the beach. The data contains two hourly snapshots with dimensions 2448x2048. The dataset spans a period from 2013 to 2016 and therefore is considered to contain a suitable representation of the variability in conditions associated with the Dutch weather conditions and beach occupation. Here images from the two central camera's (camera 3 and 4) are used, which overlook the beach in front of of the beach town (Fig. 4.3). The orange building in both images corresponds to a lifeguard station, whereas the other building represents a beach club.



(a) Camera 3 **(b)** Camera 4

Figure 4.3: Example of camera views Argus-station Kijkduin

Webcam-station Rockanje

Secondly data from a webcam deployed at a lifeguard station in Rockanje is used. The camera is installed at the first floor and the station is positioned on the beach. The data contains snapshots of a period from 21 July to 5 September in the summer of 2017. The snapshots were taken four times per hour and have a dimension of 1280*720 pixels. An example of a snapshot from the Rockanje webcam is shown below (Fig. 4.4):

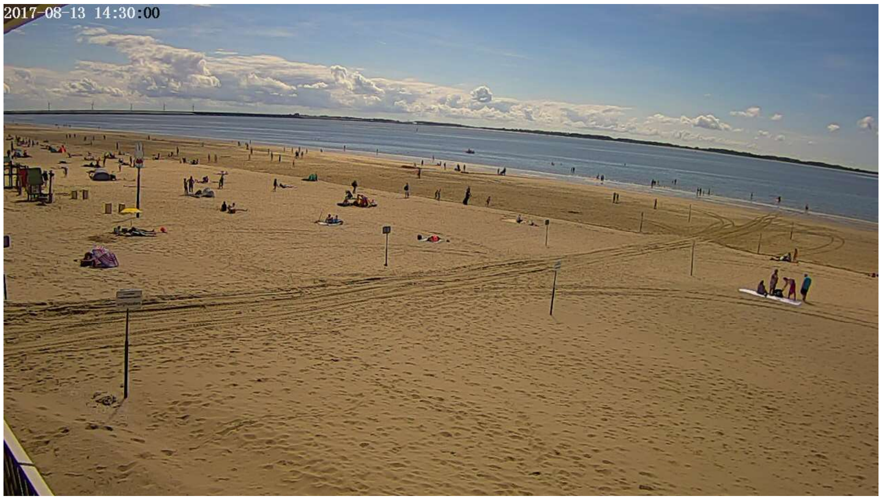
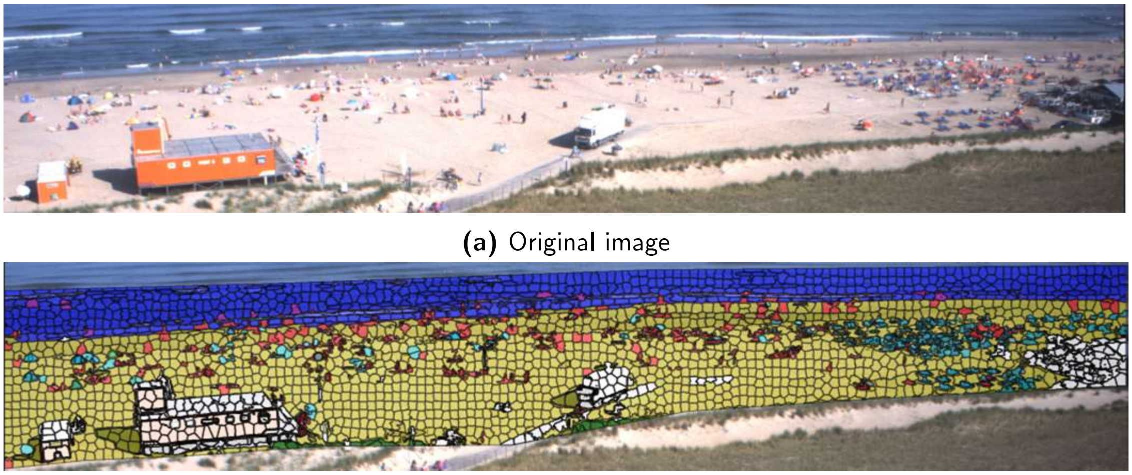


Figure 4.4: Example of snapshot from webcam Rockanje

4.2.2 Subset of data for training and validation of the oversegmented classification model and development of regression model

Part of the dataset is required for the training and validation of the oversegmented classification model and this dataset is defined as the *training dataset* in this thesis. During training the training dataset is divided in a *train partition* containing 75% of the training dataset and a *test partition* of 25%. The training dataset has also been used to develop the regression model for conversion of classified beach user superpixels to a number of beach users.

The training dataset contains images from the Kijkduin Argus source. A total of 76 images from the months June, July, August in the period 2013-2016 between 10.00h-16.30h local time, are randomly selected and visually inspected to ensure that the necessary variety in weather conditions and beach occupation is present. After sampling and inspecting the training dataset, the images are manually annotated with the classes: beach user (person on the beach), swimmer (person in the water), beach object (e.g. stretcher, beach shelter etc.), object (e.g. litter bins, beach club, lifeguard post etc.), sand, water, ROI (superpixel containing the pixels *outside* the region of interest) and vegetation (Fig. 4.5).



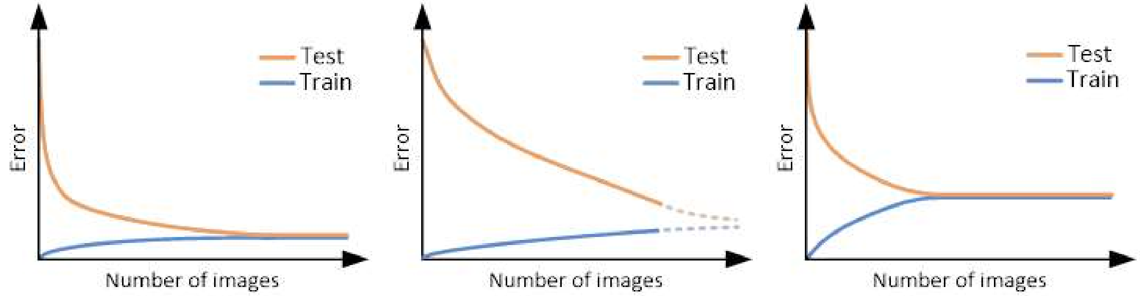
(a) Original image

(b) Annotated image

Figure 4.5: Original image together with annotated ground truth. Red superpixels correspond to beach users, yellow to sand, blue to water, green to vegetation, white to objects, cyan to beach objects and magenta to swimmers. The ROI superpixel is not indicated.

Number of images used for training of the oversegmented classification model

Besides the necessary variety in the data, the number of images that is used during training of the oversegmented classification model is important. Enough images should be used during training to learn the model the variation of image types that is present in the dataset. The suitability of the size of the training dataset can be tested by a learning curve, which is a graph that shows the behaviour of the training and test set errors for a ascending number of images (Fig. 4.6). Learning curves are classification model specific and therefore have to be determined for each oversegmented classification model separately.



(a) Optimal learning curve **(b)** High variance learning curve **(c)** High bias learning curve

Figure 4.6: Illustration on type of learning curves. A learning curve describes the distribution of the error made by the train- and test partitions For an ascending number of images.

In general the training error increases for more images due to an increased number of data points (Fig.

4.6a). The higher number of data points increases the complexity, making it more difficult for a model to divide the classes. However, increased complexity learns the model more generalised information, increasing the performance on unseen images. This results in a lower error rate of the (unseen) test images.

In an ideal situation the curves approach a horizontal line and converge at a low error rate because this indicates that the model can predict unseen images as well as earlier seen images during training at a low error rate. If the curves do not converge but a converging trend is visible, the model has a *high variance* (Fig. 4.6b). A high variance corresponds to an overfitting hypothesis, which indicates that the model has learned to specific information (e.g. to many features or to much higher order polynomial features) and is not generic enough to make good predictions on the test data (A. Ng, 2012b). If this type of learning curve is observed, the model performance might be increased by expanding the training dataset with more annotated images (containing the underrepresented conditions).

It is also possible that the curves converge but at a relatively high error rate and after a limited number of images (Fig. 4.6c). This situation corresponds to a model with *high bias* which can originate from an underfitting hypothesis that is too simple (e.g. not enough features or not enough higher order polynomial features) to divide the complex data into the classes (A. Ng, 2012b). When a high bias is observed, annotating more images will not increase the model performance.

Because the toolbox has not been applied on the quantification of beach occupation before, the amount of images required for training is unknown. The learning curve of the final oversegmented classification model is determined to obtain insight in the impact of the number of images and verify that the size of the training dataset is sufficient.

4.2.3 Subset of data for validation regression model and benchmark

The fit of the regression model developed with the training dataset, is validated by another subset of Argus-images. This subset consists of 80 randomly sampled images from the months June, July and August of the summer of 2013. After sampling, the images in this subset have visually been checked to verify that varying weather conditions and beach occupation are included. Images from 10.30h, 13.30h and 16.30h local time are included in this dataset. After validation of the regression model, the validation data is also used on an existing state-of-art model based on differences in pixel intensity. Subsequently, the performance of the oversegmented machine learning method and the differences in pixel intensity method are compared to benchmark the performance of the oversegmented machine learning method against a current state-of-the-art method.

4.2.4 Subset of data for testing applicability on images of a camera station not included during training

The subset for testing the applicability of the oversegmented machine learning method on images from a camera station that was not included during training, consists of 57 randomly sampled images from the webcam station. The images originate from the period between 21 July and 5 September in the summer of 2017.

4.3 Image Enhancement

The existing workflow does not include enhancement, whereas the literature review showed that other methods adopt procedures to enhance the images before analysing them in detail. This indicates possibilities to increase the performance of the oversegmented classification model by applying enhancement. This section elaborates on the benefits that potentially can be gained by adding image enhancement to the existing workflow (subsection 4.3.1). Moreover, several image enhancement techniques are reviewed (subsection 4.3.2) and the approach used to test the effect of the most promising technique is explained (subsection 4.3.3) .

4.3.1 Potential benefits of enhancement

Potentially, enhancement can have impact in two phases of the Flamingo workflow: during oversegmentation and during feature extraction. Both will be elaborated separately in the remainder of this subsection:

Potential positive effect on oversegmentation

The oversegmentation procedure is the main driver behind the possibility to increase the number of available features (subsection 2.1.4) and is therefore an important step in the workflow as it largely determines the information that will be learned to the model. However, the adopted oversegmentation algorithm uses a limited number of (colour-) characteristics to indicate gradients in pixel intensity (Achanta et al., 2012) and this has been identified as a weak link in the process.

Visual inspection of the training dataset showed that gradients between the classes of interest (e.g. beach users and the beach), might be vague as a result of poor image quality caused by for instance over-/underexposure (e.g. Fig. 4.7). For those type of images it is questionable whether the oversegmentation algorithm creates superpixels that contain information corresponding to a single class. If the algorithm would indeed face problems in these kind of conditions, the algorithm creates superpixels with ambiguous class information. Subsequently, the machine-learning algorithm will start to learn features of these ambiguous superpixels, which makes it difficult to distinguish between these classes during classification as the model learned features that in reality describe multiple classes.

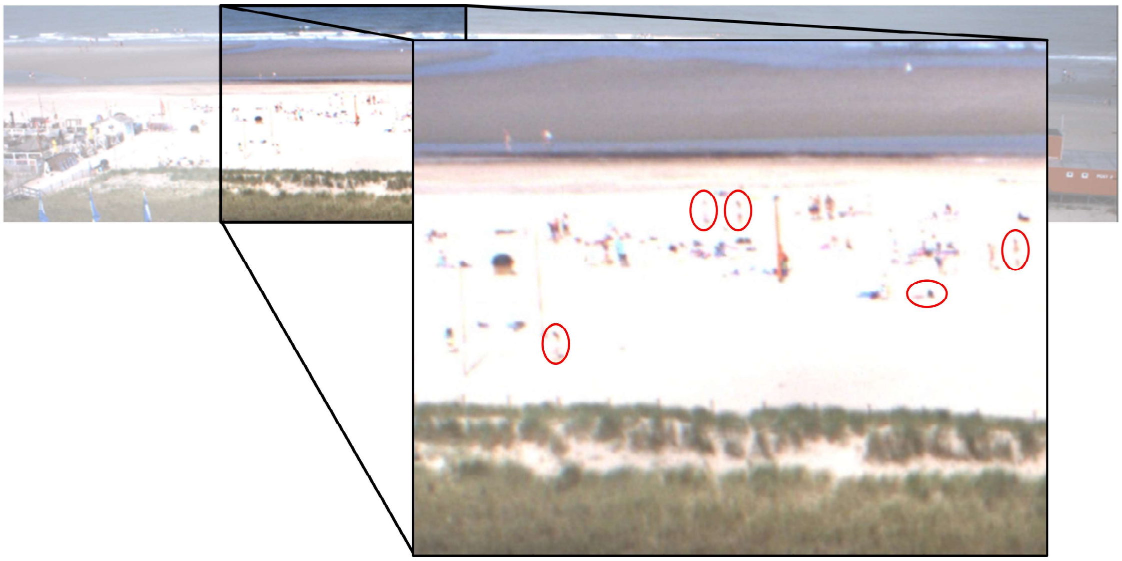


Figure 4.7: Example of image with vague gradients as a result of overexposure. The red circles indicate beach users that fade in the background caused by overexposure. The result is an unclear gradient between beach user and beach. This can complicate the superpixel formation, because superpixels are formed based on pixel gradients.

An enhancement step that increases the gradients between classes of interest, would make it easier for the oversegmentation-algorithm to draw superpixels that contain unambiguous class-information. Subsequently, superpixels containing unambiguous class-information will provide the learning algorithm with more distinctive information for classification, as feature values learned from unambiguous superpixels, correspond to a single class. Theoretically, this should lead to a better classification, because the distinction in important features per class between the classes is more clear.

Potential positive effect on extracted feature strengths

Since image enhancement is especially focused on improving the visual appearance of an image by changing its colour intensities, colour features (strengths/unary potentials) are likely to change. Enhancement could for instance result in a larger difference between the colour channels that represent the classes beach user and sand. This will be reflected in the feature strengths per class of features like mean or maximum intensity corresponding to these colour channels.

Note that this type of effect is rather similar to the effect that the original, pixel intensity based, studies tried to achieve with applying enhancement. In those studies the added value of enhancement is a (potentially) more clear distinction between the beach user and sand peak in the intensity spectrum, simplifying the subsequent thresholding.

4.3.2 Image enhancement techniques

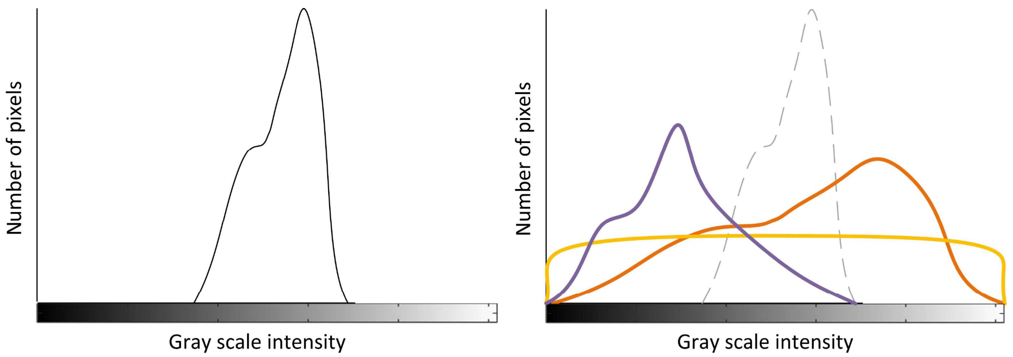
Based on the above described theoretical opportunities for enhancement in combination with the over-segmented machine learning method, a valuable enhancement technique has to increase the difference in intensity between beach user pixels and pixels from other classes. In general the cause of vague gradients can be explained by low contrast, which is characterised by an intensity spectrum that spans a limited range of the full spectrum. This indicates that a limited number of the 256-possible (colour-) intensities is available to represent the pixels. Due to the limited range, the difference in intensity between the classes of interest is limited, causing unclear gradients between the pixels.

The results are very bright or dark images in which it becomes complicated to distinguish between different classes in the image. The range is that small that classes start to interfere (e.g. beach users that have the same colour as the surrounding beach). Increasing the range of used colours/intensities can result in classes/objects that become more visible, moreover resulting in for instance a larger gradient between them. Four techniques capable of increasing the available range in the intensity spectrum are reviewed in the remainder of this section.

Image enhancement can be defined as a process in which the original pixel intensity is changed to a new value using a transformation (Hendriks, 2014b):

$$i_{new} = f(i_{original}) \tag{4.1}$$

Contrast stretching uses a linear transformation to normalise the narrow range of the original spectrum to the full range of 256 intensities for a 8-bit digital image (orange line Fig. 4.8b). Histogram equalisation redistributes the pixels equally over all the 256 intensities via a non-linear transformation, resulting in a (nearly) uniform spectrum (yellow line Fig. 4.8b). Contrast Limited Adaptive Histogram Equalisation (CLAHE) does the same as regular histogram equalisation, but instead of transforming the spectrum of the whole image, it reviews multiple local spectra. Gamma adjustment is a non-linear transformation that, dependent on the value of gamma, gives more weight to the darker ($\gamma > 1$) or lighter ($\gamma < 1$) part of the spectrum. Because beach users correspond to the darker part of the spectrum a gamma larger than one is used (violet line Fig. 4.8b). A detailed description of these techniques can be found in Appendix A.

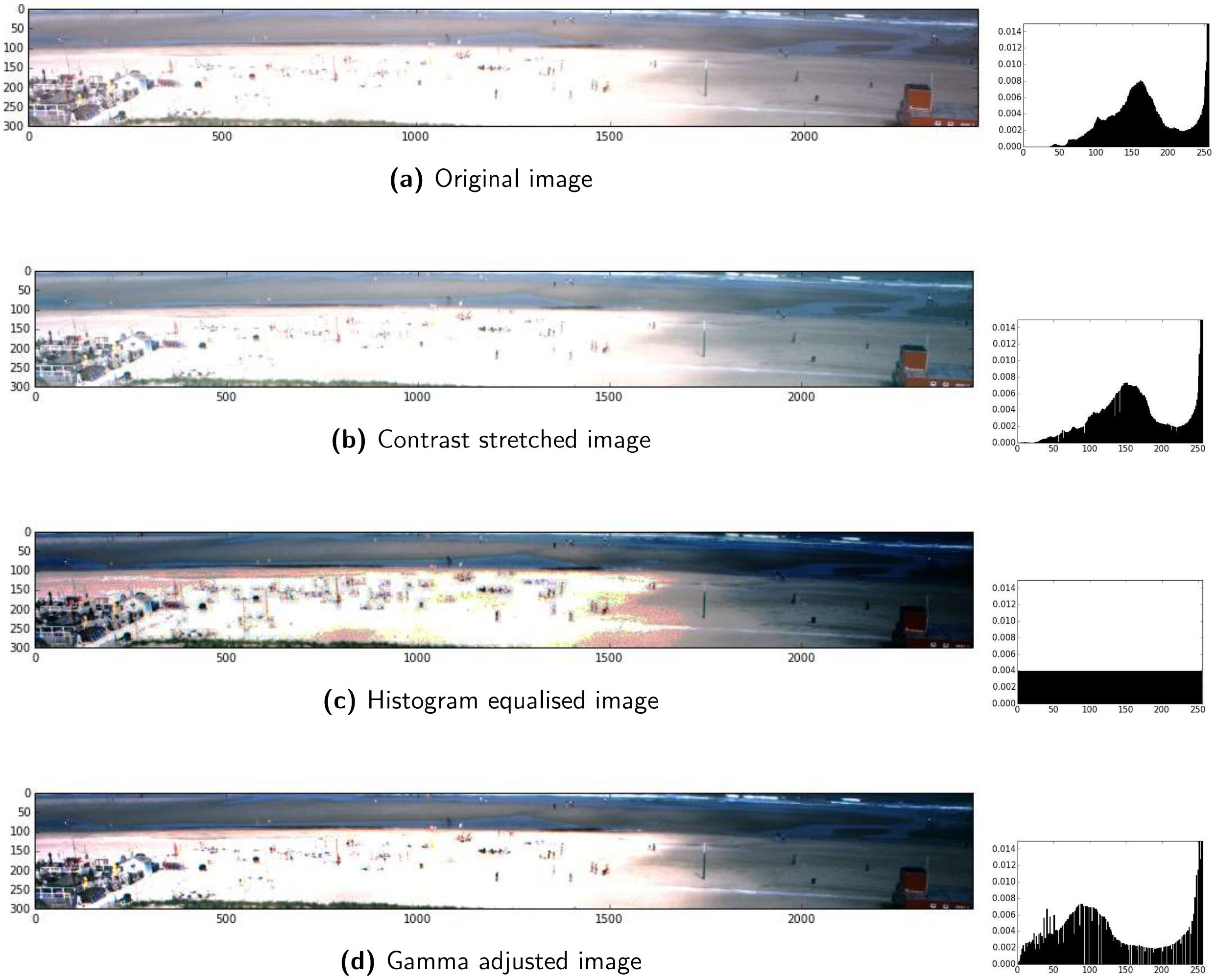


(a) Original intensity spectrum

(b) Enhanced spectra

Figure 4.8: Illustration effect enhancement on intensity spectrum. The orange line indicates contrast stretching, the yellow line histogram equalisation and the violet line corresponds to gamma adjustment with $\gamma > 1$. CLAHE does the same as histogram equalisation, but in local regions instead of globally on the spectrum of the whole image.

Enhancement can be performed on the full image or on a local region corresponding to the ROI. The latter is preferred as intensities of pixels outside the ROI can bias the intensity spectrum, limiting the effect of the enhancement transformations. Therefore, the effect of the enhancement techniques is tested on a restricted area of the original image (Fig. 4.9).

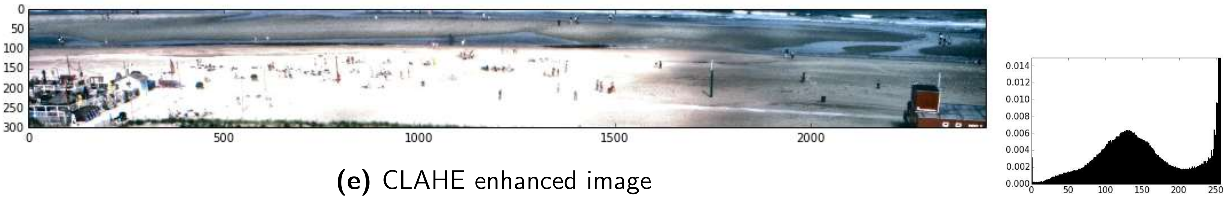


(a) Original image

(b) Contrast stretched image

(c) Histogram equalised image

(d) Gamma adjusted image



(e) CLAHE enhanced image

Figure 4.9: Effect of different enhancement techniques and the corresponding (normalised) intensity histograms. Enhancement was performed on a restricted area of the original image that approximately corresponds to the defined Region of Interest.

Based on visual inspection, the effect of contrast stretching seems to be limited (Fig. 4.9b) and this can be explained by the intensity spectrum of the original images (Fig. 4.9a). This spectrum already spans a relative large part of the available range and therefore an increase to the full spectrum via contrast stretching has a limited effect. Histogram equalisation and especially gamma adjustment perform well in the light area of the image (approximately 500-1500, horizontal axis), but decrease the visual appearance of the relative dark area (approximately 2000+, horizontal axis). CLAHE does not have this drawback as a result of the capability to enhance local parts of the image. This is an important characteristic as dark and light areas in an image are likely due to for instance passing clouds. Because of this CLAHE has been selected and tested in more detail.

4.3.3 Approach

The effect of enhancement can be expressed in multiple ways and depends on the strategy that is adopted to incorporate enhancement in the workflow. The most important choice is whether or not a new oversegmented classification model based on annotated, enhanced images is trained. Regarding the expected potential benefits of enhancement, training an enhanced model will most likely have the best results, because the non ambiguous class information is represented throughout the whole chain including training. Therefore, the original dataset has been duplicated and provided with new annotations to train a model on enhanced images. New annotations are necessary because enhancement changes the original colour gradients in the image resulting in a different superpixel grid.

Although a new model theoretically has the most benefits, the approach using enhanced images on an oversegmented classification model trained with not enhanced images can improve the result of classification. Enhancement might change images that originally were difficult to classify, into images that have a larger correspondence with images that are good to classify, resulting in better predictions and this has been tested.

4.4 Oversegmentation

The original toolbox oversegments the images with the SLIC-algorithm, which was found to out-perform other state-of-the-art algorithms (Achanta et al., 2012). SLIC determines area's of similar pixels based on three colour characteristics and the $u, v-$ coordinates of an image. The coordinates are used to limit the area in which the algorithm searches for superpixels to assign a pixel to, whereas the colour characteristics are used to determine gradients.

The variables in SLIC that can be tuned are the *number of superpixels* and the *compactness* of the grid. The number of superpixels is dependent on the relative size of a beach user in the image, because a beach user should preferably be captured by one superpixel. The compactness can be used to adjust the squareness of the pixels in order to obtain superpixels that enclose beach users as good as possible. Based on visual inspection a grid of 15.000 superpixels with a compactness of 10 is selected for the not enhanced images in the training set (Fig. 4.10). For the enhanced images a grid of 15.000 superpixels with a compactness of 20 has been used. Only the above two grids have been applied, because changes in superpixel grid directly induce another (time-consuming) annotation iteration. Therefore, limited insight in the sensitivity to the superpixel grid has been acquired.

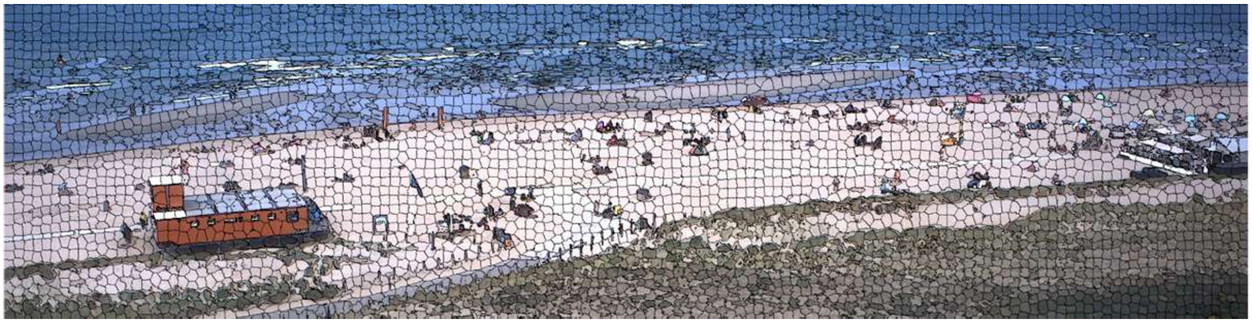
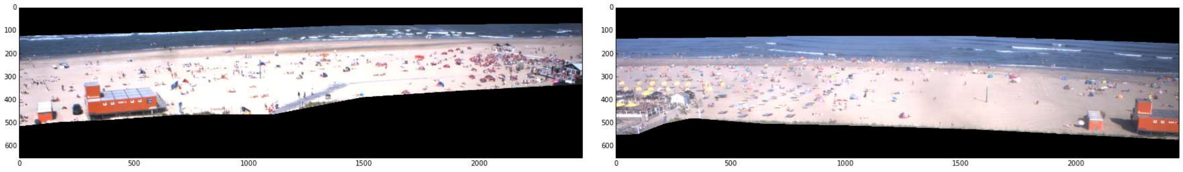


Figure 4.10: Example of oversegmented image from the training dataset

To limit the number of superpixels that had to be annotated, a Region of Interest (ROI) has been defined and included in the grid by merging all the superpixels outside the ROI to one large superpixel. The ROI stretches from the lowest low water obtained in the dataset, to the dune foot and did not exclude the inter-tidal area or fixed objects on the beach. Since use is made of images from camera 3 and 4 of the Kijkduin Argus station, two ROI's have been defined (Fig. 4.11).



(a) ROI camera 3

(b) ROI camera 4

Figure 4.11: Indication of the used Region Of Interest for the cameras 3 and 4 of the Kijkduin Argus station.

4.5 Channel and (limited) feature extraction and normalisation

Channel and feature extraction are standard procedures in the workflow, but additional channels (and therewith features), can be added to optimise the performance. Moreover, not all features have to be valuable and therefore the number of evaluated features might be reduced. The first subsection (4.5.1), explains channel and feature extraction and the possibility to add channels (and features). Subsequently, a second optimisation step to minimise the number of evaluated features is elaborated (4.5.2) and finally the standard and unchanged channel and feature normalisation step is treated (4.5.3).

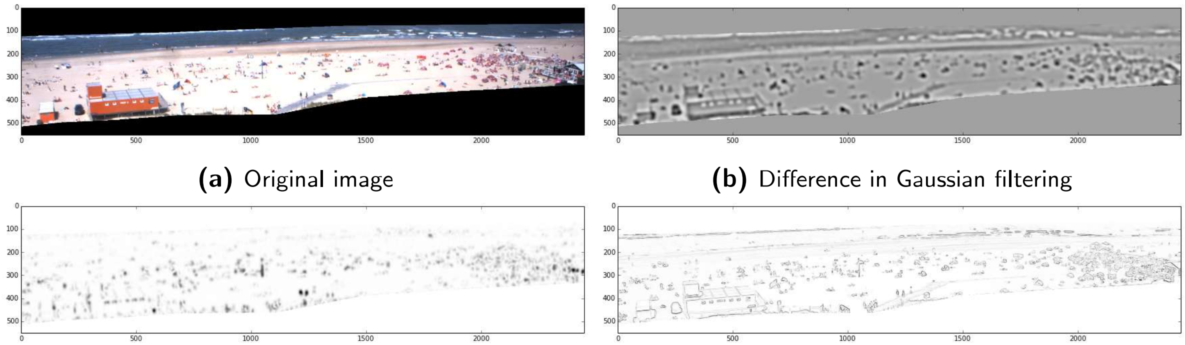
4.5.1 Channel and feature extraction

After oversegmenting an image to a grid with superpixels, the grid is used to extract superpixel information (features) from the image. Recall from section 2.1.1 that a digital image is a 2D representation of the real world, discretized by pixels that contain the intensity.The information provided by the intensity of a pixel depends on the type of digital-image that has been analysed and varies among different types of images. Therefore, different features can be extracted from different type of images, possibly providing the model with more meaningful features for classification.

The images used in this study are of the visible light type and therefore contain information from the visual (r,g,b) channels. By using more or different sensors for acquisition (e.g. infrared or multi-/hyperspectral) additional channels can be obtained. However, these type of sensors were rejected in subsection 2.1.1 based on a limited cost-effectiveness compared to the visible light sensor and therefore are not considered in this study.

Another option to attain additional channels is artificially achieved by applying filters on the original colour images (Hoonhout et al., 2015). This approach retains the benefits of data acquisition with the visible light camera, but simultaneously enables the benefits of additional channels. The drawback of an increased number of features as a result of extracting additional (artificial) channels, is an increase in the required computational time to analyse an image. Therefore, channels should only be added when model performance without additional channels is not satisfactory and channels increase the performance.

In this study the effect of (simultaneously) adding channels based on differences in Gaussian filtering, Gabor filtering and Sobel filtering is tested. Difference in Gaussian filtering (Fig. 4.12b) amplifies regions with large pixel intensity gradients compared to the surrounding (Hoonhout et al., 2015) like for instance beach users on a beach. Gabor filtering (Fig. 4.12c) is suitable for edge detection and texture classification, whereas Sobel filtering (Fig. 4.12d) determines the magnitude of edges (Van der Walt et al., 2014). Potentially these characteristics can enlarge the distinction between the beach user class and other classes and therefore all three have been tested.



(a) Original image

(b) Difference in Gaussian filtering

(c) Gabor filtering

(d) Sobel filtering

Figure 4.12: Examples of artificial channels based on the original R,G,B-image in Fig. 4.12a

4.5.2 Limiting number of extracted features

Feature extraction is identified as one of the most (computational) time consuming steps in the adopted workflow and it can be questioned to what extent all features are valuable to the oversegmented classification model. The relative importance of a feature results from the process that minimises the cost function (section 2.1.4) and is expressed by the *unary-potentials* of features in the learned parameter vector. After training, the unary potentials of a model can be retrieved and sorted in order to identify the potentials with the highest value and thus impact.

Selection of potentials with the highest value is performed for *each class separately*, because individual classes might be characterised by different features. Moreover, the magnitude of important unary potentials for particular classes can differ among the classes. Therefore, not evaluating the most important features per class, can lead to omitting the most important features of specific classes as a result of the potentials being (slightly) lower. Analysis of the effect of adding features is performed by taking predefined percentages of the highest unary potentials per class. Subsequently, a new oversegmented classification model is trained based on the limited set of features. The performance of the new classification model is evaluated and compared to the performance of a classification model including 100% of the features.

Negative potentials have been made positive by taking their modulus to prevent negative features from being excluded in the analysis. This is important because a class having a strong negative potential for a particular feature indicates that an observed superpixel having a high potential, is most likely not of the class with the negative potential and this adds distinctive value between specific classes.

4.5.3 Channel and feature normalisation

The original Flamingo toolbox incorporates a preprocessing step to make the channels and features invariant to the scale of the image and/or superpixel (Hoonhout et al., 2015). This is necessary because features exist that are related to the number of pixels in the superpixel (e.g. area) and the number of pixels in superpixels can vary as a result of difference in image size. Note that this is an important step regarding the objective to create a model that is robust on images from different camera stations, as the resolution of images can differ between camera stations.

Besides normalisation, the toolbox contains a standard step to scale the features to a standard normal distribution related to the range of feature values observed in the training data (Hoonhout et al., 2015). Rescaling is required because particular features might be characterised by relatively large values and the model could develop a preference for these features. Recall from section 4.4 that all pixels outside the region of interest are represented by one superpixel. This superpixel significantly biases the rescaling, because it blows up the observed range of feature values as a result of its size. To exclude this undesired effect, the superpixel containing all pixels outside the ROI has been removed.

The channel and feature normalisation and scaling of features to a standard-normal distribution are standard procedures in the original workflow and are retained in the workflow for the oversegmented classification model.

4.6 Model training

The preceding steps focused on preparing the information to be learned by the model and before the actual training can take place, the type of model used for training has to be selected (subsection 4.6.1). Machine learning models have so-called *hyper-parameters* that can be tweaked to obtain the highest model performance (subsections 4.6.2 to 4.6.4). Lastly the adopted work-flow to investigate the sensitivity of the model to treated hyper-parameters is explained (subsection 4.6.5).

4.6.1 Model definition

The model used in this study is Logistic Regression (LR) which, although the name suggests differently, is a classification model instead of a regression model. Dependent on the chosen combination of classes (*'class aggregation'* section 4.6.2) the problem will be binary (i.e. beach users vs. one single class representing the other classes) or multi-class. In the binary case the target value will be either 0 (negative class) or 1 (positives class), which is denoted as: $y \in \{0, 1\}$. For multi-class problems the target can have multiple values and can be denoted as: $y \in \{0, 1, .., n\}$. The binary case is the most simple situation and is therefore used to explain the fundamental functioning of Logistic Regression (LR). This explanation is largely based on the lectures 6.1 - 6.5 on the topic of Logistic Regression recorded at Stanford University in 2012 (A. Ng, 2012d).

Section 2.1.4 explained that during training a hypothesis $h_\theta(x)$ is adopted and fitted on the data by varying parameters (unary potentials) $\theta_0, \theta_1 ..., \theta_n$ until the result with the lowest cost is obtained. For the binary decision problem the hypothesis is only defined between 0 and 1 as an unclassified data-point can only be of the 0 (negative)- or 1 (positive) class. In order to establish a hypothesis that follows this behaviour, a function should be incorporated that ranges between 0 and 1. To this end Logistic Regression adopts the Logistic (or Sigmoid) function which is defined as:

$$g(z) = \frac{1}{1+e^{-z}} \tag{4.2}$$

This function has the characteristic to vary between the asymptotes 0 and 1 on the y-axis while being centred at the point z = 0 (Fig. 4.13).

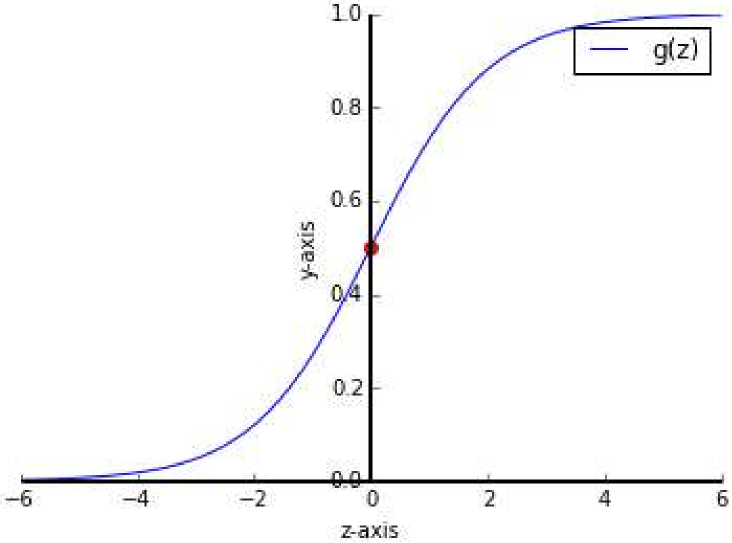


Figure 4.13: Visual representation of Logistic (Sigmoid) function. The red dot indicates the point (0,0.5) often used as threshold between the 0- and 1-class

From the above the final hypothesis used in Logistic Regression can be defined as:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{4.3}$$

Because the hypothesis is also defined in between 0 and 1 a threshold is required that defines which predictions will be classified as the positive or as the negative class. Usually the threshold $h_\theta(x) \geq 0.5$ is used to classify a data point as the positive class and as a result thereof $h_\theta(x) < 0.5$ for classification of the negative class. From Figure 4.13 it can be seen that $h_\theta(x) \geq 0.5$ is true for z-values larger than 0 and since $z = \theta^T x$ from equations 4.2 and 4.3 also $\theta^T x \geq 0$ holds for the positive class. Furthermore $\theta^T x < 0$ is true for the negative class.

To clarify the above, the example from Figure 2.9 is revisited and adapted to the use of Logistic Regression:

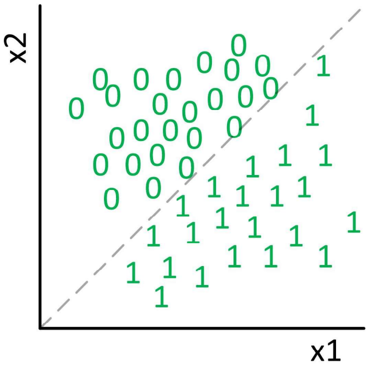


Figure 4.14: Example of classification problem.

In the case of Logistic Regression a working hypothesis would be: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ with learned parameter vector $\theta^T = [0\ 1\ {-1}]$. For a positive prediction the following holds:

$$\theta^T x \geq 0 \tag{4.4}$$

$$0 + x_1 - x_2 \geq 0 \ or \ x_1 \geq x_2 \tag{4.5}$$

Figure 4.14 confirms that 1's will be predicted in situations where x_1 is larger or equal than x_2 and that the learned parameter vector results in a good division of the classes.

As with the example in the literature review, the parameter vector containing the unary-potentials per feature, is obtained by minimising a cost function. Variations in the type of cost function are the main difference between multiple classification models and the use of another cost function (and thus model) can lead to increased model performance. Tests with other classification models have only been performed to a limited extent with a model called Support Vector Machine (SVM). The results of preliminary test with the SVM were very similar compared to the LR and therefore other models are not further explored in this study.

4.6.2 Class aggregation

The classes in the case-studies presented in Hoonhout et al. (2015) are all of interest, but this is not true for the classes in the current study. During annotation, superpixels have been separated into the classes: beach user (persons on the beach), swimmer (persons in the water), beach object (e.g. stretcher, beach shelter etc.), object (e.g. litter bins, beach club, lifeguard post etc.), sand, water, ROI and vegetation. Some of these classes can be related to the class of interest (e.g. beach user and swimmer), whereas other classes have no relation to the class of interest (e.g. water and sand). Both types of classes can be combined with the aim to create larger differences between the classes of interest and the other(s)

Combinations of classes are defined under the parameter *class aggregation*, which is input to the oversegmented classification model at the start of the model training. Variations in the aggregation result in different classification models and the effect of class aggregations on the model performance is tested. Recall from section 4.5 that the superpixel containing all pixels outside the ROI has been removed because of its negative impact on the feature rescaling. Therefore, the class ROI is absent in the tested class aggregations.

Important in combining classes is that the target class (i.e. beach users) is only combined with relevant classes. For beach user quantification this implies that beach users can be merged with swimmers as they both are humans visiting the beach, but beach users should not be combined with for instance sand as these two classes represent something totally different regarding the objective of beach user quantification.

The class beach object is slightly arbitrary as it does not represent humans. However, beach objects will in general be present on the beach when there are beach users and therefore a relation between beach users and beach objects does exist. It is expected that this relation can be determined by the conversion from classified beach users superpixels to beach users (regression model) and therefore aggregating the classes beach users and beach object is taken into account. The classes object, sand, water and vegetation are not of interest for the purpose of this study and can be combined freely.

The above mentioned considerations have been used to combine the classes in all possible configurations (Tab. 4.1). All these aggregations are tested.

Table 4.1: Tested class aggregations. The columns represent the tested aggregations. The rows tell per class to what class(es) it is merged in a particular aggregation.

	Aggregations:																		
	1:	**2:**	**3:**	**4:**	**5:**	**6:**	**7:**	**8:**	**9:**	**10:**	**11:**	**12:**	**13:**	**14:**	**15:**	**16:**	**17:**	**18:**	**19:**
A. Beach user:	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.
B. Swimmer:	B.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.
C. Beach object:	C.	C.	A.	C.	A.	C.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.	A.
D. Object:	D.	D.	D.	D.	D.	D.	D.	D.	D.	D.	D.	D.	E.	G.	D.	D.	D.	D.	D.
E. Water:	E.	E.	E.	E.	E.	E.	E.	E.	E.	D.	E.	E.	E.	E.	D.	D.	E.	E.	D.
F. Sand:	F.	F.	F.	F.	F.	F.	F.	E.	D.	F.	F.	E.	E.	G.	F.	F.	D.	E.	D.
G. Vegetation:	G.	G.	G.	D.	D.	F.	F.	G.	G.	G.	E.	E.	G.	G.	D.	F.	E.	D.	D.

4.6.3 Correction imbalanced data

Analysis of the annotated (training) dataset showed that regardless of the used class-aggregation, the classes are represented by an unequal amount of superpixels (Tab. 4.2). Especially the classes sand and water are significantly larger than the beach user related classes 'beach user', 'beach object' and 'swimmer'. Even if the latter would be merged, they will be roughly 5 to 8 times smaller than the separate water / sand classes.

Table 4.2: Distribution of superpixels over classes in training (both train- and test partitions) data of oversegmented classification model.

	Nr. of superpixels
Beach user:	8237
Beach object:	3878
Swimmer:	683
Object:	28643
Water:	63350
Sand:	99149
Vegetation:	4427

An imbalanced representation of classes in a dataset can result in a situation where it becomes favourable for a model to predict all instances as the majority class. This is because the relative penalty (cost) the model receives for classifying a superpixel with a ground-truth of the minority class as the majority class, is lower than the penalty that is risked by classifying a superpixel with ground-truth of the majority class as one of the minority class. This is undesirable and especially for the problem of beach user quantification where the class of interest is underrepresented.

The problem of learning from imbalanced data is not confined to beach user quantification and multiple techniques have been developed to cope with this phenomenon (He and Garcia, 2009). In the current study a sampling and cost sensitive method are applied to test the effect on the overall model performance. Sampling is a technique that solves the imbalance by adjusting the dataset. In general the sampling approaches can be divided in *undersampling*- and *oversampling* methods. The former removes data points from the majority class in order to get a better correspondence in the size of the minority and the majority class, whereas the latter adds artificial data points to the minority class with the same objective. A variety of techniques to remove/add datapoints is available (e.g. He and Garcia (2009)) and in this study random undersampling is reviewed. Oversampling has not been tested, because of the expected increase in computational time as a result of the added data points.

Cost sensitive methods (cost weight balancing) do not change the dataset, but account for the imbalance by adjusting the cost that the model 'pays' for classifying a particular class wrongly during training. Usually this cost is increased for the minority class and/or lowered for the majority class to ensure that it becomes less favourable for the model to classify minority superpixels as majority superpixels.

Both undersampling and the cost sensitive method can be submitted in the Python function that is used for training of the oversegmented classification model. Default settings exist that automatically determine what the most balanced situation is, but it is also possible to input values specified by the user. Both options are explored to determine the sensitivity of the model with respect to these parameters.

4.6.4 Regularisation

Regularisation can be added to the cost function and might be used to regulate the magnitude of the unary-potentials corresponding to the evaluated features (A. Ng, 2012a). Adjusting the magnitude of the unary-potentials can increase the model performance for models having a high variance or bias (see section 4.2.2). Decreasing all the feature values by a low regularisation parameter limits the influence of very high unary potentials. Potentially this solves the problem of high variance because high variance can be the result of a hypothesis containing a limited number of very important features. A higher regularisation parameter increases the values of the unary-potentials giving more weight to the higher order polynomial features. A high regularisation parameter can solve a high bias as a high bias was characterised by a hypothesis that was too simple to describe the complexity of the data.

4.6.5 Approach for testing parameters corresponding to model training

The order in which the parameters have to be changed to obtain the highest model performance is not known upfront and it is therefore not possible to follow a predefined workflow. This complicates the determination of the preferred class aggregation, as undersampling or class weight balancing might lead to other preferred class aggregations than the aggregation resulting from tests on the original data without measures to solve the imbalance in the data. Therefore, determination of the preferred class aggregation based on the original data only, includes the risk of missing aggregations that have a higher performance after undersampling or class weight balancing.

To reduce the described risk an approach has been adopted that also tests the aggregations for training instances with undersampling and class weight balancing. To this end the default settings *'auto'* (undersampling) and *'balanced'*(cost weight balancing) are used, which automatically determine the values that have be used to account for the imbalance in the original dataset. Theoretically a similar procedure should have been incorporated for the regularisation parameter, but tests in preceding model iterations showed very limited effect as a result of changes in this parameter and therefore regularisation is only tested on the final aggregation.

After the final aggregation has been determined, the effect of undersampling and class weight balancing is investigated in detail for this aggregation to obtain the final sensitivity of the oversegmented classification model to these measures.

4.7 Model scoring (oversegmented machine learning model)

A scoring method is required to express and compare the performance of oversegmented classification models, trained with different settings. In general a good classification model is characterised by a high number of true positives and true negatives, whereas the number of false positives and false negatives should be low. A common way in machine learning to express the distribution of true positives, true negatives, false negatives and false positives is the confusion matrix (Fig. 4.15).

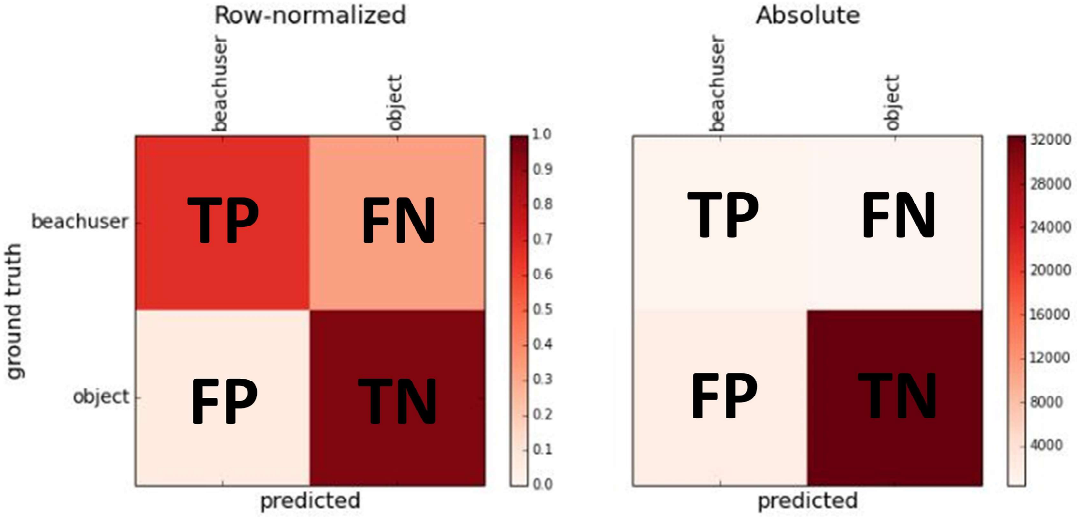


Figure 4.15: Confusion matrix of a two-class aggregation. The colours represent the number of superpixels in a quadrant. The left image represents row-normalised quadrants and indicates a lot of FN's. The right image shows the absolute quadrants indicating a large imbalance in the size of the classes.

Multiple scores can be derived from the confusion matrix by combining the quadrants in different ways. The most common scores are presented below (He and Garcia, 2009):

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} * 100\% \tag{4.6}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.8}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.9}$$

In which:

TP = True positives; correctly predicted super-pixels of the class of interest
TN = True negatives; correctly predicted super-pixels of a class that is not of interest
FN = False negatives; super-pixels of the class of interest, predicted as a class that is not of interest
FP = False positives; super-pixels of the classes that are not of interest, predicted as the class of interest

In general a suitable metric should at least contain information on the TP, FN and FP, because these reflect the performance related to the class of interest. From the listed scores above, only the *accuracy* and *F1* full-fill this requirement. However, the presence of an imbalance in the data, puts restrictions on the applicability of the accuracy. In the beach user dataset, the negative class is much larger than the positive class (Tab. 4.2) and combing this with the definition of the accuracy (Eq. 4.6) shows that the negative class can average out bad performance on the positive class. This can result in decent accuracy scores whereas the performance on the class of interest is actually very limited. Therefore, accuracy provides limited information on the performance of models for beach user classification.

The F1-score does not have this drawback because the underlying precision and recall scores are not normalised by the TN quadrant. However, the F1-score is expressed in a single value omitting the relative contribution of precision and recall to the composition of the final value. The Precision-Recall (PR)-curve (He and Garcia, 2009) is a metric that visualises these contributions by plotting the score of a particular model in a recall (x-axis), precision (y-axis) plane (Fig. 4.16).

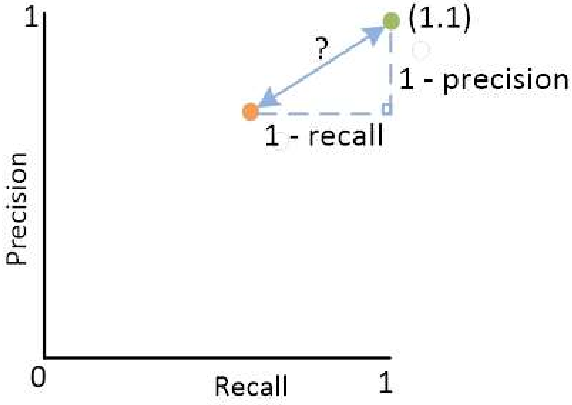


Figure 4.16: Illustration on the precision-recall curve. The green dot indicates the optimal situation in which both the recall and precision equal one. The distance from other points to this optimum can be determined by Pythagoras theorem and can be used to express the score in a single value.

Besides the relative distribution of precision and recall visible in the PR-curve, the PR-score can be quantified by the distance from a model to the optimal score of 1,1. In this situation both the recall an precision equal one, indicating that the model does not predict false positives and negatives and only true positives and true negatives. This distance can be obtained via Pythagoras theorem and the closer this distance is to zero, the better. The PR-score is used for the scoring of classification models in the remainder of this study.

4.8 Conversion classified beach user superpixels to beach users by regression model

The result of the preceding steps is an oversegmented classification model that has the best performance on classifying superpixels given the annotation. The original work of (Hoonhout et al., 2015) focused on classes that directly correspond to the (area) of the classified superpixels, but this does not hold for beach user classification.

During formation of the superpixel grid (section 4.4), superpixels are created that capture individual beach users as much as possible but it is not inconceivable that superpixels contain multiple beach users due to for instance occlusion. Moreover, the possibility to aggregate beach objects to the beach users class, lowers the direct relation between classified beach user superpixels and actual beach users.

The above clarifies the necessity of an additional regression model to convert the classified beach user superpixels from the oversegmented classification model to the final number of beach users. The conversion can be made based on the *number* of classified superpixels or the *area* corresponding to the classified beach user superpixels. Both options are elaborated in subsection 4.8.1 and finally the used approach to test the relations is worked out in subsection 4.8.2.

4.8.1 Conversion relations

To make a conversion from the predicted number/area of beach users to the actual number of beach users, a manually counted ground has been used. To this end the images of the training dataset of the oversegmented classification model 4.2.2 are provided with a ground truth by manually counting the beach users in the images. With the availability of a ground truth, a relation between the predicted number/area of beach user superpixels and the counted number of beach users can be derived by determining the regression line.

Direct relation

The approach using the direct relation, plots the counted ground truth of images in the training dataset against the corresponding predicted *number* of beach users resulting from the oversegmented classification model. Subsequently, a regression line is drawn through the data points in order to derive the required relation. The classified number of beach user superpixels can be inserted in this relation to to obtain the number of beach users.

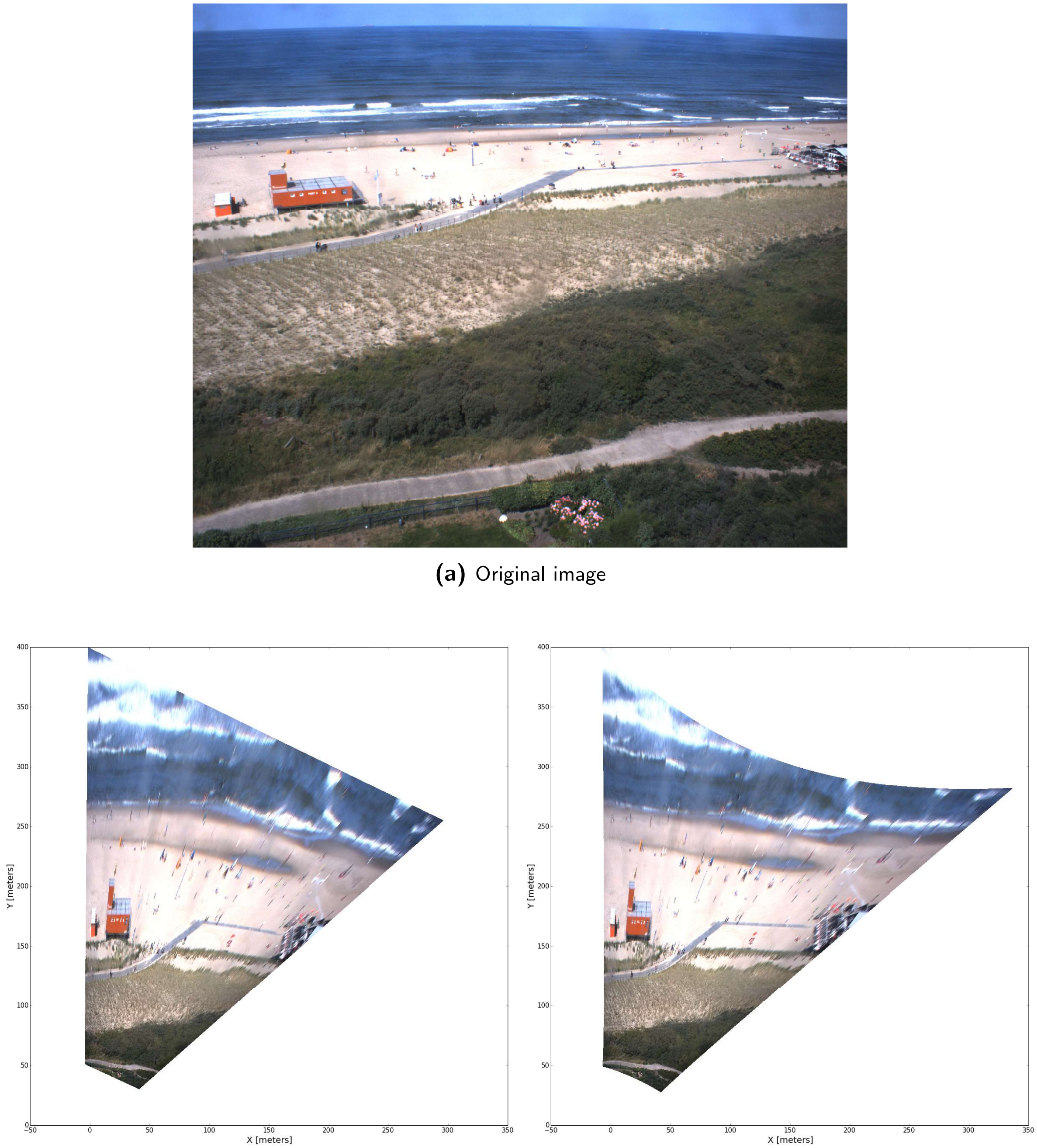
Relation after rectification of superpixels

The second approach translates the superpixels defined in u, v-image coordinates to real world $x, y-$ coordinates before deriving a relation between classified beach user superpixels and the counted number of beach users. The procedure of transferring image coordinates in real world coordinates is called rectification and is common in coastal imagery (e.g. De Vries et al., 2011).

In general beach users relatively close to the camera will have a larger size in the image than beach users further away from the camera and rectification might account for this. However, beach users relatively close to the camera are captured more from above and therefore might block a smaller area. Moreover, sitting/lying beach users are expected to block a smaller area than standing beach users and this can complicate the relation between rectified beach user areas and beach users. The relative impact of the above mentioned potential advantage /drawbacks is expected to determine the suitability of rectification.

Rectification requires information on the intrinsic and extrinsic camera parameters and these parameters are known for the Argus camera station that has been used for the development of the oversegmented machine learning method (Deltares, 2013). The intrinsic and extrinsic camera parameters are combined in a homography that is used to project $u, v-$ pixel coordinates into real world $x, y-$ coordinates. This procedure is part of the original Flamingo toolbox (Hoonhout and Radermacher, 2014a) and therefore the toolbox id used for the transformation of pixel coordinates into real-world coordinates.

However, images can be distorted due to distortion of the camera lens (comparison Fig. 4.17b and Fig. 4.17c) and the toolbox does not account for this. Therefore, the images are first undistorted with the help of the code in Hoonhout (2016) and afterwards the pixel coordinates are transformed into real-world coordinates with the toolbox.



(a) Original image

(b) Distorted, rectified image

(c) Undistorted, rectified image

Figure 4.17: Visualisation of effect rectification. The length of the y-axis in the rectified images is limited at 400 meters.

Rectification transforms the rectangular $u, v-$ pixel grid into a staggered $x, y-$ grid. The areas of the staggered $x, y-$ grid cells represent the real-world areas of the corresponding pixels and are obtained with the *XB-stagger* algorithm (Hoonhout, 2011). This procedure results in a matrix with the shape of the original image containing the real world area's of the pixels in the images. The matrix is camera-specific and therefore two matrices are created corresponding to Kijkduin Argus camera 3 and for 4. By summing the pixel area's belonging to superpixels that are classified as beach user, the total area in an image corresponding to beach users is obtained. The area is subsequently plotted against the counted number of beach users for that particular image and a regression line is fitted in order to determine the relation between the predicted area of beach users and the counted number of beach users.

4.8.2 Approach

The performance of the two conversion approaches is evaluated with the R^2-score of the fitted regression lines. This metric is a good representation of the absolute error that is made by a model because it is based on the squared residuals between the actual point (classified number of beach users vs. counted number of beach users) and the number of beach users that follows from inputting the number of classified beach user superpixels into the proposed relation.

After selection of the most suitable conversion approach, the performance of the combined oversegmented classification- and regression model on the training data is quantified. The R^2-score corresponding to the preferred type of relation is the first indication of the performance but the R^2-score is a dimensionless score providing limited insight in the actual number of beach users that is predicted incorrect. Therefore, also the Root-Mean-Squared-Error (RMSE) and bias are evaluated. The RMSE and bias have been determined for predefined bins of images with respectively 0-25, 25-100 and 100+ counted beach users, to place the RMSE/bias in perspective with the counted ground truth. The ranges of the bins have been defined such that each bin contains approximately the same number of images.

4.9 Validation of regression model

The combined oversegmented classification and regression model is tested on a new dataset to validate the derived regression relation. Validation is performed with the in section 4.2.3 introduced validation dataset. The oversegmented classification model is used to classify the images in the validation dataset. Subsequently, the regression relation (i.e. direct/rectified) obtained in the previous section, is fitted on the validation data to obtain the R^2 of the regression model on data that was not used during development of the model (R^2(validation) Fig. 4.18). Moreover, the R^2 corresponding to a new regression model corresponding to the optimal fit on the validation is obtained (R^2(validation *optimal*) Fig. 4.18).

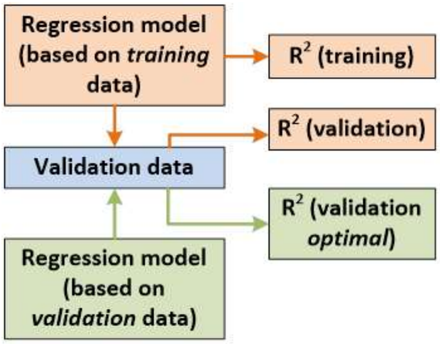

Regression model (based on *training* data)
R² (training)
R² (validation)
Validation data
R² (validation *optimal*)
Regression model (based on *validation* data)

Figure 4.18: Illustration of the evaluated R^2 scores. The colour of the arrow/box of a R^2 score indicates the regression model that was used to obtain the R^2 score.

Comparison of R^2(validation) with R^2(training) validates the performance of the regression model based on the training data on images that have not been used in the development of the regression model. The difference between R^2(validation) and R^2(validation optimal) indicates to what extend the regression model developed with the training data approaches the optimal fit based on the validation data. After evaluation of the R^2-scores the bias and RMSE corresponding to application on the training- and validation data are compared.

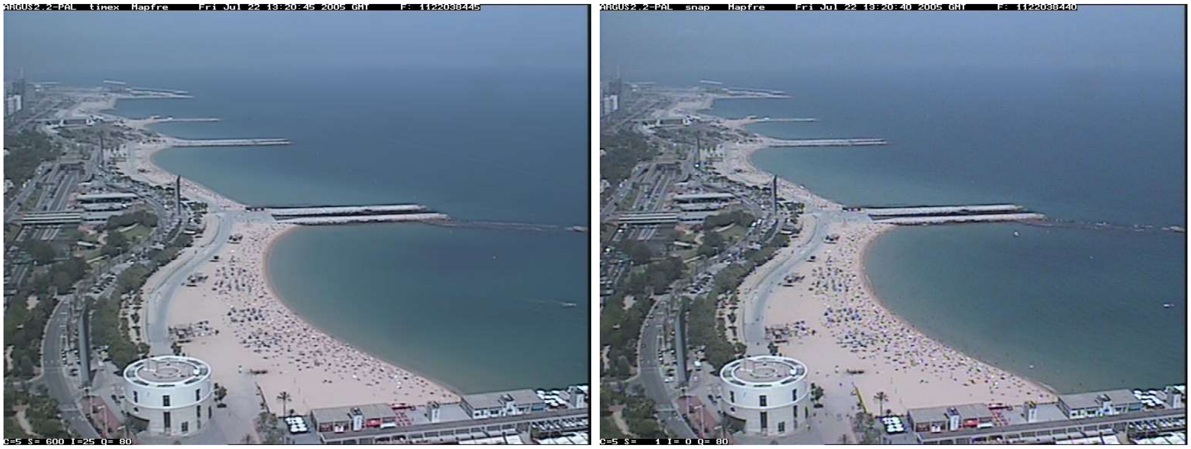
4.10 Benchmarking oversegmented machine learning method against current state-of-the-art

The desire to develop the oversegmented, machine learning method, originates from the limited predictive ability of the method based on differences in pixel intensity (current state-of-the-art) documented by previous studies. To prove the added value of the oversegmented machine learning method, it is benchmarked against one of the methods based on differences in pixel intensity (section 2.1.2). This section starts with the details of the adopted differences in pixel intensity method and subsequently explains the used approach for benchmarking.

4.10.1 Details current state-of-the-art difference in pixel intensity method

The benchmark has been performed against the model as presented in Guillén et al. (2008). This algorithm applies contrast stretching on the red colour-band and subsequently applies gamma adjustment with $\gamma > 1$ giving preference to low-intensity values (section 4.3.2). A constant ROI covering the maximum beach area is extracted and the image is converted to a binary format by applying a tresholding process that segments objects (i.e. beach users) from the beach based on a two-class histogram. After segmentation objects are characterised as white regions whereas sand is black and the number of objects is obtained by automatically counting the white regions. Finally, the number of white regions is translated to a number of beach users by applying linear regression on a comparison of the predicted number of objects and a counted ground truth.

The study of Guillén et al. (2008) evaluated two beaches in Barcelona with an Argus video system acquiring ten minute averaged, so-called, Timex images, whereas the current study is focused on snapshots. Theoretically the averaging causes moving beach users to be absent, resulting in slightly different images. However, moving beach users are included in the ground truth of both the training and validation data and the nature of a timex and snapshot are very similar (Fig. 4.19). Therefore, it is assumed reasonable to test the original method on snapshots instead of timex images.


ARGUS2.2-PAL timex Mapfre Fri Jul 22 13:20:45 2005 GMT F: 1122038445
C=5 S= 600 I=25 Q= 80
ARGUS2.2-PAL snap Mapfre Fri Jul 22 13:20:40 2005 GMT F: 1122038440
C=5 S= 1 I= 0 Q= 80

(a) Timex (b) Snapshot

Figure 4.19: Indication of the difference between Timex and snapshots on the original beach in Barcelona (images obtained from Coastal Morphodynamics (UPC) & Coastal Ocean Observatory (ICM-CSIC) (2005)).

4.10.2 Approach

The differences in pixel intensity model is tested on the same dataset as was used for validation of the regression model (subsection 4.2.3). The oversegmented machine learning method has been applied with a ROI covering more than only the beach area (e.g. Fig 4.11), whereas the differences in pixel intensity method was restricted to the beach area. To exclude errors as a result of the definition of the ROI, a tailored ROI only covering the beach area is defined. The R^2 of the differences in pixel intensity is selected based on a comparison of the R^2-scores.

The R^2 corresponding to the best performing ROI of the differences in pixel intensity method is compared to the R^2 of the oversegmented machine learning method to obtain insight in the performance of the oversegmented machine learning method with respect to the current state-of-the-art. Moreover, the differences in bias and RMSE are evaluated.

4.11 Applicability of the oversegmented machine learning method on a camera station not used during training

Execution of the preceding steps results in a oversegmented machine learning method that has a certain performance on the type of *data used for training*. Evaluation of this performance can provide information on the robustness in different type of conditions (i.e. weather and beach occupation), but does not entail insight in the applicability of the oversegmented machine learning method on data from other stations. The applicability of the oversegmented machine learning method on a camera station not used during training, is tested by re-evaluating the workflow for the webcam dataset (section 4.2.1).

4.11.1 Approach

Based on the workflow for the oversegmented classification model (Fig. 4.1), two different approaches can be distinguished. The first approach incorporates an additional annotation and training phase with images from the new camera station, whereas the second approach directly tries to classify images from the new camera station based on the learned information of the original camera station that was used during training. The former is more likely to return satisfying results, because a new oversegmented classification model is trained. However, annotation is a time consuming task and therefore it is desirable to have a method that is capable to classify images of new camera stations without additional annotation. Therefore, the second approach is tested (Fig. 4.20).

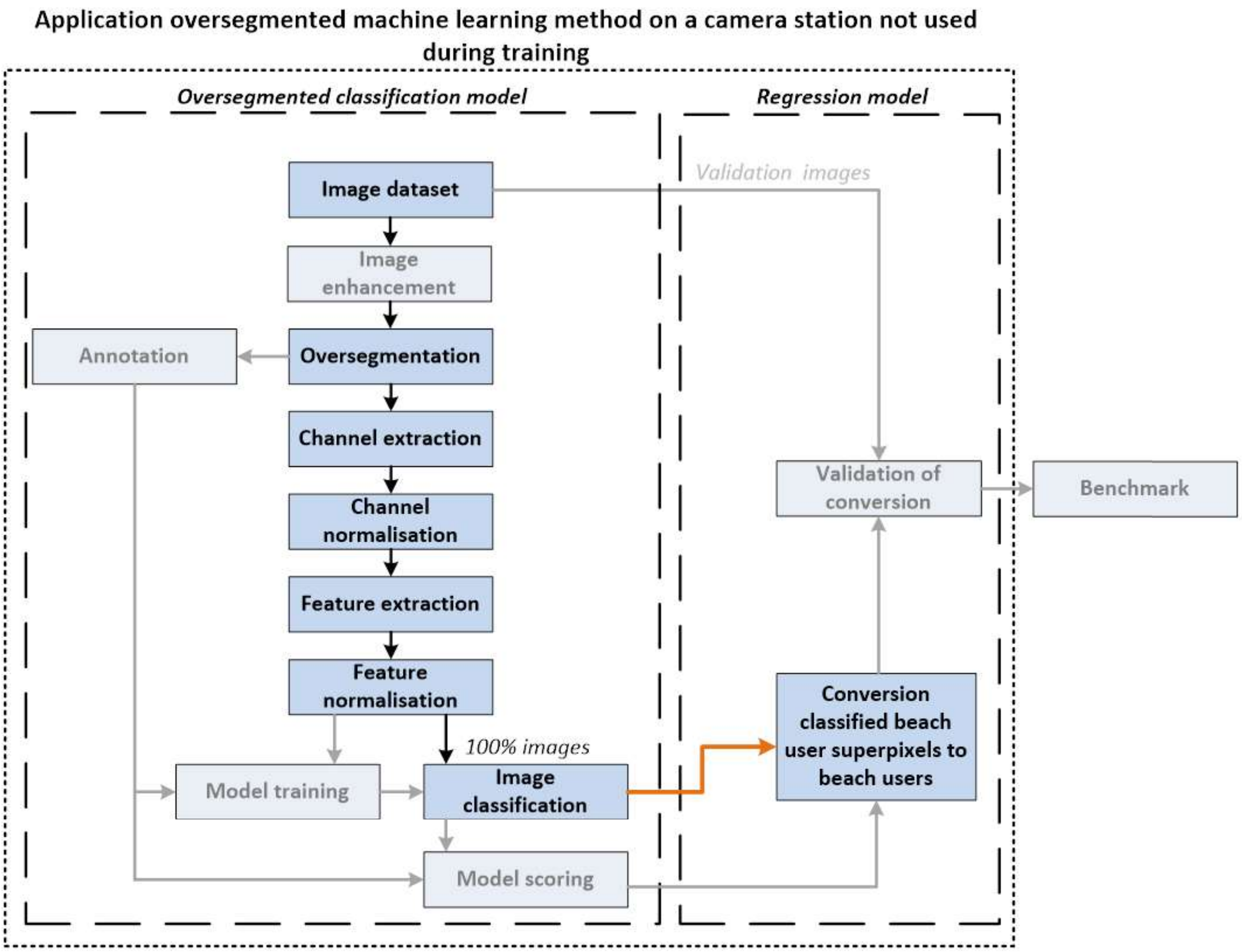


Figure 4.20: Workflow of the approach to test the applicability of the oversegmented machine learning method on a camera station that was not used during training. The dark blue boxes indicate the steps that are revisited.

The workflow shows that annotation and subsequent model training and scoring have not been revisited. Moreover, image enhancement has not been applied on images of the camera station that was not used during training. After oversegmentation and channel- and feature extraction and normalisation all images of the webcam dataset are classified by the oversegmented classification model trained by the Argus images from the training dataset. Subsequently, the classified beach user superpixels are converted to the number of beach users.

4.12 Summary

This section introduced the original workflow of the oversegmented machine learning method, indicated and elaborated the possibilities to adapt/ extend this workflow for quantification of beach user occupation and explained the correspondence between steps in the workflow and the research sub-questions. The following steps/parameters that might be changed/added have been treated:

- **Number of images:** The dataset for training of the oversegmented classification model should be large enough to learn the model generalised information. To test the size of the dataset, the learning curve of the final model will be determined and evaluated.

- **Image enhancement:** Image enhancement can potentially result in superpixels containing less ambiguous and/or stronger feature values that might increase model performance. The effect of enhancement will be tested with a new oversegmented classification model trained with enhanced images, and application of the not enhanced oversegmented classification model trained on enhanced images.
- **Superpixel grid:** Based on visual inspection a superpixel grid has been selected for images captured by the Argus camera stations used for training and validating the oversegmented machine learning method. This grid is retained in subsequent sections implicating that the effect of variations in the grid is not further investigated.
- **Artificial channels:** The oversegmented classification model can be provided with more (potentially useful) features by the addition and extraction of artificial channels. The effect of the combined addition of differences in Gaussian filtering, a Gabor filter and Sobel filter will be tested.
- **Possibility to lower number of features:** The oversegmented classification model evaluates hundreds of features, but it is questionable whether all features are important. The effect of reducing the number of features on the performance will be investigated.
- **Model training:** During training of the oversegmented classification model the following parameters have been changed/added:
  - **Model type:** The type of oversegmented classification model is Logistic Regression. This model type is retained in subsequent sections and therefore the effect of different types of classification models is not further explored.
  - **Class aggregation:** The class of interest might be represented by three of the annotated classes ('beach user', 'beach object', 'swimmer'), whereas the other classes ('object', 'water', 'sand', 'vegetation') are not of interest. Therefore, combinations of classes can be made that possibly result in a higher performance of the oversegmented classification model. This will be tested for all possible aggregations.
  - **Undersampling and cost weight balancing:** The annotated classes are unequally present in the training dataset and the measures undersampling and cost weight balancing are proposed to solve this imbalance and potentially increase the performance of the oversegmented classification model.
  - **Regularisation:** The oversegmented classification model might be over-/underfitted and changes in the regularisation parameters can be used to account for this. The performance of oversegmented classification models with different regularisation values will be tested to obtain insight in the effect on the model performance.
- **Type of conversion:** The conversion from classified beach user superpixels to a number of beach users can be performed based on the the classified *number* of beach user superpixels or the *rectified area* corresponding to classified beach user superpixels. Both approaches will be evaluated.

The steps have been structured by a sequential workflow that indicates the position of steps in the oversegmented machine learning method. However, testing the impact of changed/added steps cannot be performed in the same order because 'baseline' models (both oversegmented classification- and regression model, green box Fig. 4.21) are required before the effects of possible optimisation by the number of images, image enhancement, artificial channels and a lower number of features can be quantified (violet box Fig. 4.21). Therefore, first baseline classification- and regression models are developed based on the default/fixed settings of the steps prior to *model training* (light boxes Fig. 4.21). During development of the baseline models the variables defined under *model training* (subsection 4.6) and *type of conversion* (subsections 4.8 and 4.9) are investigated to obtain a baseline oversegmented classification model with corresponding baseline regression model. Subsequently, the effects of addition/changes of the optimisation steps will be evaluated separately by applying them on the baseline models.

Due to practical reasons the artificial channels were already present during development of the baseline model and Therefore, the effect can only be expressed by results obtained in preceding model iterations. This moreover implies that results on the other parameters are obtained on images with added artificial channels.

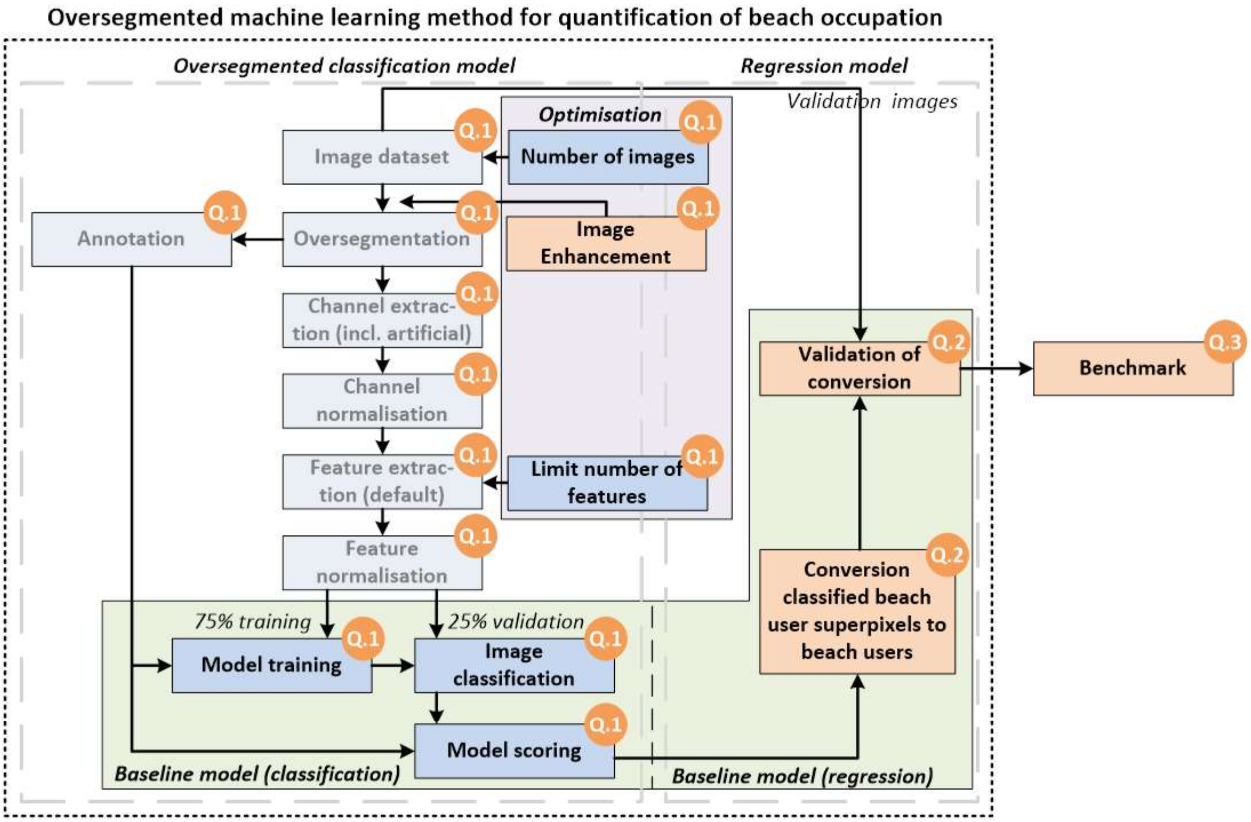


Figure 4.21: Workflow indicating the order in which changes/additions of steps/parameters are reviewed. The light boxes indicate steps with default/fixed settings. The darker boxes correspond to steps that will be investigated.

Validation of the regression model and the benchmark of the oversegmented machine learning method against a state-of-the-art model are performed after the conversion from classified beach user superpixels to the number of beach users. Lastly the applicability of the oversegmented machine learning method on a camera station that was not used during training is tested.

The introduction to the next chapter provides a detailed description of the sections in which the parameters will be treated and the correspondence of sections to the research sub-questions.

5 Results

This chapter presents the results to the research sub-questions defined in chapter 3. In the summary of the previous chapter a distinction has been made between parameters that are elaborated during development of the baseline classification and regression models and parameters that might result in optimisation of these baseline models. The parameters that are adjusted during development of the baseline models are: *class aggregation, undersampling, cost weight balancing and regularisation* (baseline classification model) and the *conversion* from a predicted number/area of beach user superpixels to a number of beach users (baseline regression model). The parameters *number of images, image enhancement, artificial channels and a lower number of features* have been identified as possible optimisation steps of the baseline classification model.

The division in parameters elaborated during development of the baseline models and parameters that potentially optimise the baseline models has been retained in the presentation of the results. Therefore, the first section (5.1) starts with the results obtained during development of the baseline, oversegmented classification model followed by a section (5.2) about the development of the baseline regression model. Subsequently, the results on application of the proposed optimisation steps are treated in section 5.3. Next the optimal oversegmented machine learning method is benchmarked against a current state-of-the-art model in section 5.4. Finally the results of applying the oversegmented machine learning method on a camera station that was not used during training are treated in section 5.5.

Sections 5.1 and 5.3 contribute to research sub-question 1. The results of section 5.2 provide an answer to sub-question 2, whereas sections 5.4 and 5.5 correspond to respectively research sub-questions 3 and 4.

5.1 Development baseline, oversegmented classification model

This section focuses on the results corresponding to changes in the parameters *class aggregation*, *undersampling*, *cost weight balancing* and *regularisation*. First the effect of different class aggregations is treated and an optimal aggregation is selected. Subsequently, for this optimal aggregation the results of changes in undersampling, cost weight balancing and regularisation are elaborated separately.

5.1.1 Class aggregation

The aggregations presented in Tab. 4.1 have been used to train different oversegmented classification models to obtain insight in the combination of classes that results in the highest performance. The aggregations have been tested without measures to account for the imbalance in the data and with the default 'auto' and 'balanced' settings for respectively undersampling and cost weight balancing. The aggregations have been combined with default undersampling and cost weight balancing to test whether the distribution of well performing aggregations is constant.

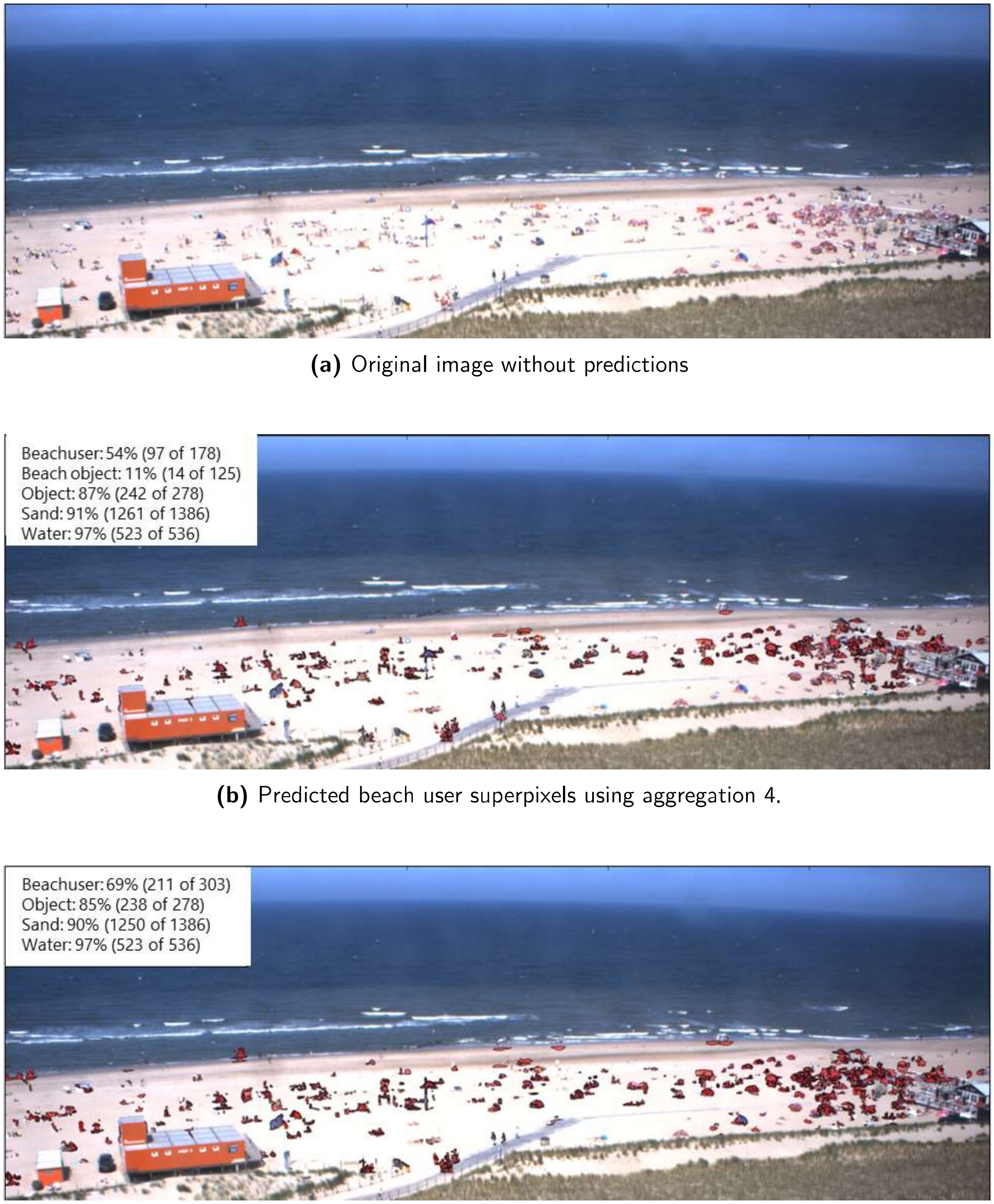
Comparison of aggregation 1 and 2 shows that the swimmer class is the only difference between both aggregations (Tab. 4.1). In the first aggregation this class is present as a separate class, whereas it is merged to the beach user class in the second aggregation. The performance for models with no measures or coast weight balancing is similar, but remarkable is the limited performance of the first aggregation in combination with auto-undersampling (Tab. 5.1). A possible explanation of this result is the size of the swimmer class (Tab. 4.2). Auto-undersampling reduces the majority classes to the size of the minority class and in case swimmers are represented by an own class, this swimmer class is the minority class. Because the swimmer class is very small, the majority classes are significantly reduced which limits the data for training and therefore potentially the performance of the classification model. Due to the observed decrease in performance of a separate swimmer class in combination with undersampling, the swimmer class is aggregated to the beach user class in this study.

A comparison of class aggregations 2-7 with corresponding scores, indicates that the oversegmented classification model has a higher model performance for aggregations that merge the classes *beach users*, *swimmer* and *beach object* (Tab. 5.1). The only variation between aggregations 2 & 3, 4 & 5 and 6 & 7 is the aggregation of the class beach object. For aggregations 2, 4 and 6 the class beach object is aggregated as a separate class, whereas it is merged to the class beach user in aggregations 3, 5, 7. The results show that aggregating the class beach object with beach user, outperforms aggregations that maintain a separation between the classes.

Table 5.1: PR-scores per aggregation for the situation without undersampling/adapted cost weight, the situation with undersampling = *'auto'* and the situation cost weight = *'balanced'*. Lower PR-scores correspond to better model performances and the bold aggregations indicate the best performing aggregations.

	PR-scores:		
	No measures:	Undersampling:	Cost weight:
Aggregation 1:	0.63	0.81	0.68
Aggregation 2:	0.64	0.69	0.64
Aggregation 3:	**0.48**	**0.54**	**0.49**
Aggregation 4:	0.63	0.70	0.64
Aggregation 5:	**0.48**	**0.54**	**0.49**
Aggregation 6:	0.64	0.68	0.64
Aggregation 7:	**0.48**	**0.53**	**0.49**
Aggregation 8:	0.52	0.57	0.50
Aggregation 9:	0.53	0.58	0.50
Aggregation 10:	0.52	0.57	0.51
Aggregation 11:	**0.49**	**0.54**	**0.49**
Aggregation 12:	0.52	0.56	0.51
Aggregation 13:	0.54	0.63	0.51
Aggregation 14:	0.52	0.57	0.48
Aggregation 15:	0.52	0.57	0.50
Aggregation 16:	0.52	0.57	0.51
Aggregation 17:	0.53	0.57	0.50
Aggregation 18:	0.51	0.56	0.50
Aggregation 19:	0.52	0.61	0.60

The impact of the observed difference in PR-score is visualised by printing the predicted beach user superpixels for aggregations 4 (beach user and swimmer) and 5 (beach user, swimmer *and* beach object) on an example image (Fig. 5.1). Visually the differences are minor but the percentages of correctly predicted superpixels show that aggregation 5 indeed predicts a higher percentage of beach users correctly. However, aggregation 5 added the beach object class to the beach user class of aggregation 4 and therefore the higher performance can also result from an increased performance on predicting beach objects. Separation of the beach user class of aggregation 5 in superpixels annotated as beach user/swimmer and beach objects, showed that the percentage of annotated beach user/swimmer superpixels predicted as beach user increased from 54% (aggregation 4) to 67% (aggregation 5). This is considered a significant increase in performance of the oversegmented classification model.



(a) Original image without predictions

(b) Predicted beach user superpixels using aggregation 4.

(c) Predicted beach user superpixels using aggregation 5

Figure 5.1: Example of predictions made by aggregation 4 and the better performing aggregation 5. The red area's correspond to superpixels predicted as beach user. The box in the top left indicates per class the percentage of the annotated superpixels that is predicted correct (indication true positives).

The similarity in performance between aggregations 3, 5, 7 and 11 indicates that the *vegetation* class is of minor importance because the only difference between these aggregations is the class to which vegetation is aggregated. In aggregation 3 vegetation has its own class, the aggregations 5 and 7 combine vegetation with the classes *object* and *sand* respectively, whereas in aggregation 11 vegetation is merged with the class *water*.

The difference between aggregations 3, 5, 7 and 11 and the aggregations 8 - 19 is the combination of the classes *object*, *water* and *sand*. In aggregations 3, 5, 7 and 11 these classes are separated, whereas combinations of the classes are made in aggregations 8 - 19. The results show that aggregations 3, 5, 7 and 11 with separated object, water and sand classes have a better performance but the differences are limited. The implications of these minor changes are revisited in the discussion (section 6.1.1).

Lastly, the distribution of well- and bad performing aggregations is relatively constant among the three different training instances. This supports the expectation that the optimal result can be achieved by proceeding with the one of the, at this stage, well performing aggregations 3, 5 ,7 or 11. The differences between these aggregations are negligible and due to practical reasons aggregation 5 is selected.

5.1.2 Undersampling

This subsection discovers the effect of undersampling aggregation 5 in more detail. Undersampling removes data points of the majority classes to obtain classes closer to the size of the minority class. The first row in Tab. 5.2 shows that without undersampling the classes water, sand and object are respectively 5, 7.7 and 2.5 times larger than the beach user class. The default setting *'auto'* reduces the classes water, sand and object to the size of the beach user class to completely solve the imbalance. It is also possible to reduce the majority classes with user specified ratio's and the effect of adjustments herein have been reviewed (Tab. 5.2).

Table 5.2: Results of undersampling with different values and the same aggregation (number 5). The sampling values correspond to the classes (f.l.t.r.): water, sand, object and beach user. The bold fond indicates the sampling settings with the optimal PR-score

	Sampling:	**PR-score:**
None:	5, 7.7, 2.5, 1	0.48
A	*auto*	0.54
B	1 2 1 1	0.50
C	1 2 1 2	0.56
D	1 3 1 1	0.48
E	1 4 1 1	0.47
F	1 5 1 1	0.47
G	1 5 2 1	0.47
H	1 6 1 1	0.47
I	1 6 2 1	0.47
J	1 6 2 2	0.49
K	**2 6 1 1**	**0.46**
L	3 6 1 1	0.47
M	1 7 1 1	0.47

The results show that variation in sampling can improve the PR-score, but the changes are minor. In general the model performance increases for a higher contribution of the sand class until a value of 7 (comparison of F,H and M). An increased contribution of the beach user class does not result in higher PR-scores (comparison B & C and I & J) and the same can be concluded for the object class (comparison F & G and H & I). A slightly larger contribution of the water class until a value of 3 does increase the model performance (comparison H, K & L). The optimal undersampling values based on the PR-score are: water: 2, sand: 6, object:1 and beach user: 1.

It is remarkable that auto-undersampling results in an oversegmented classification model with a lower performance than the model without measures, because auto-undersampling theoretically solves the imbalance between the classes. Moreover, sampling combinations that retain an imbalance (models B-M) score better and this is unexpected. A possible explanation for this observation is the removal of a substantial part of the dataset corresponding to the majority classes. The model without measures does not remove any data of the majority classes, whereas the sampling combinations B-M remove lesser data and this can cause the higher performance of these models. To this end, random oversampling might be advantageous because this type of sampling remains the majority class and achieves a balance by enlarging the minority class. However, the drawback of oversampling is an increase in the computational time due to the larger dataset.

Inspection of the precision and recall metrics corresponding to the combined PR-scores in Tab. 5.2 shows that variation in undersampling can cause relatively large changes in the distribution of precision and recall (Fig. 5.2). It can be seen that model K (lowest PR, Tab. 5.2), is a relatively balanced model regarding precision and recall. The model without measures ('none', Fig. 5.2) has the highest observed precision, whereas models A (auto-undersampling) and C have a relatively high recall. The importance of these differences is revisited in section 5.2.

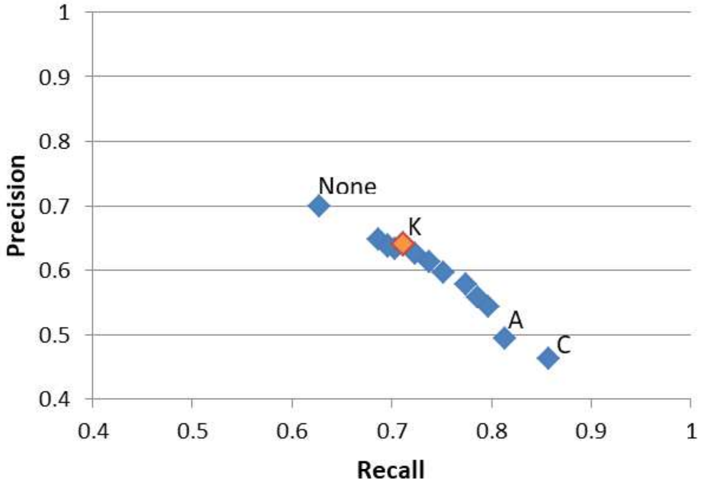


Figure 5.2: Distribution of precision and recall corresponding to the sampling combinations in Tab. 5.2.

5.1.3 Cost weight balancing

Oversegmented classification models with different cost weights have been trained to obtain insight in the impact of adjustments in the weight of the different classes in the cost function. The results show that cost weight balancing has a constant or minimal negative effect based on the PR-score (Tab. 5.3). The first row corresponds to a situation without cost weight adjustments and all errors are equally penalised. The third row is a manual reproduction of the *'balanced'* default setting (second row) to verify understanding of the cost weights corresponding to the classes.

Mistakes in the beach user class have a relatively strong penalty to decrease the situation in which it becomes favourable for the model to classify all superpixels as the majority (sand or water) class. Subsequent iterations (IV, V, VI, VII) have an increased penalty on mis-classifying the sand class, because this class has a relatively large amount of false positives in the balanced situation. A large amount of false positives may indicate that it becomes relatively too favourable to classify superpixels as beach users and increasing the penalty of the class representing this false positives might solves this.

However, the results show limited effect of this measure.

Table 5.3: Results of cost weight balancing with different values and the same aggregation (number 4). The cost weight correspond to the classes (f.l.t.r.): water, sand, object and beach user.

	Cost weight:	PR-scores:
I (none)	0.25 0.25 0.25 0.25	0.48
II	*balanced*	0.49
III	0.121 0.072 0.271 0.535	0.49
IV	0.095 0.150 0.245 0.510	0.48
V	0.078 0.200 0.229 0.493	0.48
VI	0.062 0.250 0.212 0.476	0.48
VII	0.045 0.300 0.195 0.460	0.48
VIII	0.065 0.200 0.215 0.520	0.48

Comparison of the distributions of precision and recall corresponding to the combinations in Tab. 5.3 shows that the tested cost weight combinations lower the precision and increase the recall relative to a oversegmented classification model without measures (Fig. 5.3). In contrast to undersampling, the variation between cost weighted models is limited.

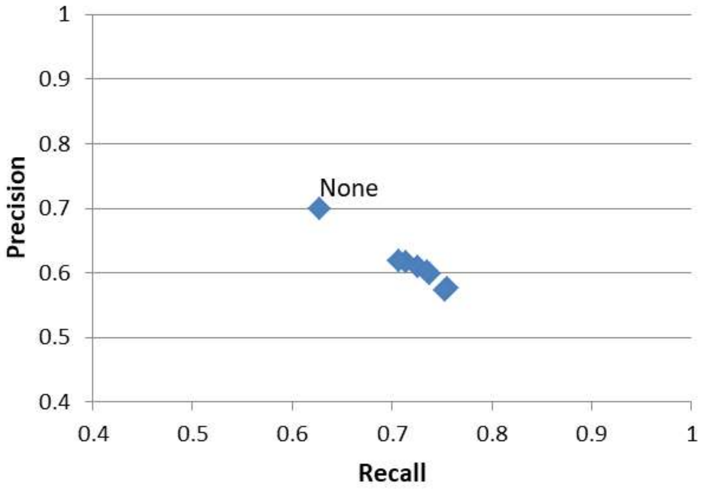


Figure 5.3: Distribution of precision and recall corresponding to the cost weight combinations in Tab. 5.3.

5.1.4 Regularisation

The regularisation of the oversegmented classification model with aggregation 5 has been varied to exclude the possibility of the model being over-/ underfitted. Analysis of the different PR-scores obtained by adjusting the regularisation parameter C shows that the moderate regularisation corresponding to the default setting C = 1, results in the optimal PR-score (Tab. 5.4). Increased (C<1) respectively decreased (C>1) regularisation parameters have no effect, indicating that the model is not sensitive to changes in the regularisation parameter.

Table 5.4: PR-scores for different value of the regularisation parameter C. The default is C=1 and this corresponds to the regularisation of aggregation 5.

Regularisation:	PR-score:
0.1	0.48
0.5	0.48
1 (default)	0.48
5	0.48
10	0.48

The precision-recall curve describing the distribution of precision and recall, shows that the oversegmented classification models with varying regularisation are very similar (Fig. 5.4). This supports the above result that changes in the regularisation parameter have no effect on the oversegmented classification model.

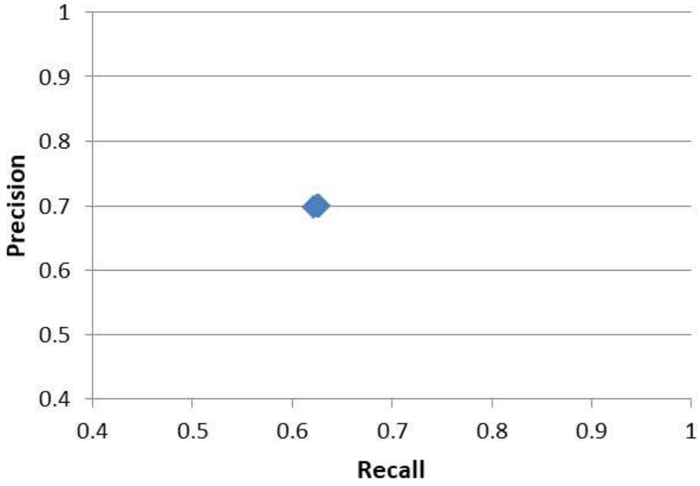


Figure 5.4: Distribution of precision and recall corresponding to the regularisation parameters in Tab. 5.4.

5.1.5 Conclusions oversegmented classification model (baseline)

In this section the results of changes in the parameters: *class aggregation*, *undersampling*, *cost weight balancing* and *regularisation* are evaluated and this contributes to the first research question. The following can be concluded from the results:

1. Changing the class aggregation can significantly increase the model performance.

2. The optimal aggregation for beach user quantification merges the classes 'beach users', 'beach object' and 'swimmer' into one class. Higher PR-scores are obtained for aggregations with separate 'object', 'sand' and 'water' classes but the differences are small compared to aggregations that merge these classes. The class 'vegetation' can be merged to one of these classes without major differences.

3. The distribution of best performing aggregations is constant over classification models trained without measures to account for the imbalance in the data, models adopting undersampling

and models using cost weight balancing. This implies that during determination of the optimal aggregation it is sufficient to consider one of the three.

4. Dependent on the sampling values, undersampling can be used to train oversegmented classification models with a lower PR-score than the model without measures to account for the imbalance. However, the differences are minor.

5. Application of cost weight balancing does not result in oversegmented classification models with a lower PR-score than the model without measures to account for the imbalance.

6. The oversegmented classification model is not sensitive to changes in the regularisation parameter.

Inspection of the precision and recall curves corresponding to the evaluated parameters, showed that especially undersampling and cost weight balancing can cause relatively large changes in the distribution of precision and recall. To obtain more insight in the importance of changes in the relative distribution of precision and recall with respect to the objective of beach user quantification, multiple regression models corresponding to different oversegmented classification models are determined in the next section.

The regression models corresponding to oversegmented classification models A, C, K, II and 5 are compared in the next section. These models have different precision and recall characteristics and are all based on aggregation 5 (Tab. 5.1). Model A, C and K are undersampled, whereas model II is cost weight balanced by the default setting *'balanced'*. Model C has the highest recall, whereas model K has the lowest PR-score. Model A is evaluated because this is the auto-undersampled model that theoretically should give good results. Model 5 corresponds to the original aggregation 5 without measures and is treated to obtain insight in the added value of the measures. Model 5, moreover, has the highest precision.

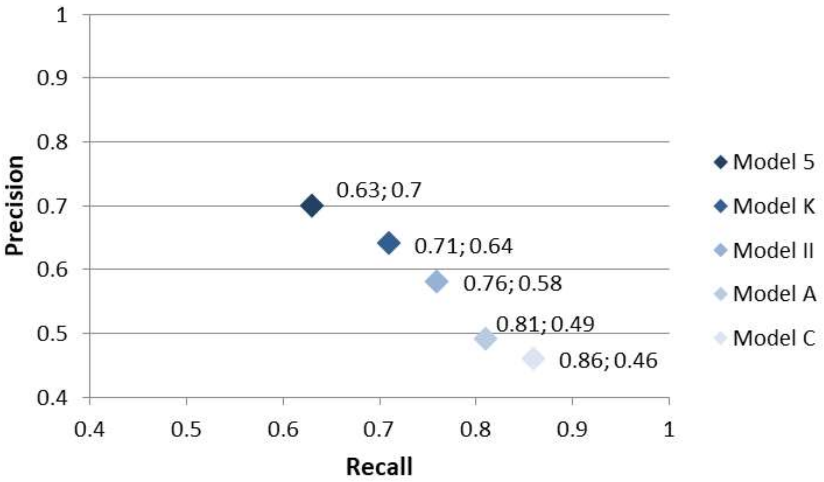


Figure 5.5: Distribution of precision and recall for a selection of models. Model 5 corresponds to aggregation 5 in Table 5.1, models A, C and K correspond to models from Table 5.1.2 and model II corresponds to the model from Table 5.3.

5.2 Development and validation of baseline, regression model

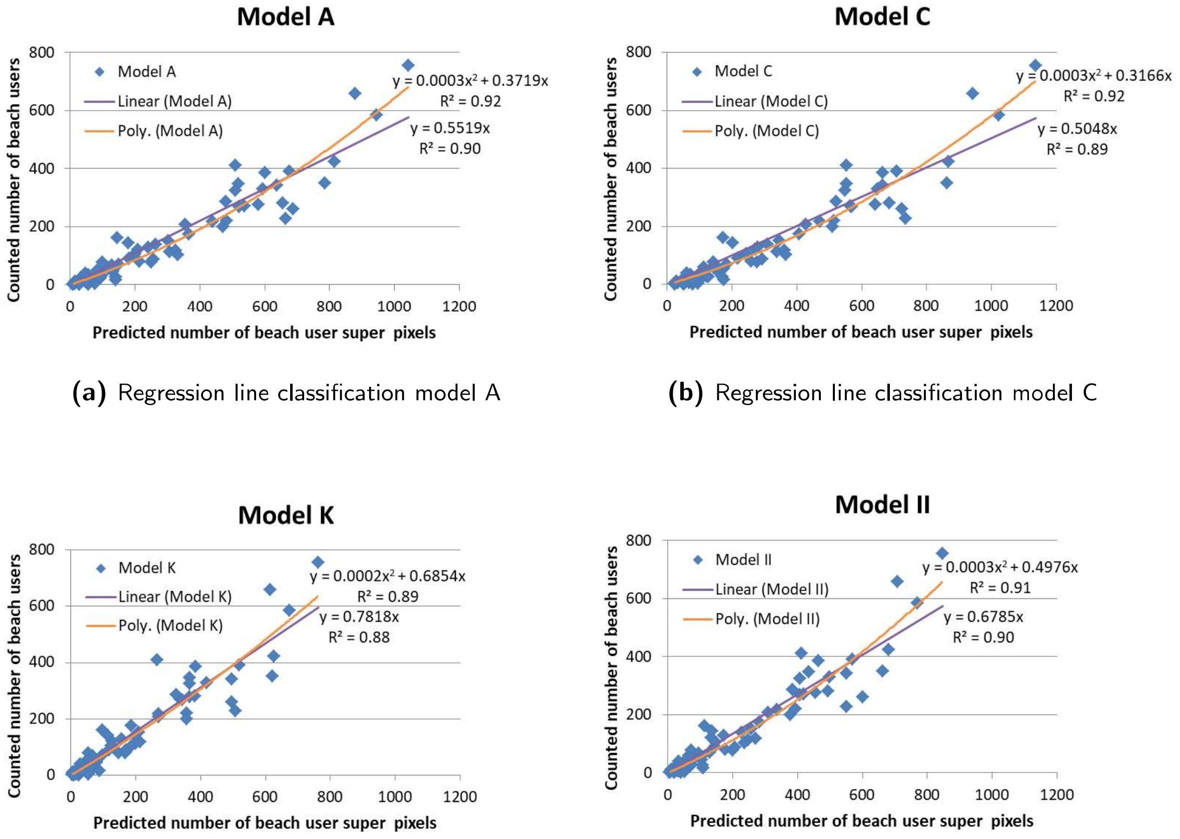
The regression model describes the conversion from the predicted number/area of beach user superpixels to the number of beach users. This section starts with the results obtained by directly fitting regression lines on the predicted *number* of beach user superpixels and the counted ground truth. Regression lines are fitted on predictions made with oversegmented classification models A, C, K, II and 5 to obtain insight in the effect of differences in precision and recall on the R^2.

Based on comparison of the direct fits of oversegmented classification models A,C, K, II and 5 an oversegmented classification model is selected for the remainder of this study. For this oversegmented classification model a new fit is determined based on the rectified areas corresponding to the predicted beach user superpixels. Subsequently, the R^2 corresponding to the direct- and rectified relations are compared to obtain the best approach for conversion of predicted beach user superpixels into beach users.

Lastly, the performance of the combined oversegmented classification and regression (baseline) model is expressed based on application of both models on the dataset used for training/development of both models. This procedure is repeated on a validation dataset with images that have not been used during training/development to validate the fit of the baseline regression model. The results obtained in this section correspond to the second research question.

5.2.1 Direct relation

Linear and second order regression lines with a zero intercept are fitted to the classified number of beach user superpixels and the counted ground truth (Fig. 5.6). Regression lines with a zero intercept have been used, because preliminary tests conducted with free intercepts resulted in physically impossible results (e.g. negative number of beach users on days with limited beach occupation). The difference in R^2 between fits with a free or fixed intercept is found to be negligible.

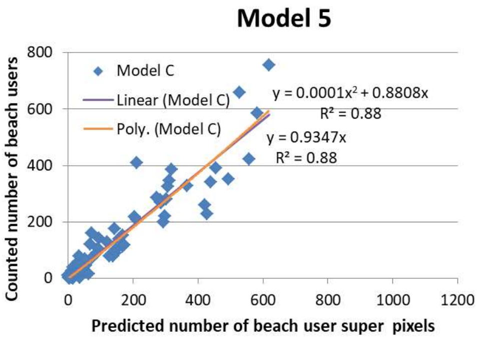


(a) Regression line classification model A

(b) Regression line classification model C

(c) Regression line classification model K

(d) Regression line classification model II



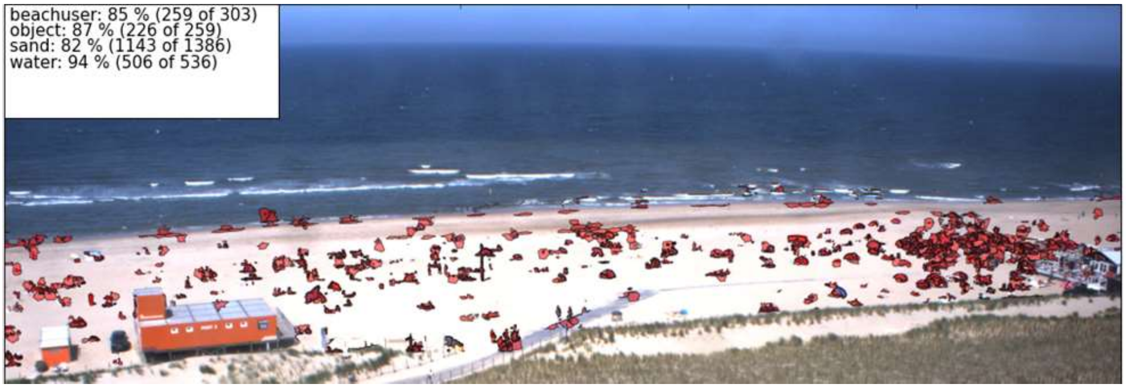
(e) Regression line classification model 5

Figure 5.6: Fitted regression lines on data corresponding to predictions made by oversegmented classification models A, C, K, II and 5 and a counted ground truth.. Both a linear and second order polynomial line are fitted. The intercept is set to zero in order to prevent physically impossible (negative) beach user predictions. Model 5 is the original aggregation 5, models A, C and K are undersampled and Model II is cost weight balanced.

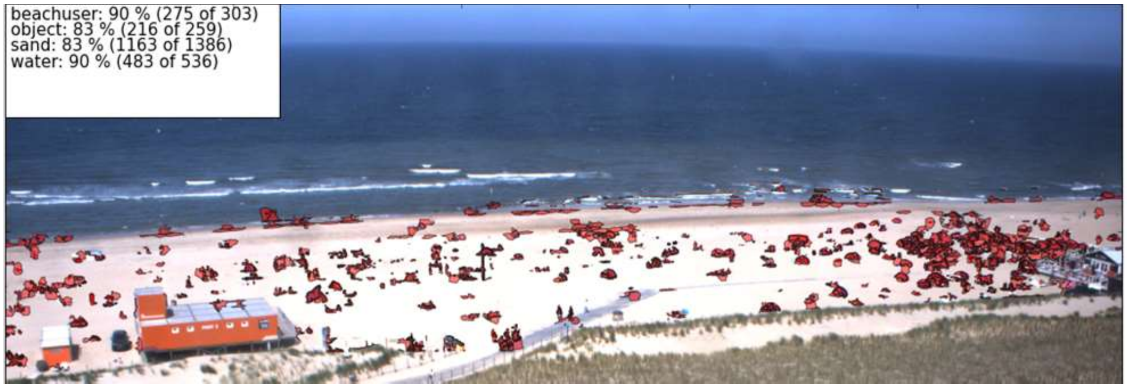
Analysis of the linear and second order regression lines, indicates that a polynomial fit results in a higher R^2-score on all models and therefore is a better fit (Fig. 5.6). The used polynomial fit predicts an increased number of beach users per superpixel for images with a high number of beach user superpixels because of its concave behaviour. This implies that in images with high beach occupation, a superpixel corresponds to more beach users. Physically this can be explained by occlusion of beach users by other beach users, which is likely to occur more often on days with high beach occupation. A similar trend is observed in Kammler and Schernewski (2004) where a number of pixels is related to the number of beach users. Eventually a linear fit was applied, but this turned out to be a bad fit on images with high beach occupation. It was found that in these images beach users are represented by a reduced number of pixels compared to images with limited beach occupation and this corresponds to the above described observations in this study. However, the difference between the linear and polynomial fit is most likely originating from three outliers in the top right and it is therefore recommended to verify the polynomial fit by evaluating more days with high beach occupation.

Comparison of the R^2-scores corresponding to the polynomial fit on the predictions of the different oversegmented classification models, indicates that models A and C have the best fit. Models II, K and model 5 perform less based on their R^2-score, but the differences are limited. The lower score of model 5 is expected, because this is the original model without measures to account for the imbalance in the data. However, the lower performance of model K is a remarkable result regarding the PR-score, because model K is characterised by the best observed PR-score. Moreover, models A and C have a relatively bad PR-score but have the highest performance regarding the R^2-score. Noticeable is that both models A and C have a relatively high recall and this indicates that the underlying distribution of precision and recall might be of larger importance than the combined PR-score. To obtain more insight in the characteristics of the oversegmented classification models, the models are used to make predictions on the same image (Fig. 5.7).

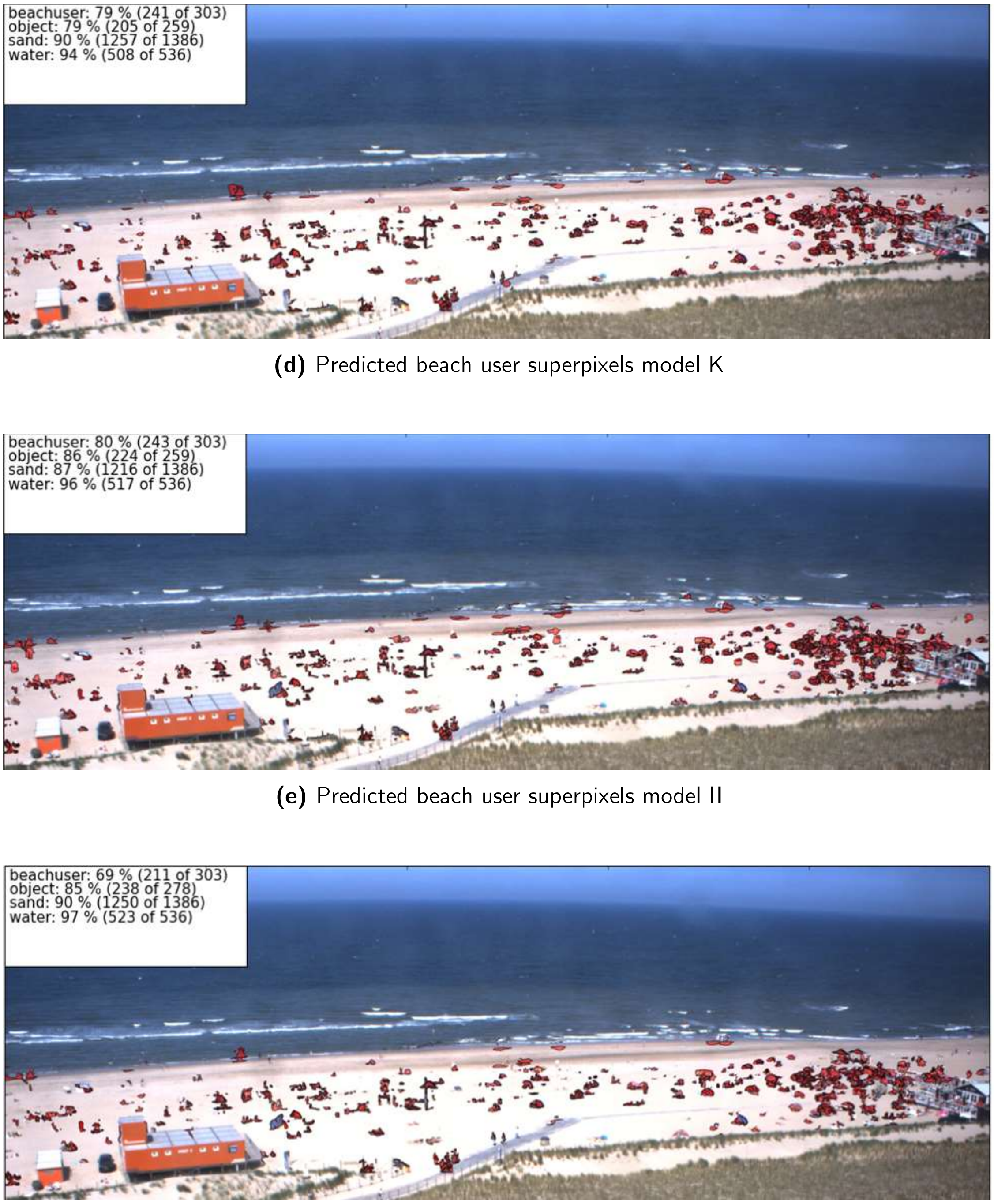
(a) Original image without predictions


beachuser: 85 % (259 of 303)
object: 87 % (226 of 259)
sand: 82 % (1143 of 1386)
water: 94 % (506 of 536)

(b) Predicted beach user superpixels model A


beachuser: 90 % (275 of 303)
object: 83 % (216 of 259)
sand: 83 % (1163 of 1386)
water: 90 % (483 of 536)

(c) Predicted beach user superpixels model C


beachuser: 79 % (241 of 303)
object: 79 % (205 of 259)
sand: 90 % (1257 of 1386)
water: 94 % (508 of 536)
beachuser: 80 % (243 of 303)
object: 86 % (224 of 259)
sand: 87 % (1216 of 1386)
water: 96 % (517 of 536)
beachuser: 69 % (211 of 303)
object: 85 % (238 of 278)
sand: 90 % (1250 of 1386)
water: 97 % (523 of 536)

(d) Predicted beach user superpixels model K

(e) Predicted beach user superpixels model II

(f) Predicted beach user superpixels model 5

Figure 5.7: Example of predictions made by models A, C, K, II and 5 on the same image. The red area's correspond to superpixels classified as beach user. The box in the top left indicates per class the percentage of annotated superpixels that is predicted corrected (indication true positives).

Comparison of the predicted beach user superpixels and percentages of correctly predicted superpixels between the models, shows that models A and C detect the highest number of annotated beach user superpixels (percentage of beach users in top left boxes Fig. 5.7), but these models also have a relatively high number of false positives (red areas in Fig. 5.7b and Fig. 5.7c not corresponding to a beach user visible in Fig. 5.7a). The percentages of the models K, II and 5 show a lower detection rate of the annotated beach user superpixels compared to models A and C, but the number of false positives is also limited.

The above observations based on comparison of predictions on a single image are supported by the regression relations (Fig. 5.6). The regression relations indicate that all models predict more beach user superpixels than beach users are counted. In general the number of predicted beach user superpixels is expected to be higher than the counted ground truth as a result of the final aggregation that merged beach objects to the beach user class. The oversegmented classification models predict beach objects as beach users, whereas beach objects are not counted as beach user. However, comparison of the relations shows that models A and C require a relatively large reduction of predicted beach user superpixels to obtain the actual number of beach users and this indicates a relatively large number of false positives.

The observed differences can be explained by the separate precision and recall scores corresponding to models A, C, K, II and 5 (Fig. 5.5). Models A and C have a relatively high recall combined with a limited precision. Recall is a metric used to express the relation between true- and false negatives (eq. 4.8) and a higher recall indicates that a model has a relatively high number of true positives compared to the number false negatives. Precision expresses the relation between the true- and false positives (eq. 4.7) and limited precision indicates that a model has a relatively high number of false positives. Models K, II and 5 have a lower recall and higher precision resulting in a lower detection rate of the target beach user class, but also a lower number of false positives.

Dependent on the required application, the most suitable model may change and it is therefore difficult to select an optimal model. The differences in R^2 are limited indicating that the residual errors made by the models are similar. The differences in distribution of precision and recall results in models with different characteristics regarding the true- and false positive rates and depended on the application the preferred classification model may change. Due to practical reasons this study focuses on model A.

5.2.2 Relation based on rectified area of superpixels

To test the effect of rectification, the area's of the beach user superpixels predicted by oversegmented classification model A have been determined. Subsequently, a regression line is fitted on the area corresponding the to superpixels and the counted ground truth (Fig. 5.8). The R^2-score of both the linear- and second order polynomial fit indicates that rectification results in a less optimal fit. Because of this the direct relation is considered optimal and is maintained in the remainder of this report.

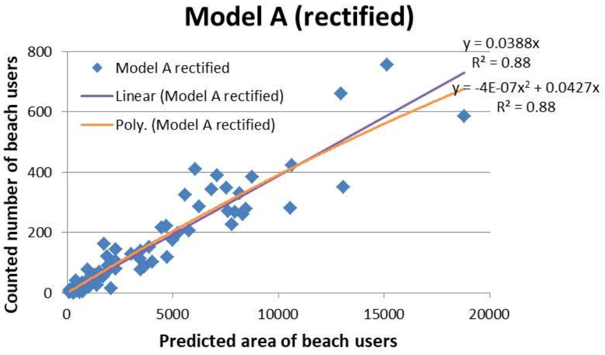


Figure 5.8: Calibration curve rectified model A

The limited performance of the rectification approach can be explained by especially the distance of beach users from the camera. Beach users close to the camera are captured more from above, whereas beach users near the water line are captured more from the side. The results show that beach users at larger distance from the camera are disproportionately enlarged by the rectification procedure (Fig. 5.9).

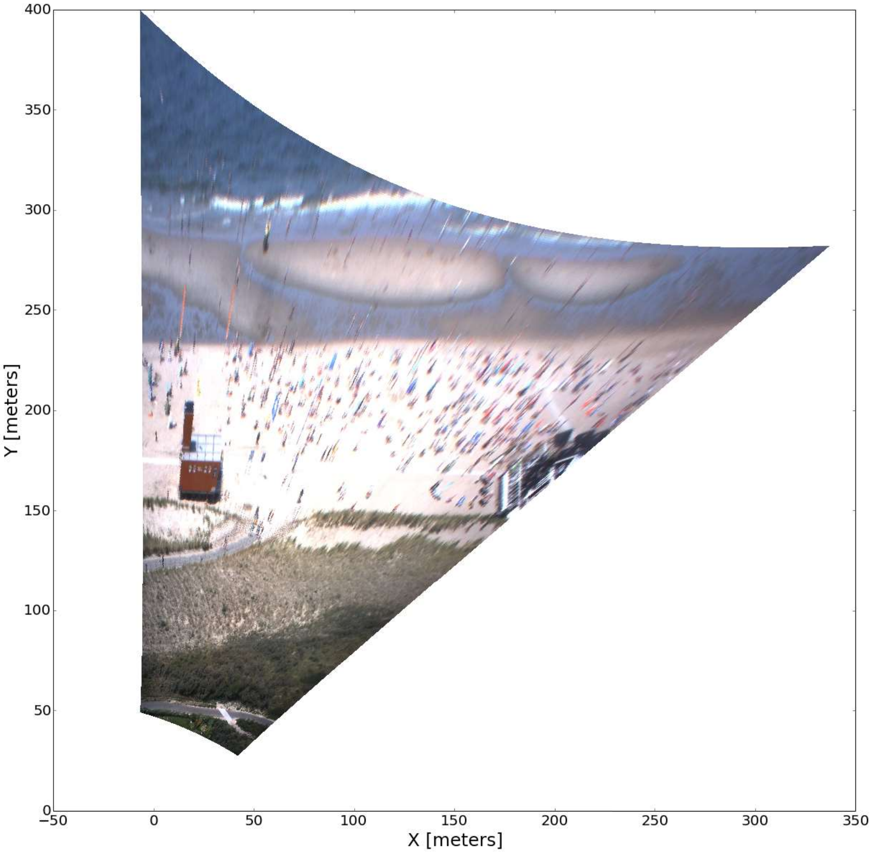


Figure 5.9: Example of rectified Argus image. The image shows that beach users at a relatively large distance from the camera become disproportionately large after rectification.

5.2.3 Performance of combined oversegmented classification and regression baseline model

The final baseline model is a combination of an oversegmented classification model with class aggregation 5 (Tab. 5.1) that has been auto-undersampled and a regression model with a zero intercept and second order polynomial fit. The conversion from the classified number of beach user superpixels to the number of beach users is expressed by the following regression relation:

$$N_{bu} = 0.0003N_{pix,bu}^2 + 0.3179N_{pix,bu} \tag{5.1}$$

In which:

N_{bu} = Predicted number of beach users in analysed image
$N_{pix,bu}$ = Number of superpixels assigned to the class beach user in analysed image

The performance of the combined oversegmented classification and regression model is expressed by the R^2-score. Although the adopted R^2-scores is a representation of the errors that are made by the oversegmented classification model, it is difficult to relate the R^2 to an error expressed in beach users. Moreover, the R^2 has been determined based on the errors of *all* images in the dataset and does not provide information on the magnitude of the error with respect to the counted ground truth. This type of information is desirable due to the large variability in beach occupation and corresponding errors between different images in the dataset (Fig. 5.10).

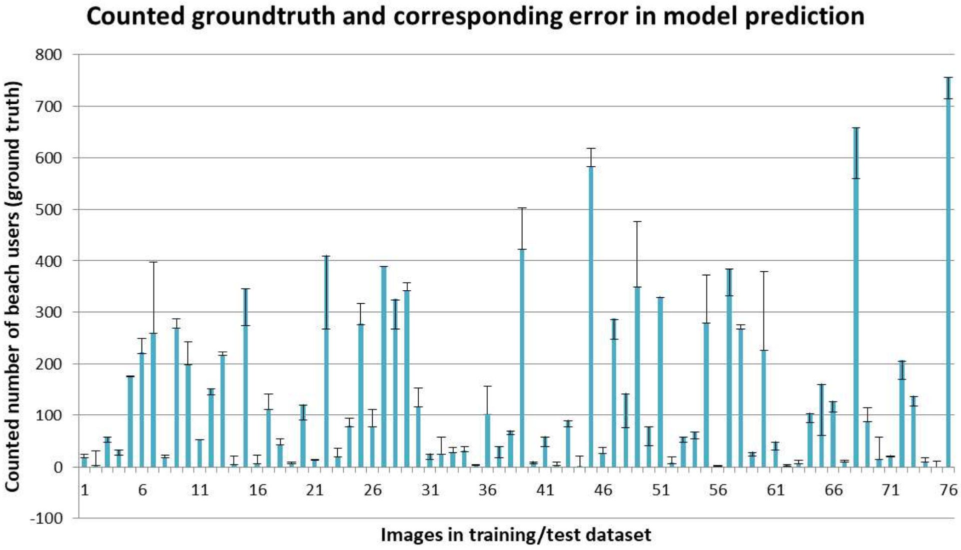


Figure 5.10: Counted ground truth per image. The error bars indicate the error between the predicted number of beach users by the combined oversegmented classification and regression model A and the manually counted ground truth.

To quantify the errors with respect to the counted ground truth a division in three bins based on the counted beach occupation has been made (Tab. 5.5). For each bin the bias and RMSE are determined and these metrics are expressed in the number of beach users.

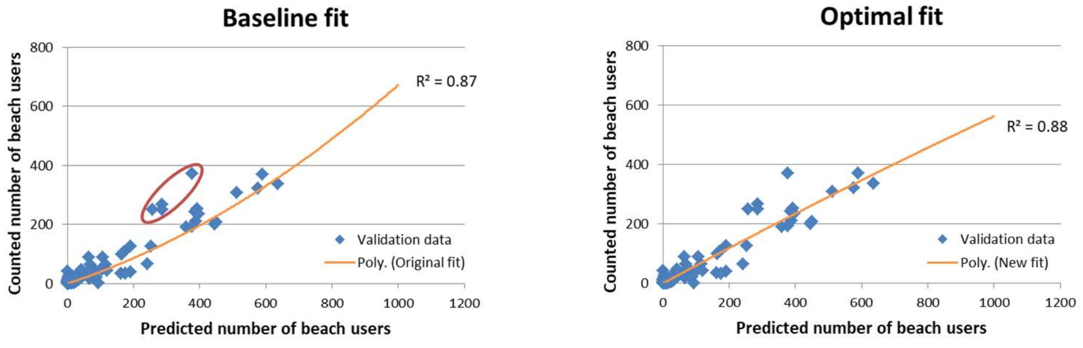
Table 5.5: Bias and RMSE distributed over bins based on the counted number of beach users

	0-25:	**25-100:**	**100+:**
Number of images:	20	22	33
Bias:	8.7	-1.7	5.5
RMSE:	11.1	17.7	64.3

In general the bias is limited with respect to the ground truth of the corresponding bins. This implicates that application of the model for long term monitoring to obtain for instance insight in the number of beach users on a beach, will give a fairly accurate representation of the truth because negative and positive errors average out. The RMSE gives an indication of the error that is on average made in individual images and this error is large, especially the 100+ bin (RMSE of 64.3). However, Fig. 5.10 shows that larger errors in this bin are especially made on images with a ground truth that is far above 100. Therefore, the relative error is limited and the predictions can still be indicative. During elaboration of the objective of this research (subsection 1.3) accuracy was defined as the capability to predict the beach occupation in the correct order of magnitude. Based on the results obtained by application of the oversegmented machine learning method on the training data, the method is considered accurate for the purposes of beach user quantification. Possible limitations/sources of errors, are treated at the end of the next section.

5.2.4 Validation

The fit of the in the previous section determined baseline regression model is validated by application of the combined (baseline) oversegmented classification and regression model on the validation dataset. The validation dataset contains 80 'new' Argus images that have not been used during training and development of respectively the oversegmented classification- and regression models. Analysis of the baseline fit on the validation data shows that the validation data is described reasonably well by the baseline fit except for four outliers indicated by the red circle (Fig. 5.11a).



(a) Original fit on validation data **(b)** Specific fit on validation data

Figure 5.11: Validation data together with the original fit of model A based on the training data and the optimal fit based on the validation data. The red circle indicates the four largest outliers.

Comparison of the R^2-scores corresponding to baseline fit on the training data and the baseline fit on the validation data shows that the R^2-score on the validation data is lower, but that the difference is limited (Tab. 5.6). This indicates that the fit obtained during development of the baseline regression model has generic applicability as it is capable to make reasonable accurate predictions on data points that have not been used during development of the model. The foregoing is supported by the R^2-score of a fit that is obtained by applying regression on the validation data (Fig. 5.11b). This fit is the optimal fit on the validation data and the R^2 of the baseline fit on the validation data is very similar. The correspondence between the optimal and baseline fit on the validation data, indicates that the baseline fit is very close to the optimal fit on the validation dataset.

Table 5.6: Comparison of R^2-scores

	R^2-score
Original fit (training data):	0.92
Original fit (validation data):	0.87
Optimal fit (validation data):	0.88

Analysis of the fits in Figure 5.11 shows that, although a second order polynomial fit has been used, the optimal fit is nearly linear and even has a slightly convex behaviour. This is a remarkable result because the baseline fit showed clearly concave behaviour. Recall from section 5.2.1 that the concave curve of the baseline fit is predominantly the result of three data points in the top right corner of the plot. Data points corresponding to this magnitude of beach occupation are absent in the validation data, reducing the necessity of a concave curve. Moreover, the validation data shows four outliers (red circles Fig.

5.11), that even further reduce a potential concave behaviour and contribute to the (slightly) convex result.

The counted ground truth of the validation dataset is presented together with the errors made by the baseline fit to obtain insight in the relative errors that are made in individual images (Fig. 5.12). The four largest peaks are indicated with red and correspond to the outliers indicated with red circles in Figure 5.11.

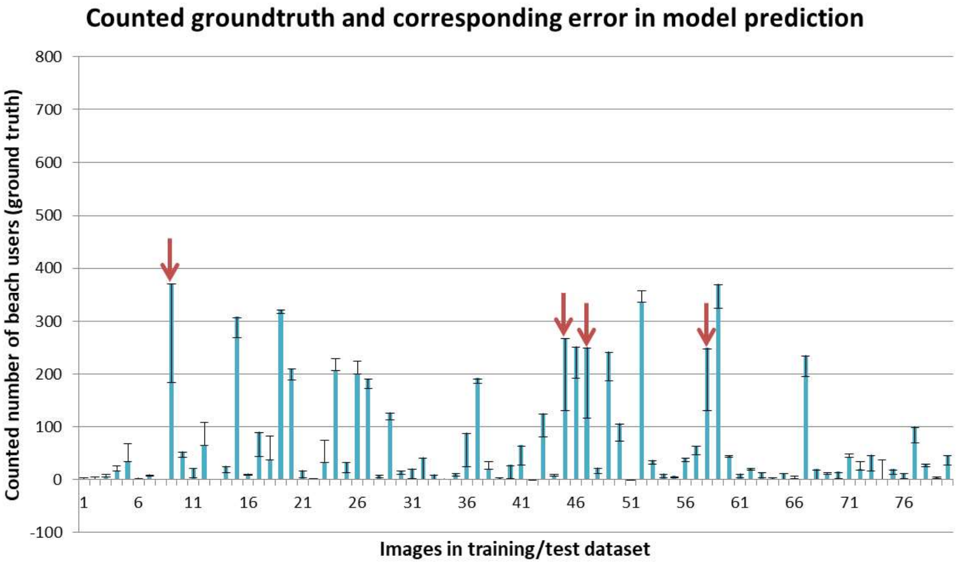


Figure 5.12: Counted ground truth per image in the *validation dataset*. The error bars indicate the error between the predicted number of beach users by the baseline oversegmented classification and regression model on the validation data. The red arrows correspond to the outliers indicated with red circles in Fig. 5.11.

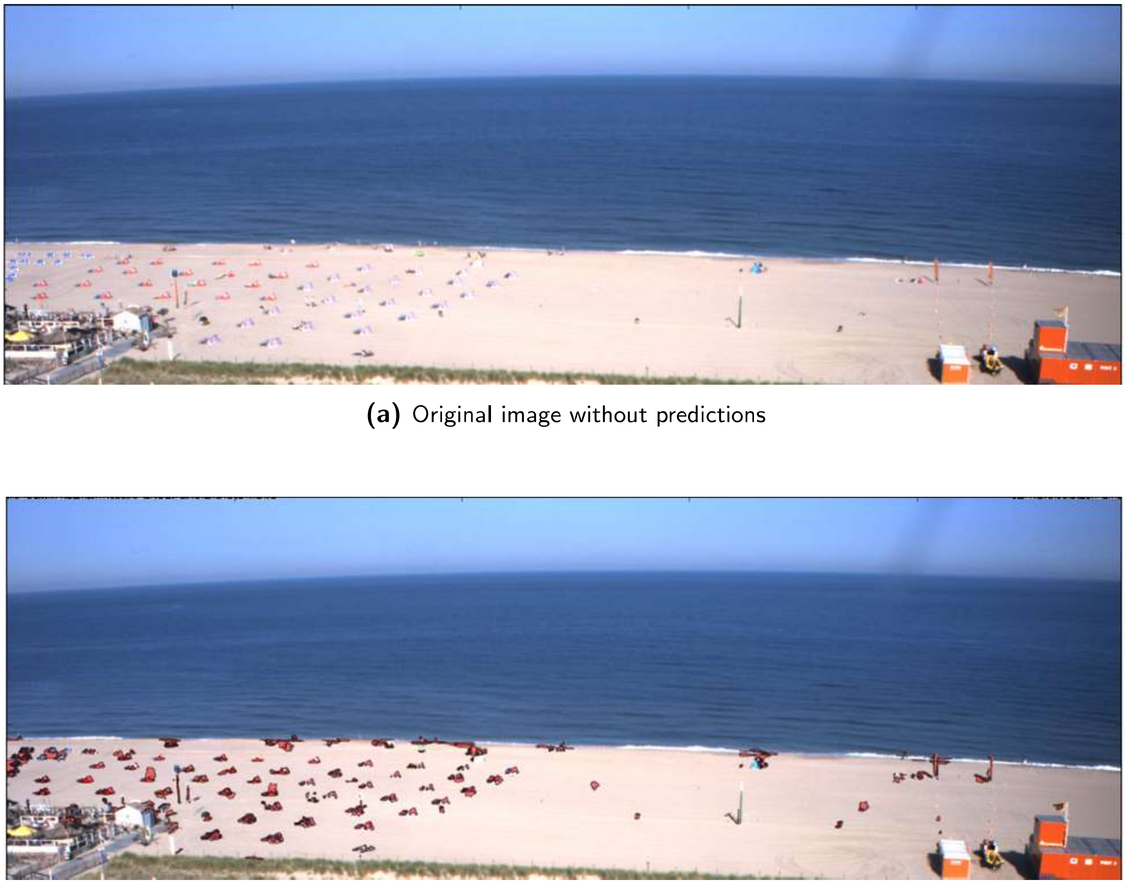
Quantification of the errors shows that the performance on the validation data in the bin 0-25 is better than on the original training data (Tab. 5.7). The performance in the bin 25-100 and 100+ is significantly worse. The limited performance in the 100+ bin can largely be attributed to the before mentioned outliers in this range. Too little beach users are classified in these images resulting in a negative error. Inspection of the outliers showed that they are responsible for approximately 30 of the 47.5 on average mis-classified beach users indicated by the bias. Eliminating this number from the bias in the 100+ bin still results in a higher number than for the training data, but the error becomes acceptable regarding the high number of beach users represented by this bin.

Table 5.7: Bias and RMSE distributed over bins based on the counted number of beach users. The values between brackets correspond to the values obtained by application of the baseline fit on the training data (Tab. 5.5).

	0-25:	**25-100:**	**100+:**
Number of images:	38 *(20)*	21 *(22)*	19 *(33)*
Bias:	-3.2 *(8.7)*	-10.3 *(-1.7)*	-47.5 *(5.5)*
RMSE:	10.2 *(11.1)*	29.0 *(17.7)*	56.8 *(64.3)*

The bias of the 25-100 bin might still be acceptable, but the RMSE is in the order of magnitude of the lower bound of this bin (25 beach users). Occurrence of these type of errors on images with a counted ground truth close to the lower bound of this bin, can lead to significant errors with respect to the counted ground truth. Visual inspection of the images and their errors corresponding to this bin, showed two main causes for the large errors:

Firstly, significant positive errors are induced by (rental) beach stretchers on the beach (Fig. 5.13). The beach clubs place these stretchers on the beach already early in the morning on days with nice weather. During this time of day the beach occupation is minor and the stretchers are limited occupied. Therefore, no direct relation between the number of stretchers and beach users exists. The existence of this relation was the reason to allow the combined aggregation of beach users and beach objects (e.g. stretchers) in section 4.6.2. The absence of a relation between beach objects and beach users results in beach user classification while in reality no beach users are related to the stretchers. Images containing this type of situation are relatively often present in the validation data due to the high number of image corresponding to 10.30h local time.

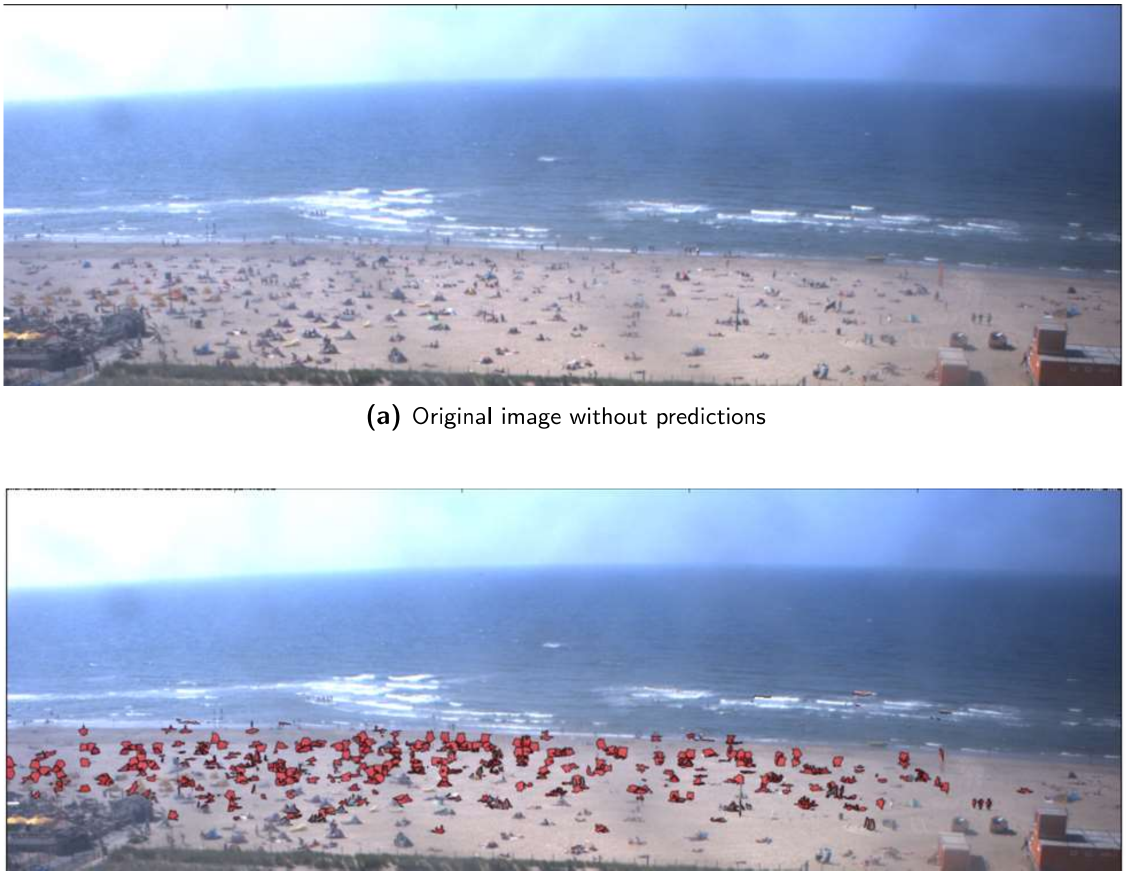


(a) Original image without predictions

(b) Image with predicted beach user superpixels

Figure 5.13: Example of image with a lot of unoccupied (rental) beach stretchers around the beach club. This picture was taken at 10.30 local time and has a counted ground truth of 33 beach users. The combined (baseline) oversegmented classification and regression model predicted 75 beach users.

Secondly, visual observation of the images returning relatively large negative errors, indicates that these errors are made on images caputred by an unclean lens. The visibility through the lens is lowered by sand/salt on the lens and therefore the oversegmented classification model classifies a limited number of beach users. This can be explained by the gradient between beach users and the surrounding pixels. The unclean lens lowers this gradient causing beach users to merge into the background and reducing the probability that a distinction between beach user and the surrounding is made during superpixel formation since superpixels are formed based on gradients (subsection 4.4). In images with a lot of beach users near the water line, the error seems to be amplified because the gradient between beach users near the waterline and the surrounding water pixels is even lower than between beach users and the (dry) beach. This type of error is also observed in the training dataset.



(a) Original image without predictions

(b) Image with predicted beach user superpixels

Figure 5.14: Example of image with an unclean lens. This picture has a counted ground truth of 250. The combined (baseline) oversegmented classification and regression model predicted 130 beach users.

The above considerations indicate that the oversegmented machine learning method has limitations on images with a lot of unoccupied (rental) beach stretchers. Moreover, the method performs bad on images captured with an unclean lens. Both limitations are elaborated in more detail in section 6.4.

5.2.5 Conclusions regression model (baseline)

This section presented the results obtained with different fits/relations to convert classified beach users superpixels by the oversegmented classification model to the number of beach users. These results answer the second research sub-question.

The approach directly relating the *number* of beach user superpixels to the number of beach users led to better results compared to the approach relating the *area* of beach user superpixels to the number of beach users. A second order polynomial fit predicting more beach users per beach user superpixel for images with a relatively high beach occupation has been used. This type of fit can physically be explained by occlusion of beach users on busy days, causing the presence of more than 1 beach user in a superpixel.

The fits of multiple oversegmented classification models have been determined and the corresponding R^2-scores show similar performance. However, the relations describing the obtained fits differ between the models and show that the number of classified beach user superpixels of some models has to be reduced significantly more than for other models to finally obtain the number of beach users. This is explained by the underlying distribution of precision and recall and the relation of these metrics to the number of true positives, false positives and false negatives. The models show relatively large differences in especially the number of true positives and false positives and dependent on the application the most suitable oversegmented classification model should be selected. In this study the 'auto'-undersampled model has been selected due to practical reasons.

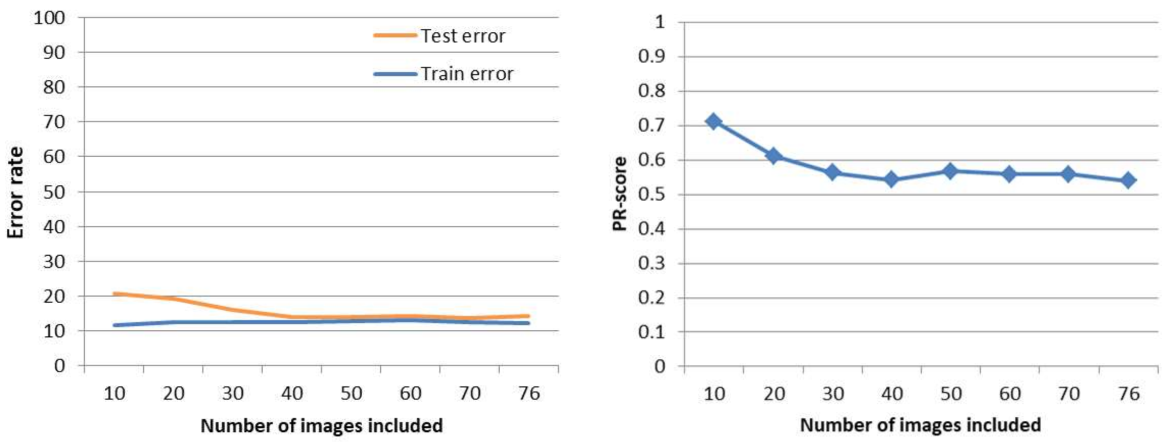
Lastly the applicability of the regression model corresponding to the selected oversegmented classification model is validated. Based on the R^2-score the fit performs well on images that are not used during development of the regression model. However, analysis of the error on individual images showed that the model over-predicts the number of beach users in the situation with a lot of beach stretchers and limited beach occupation. Moreover, the validation dataset contained images with an unclean lens that led to significant under-predictions of the number of beach users.

5.3 Optimisation of baseline model

In section 5.1 a baseline oversegmented classification model has been developed that, in combination with the baseline regression model of section 5.2, completes the oversegmented machine learning method for quantification of the beach occupation. This section treats the results of four possible optimisation steps: *the number of images*, *images enhancement*, *artificial channels* and *limiting the number of features*. The obtained results contribute to the answer to the first research question.

5.3.1 Number of images

The oversegmented classification baseline model has been trained with a dataset of 76 images and the learning curve is reviewed to verify that this number of images is sufficient (Fig. 5.15a). The learning curve shows that the train- and test error converge at approximately 45 images. The convergence of the train- and test error indicates that the model does not have a high variance. Recall from subsection 4.2.2 that high variance is typical for classification models that could benefit from a larger training dataset and therefore a limited variance indicates that annotating more images will not result in an increased performance of the oversegmented classification model.



(a) Learning curve **(b)** Distribution of PR-score for ascending # of images

Figure 5.15: Example of image with an unclean lens. This picture has a counted ground truth of 250. The combined (baseline) oversegmented classification and regression model predicted 130 beach users.

Besides information on the variance of the model, subsection 4.2.2 explained that the learning curve can be used to identify the bias of a classification model. A high bias was characterised by a train- and test error that converge but at a relatively high error rate and after a limited number of images. Both conditions are qualitative and therefore it is difficult to exclude a high bias. The learning curve shows that the final error rate is approximately 12% and the errors converge at 45 images. Especially the number of 45 images might be considered limited, but each image contains approximately 3000 superpixels resulting in a total of circa 135.000 training instances. From this perspective the number of images might be sufficient.

The results obtained by varying the regularisation parameter (subsection 5.1.4), can be used to support/obtain considerations regarding the variance and bias of the oversegmented classification model. The results on regularisation indicated that both high and low regularisation values have very limited effect. The limited effect of low regularisation values supports the low variance hypothesis, because low regularisation values counteract overfitting and overfitting is typical for classification models with a high variance (subsection 4.6.4). Moreover, the limited effect of a relatively high regularisation parameter indicates that the model does not have a high bias. This is because an increased regularisation parameter is one of the measures to counteract a high bias.

The learning curve in Fig. 5.15a is based on the *overall* error rate. However, the error rate is defined as 100 minus the *overall* accuracy and the accuracy was considered to be a less suitable scoring method in section 4.7. Therefore, the distribution of the PR-score for an ascending number of images is determined (Fig. 5.15b). This curve supports the requirement of 40-50 images for training of the oversegmented classification model. The underlying precision and recall also stabilise after this number of images. The fluctuations in PR-score after 40 images can be explained by the random sampling procedure that corresponds to the undersampling approach that has been used for the reviewed oversegmented classification model. Note that too little images can significantly lower the performance based on the PR-score and this shows that it is relatively important to train the classification model with enough images.

5.3.2 Effect of image enhancement

The original images in the training dataset have been enhanced with Contrast Limited Adaptive Histogram Equalisation (CLAHE) to potentially increase the gradients between beach users and the other classes. An increased gradient might result in less ambiguous superpixels and superpixels with larger differences between the beach user class and other classes. Two enhancement approaches have been tested: in the first approach the baseline oversegmented classification model trained on *not enhanced* images has been used to predict the number of beach user superpixels in *enhanced* images. Subsequently, the predicted number of beach user superpixels has been converted to the number of beach users by the baseline regression model. The second approach trains a new oversegmented classification model with *enhanced* images retaining the settings for class aggregation, undersampling and regularisation. Moreover, the validated baseline regression model has been used for the conversion from the classified number of beach user superpixels to the number of beach users. This section presents the results of both approaches separately.

Effect of enhanced images on the baseline classification model trained with not enhanced images

The results show that application of the baseline models on CLAHE enhanced images reduces the R^2 from 0.92 (baseline models in combination with not enhanced images, Fig. 5.7b) to 0.20 (baseline model in combination with enhanced images, Fig. 5.16). This indicates that the approach using enhanced images in combination with the baseline models results in larger errors compared to application of the baseline models in combination with not enhanced images.

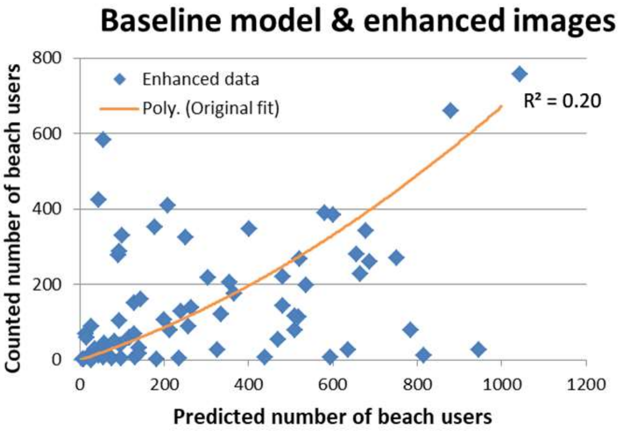


Figure 5.16: Regression line on predictions made with not enhanced classification baseline model with and enhanced images

The results confirm the expectation that application of the baseline models in combination with enhanced images is undesirable, because the theoretical benefits of enhancement are only explored in a limited part of the total work flow. Due to the significant increase in the error made by this approach, the approach is not treated in more detail in the remainder of this study.

Effect of enhanced images on a classification model trained with enhanced images

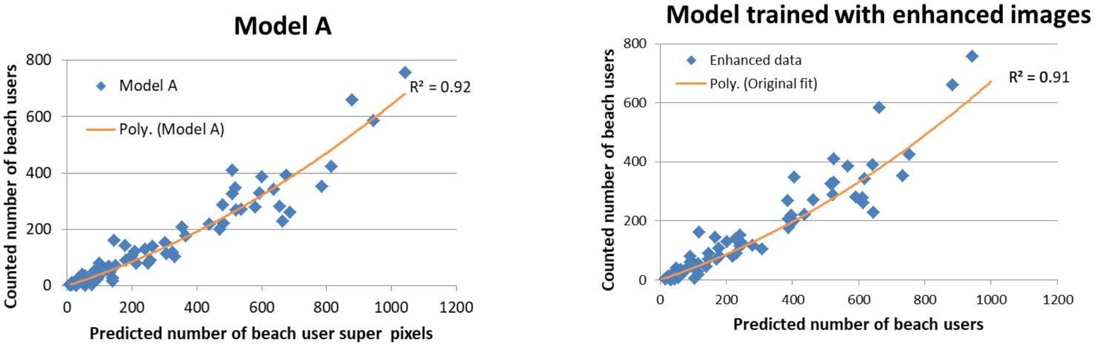
This subsection starts with a comparison of the PR-score corresponding to the baseline classification model and the new classification model trained with enhanced images. Note that this comparison based on PR-score is possible for this approach, because a new model has been trained. Subsequently, the R^2, bias and RMSE corresponding to the baseline models and the classification model trained with enhanced images in combination with the baseline fit will be treated.

Effect based on PR-score

The classification model trained with enhanced images can be scored by the same PR-score that was used in section 5.1 to develop the baseline classification model. The PR-score of both models is 0.54 and moreover the underlying precision and recall are very similar. This result indicates a limited effect of CLAHE-enhancement.

Effect based on R^2

A comparison of the R^2-score of the baseline models and a combination of a classification model trained with enhanced images and the baseline regression model shows that the R^2 of both models is very close although the R^2 of the model trained with enhanced images is lower. Moreover, visual comparison of the original data points (Fig. 5.17a) with the enhanced data points (Fig. 5.17b) shows variation in the distance of data points to the regression line and this indicates fluctuations in the errors per image between both models.

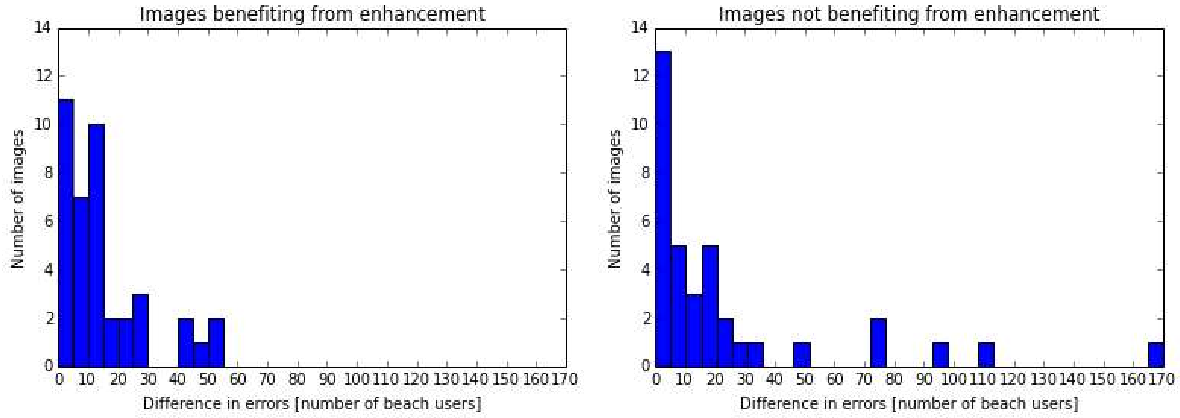


(a) Fit of not enhanced model A

(b) Fit of enhanced model A

Figure 5.17: Visualisation of the optimal fits on the original training data and the enhanced training data.

Quantitative comparison of the errors shows that 40 of the 76 images in the evaluated data set have a lower error after being enhanced and classified by the oversegmented classification model trained on enhanced images. The remaining 36 images have a lower error when not enhanced and classified by the baseline model trained with not enhanced images. It is remarkable that over 50% of the images benefit from enhancement while the R^2 was found to be less and this indicates that the model trained with enhanced images makes relatively large errors on (some of) the images that do not benefit from enhancement. This is supported by comparison of Figures 5.18a and 5.18b, which shows that the enhanced model has some relatively large outliers.



(a) Images benefiting from enhancement **(b)** Images not benefiting from enhancement

Figure 5.18: Histograms on the difference between errors for images that benefit from enhancement and images that do not benefit from enhancement. In the former situation the enhanced errors are subtracted by the not enhanced errors and for the latter the opposite holds.

The observation that enhancement is not beneficial to all images and moreover causes large outliers in some images, motivates the search for a performance indicator to distinguish images that will benefit from enhancement and images that do not before classification. During preliminary attempts it was tried to create a performance indicator that identifies all images that benefit from enhancement based on correlating image histograms. Unfortunately, the obtained results were not satisfactory and this indicates that not one type of image exists that benefits from enhancement. Visual inspection of the data supports this.

5.3.3 Addition of artificial channels

Differences in Gaussian filtering, Gabor filtering and Sobel filtering have been applied to the original r,g,b images to create additional artificial channels next to the original r,g,b channel. Extraction of these channels and the corresponding features increased the number of available features from 865 to 1341 features after addition of the artificial channels. Due to practical reasons these channels were already extracted before development of the final baseline model and the effect can therefore only be expressed by a limited number of models trained in preliminary model iterations. Together with the addition of the channels, the 'ROI-superpixel' (superpixel containing all pixels outside the ROI) was removed to limit the range of features values during feature re-sizing (subsection 4.5.3). Therefore, the below presented results cannot solely be attributed to the added artificial channels, since removal of the ROI-superpixel can be of influence.

In previous model iterations the oversegmented classification model was trained by 52 images and with two-class models (i.e. aggregation 19, Tab. 5.1). To get a rough estimate of the merits of adding artificial channels two of these models are compared (Tab. 5.8). The first model has been undersampled by the default *'auto'* setting whereas the cost weight has been adjusted with the default *'balanced'* setting for the second model.

Table 5.8: Effect of adding artificial channels on PR-score.

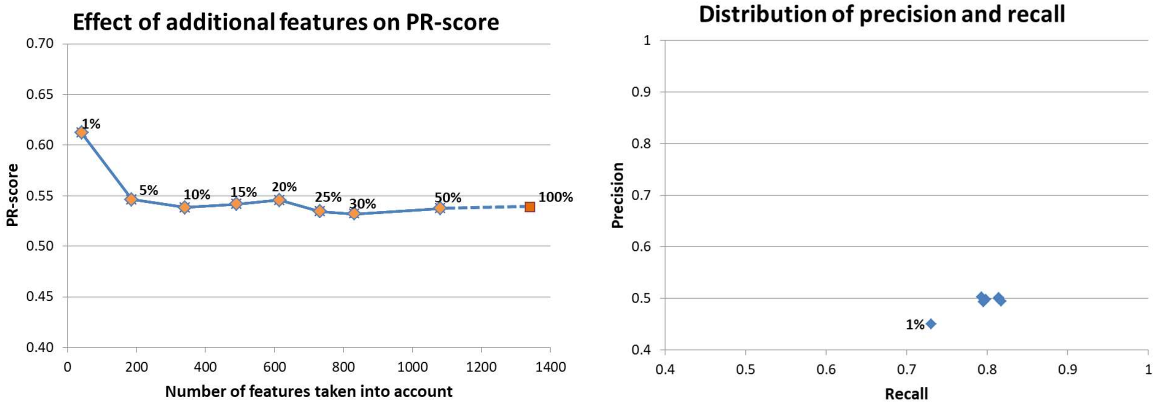
	PR-scores:	
	-Channels	*+Channels:*
Aggregation 19 (*auto*-undersampling):	0.70	0.64
Aggregation 19 (*balanced* cost weight balancing):	0.70	0.62

The difference in PR-score (Tab. 5.8) shows for both models a relatively large increase in model performance as a result of adding artificial channels. This indicates that the artificially added channels provide the classification model with additional information that is useful to distinguish between the classes of interest in beach user quantification. Therefore, artificial channels have been applied during development of the final oversegmented classification baseline model.

5.3.4 Effect limiting number of features

The final oversegmented classification (baseline) model has been trained with 1341 features and the importance of these features is investigated by providing the oversegmented classification model with a limited number of features during training.

Analysis of the performance shows that remaining approximately 5% of the most important features *per class* already results in the same model performance as training with 100% of the features (Fig. 5.19a). The minor variations in PR-score observed for models with 5% or more features, is expected to be caused by the applied random undersampling. The distribution of precision and recall shows limited variation among the different percentages, supporting the indication that the models have the same performance (Fig. 5.19b). The time required for training of the reduced oversegmented classification model reduces from in the order of an hour to a few minutes for the lower percentages.



(a) Effect of additional features

(b) Distribution of precision and recall

Figure 5.19: The left graph shows the effect of taking more features into account expressed by the PR-score. The right graph shows the underlying distribution of the precision and recall corresponding to the points in the left graph.

The number of features corresponding to a percentage of most important features *per class* is higher than the number of features corresponding the same percentage of the total number of features (Tab. 5.9). This indicates that the classes are characterised by different features.

Table 5.9: Number of features corresponding to percentages of most important features *per* class. The numbers between brackets present the minimal number of features that would have been evaluated in case the most important features would have been equal for all classes. The third column represent the ratio between the evaluated and minimal number of features.

	Number of features:	**Ratio:**
1%:	41 *(13)*	3.1
5%:	186 *(67)*	2.8
10%:	341 *(134)*	2.5
15%:	490 *(201)*	2.4
20%:	614 *(268)*	2.3
25%:	733 *(335)*	2.2
30%:	832 *(402)*	2.1
50%:	1081 *(671)*	1.6
100%:	1341 *(1341)*	1.0

The limited training time is beneficial, but this measure would be really valuable if it could be used to limit the number of features that has to be extracted during feature extraction. Feature extraction has been identified as one of the most time consuming steps in the process and the results show that it is not necessary to extract all 1341 features in order to obtain adequate model scores. At this moment the channel and feature extraction of one images takes approximately 30 minutes. A reduction of this time in the order of the reduction of the training time, would increase the applicability of the oversegmented machine learning method.

However, the Python package *scikit-learn* (Pedregosa et al., 2011) used for feature extraction, extracts the features in predefined blocks and it is difficult to select a limited feature set without first extracting all blocks. The final result is therefore that, theoretically the oversegmented machine learning method can be used with a limited set of features, but that it currently is not possible due to practical limitations.

5.3.5 Conclusions optimisation steps

This section evaluated the increase in performance of the oversegmented machine learning method that can be expected after optimising the baseline classification model with: a training dataset containing more images, applying image enhancement, adding artificial channels and limiting the number of extracted features. The results contribute to the answer to research sub-question 1.

Analysis of the learning curve showed that the baseline oversegmented classification model has been trained with enough images and therefore adding more images to the training dataset will not increase the model performance. Moreover, the learning curve indicated that training with too little images can significantly lower the model performance and from this it is concluded that the number of images in the training data is an important parameter.

Application of a classification model trained and applied on CLAHE-enhanced images showed that image enhancement resulted in lower errors for approximately 50% of the analysed images. However, comparison of the R^2-scores indicated that the overall errors are similar. Moreover, the classification model trained with enhanced images showed large outliers. From this it is concluded that enhancement cannot be implemented as a generic step in the workflow. An indicator capable to distinguish between images that can and cannot benefit from enhancement can be a solution but is not available at this point.

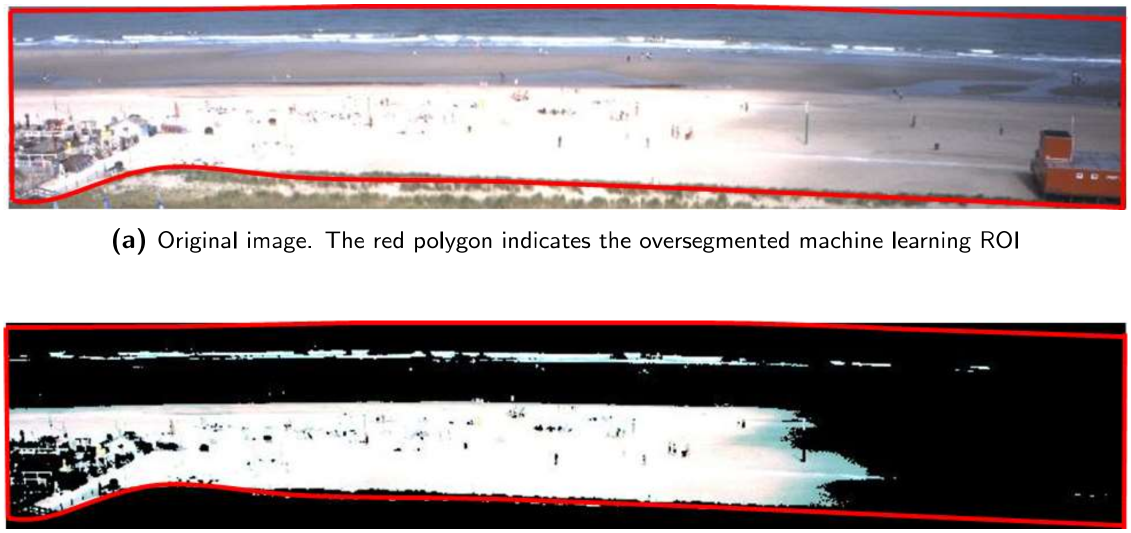
The addition of artificial channels resulted in an increase of the performance. However, artificial channels were added together with removal of the superpixel containing all pixels outside the ROI. Therefore, the increase in performance can also (partly) be caused by the removal of the superpixel.

Lastly the effect of limiting the number of features has been explored. The results show that it is possible to significantly lower the number of evaluated features without limiting the performance and this resulted in a reduced time required for training of the classification model. However, due to practical limitations of the algorithm used for feature extraction, it is at this point not possible to limit the number of extracted features.

5.4 Benchmark oversegmented machine learning method against current state-of-the-art model based on differences in pixel intensity

This section elaborates the results of a comparison between the performance corresponding to a current state-of-the-art (differences in pixel intensity) model and the oversegmented machine learning method (combined oversegmented classification- and regression model) to answer the third research sub-question. The current state-of-the-art method has been used to predict the number of *objects* in the images corresponding to the validation dataset as used during validation of the baseline regression model (sections 4.2.3 and 5.2.4).

The model has been applied with the oversegmented machine learning ROI *including* the inter-tidal area and a tailored ROI that is restricted to the dry beach and *excludes* the inter-tidal area following the definition of Guillén et al. (2008). Regression lines are fitted on the data corresponding to the predicted number of objects by the state-of-the-art model and the manually counted ground truth (Fig. 5.21). A non-zero intercept is allowed because the fits in the original work of Guillén et al. (2008) had a negative intercept. Moreover, a non-zero intercept resulted in a higher R^2 for the state-of-the-art method and therefore applying a zero intercept might underestimate the performance of this method.

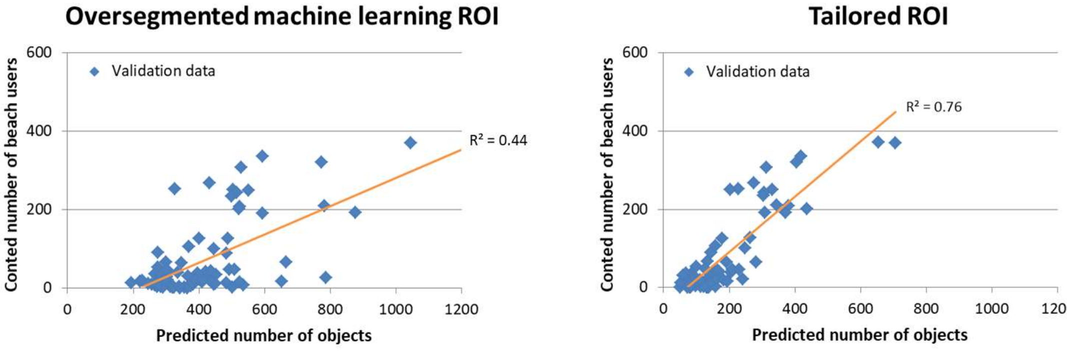


(a) Original image. The red polygon indicates the oversegmented machine learning ROI

(b) Predictions state-of-the-art model. All black pixels inside the ROI are classified as objects.

Figure 5.20: Indication of performance state-of-the-art method in combination with a ROI including the inter-tidal area.

Comparison of the R^2-scores corresponding to predictions made with the two different ROI's, indicates that the state-of-the-art model has a significantly higher performance for application in combination with the tailored ROI (comparison Fig. 5.21a and Fig. 5.21b). This is an indication that the state-of-the-art model performs bad in situations with a large inter-tidal area since the inter-tidal area is the only difference between the ROI's. Analysis of Fig. 5.20 confirms that the state-of-the-art model classifies a large part of the inter-tidal area as object limiting the performance of this ROI. Therefore, the state-of-the-art model is applied with the tailored ROI in the remainder of this section.

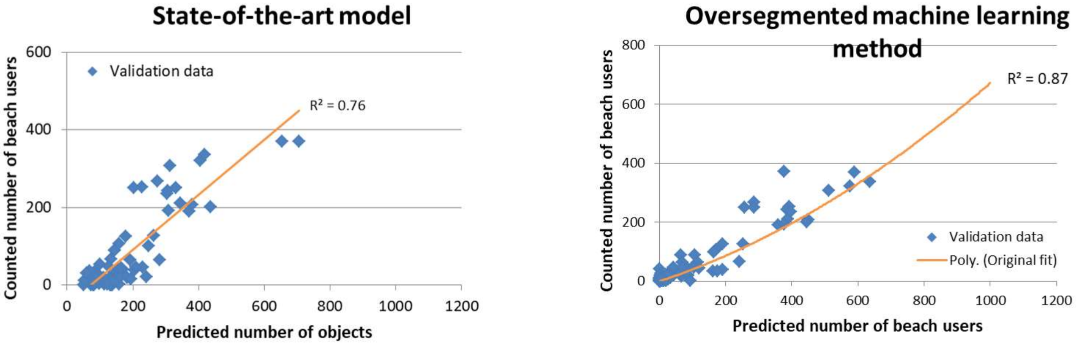


(a) Fit on oversegmented machine learning ROI

(b) Fit on tailored ROI

Figure 5.21: Fits of the state-of-the-art differences in pixel intensity on the snapshots of the validation dataset.

Comparison of the R^2-score obtained with the state-of-the-art model (tailored ROI) and the oversegmented machine learning approach (oversegmented machine learning ROI) shows that the machine learning approach has a significantly better fit on the validation dataset (Fig. 5.22). From this it directly follows that the errors made in the oversegmented machine learning method are smaller than the errors made with the current state-of-the-art model.



(a) Fit state-of-the-art model.

(b) Fit oversegmented machine learning method.

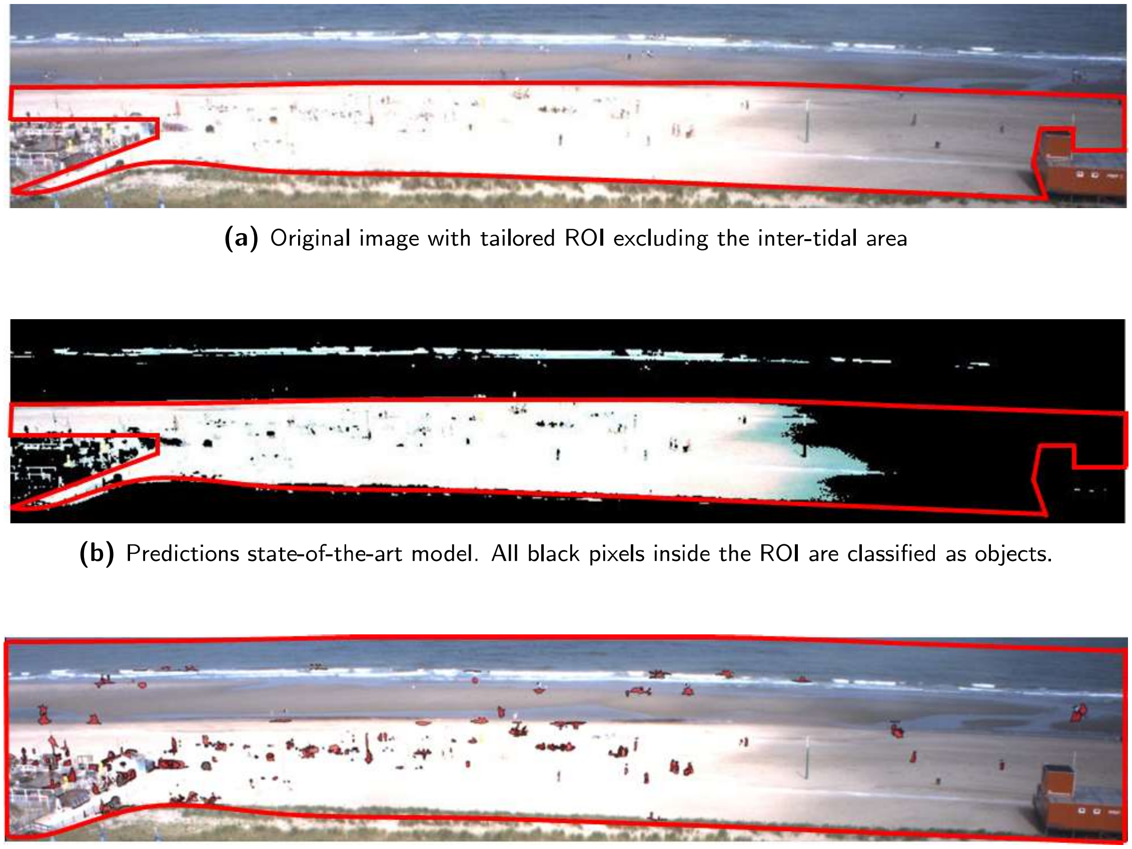
Figure 5.22: Comparison of state-of-the-art and oversegmented machine learning methods fits on the validation dataset that is used for the benchmark.

The increased performance of the oversegmented machine learning method with respect to the current state-of-the-art is supported by the bias and RMSE corresponding to the previously defined bins (Tab. 5.10). The oversegmented machine learning method has significantly lower errors for especially the 0-25 and 100+ bins. The performance on the 25-100 bin is relatively similar indicating that the new model does not necessarily provides better results than the old method in this range. Note that the performance of the oversegmented machine learning method on this bin was limited for the validation dataset due to the unclean lens and unoccupied beach stretchers and was higher on the training dataset.

Table 5.10: Overview of the performance of the state-of-the-art model on the validation dataset adopting the tailored ROI. The values between brackets correspond to the performance of in this study developed oversegmented machine learning method.

	0-25:	25-100:	100+:
Number of images:	38 (38)	21 (21)	19 (19)
Bias:	31.2 (-3.2)	12.5 (-10.3)	-126.6 (-47.5)
RMSE	17.9 (10.2)	26.9 (29.0)	53.9 (56.8)

Analysis of predictions made by the state-of-the-art method shows that this method has a limited performance in darker areas. For the darker area corresponding to the inter-tidal area this can be solved by defining a ROI that excludes the inter-tidal zone, but for dark areas caused by passing clouds (e.g. right side Fig. 5.23b) or bad weather this is not possible and the model over-predicts the number of beach users. The oversegmented machine learning method is less sensitive to darker areas resulting in better predictions (Fig. 5.23c).



(a) Original image with tailored ROI excluding the inter-tidal area

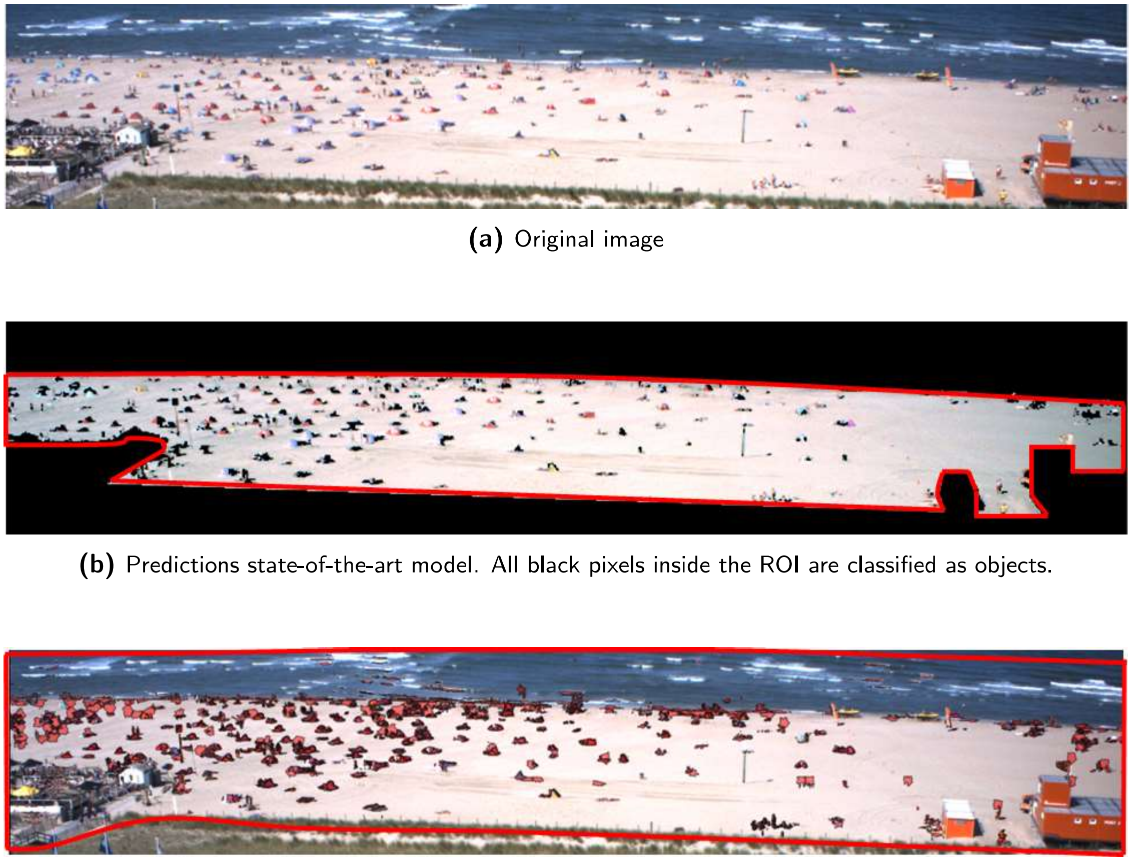
(b) Predictions state-of-the-art model. All black pixels inside the ROI are classified as objects.

(c) Predictions oversegmented machine learning method using the ROI including the inter-tidal area. The predicted beach users are indicated by the red areas inside the ROI.

Figure 5.23: Comparison of state-of-the-art and oversegmented machine learning methods.

The over-predictions of the state-of-the-art model in cloudy conditions are supported by the positive bias in the 0-25 and 25-100 bins. However, the 100+ has a large negative bias and this cannot be explained by the weak performance of the state-of-the-art model in cloudy conditions. Analysis of images from the 100+ bin showed that the state-of-the-art model has a limited detection rate (e.g. Fig. 5.24).

The detection rate in images corresponding to the 0-25 and 25-100 bins is also limited, but in these images the limited detection rate is compensated by the over-predictions due to clouds/bad weather. Clouds/bad weather are less likely on days with relatively high beach occupation possibly causing the negative bias of the 100+ bin. The oversegmented machine learning method also has a negative bias for this bin, but this bias is approximately 3 times smaller (Tab. 5.10)



(a) Original image

(b) Predictions state-of-the-art model. All black pixels inside the ROI are classified as objects.

(c) Predictions oversegmented machine learning method using the ROI including the inter-tidal area. The predicted beach users are indicated by the red areas inside the ROI.

Figure 5.24: Comparison of state-of-the-art and oversegmented machine learning methods.

5.4.1 Conclusions benchmark oversegmented machine learning method with current state-of-the-art

This section treated the result of the comparison between a current state-of-the-art method and the in this study developed oversegmented machine learning method to answer the third research sub-question.

Comparison of the R^2, bias and RMSE corresponding to both methods showed that the oversegmented machine learning method has a higher performance on images corresponding to the Dutch coast. The current state-of-the-art method significantly over-predicts the number of beach users in darker areas caused by for instance clouds or the inter-tidal area, whereas the oversegmented classification method does not. Moreover, the current state-of-the-art method in general under-predicts the number of beach users and this leads to a large negative bias for the 100+ bin.

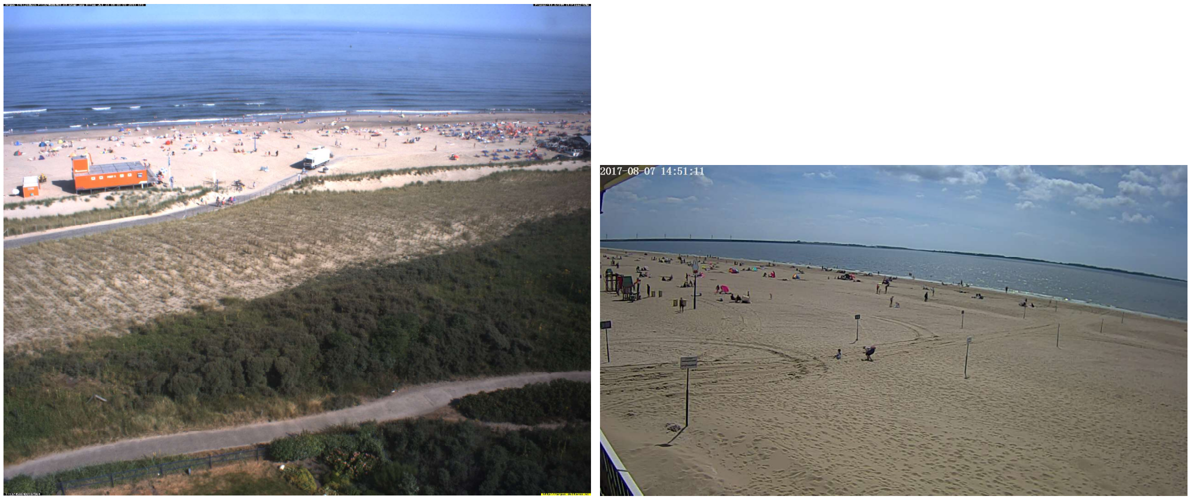
5.5 Performance of oversegmented machine learning method on new camera station

The oversegmented machine learning method trained with Argus images has been applied on images from the webcam camera station in Rockanje to test the applicability of the method on data from a camera station that has not been used during training (research sub-question 4). This section is structured by a subsection that revisits the oversegmented classification model (subsection 5.5.1) and a subsection that treats the regression model (subsection 5.5.2).

5.5.1 Oversegmented classification model applied on a new camera station not used during training

The selected approach for application of the oversegmented machine learning model on images originating from a camera station that was not used during training, did not include additional training with images from the camera station that was not used during training. Therefore, the parameters *'class aggregation'*, *undersampling*, *cost weight balancing* and *regularisation* evaluated during training of the classification model are not revisited. Due to the practical limitations of image enhancement and limiting the number of extracted features, these steps are also not applied. Artificial channels have been added and extracted from the webcam images. The remaining steps that are treated in this subsection are the characteristics of the superpixel grid and feature and channel normalisation.

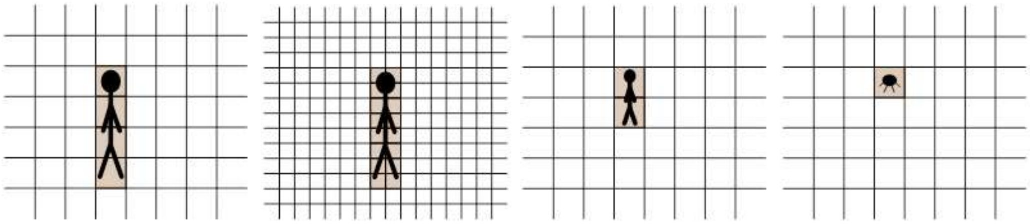
Comparison of the Argus (training) images and the webcam images corresponding to the station not used during training shows that the webcam images are acquired at a relatively low height, closer to the ROI (i.e. beach) and have a relatively oblique angle with respect to the cross-shore (Fig. 5.25). Moreover, the webcam images have a lower resolution than the Argus images.



(a) Example Argus image **(b)** Example webcam image

Figure 5.25: Example of differences between the Argus images used for training of the oversegmented classification model and a webcam image of the Rockanje camera station that was not used during training.

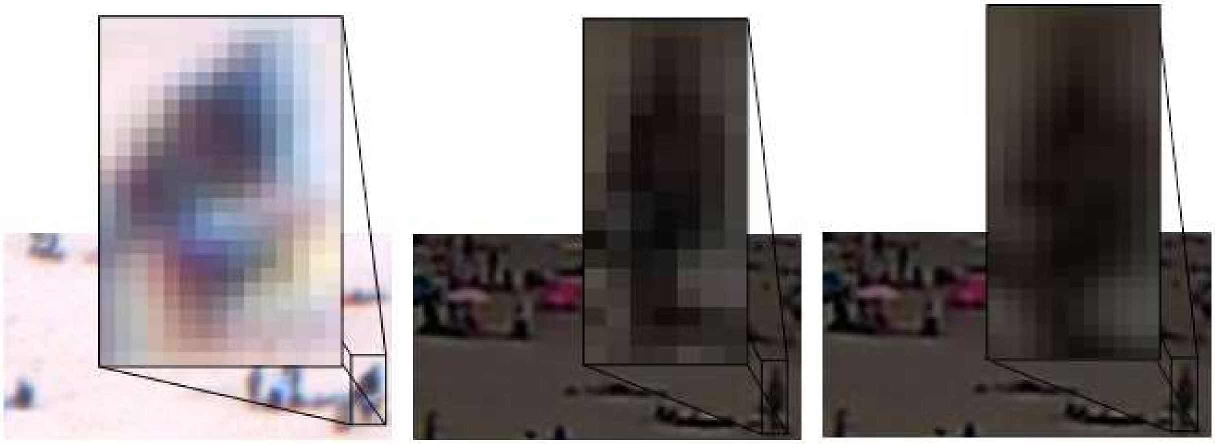
The above described differences can affect the number of pixels that correspond to a beach user (Fig. 5.26). If subsequently a superpixel grid is determined that contains approximately one beach user per superpixel, the number of pixels inside a superpixel differs. Section 4.5.3 indicated that the workflow includes a default normalisation step to account for differences in the number of pixels in a superpixel. However, detailed analysis of this procedure showed that the implemented normalisation is not applied to all features that depend on the number of pixels in a superpixel, possibly resulting in scale-variant features.



(a) Low resolution **(b)** High resolution **(c)** Large distance **(d)** High camera

Figure 5.26: Illustration on impact of image size (resolution), distance of camera to ROI and height of camera on number of pixels that correspond to a beach user.

Scale-variant features can limit the performance of the oversegmented machine learning method on images from a new camera station as a result of differences in the number of pixels in a superpixel. A different number of pixels in a superpixel results in feature relations that are not learned by the model and therefore not recognised during classification. Comparison of the number of pixels that corresponds to an average beach user in both an Argus and webcam image, shows that a beach user in a webcam image consists of a lower number of pixels (Fig. 5.27). This observation can be explained by the difference in image size; images corresponding to the webcam are significantly smaller (1280*720 vs. 2448*2048, subsection 4.2.1).



(a) Argus image. Beach user: ± 15x25 pixels **(b)** Webcam image. Beach user: ± 7x18 pixels **(c)** Re-sized webcam image. Beach users ± 11x33 pixels

Figure 5.27: Indication of number of pixels corresponding to one beach user

The webcam images are re-sized to bring the number of pixels corresponding to a beach user in webcam images in accordance with Argus images. Re-sizing of the webcam is performed based on the (horizontal) size of the Argus images and the aspect-ratio of the webcam images is kept constant. After re-sizing an average beach user in the webcam image corresponds to a number of pixels in the same order of magnitude as the Argus image (Fig. 5.27c).

It should be noted that re-sizing based on image sizes omits the observation that the distance/angle of another camera station can differ from the original camera station causing relatively smaller/larger beach users in images corresponding to the camera station that was not used during training. Therefore, re-sizing based on image dimensions and subsequent application of the same superpixel grid, can still result in superpixels containing a different number of pixels (Fig. 5.28). To this end re-sizing based on the ratio between the number of pixels capturing a beach user in images corresponding to the original and new camera station is considered a more suitable approach and it is recommended to follow this approach in future research. For the current study the ratios between the horizontal image size and beach user size are similar and therefore re-sizing based on the horizontal image size is considered sufficient.

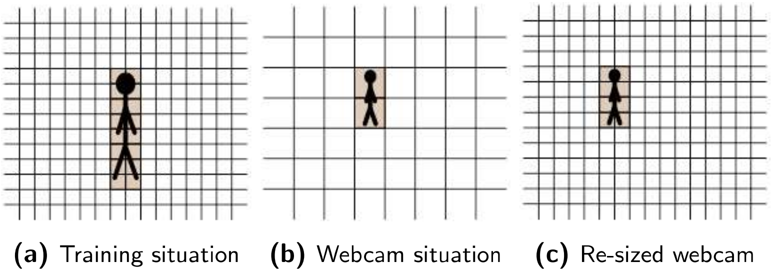

(a) Training situation **(b)** Webcam situation **(c)** Re-sized webcam

Figure 5.28: Illustration on effect of re-sizing based on image dimensions.

The re-sized webcam images have a resolution of 2448*1377 pixels, whereas the Argus images have a resolution of 2448*2048 pixels. This implies that the Argus images contain approximately 1.5 times more pixels than the webcam images and this is important in the specification of the superpixel grid for the webcam images. The ratio between the number of pixels has to be retained in the number of superpixels to obtain superpixels containing approximately the same number of pixels. The superpixel grid corresponding to the Argus images contains 15.000 superpixels and therefore a webcam superpixel grid of 10.000 superpixels is selected.

After the preceding steps, the channels and features corresponding to the superpixels of the webcam images are extracted and classified based on the classification model trained with Argus images. The results show that, although the pixel and superpixel ratios are kept constant, the oversegmented classification model performs unsatisfactory on webcam images (Fig. 5.29). This indicates that at this point the oversegmented machine learning method is not applicable on images from a camera station that was not used during training of the classification model. Moreover, pixel- and superpixel ratios might be of importance but at this point another factor significantly limits the performance. Other factors that can be of importance are treated in the discussion.

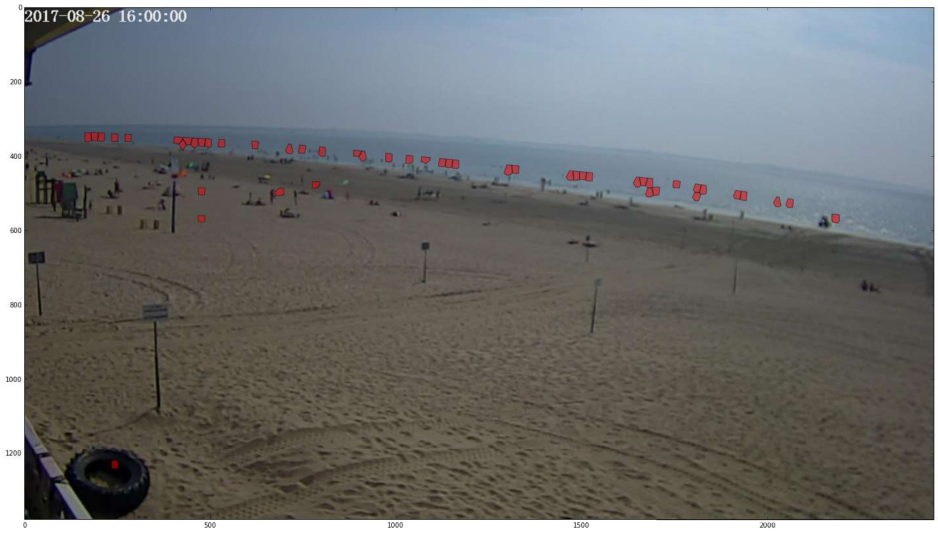


Figure 5.29: Example of classification of webcam image. The red areas correspond to superpixels classified as beach user.

5.5.2 Regression model for a camera station not used during training

The oversegmented classification model did not produce satisfactory classifications and therefore it is at this point not possible to test the effect of a different camera station on the regression relation used to convert the number of classified beach user superpixels to beach users. The variations in camera height, orientation with respect to the cross-shore and distance to the ROI are expected to limit the applicability of the regression model developed with Argus images on webcam images. This indicates that it is expected to be necessary to count a ground truth corresponding to images of the webcam and subsequently determine a new regression relation.

5.5.3 Conclusions

This section treated the application of the oversegmented machine learning method on a camera station not used during training to answer research question 4. It is explained that differences in camera height, camera orientation with respect to the cross-shore, distance from the camera to the ROI and the image size can cause differences in the number of pixels corresponding to a beach user between different stations. This is important because scale-variant features exist and it is showed that image re-sizing and a different superpixel grid can account for this.

Despite re-sizing of the webcam images and a superpixel grid retaining the pixel-superpixel ratios, the classifications made by the oversegmented classification model on webcam images are unsatisfactory. This indicates that the oversegmented machine learning method is at this point not applicable on camera stations not used during training.

6 Discussion

The discussion starts with a section on the scoring of the oversegmented machine learning method, reflecting the implications of the PR- and R^2-scores observed in the results on the first two research questions. Subsequently the results of the benchmark are discussed followed by a section on the possible causes of the limited applicability of the oversegmented machine learning method on a camera station that has not been used during training. Next the observed limitations of the oversegmented machine learning method are discussed and this chapter is finalised by a case-study that indicates the possibilities of the oversegmented machine learning method.

6.1 Scoring of oversegmented machine learning method

In section 4.7 the PR-score was selected for the scoring of the oversegmented classification model based on its theoretical suitability on imbalanced datasets. Initially the impact of small changes in PR-score was not known and therefore regression models and corresponding R^2-scores were determined for multiple classification models. This section elaborates on the impact of small changes in PR-score and subsequently discusses the relation between the observed PR- and R^2-scores. Lastly, the proposed use of the PR- and R^2-scores for quantification of the performance of the oversegmented machine learning method is treated.

6.1.1 Implication of small changes in PR-score

The results corresponding to the first research questions show that especially changes in the aggregation (e.g. beach object aggregated separately or together with beach user and swimmer), adding additional channels (and/or removing the 'ROI' superpixel) and the number of images can result in relative large differences in the performance of the oversegmented classification method. Application of undersampling and cost weight balancing to account for the imbalance in the dataset had limited effect although the results show that it is possible to lower the PR-score.

Regarding these small changes in PR-score, it can be questioned to what extent it is valuable to optimise the score by making small adjustments to the parameters. An indicator of the relative importance of small changes is the variation in the PR-score as a result of changing the train- and test partitions of the *same* model (Fig 6.1). The figure indicates that dependent on the chosen train- and test partitions the same model can have a PR-score varying from approximately 0.55 to 0.49 with an average around 0.52. Based on this result it is doubt-full whether changes of the PR-score in the order of 0.03 are valuable, because these can already been obtained by defining the train- and test partitions differently. However, note that despite small changes in PR-score the distribution of precision and recall can differ and this might be of importance.

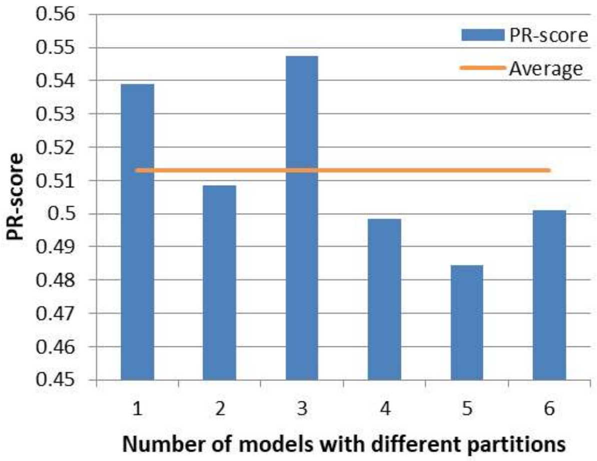


Figure 6.1: Graph indicating the spread in PR-score among different train- and test partitions of the same oversegmented classification model.

6.1.2 Relation PR-score and R^2

To obtain more insight in the impact of minor differences in PR-score on the performance of the oversegmented classification models, regression models corresponding to multiple oversegmented classification models with different characteristics were determined. Comparison of the R^2-scores corresponding to the developed regression models indicated that the oversegmented classification model with the lowest (and thus optimal) PR-score does not necessarily lead to the highest R^2 (section 5.2.1). Moreover, remarkable was that the highest R^2-scores have been observed for classification models with a high recall and this indicates that the distribution of precision and recall is of larger importance than the combined PR-score.

The difference in best performing oversegmented classification model based on respectively the PR-score or R^2 can be explained by the characteristics of these metrics. The PR-score provides information on the relative number of annotated beach user superpixels in the *whole* test dataset that have been classified correctly. The R^2 resulting from the regression model is a measure of the minimised error on *individual* images.

For classification problems in which the class of interest is almost equally present in each image (e.g. the classification of sand in a coastal image) the PR-score and R^2 are likely to be in accordance, because the underlying changes in true positives, false positives and false negatives causing the PR-score to change, hold more or less for all images resulting in a similar effect on the R^2. However, in the case of beach user quantification this is no longer valid, because the number of beach users shows significant variations among different images. A higher PR-score can therefore result in a situation where specific images gain a lot of correctly classified superpixels, whereas other images perform less. The PR-score is not sensitive to the performance on individual images, whereas the R^2 is and this might cause the observed discrepancy between the PR-score and R^2.

6.1.3 Proposed use PR-score and R^2 for quantification of performance oversegmented machine learning method

The previous section indicates that the R^2 provides more information on the error corresponding to individual images with respect to the PR-score. This is desirable due to the large variability in the number of beach users between different images. However, the R^2 does not provide information on the physical correctness (e.g. true positives, false positives etc.) of a prediction and this can be important. The distribution of precision and recall corresponding to the PR-score can provide this information and therefore it is recommended to score models based on a combined evaluation of R^2 and PR-score with its underlying distribution of precision and recall.

6.2 Benchmark oversegmented machine learning method with current state-of-the-art method

The results on the third research question clearly show an increased performance of the oversegmented machine learning method with respect to the current state-of-the-art model expressed by a higher R^2-score and lower error statistics. The R^2-score of the original study of Guillén et al. (2008) was in the range of the newly developed oversegmented machine learning method and the difference in R^2-score between the original study and the benchmark performed in this study, can be explained by the higher diversity of conditions that is taken into account in the current study.

The results showed that in general the current state-of-the-art model under-predicts the number of beach users, but that this can be over-compensated due to dark areas in an image. The under-predicting characteristic of this model is supported by the original regression relations used in Guillén et al. (2008) to convert the predicted number of objects to the number of beach users. Two relations corresponding to two different beaches are presented and both increase the number of predicted objects (coefficients 2.94 and 4.67) to obtain the number of beach users.

Due to the presence of a relatively high number of images with less favourable conditions in this studies validation dataset, the regression model compensates for the over-predicts by a coefficient smaller than 1. This limits the errors made during classification of images with dark areas, but amplifies the error in images without dark areas. The contradiction between these situations indicates limited robustness of the current state-of-the-art method on the variable conditions associated with the Dutch coast. The oversegmented machine learning method performs well in these conditions, indicating that the oversegmented machine learning method can cope with the diversity observed on the Dutch coast and has the required robustness.

However, the benchmarked state-of-the-art method is developed based on Timex images, whereas snapshots have been used in this study. Although the difference is expected to be minor, a benchmark on Timex images is recommended to exclude the possibility that the limited performance of the old model is induced by the difference between Timex and snapshots.

6.3 Possible causes limited applicability on camera stations not used during training

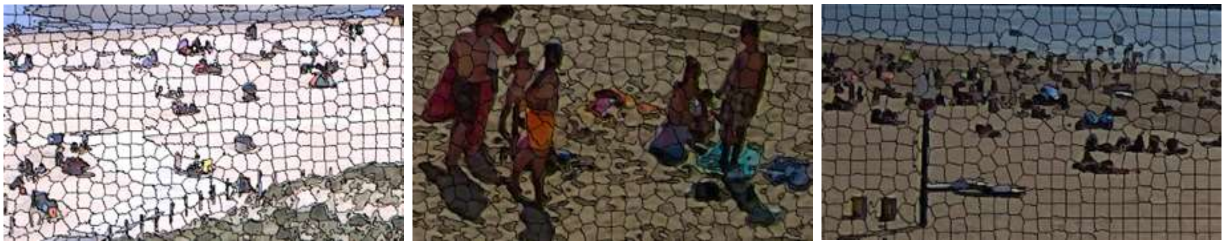
The results on research sub-question 4 showed that at this point it is not possible to apply the oversegmented machine learning method on images from a camera station that was not used during training. Re-sizing of the (webcam) images from the camera station not used during training did not lead to satisfying results and therefore other aspects have to be of importance. This section elaborates two possible causes:

- Variation in relative size beach users
- Variation in feature strengths/ranges

The possible causes are treated separately.

6.3.1 Variation in relative size beach users

In the results a re-sizing procedure has been described to bring the number of pixels corresponding to an *average* beach user in the training (Argus) images and (webcam) images from a new camera station in accordance. However, due to differences in camera height, orientation with respect to the cross-shore and distance from the ROI, the size of beach users varies relatively more in the webcam images. Especially beach users close to the webcam can be relatively large and described by multiple superpixels (comparison Fig. 6.2a and 6.2b).



(a) Snap Argus (training data) **(b)** Snap webcam (close to camera) **(c)** Snap webcam (larger distance)

Figure 6.2: Indication of number of superpixels describing a beach user.

Beach users in the training data have a relatively constant size and large beach users captured by multiple superpixels are not present. This might cause errors during classification of webcam images, because the oversegmented classification model has to classify a situation that was not learned. However, Fig. 6.2c shows that webcam images also contain areas with beach users of a size comparable to the Argus images used for training of the oversegmented classification model. Because of the similarity in beach user size these type of areas are expected to be classified correctly, but the results are unsatisfactory (e.g. Fig. 5.29).

The above indicates that the variation in size of beach users is unlikely to be the main cause of the limited performance of the oversegmented machine learning method on images originating from camera stations not used during training. Nevertheless, the variation in beach user size might complicate the conversion from classified beach user superpixels to beach users, because the number of superpixels capturing a beach user can vary between camera stations. This moreover supports the the expectation that new regression relations might be necessary for different camera stations as was already mentioned in section 5.5.2.

6.3.2 Variation in feature strengths/ranges

Besides variation in camera height, orientation with respect to the cross-shore, distance to the ROI and images size, the type of camera and corresponding quality/characteristics may vary between stations. This variation can induce differences in feature strengths/ranges for different camera stations. Because the oversegmented classification model has been trained by the feature strengths/ranges corresponding to one camera station, it can have a limited performance on images from a camera station with different feature strength/ranges that are not included during training.

Comparison of the Argus images used during training and the webcam images corresponding to a station not used during training indicates that in general the webcam images are darker (Fig. 6.3). Moreover, closer analysis of the superpixel shapes shows that the shape of the Argus superpixels are relatively jagged compared to the webcam superpixels. These aspects can differ the feature values that are extracted from the superpixel grid and this might cause the unsatisfactory results on webcam images.

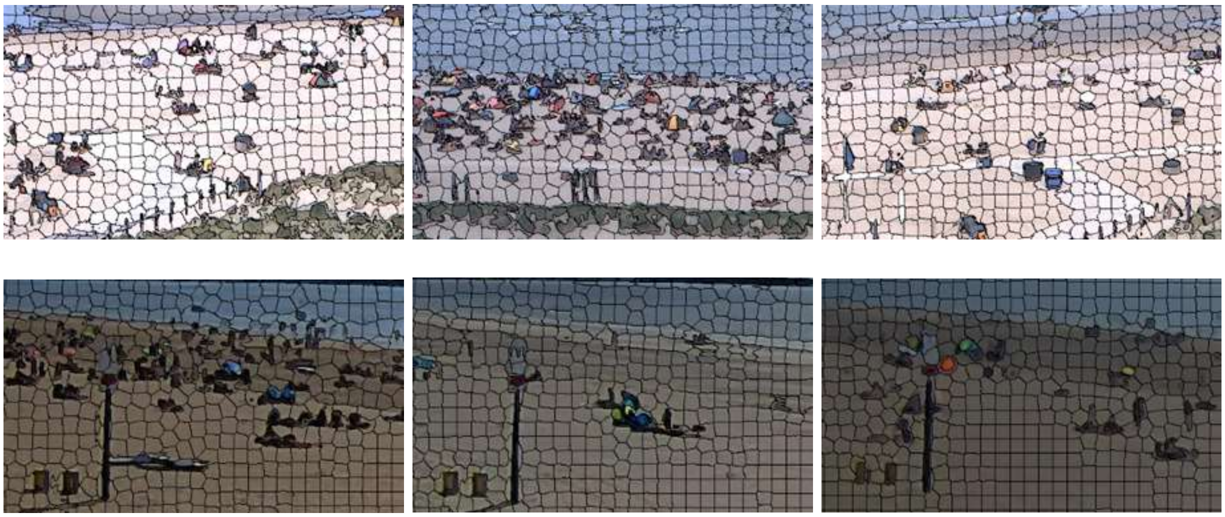


Figure 6.3: Comparison of Argus images (top row) used for training and webcam images (bottom row) of the camera station not used during training. The images correspond to different light conditions.

Two approaches are proposed to account for the differences between the Argus and webcam images. The first approach entails training of a new oversegmented classification model based on an extended training dataset including images originating from mulitple camera stations. This approach assumes that the existing training dataset does not include enough variety regarding differences caused by variation in camera stations and focuses on including more variety. The second approach applies image enhanced with Contrast Limited Histogram Equalisation (CLAHE) to the images of both camera stations, to possibly increase the accordance regarding colour and superpixel shapes (Fig. 6.4).

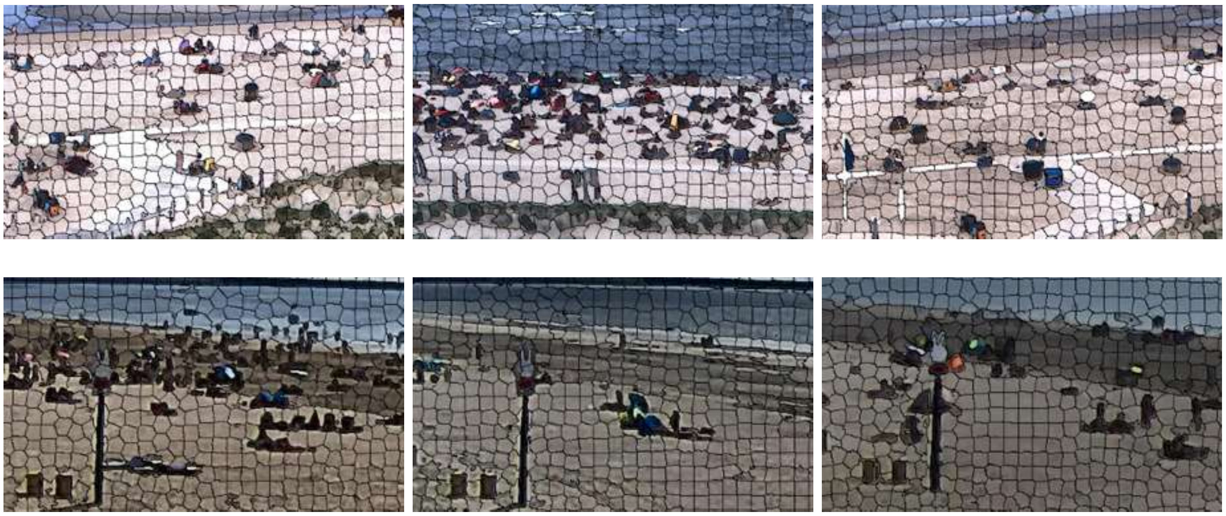


Figure 6.4: Comparison of enhanced Argus images (top row) used for training and enhanced webcam images (bottom row) of the camera station not used during training. The images correspond to different light conditions.

Visual comparison of Fig. 6.3 and Fig. 6.4 indicates that the differences between Argus and webcam images reduce after enhancement. However, most feature relations are abstract and it is therefore difficult to predict the final performance based on visual inspection. Moreover, section 5.3.2 showed that the effects of CLAHE-enhancement are not necessarily positive. Beneficial of the application of enhancement compared to extension of the training dataset is that no additional superpixel annotation is required. Therefore, it is recommended to first test the effect of applying enhancement and subsequently investigate extension of the training dataset by annotating webcam images.

6.4 Limitations of oversegmented machine learning method

Application of the oversegmented machine learning method on the validation data showed two limitations: images with a lot of unoccupied (rental) beach stretchers and images captured by an unclean lens. Both are treated separately in this section and moreover a third limitation is elaborated.

6.4.1 Unoccupied beach stretchers

Analysis of the positive errors in the dataset used for validation of the regression model showed that the positive errors are especially caused by unoccupied (rental) beach stretchers on the beach and this is considered a limitation of the oversegmented machine learning method. An explanation of this limitation is the combination of the selected class aggregation and the choice for regression models with a zero intercept. To explain this first a possible cause of the negative intercept is elaborated. Subsequently, the effect of a zero intercept in combination with the selected class aggregation is treated and two possible solutions are proposed.

Possible cause of negative intercept
In this study regression models with a zero intercept have been used to exclude negative beach user predictions and this has effect on especially days with minor beach attendance. In general a negative intercept will be the result of superpixels that are systematically classified as beach user while they are not (e.g. wrongly classified litter bins). Visual inspection of the images used for training showed that the appearance of a negative intercept in the regression relations might be caused by the presence of rental beach stretchers. The class beach object (e.g. stretchers) is aggregated with the class beach user in the final aggregation and therefore stretchers are predicted as beach users. As a result of the foregoing, the model classifies too much beach users and the regression model has to correct for this.

As long as the number of stretchers is dependent on the number of beach users, this can be fitted by a regression line without an intercept because a direct relation exists. However, the images used for training contain a beach club that is usually surrounded by numerous rental beach stretchers. Visual analysis showed that the presence of this type of stretchers is not necessarily dependent on the number of beach users because the stretchers are often limited occupied. The limited dependency between rental beach stretchers and beach users requires a negative intercept.

While on most days the rental stretchers are present, they are not placed on the beach for days with less nice weather. In general days with less nice weather correspond to a limited beach occupation. The combination of no rental stretchers and a limited beach occupation can cause a negative intercept to result in a negative number of classified beach users and this is physically impossible. Therefore, a zero intercept has been selected in this study.

Effect of zero intercept in combination with selected class aggregation
Although a zero intercept ensures that the oversegmented machine learning method will not return a negative number of beach users on days with less nice weather, it can also lead to a method that classifies too much beach users. Minor beach occupation is not only found on days with less nice weather as this situation also occurs at the before mentioned early morning of nice days (section 5.2.4). At these days the rental stretchers are already placed on the beach at this time of day, whereas the

beach occupation is still limited. In those conditions the model will classify most of the stretchers as a result of the selected aggregation that merges the beach object class to the beach user class. However, in reality only a limited number of beach users is present on the beach at this early time of day, resulting in a high (positive) error. Dependent on the magnitude of the negative intercept, a negative intercept could have corrected this error. On days with moderate to high beach occupation the zero intercept is considered to have a relatively limited impact due to the higher number of beach users. Moreover, the rental stretchers start to be occupied, restoring the relation between beach users and stretchers.

The trade-off between negative predictions (less nice days, negative intercept) and relatively high positive errors (early morning nice days, zero intercept) is considered to be the result of the desire to develop and apply a model in a variety of conditions. The variety in conditions causes the number of unoccupied stretcher to differ among different images, complicating the regression relation because the relation between stretchers and beach users is not constant. Deriving a calibration relation that is (better) capable to express this variability would directly increase the model performance and it is recommended to investigate this in further research.

Another option is to train an oversegmented classification model with separate beach user- and beach object classes. This approach theoretically limits the classification of beach stretchers as beach users in images corresponding to the mornings of days with nice weather. However, separate beach user and beach object classes can cause other problems like for instance difficulties to distinguish between beach users and beach objects or occlusion of beach users by beach objects. This is supported by the results on the performance of class aggregations with combined/separate beach user and beach object classes (section 5.1.1. It is recommended to investigate the effect of combining/separating the beach user and beach object classes in more detail.

6.4.2 Unclean lens

Another limitation observed during analysis of the performance of the oversegmented machine learning model on the validation dataset was the performance on images captured by an unclean lens. Regarding this limitation it becomes questionable whether coping with an unclean lens can be considered part of robustness or that the lens should be treated as a boundary condition that can be expected clean. The former could hold as long as individual beach users can be indicated in an image. Moreover, an important aspect is that enough images captured by an unclean lens should be present in the training data. Otherwise the model will not learn how to distinguish beach users in images captured by an unclean lens.

Comparison of the images in the training dataset with the images of the validation dataset showed that images with a unclean lens are also present (and causing large errors) in the original data but that the relative contribution to the total is limited. In the validation data a large part of the images is captured by an unclean lens causing bad predictions. Because the number of images captured by an unclean lens is limited in the training data the results on the relatively unclean validation images should be put in the right perspective. Regarding this it is recommended to either adjust/extend the training data with images captured by an unclean lens or create a new validation set with relatively clear images. If the latter option is selected, it is moreover recommended to investigate the possibilities to develop a quality control algorithm capable of detecting images captured by a unclean lens

6.4.3 Situations not included during training of the oversegmented classification model

The limitation as a result of an unclean lens showed that the performance of the oversegmented machine learning method can reduce for situations that were not/limited present during training of the oversegmented classification model. This type of limitation is considered exemplary for a method using a trained classification model; the classification model is capable to classify instances that it is trained on but will have a limited performance on situations that were not/ limited present during training.

Besides the limitations on images with an unclean lens, the limited applicability on camera stations not used during training might be an example of this type of limitation (section 6.3.2). Moreover, the training data contains only data spanning from 10.30h to 16.30h local time and this implies that no images during sun rise and sun set are included. The absence of images around sun rise/sun set in the training data might cause troubles when analysing images around these time periods as a result of changing light conditions not observed during training.

6.5 Case-study

The oversegmented machine learning method has been applied on a case-study to show the possibilities of the oversegmented machine learning method in practice. The case-study treats a dataset containing images corresponding to the months June, July and August of the summer of 2013. The images in the dataset are acquired at 10.30h, 13.30h and 16.30h local time with camera 4 of the Kijkduin Argus station (subsection 4.2.1). Note that the case-study dataset contains images that are also present in the dataset used for validation of the regression model. Evaluation of the validation data showed multiple outliers due to an unclear lens and unoccupied beach stretchers and therefore the same outliers can be expected in the case-study dataset.

The case-study treats application of the oversegmented machine learning method to quantify the *number* of beach users and to indicate the *intensity* of beach users.

6.5.1 Application number of beach users

Use of the oversegmented machine learning method to quantify the number of beach users entails classification of the images with the oversegmented classification model and subsequent conversion of the number of classified beach user superpixels to beach users with the regression model. The obtained number of beach users correspond to a beach area of approximately 13.000m^2. The remainder of this subsection treats the variation in the number of beach users on a monthly-, weekly- and daily- timescale.

Monthly variation

The monthly variability is determined by evaluation of the number of beach users in images captured at 13.30h (Fig. 6.5).

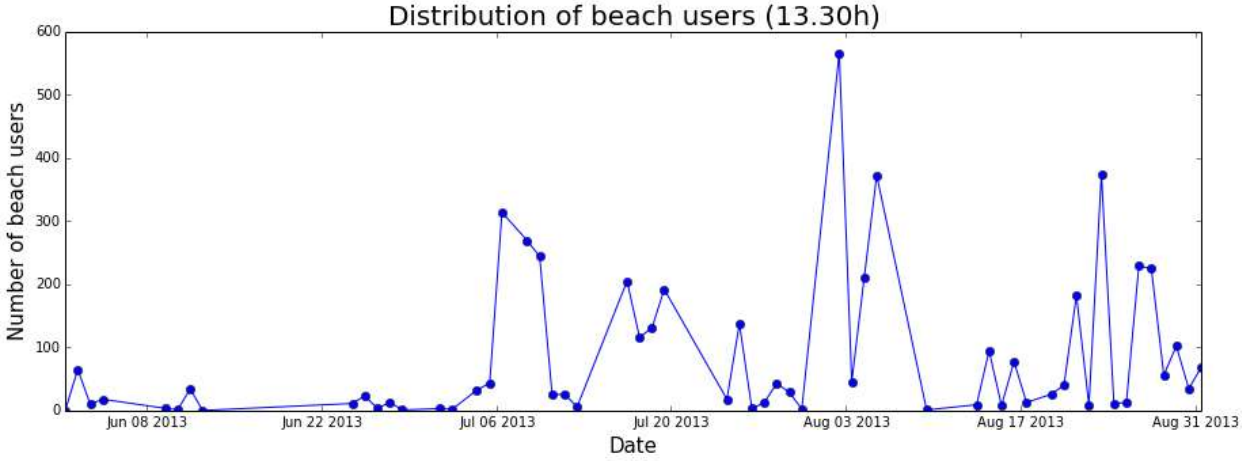


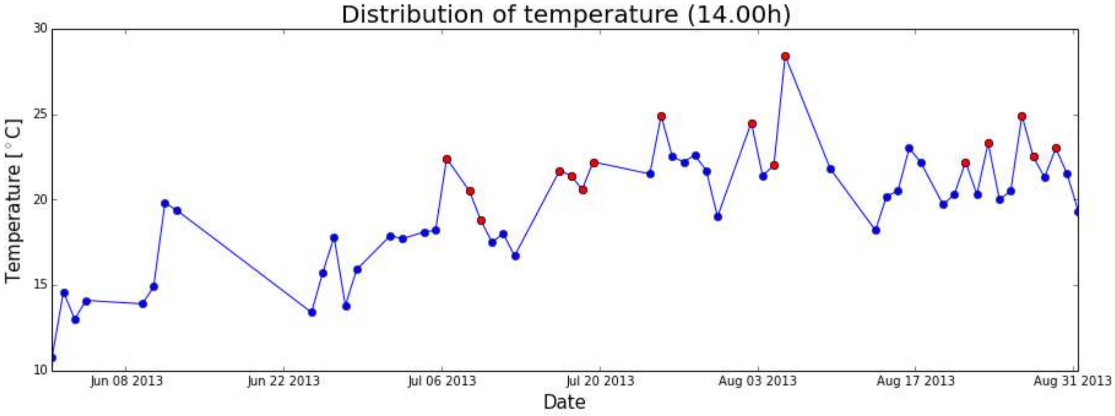
Figure 6.5: Distribution of the number of beach users.

The graph shows that the beach occupation fluctuates significantly during the observed period and this directly indicates the advantage of a method with a high temporal resolution. Unfortunately a relatively large part of June is absent in the dataset and this explains the limited number of data points in this month (Fig. 6.5). Quantitative analysis of the data shows that the maximum number of beach users in the summer months of 2013 is 566, whereas the minimum number of beach users is 0 (Tab. 6.1). Comparison of the maximum- and average number of beach users indicates an increase in beach occupation in July and August with respect to June.

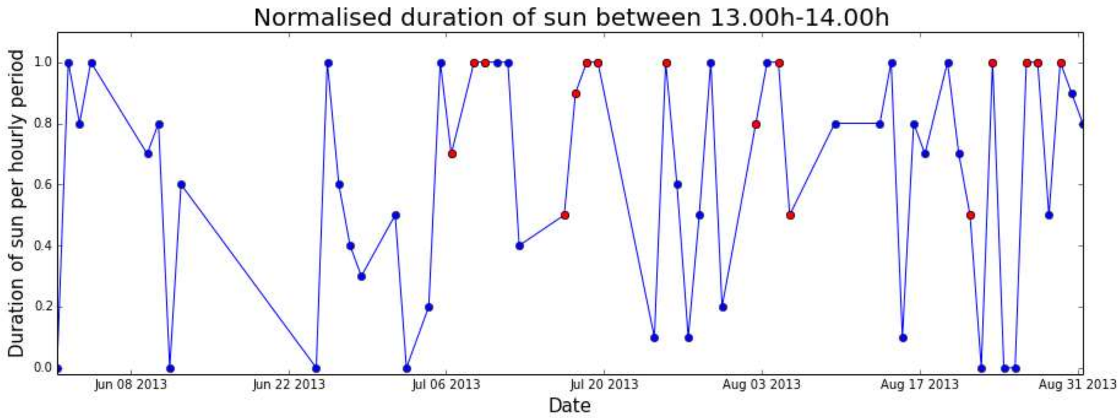
Table 6.1: Variation of the number of beach users corresponding to the months June, July and August of the summer of 2013

	June:	July:	August:
Number of evaluated images:	13	21	23
Maximum number of beach users:	65	314	566
Minimum number of beach users:	0	2	1
Average number of beach users:	14	88	120

The observed differences in beach occupation between June, July and August might be explained by weather data corresponding to these months (Fig. 6.6). Analysis of the temperature shows that the temperature was relatively low in June and higher in July and August possibly explaining the increased beach occupation in July and August (Fig. 6.6a). Moreover, the red dots indicate that days with relatively high beach occupation (i.e. 100+ beach users) correspond to peaks in the temperature distribution.



(a) Distribution of temperature. The red dots correspond to images with 100+ beach users.



(b) Duration of sun per hour. The red dots correspond to images with 100+ beach users.

Figure 6.6: Weather conditions during June, July and August in the summer of 2013. Data of the weather conditions correspond to the Hoek van Holland weather station and are obtained from Koninklijk Nederlands Meteorologisch Instituut (2013).

It is remarkable that days with the same temperature do not necessarily result in similar beach occupation and this might indicate that temperature is not the only important factor. Besides temperature, sun duration (i.e. clear sky without clouds) can for instance be of importance because beach occupation is expected to be higher for days with a relatively high sun duration. However, as for temperature, days with comparable sun duration do not necessarily result in similar beach occupation and this supports the hypothesis that beach occupation depends on multiple factors (Fig. 6.6b). Analysis of the correspondence between days with a relatively high beach occupation and combined high temperature and/or high sun duration, indicates that high beach occupation is indeed dependent on both temperature and sun duration (Fig. 6.7). Nevertheless, also for the combination of temperature and sun duration, similar conditions do not necessarily result in comparable beach occupation and this indicates that temperature and sun duration are not the only important factors.

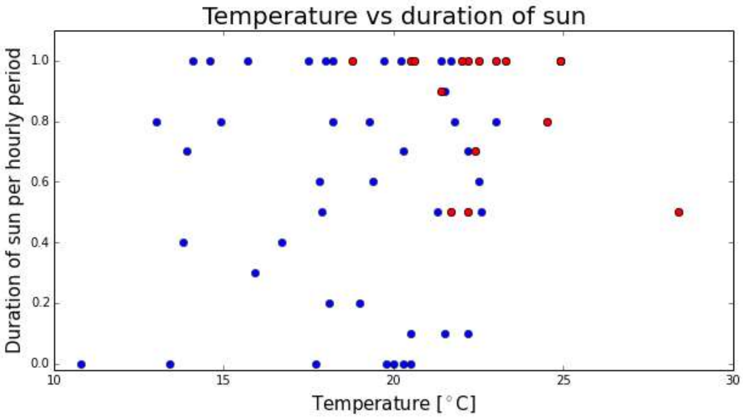


Figure 6.7: Plot of temperature at 14.00h versus the normalised sun duration in the preceding hour (13.00h-14.00h). The red dots correspond to images with 100+ beach users.

Weekly variation

Besides variations in weather conditions another aspect that might be of importance are the days of the week. Beach occupation is expected to be higher during the weekends because working people do not have to work. Evaluation of the (limited number of) Saturday and Sunday (13.30h) images present in the case-study data shows that high beach occupation does not necessarily correspond to weekend days (Fig. 6.8). This can be an indication that beach users during the evaluated period of the year are predominantly tourists, because high beach occupation during the week is unlikely to be caused by working people.

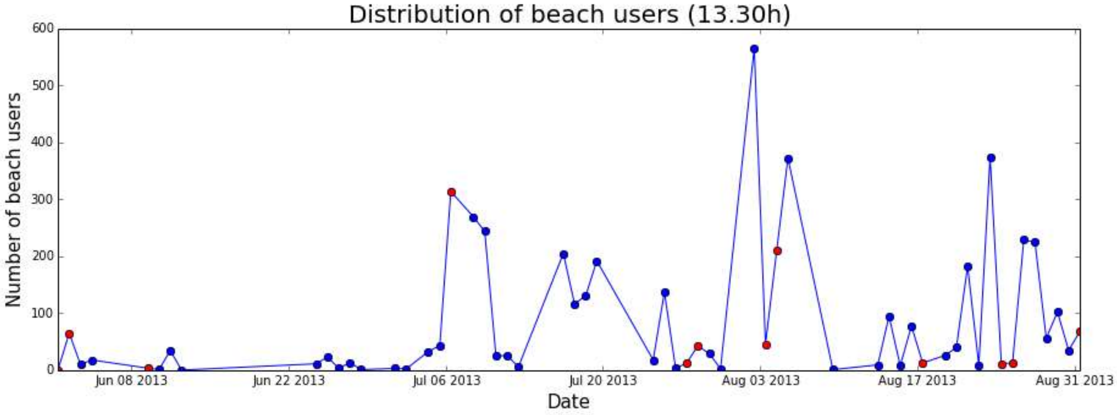


Figure 6.8: Distribution of beach users. The red dots indicate Saturdays and Sundays.

Daily variation

The beach occupation at 10.30h, 13.30h and 16.30h local time can be evaluated to obtain insight in the distribution of the number of beach users during the day (Fig. 6.9). Only a limited number of days are evaluated due to missing data points on either the 10.30h or 15.30h point of time in the other days.

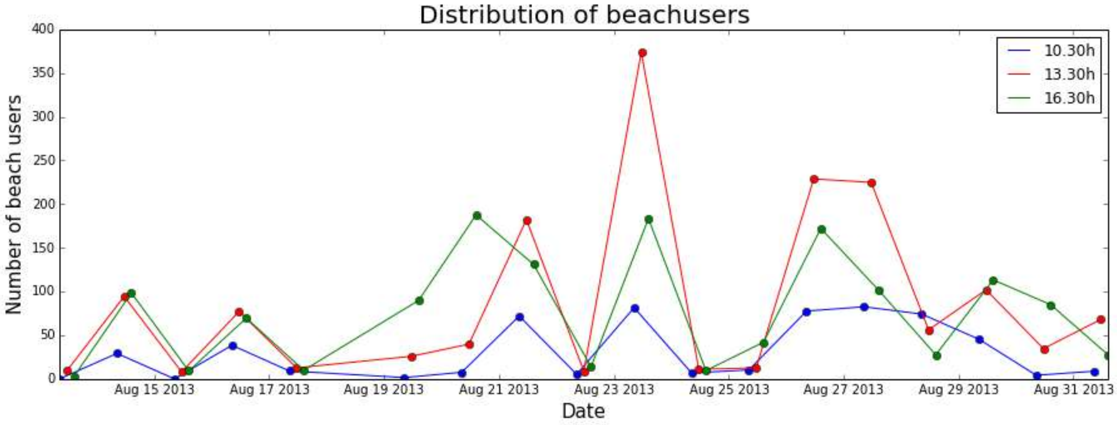


Figure 6.9: Distribution of beach users over the day

Evaluation of the distribution of the number of beach user during the day shows that in general the beach occupation is the lowest at 10.30h, peaks at 13.30h and decreases again at 16.30h (Fig. 6.9). Moreover, the difference in beach occupation shows larger variations on days with relatively high beach occupation compared to days with low beach occupation. Quantification of the the number of beach users supports the increase of beach users between 10.30h-13.30h and the decrease between 13.30h-15.30h (Fig. 6.2). The data moreover shows that the increase between 10.30h-13.30h is relatively large compared to the decrease between 13.30h and 15.30h.

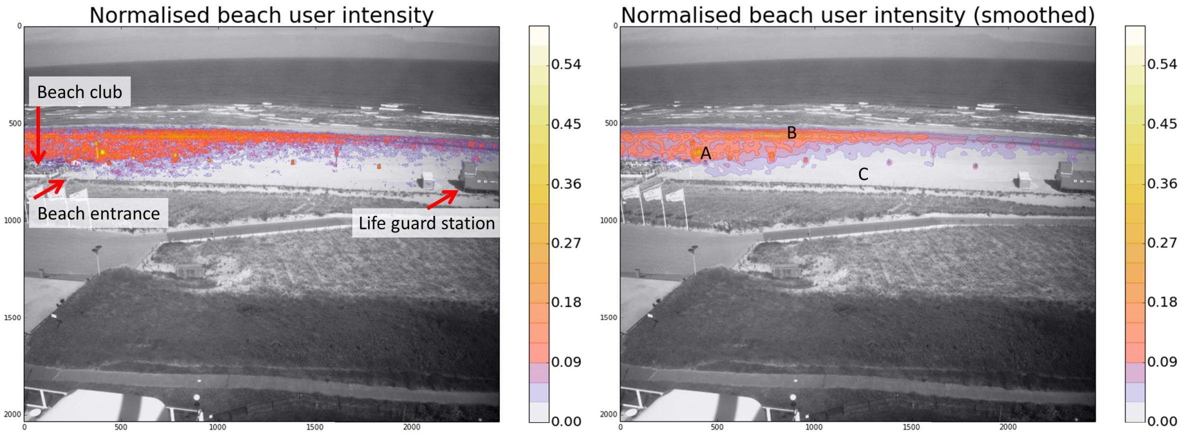
Table 6.2: Variation of the number of beach users during the day

	10.30h:	13.30h:	15.30h:
Number of evaluated images:	18	18	18
Maximum number of beach users:	82	374	188
Minimum number of beach users:	0	8	3
Average number of beach users:	31	87	76

6.5.2 Application intensity of beach users

The oversegmented machine learning method can be used to determine and visualise the intensity of beach occupation on a beach by a heat map. A heat map expresses the differences in intensity by different colours. In this study heat maps are constructed by classifying images with the oversegmented classification model. Subsequently, for each pixel the number of images in which it is part of a classified beach user superpixel is counted. Thereafter the counted number of times a pixel corresponds to a classified beach user superpixel, is divided by the number of evaluated images resulting in the percentage of evaluated images in which a pixel is classified as beach user. A colour bar provides information on the relation between colours in the map and the percentage of evaluated images in which a particular pixel was part of a beach user superpixel.

The resulting heat map is relatively scattered due to the appearance of coloured superpixels for especially the lower percentages (Fig. 6.10a) and therefore the final heat map is smoothed (Fig. 6.10b).



(a) Heat map without smoothing

(b) smoothed heat map

Figure 6.10: Spatial distribution of beach users constructed from all 160 images in the case-study data. The colour bar indicates the relation between a colour and the corresponding percentage of images a pixel was part of a classified beach user superpixel. The values on the axis correspond to image coordinates.

The smoothed heat map shows that most beach users are present around the beach club and near the waterline (Fig. 6.10b). The intensity peaks at locations A (close to beach club) en B (close to waterline) with classified beach users in approximately 36% of the evaluated images, whereas in region C (further from beach club and waterline) classified beach users are only encountered in 0-3% of the images. The relatively high beach user intensity near the beach club might indicate that beach users prefer to recreate close to facilities associated with a beach club. The observation of high beach occupation near facilities is supported by the study of Trygonis et al. (2015) who describe a disproportionate spatial distribution of beach users as a result of closely located tourist facilities.

Another explanation of the difference in spatial distribution of beach users in along shore direction (difference A and C) can be the location of a beach entrance next to the beach club and the tendency of beach users to recreate close to where they enter the beach. This explanation is supported by Jiménez et al. (2007) who observed relatively high beach occupation close to the beach entrance.

The relatively high beach occupation close to the waterline (location A), with respect to the beach occupation at larger cross-shore distances from the waterline (location C), indicates that beach users prefer to recreate close to the waterline. Similar observations are mentioned in other studies on the quantification of the beach occupation (e.g. Kammler and Schernewski (2004)).

6.5.3 Conclusion case-study

The case-study applied the oversegmented machine learning method to analyse the number- and intensity of beach users during the months June, July and August of the summer of 2013. Evaluation of the number of beach users showed that the temporal distribution of beach users varies between the months but also during the day. Possible explanations for the variation in beach occupation are changes in temperature and sun duration, but these factor cannot explain all the variation.

Analysis of the beach user intensity showed that the spatial distribution of the beach user intensity can differ in both the cross-shore and along-shore direction. Possible causes of these changes can be the location of facilities, beach entrances and the preference of beach users to recreate close to the shore line.

In all, the case-study shows that beach occupation varies relatively much in both the temporal and spatial distribution. The case-study moreover demonstrates the capability of the oversegmented machine learning method to quantify the beach occupation and express this variation on a beach along the Dutch coast. The combination of the captured variation in beach occupation and possible explanations for these variations, indicates aspects that can be used/changed by coastal zone managers to maintain/improve their beaches.

7 Conclusions

The objective of this study was to develop/improve and test a method for accurate, robust and automatic quantification of the spatial- and temporal distribution of beach users on the Dutch coast. After a literature review into possibilities to detect humans, an oversegmented machine learning method was selected.

The oversegmented machine learning method was not applied on beach user quantification before and therefore an extensive analysis of possibilities to adapt or add steps to the original workflow is performed. **The oversegmented machine learning method proved to be a suitable method for the quantification of beach occupation on the Dutch coast indicated by a R^2 of 0.87**. The method is developed with a training dataset containing 76 images and subsequently validated with 80 new and unseen images to derive the final performance. The development and determination of performance of the oversegmented machine learning method has been structured by four research sub-questions. The answers to these sub-questions are presented in the next four paragraphs:

Q1. Which parameters have to be adapted or added to make the oversegmented machine learning method capable to classify beach users?

The analysis of parameters that have to be adapted or added to the original workflow showed that *class aggregation*, *artificial channels* and *number of images in the training dataset* are the parameters with the highest impact. Based on the adopted score for quantification of the classification model (PR-score) the effects of the measures *undersampling* and *cost weight balancing* to account for the imbalanced dataset are limited. However, analysis of the corresponding Precision - Recall curve showed that both measures can change the distribution of precision and recall which respectively express the ratio between true positives/false negatives true positives/false positives.

Analysis showed that differences in precision and recall result in classification models with different characteristics regarding R^2-score and true- and false positive rates and therefore different classification models might be preferred for different applications. Optimisation with CLAHE *image enhancement* decreased the error corresponding to approximately 50% of the tested images but led to a number of large outliers and therefore image enhancement cannot be implemented as a generic step in the workflow. Lastly, it proved to be possible to significantly *reduce* the *number of extracted and evaluated features* without limiting the performance, but due to practical limitations this cannot be implemented at this point.

Q2. What is the most suitable approach to convert classified beach user superpixels to the number of beach users?

In this study the most suitable approach to convert classified beach user superpixel to beach users is investigated. From the results it is concluded that a regression model directly relating the classified *number* of beach user superpixels to the manually counted ground truth is preferred over a regression model relating the classified *area* of beach user superpixels to the counted ground truth. After development of the regression model, it is applied on a dataset that was not used during development to validate the fit. Comparison of the R^2 corresponding to the dataset used for development of the regression model and a validation dataset, indicated that the regression model has general applicability on images that are not used during development of the regression model (R^2=0.87). Evaluation of the validation dataset showed limitations of the oversegmented machine learning method on images with unoccupied (rental) beach stretchers and an unclean lens.

Q3. How does the oversegmented machine learning method perform in comparison to the current state-of-the-art in the variable conditions associated with the Dutch coast?

A benchmark between the performance corresponding to the oversegmented machine learning method and a differences in pixel intensity model representative for the current state-of-the-art, showed that the R^2 of the oversegmented machine learning method is higher (0.87 vs 0.76). Analysis of the predictions showed that the current state-of-the-art method over-predicts the number of beach users on dark beach areas (e.g. due to clouds or the inter-tidal area), whereas the oversegmented machine learning method does not. Moreover, the current state-of-the-art method under-predicts the number of beach users on days with high beach occupation resulting in a large negative bias. The oversegmented machine learning method also has a negative bias for this bin, but this bias is approximately 3 times lower. From this results it is concluded that the oversegmented machine learning method has a higher performance compared to the current state-of-the-art in the type of conditions associated with the Dutch climate.

Q4. What is the performance of the oversegmented machine learning method on images from a camera station that was not included in the training procedure?

Research to the fourth research sub-question showed that the camera height, orientation with respect to the cross-shore, distance to the ROI and image size can differ between camera stations. These differences cause variation in the number of pixels that correspond to a beach user between camera stations and this is important because scale-variant features exist. This study showed that image re-sizing can be used to restore the pixel-superpixel ratio, but despite re-sizing the performance of the oversegmented machine learning method on images from camera stations not used during training is unsatisfactory a this point.

In all, it is concluded that the oversegmented machine learning method is a promising method for beach user quantification. Furthermore, the oversegmented machine learning method proved to be more accurate and less sensitive to the variable conditions (e.g. clouds and inter-tidal area) associated with the Dutch coast, in comparison to the current state-of-the-art method for automatic quantification of beach occupation. However, unoccupied beach stretchers and an unclean lens can significantly lower the accuracy and the method needs further development to make it applicable on camera stations that have not been used during training.

8 Recommendations

This thesis described the development, validation and subsequent testing of an oversegmented machine learning method for the quantification of beach occupation. This section suggests recommendations to solve the observed limitations and further improve the performance and applicability of the oversegmented machine learning method on a wide range of beaches. The order in which recommendations are treated, indicates the priority of the proposed recommendation.

1. At this point, application of the oversegmented machine learning method on images originating from a new camera station not used during training did not lead to satisfying results. This implies that new classification models based on annotated images of new camera stations have to be trained. Annotating is time consuming and therefore the applicability of the oversegmented machine learning method is limited. It is recommended to investigate the possibilities to make the oversegmented machine learning method applicable on camera stations not used during training without (or only limited) additional annotating.

2. Detailed analysis of the individual images in the validation data set showed that a relatively large number of images was captured by an unclean lens and that the largest errors are made on these images. This type of images was less present in the training data and dependent on the choice to include/exclude a dirty lens as an aspect of robustness, the training data should be extended with dirty images/a new validation set should be evaluated. If the latter option is chosen it is moreover recommended to investigate the possibilities of a quality control algorithm to detect images that are captured by an unclean lens.

3. The problem of positive errors as a result of unoccupied (rental) stretchers is expected to be linked to the adopted zero intercept in combination with the aggregation of the beach object class to the beach user class. A free intercept led to negative beach user predictions on calm days with less nice weather. This indicates that the variety in observed conditions is limited expressed in the final regression relation and it is recommended to search for better ways to take this variability into account. Another option is to test classification models with separate beach object and beach user classes in more detail.

4. The results indicate that it is possible to obtain the same precision and recall scores with a limited number of features and this can significantly decrease the required computational time because feature extraction is identified as one of the most demanding steps in the workflow. Currently this is not possible because of features being loaded in predefined blocks, but is recommended to search for a work-a-around due to the large expected decrease in required computational time.

5. Auto-undersampling resulted in the optimal model regarding the evaluated R^2-scores. This model has the highest recall, but a relatively low precision indicating a lot of false positives. A possible explanation of this high number of false positives is the removal of data-points of the majority class as a consequence of undersampling. Oversampling solves the imbalance by artificially adding data points of the minority class while remaining the data points of the majority class. This approach might result in the same high recall but also an higher precision due to retaining data points corresponding to the majority class. It is therefore recommended to test oversampling.

6. Enhancement showed to be beneficial for a large part of the tested training set but not on all images and therefore enhancement cannot be implemented as a standard procedure on all images. In order to implement enhancement, a performance indicator capable to distinguish between images that benefit from enhancement and images that do not benefit from enhancement is required. It is

recommended to search for this indicator because of the relatively large number of images that could benefit from enhancement.

7. A concave second order polynomial fit was found to give the best fit on the data. However, this concave behaviour seems to be induced by a limited number of images with a (very) high beach occupation. It is recommended to evaluate more of these type of images to verify the polynomial fit.

8. The benchmark is performed based on snapshots, whereas the original model was developed with the use of Timex-images. It is recommended to re-do the benchmark with Timex-images to verify the increased performance of the newly developed oversegmented machine learning approach.

The first three recommendations correspond to limitations of the current oversegmented machine learning method that constrain the applicability of the method. Therefore, the largest improvements are expected from elaborating the first three recommendations. Recommendations four to six correspond to optimisation steps, whereas seven and eight merely focus on verification of the previously treated results.

References

A. Ng (2012a). Lecture 07.2 - regularization-costfunction-machine learning. https://www.youtube.com/watch?v=C79kIYkKZ1g.

A. Ng (2012b). Lecture 10.6 - advice for applying machine learning — learning curves. https://www.youtube.com/watch?v=ISBGFY-gBug&t=1s.

A. Ng (2012c). Lecture 2.5 - linear regression with one variable — gradient descen. https://www.youtube.com/watch?v=F6GSRDoB-Cg.

A. Ng (2012d). Lectures 06.1 - 06.5 - logisticregression-machine learning. https://www.youtube.com/watch?v=LLx4diIP83I&t=8s.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on pattern analysis and machine intelligence 34*, 34:2274–2280.

Balouin, Y., Rey-Valetteb, H., and Picand, P.-A. (2014). Automatic assessment and analysis of beach attendance using video images at the Lido of Sète beach, France. *Ocean & Coastal Management 102*, pages 114–122.

Coastal Morphodynamics (UPC) & Coastal Ocean Observatory (ICM-CSIC) (2005). Barcelona beaches. http://cooweb.cmima.csic.es/video-coo/images.jsp?site=mapfre-argus&opt=2&fecha=22/07/2005.

Davidson, M., Koningsveld, M. V., de Kruif, A., Rawson, J., Holman, R., Lamberti, A., Medina, R., Kroon, A., and Aarninkhof, S. (2007). The CoastView project: Developing video-derived Coastal State Indicators in support of coastal zone management. *Coastal Engineering 54*, pages 463–475.

De Vries, S., Schipper, M. A. D., Hill, D. F., and Stive, M. (2011). Remote sensing of surf zone waves using stereo imaging. *Coastal Engineering 58*, pages 239–250.

Dekking, F., Kraaikamp, C., Lopuhaä, H., and Meester, L. (2005). *A Modern Introduction to Probability and Statistics*. Springer.

Deltares (2013). Argus api. http://argus-public.deltares.nl/db/table.

Gonzalez, R. and Woods, R. E. (2008). *Digital Image Processing (Third Edition)*. Pearson Education Inc.

Goubet, E., Katz, J., and Porikli, F. (2006). Pedestrian Tracking Using Thermal Infrared Imaging. *Proceedings of SPIE - The International Society for Optical Engineering 6206*.

Guillén, J., García-Olivares, A., Ojeda, E., Osorio, A., Chic, O., and González, R. (2008). Long-Term Quantification of Beach Users Using Video Monitoring. *Journal of Coastal Research 24*, pages 1612–1619.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEE Transaction on knowledge and data engineering 9*, 21:1263–1284.

Hendriks, E. A. (2013-2014a). Lecture 1 of course ti2715-b. https://collegerama.tudelft.nl/Mediasite/Play/2169e277a7f24f3bbc6e3ac9e0fdf4c91d?catalog=528e5b24-a2fc-4def-870e-65bd84b28a8c&playFrom=38414&autoStart=true. Accessed 10-04-2017.

Hendriks, E. A. (2013-2014b). Lecture 2 of course ti2715-b. https://collegerama.tudelft.nl/Mediasite/Play/d4536ec51dd04a7bbfff803bf8dd07de1d?catalog=100392a7-9870-4935-8457-5f9dd14f7644. Accessed 08-10-2017.

Hoonhout, B. (2011). Xb stagger. https://svn.oss.deltares.nl/repos/openearthtools/trunk/matlab/applications/xbeach/xb_lib/xb_stagger.m.

Hoonhout, B. (2016). Distortion. https://github.com/openearth/argus-python/blob/master/argus/argus/distortion.py.

Hoonhout, B. and Radermacher, M. (2014a). Documentation rectification implemented in flamingo toolbox. http://flamingo.tudelft.nl/docs/rectification.html#id1.

Hoonhout, B. and Radermacher, M. (2014b). Flamingo toolbox. http://flamingo.tudelft.nl/.

Hoonhout, B., Radermacher, M., Baart, F., and van der Maaten, L. (2015). An automated method for semantic classification of regions in coastal images. *Coastal Engineering 105*, pages 1–12.

Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. *Computer Vision and Pattern Recognition*, pages 1037–1045.

Jiménez, J., Osorio, A., Marino-Tapia, I., M.Davidson, R.Medina, Kroon, A., R.Archetti, Ciavola, P., and Aarninkhof, S. (2007). Beach recreation planning using video-derived coastal state indicators. *Coastal Engineering 54*, pages 507–521.

Kammler, M. and Schernewski, G. (2004). Spatial and temporal analysis of beach tourism using webcam and aerial photographs. *Coastline Reports 2*, pages 121–128.

Koninklijk Nederlands Meteorologisch Instituut (2013). Uur gegevens van het weer in nederland. https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A Survey of Mobile Phone Sensing. *IEEE Communications Magazine (September)*, pages 140–150.

Müller, A. C. and Behnke, S. (2014). pystruct - learning structured prediction in python. *Journal of Machine Learning Research*, 15:2055–2060.

O'Neill, E., Kostakos, V., Kindberg, T., gen. Schiek, A. F., Penn, A., Danae, Fraser, S., and Jones, T. (2006). Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape. *Computer Science 4206*, pages 315–322.

OpenCV documentation (2015). Histograms - 2: Histogram equalization. http://docs.opencv.org/3.1.0/d5/daf/tutorial_py_histogram_equalization.html. Accessed 09-10-2017.

OpenCV.org (2015). Open source computer vision library. https://github.com/itseez/opencv.

Osorio, A. F., Medina, R., Garcia, N., and Labégorre, M. (2006). Utilisation de l'imagerie vidéo pour la gestion intégrée des côtes: application á la gestion touristique de la plage du Puntal à Santandar (Espagne). *European journal of environmental and civil engineering*, pages 547–554.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Python Software Foundation (2017). Python. `https://www.python.org/`. Accessed 07-09-2017.

Trygonis, V., Ghionis, G., Andreadis, O., Vousdoukas, M., Ntemogiannis, I., Rigos, A., Psarros, F., Velegrakis, A., Hasiotis, T., and Poulos, S. (2015). Monitoring beach usage with a coastal video imaging system: an application at Paralia Katerinis, Greece. *Conference Paper: 11th Panthellenic Symposium on Oceanography and Fisheries*, pages 737–740.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors (2014). Scikit-image: image processing in Python. *PeerJ 2:e453*.

Voskamp, A. (2012). Measuring the influence of congested bottlenecks on route choice behaviour of pedestrians at Utrecht Centraal. Master's thesis, Delft University of Technology, the Netherlands.

Young, I. T., Gerbrands, J. J., and van Vliet, L. J. (2007). *Fundamentals of Image Processing (version 2.3)*.

Zhao, H. and Shibasaki, R. (2005). A Novel System for Tracking Pedestrians Using Multiple Single-Row Laser-Range Scanners. *IEEE transactions on systems, man and cybernetics Part A: Systams and Humans 35*, pages 283–291.

Appendix A Image enhancement techniques

Multiple studies based on the method using differences in pixel intensity, incorporate image enhancement steps. This appendix treats the theoretical functioning of these methods, finally resulting in an enhancement step that is tested on the dataset for training. The remainder of the appendix starts with an introduction on image enhancement, followed by the theoretical functioning of the reviewed enhancement techniques.

A.1 Introduction image enhancement

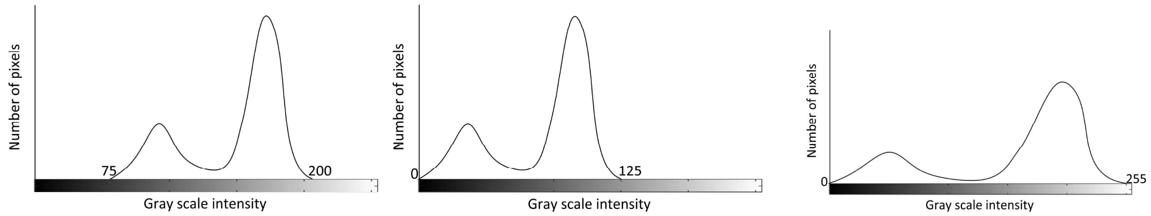
Image enhancement can be defined as a process in which the original pixel intensity is changed to a new value using a transformation (Hendriks, 2014b):

$$i_{new} = f(i_{original}) \tag{A.1}$$

The reviewed enhancement techniques predominately differ in the function that is used for the transformation. As these functions change the original intensity of an image, the effect of an enhancement technique can be explained by the difference in intensity spectrum/histogram corresponding to the original/enhanced image. The following subsections explain the reviewed techniques based on the effect on the intensity spectrum.

A.2 Contrast stretching

In general over-/underexposed images have low contrast, which results in a narrow (high) intensity peak in the intensity spectrum. The existence of a narrow peak indicates that only a limited number of the 256-possible colours is represented by the pixels in an image.Contrast stretching maps the narrow band of intensities as a result of over-/underexposure to a wider (or full) range (Fig. A.1) via a linear transformation: $i_{new} = 256\frac{(i_{original}-i_{min})}{(i_{max}-i_{min})}$ (Hendriks, 2014b).



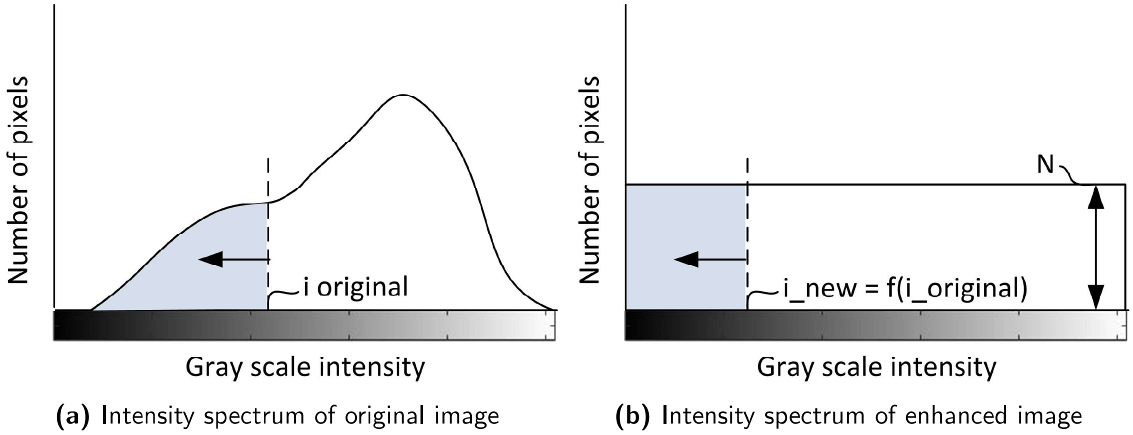
(a) Original intensity spectrum **(b)** Translated intensity spectrum **(c)** Normalised intensity spectrum

Figure A.1: Illustration of contrast stretching. The intensity spectrum is moved to the origin by a translation of -75 (i_{min}). Subsequently the original intensities are multiplied with 256 (all possible intensities in a 8-byte gray-scale image) and divided by 125 ($i_{max} - i_{min}$, the original range) resulting in a spectrum that covers the whole available range of 256.

Contrast stretching was performed in Python with the function *rescale_intensity*, which is part of the *exposure*-module implemented in the package Skimage. The intensities of the red colour-band were stretched as this was the band that was also enhanced in Guillén et al. (2008). The range of the new intensity spectrum was set from 0 to 255, spanning the whole available range for a 8-byte image.

A.3 Histogram equalisation

Histogram equalisation aims at using a function that transforms the original intensity spectrum in a uniform spectrum where all intensities are equally represented by: $N = \frac{n_{pixels}}{256}$ (Hendriks, 2014b). To derive a function that has this behaviour, use is made of the area under the intensity spectrum. This area represents the number of pixels present in the image and is the same between the original and enhanced intensity spectrum (see blue area Fig A.2).



(a) Intensity spectrum of original image **(b)** Intensity spectrum of enhanced image

Figure A.2: Illustration histogram equalisation. The blue areas in both images are equal and represent the number of pixels that correspond to the intensities in the range of interest. The blue area in the left image can be calculated and subsequently used to determine the intensity in the right image, because this is the only unknown.

For the original intensity spectrum the number of pixels left of the intensity of interest are known and can be obtained by summing the pixels in the intensities left of the intensity of interest. The new intensity is unknown, but the number of pixels corresponding to that intensity is known ($= N$) and moreover the number of pixels to the left should be equal as the areas should be equal. This leads to the following equations with $i_{new} = f(i_{original})$ being the only unknown (Hendriks, 2014b):

$$Pixels\ to\ left\ (original) = \sum_{j=0}^{i} H(j) \tag{A.2}$$

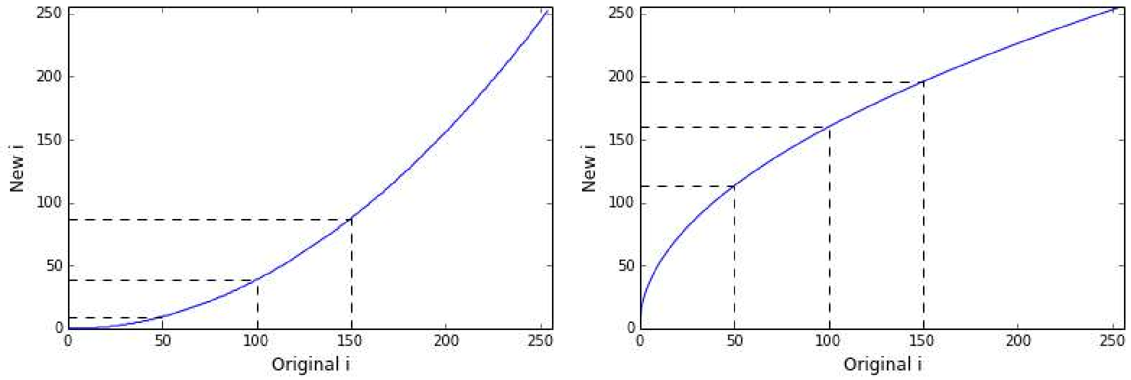
$$Pixels\ to\ left\ (new) = N f(i_{original}) \tag{A.3}$$

$$f(i_{original}) = \frac{1}{N}\sum_{j=0}^{i} H(j) \tag{A.4}$$

Histogram equalisation is performed with the function *equalize_hist* from the module *exposure* in the package Skimage. The number of bins was set to 256, resulting in a range of 0 - 255 for the enhanced intensity spectrum. Enhancement was performed on the whole image and therefore no mask, reducing the number of pixels that are taken into account, was used.

A.4 Gamma adjustment

Gamma adjustment is a technique that can either give preference to darker intensities ($\gamma > 1$), or lighter intensities ($\gamma < 1$) via a non-linear transformation: $i_{new} = Ci_{original}^{\gamma}$ (Hendriks, 2014b). In this equation C represents a constant for normalisation and can be defined as: $C = 256^{(\gamma - 1)}$. For $\gamma > 1$, the gamma function is convex (Fig. A.3a), causing a relative large amount of intensities to become lower and thus darker. The opposite holds for $\gamma < 1$, which results in a concave function (Fig. A.3b) that maps the original intensity to higher intensities and thus lighter.



(a) $\gamma = 2$

(b) $\gamma = 0.5$

Figure A.3: Examples of effect gamma adjustment on intensities. The horizontal axes represent the original intensity, whereas the vertical represent the new intensity as a result of the gamma adjustment. Te dotted lines indicate that for $\gamma > 1$, input intensities are lower after teh transformation, whereas the opposite holds for $gamma < 1$.

Gamma adjustment was observed in the study of (Guillén et al., 2008), were a gamma of 3 was used to give more preference to the darker colours. The increased preference in darker colours can be explained by the interest in beach users, as these are represented by the darker colours in the spectrum. Gamma adjustment was performed with the function *adjust_gamma* from the *exposure*-module in Skimage.

A.5 Contrast Limited Adaptive Histogram Equalisation

Contrast Limited Adaptive Histogram Equalisation (CLAHE), is a rather similar technique as histogram equalisation, but instead of adjusting a pixel intensity based on histogram of the whole image, it uses the histogram from a local region around the pixel that is adjusted. This approach allows the enhancement of local areas, which is preferable for images with intensity spectra that are not confined to a small range (OpenCV documentation, 2015). The degree of contrast enhancement is limited in order to limit over-amplification of noise. CLAHE was performed with the function *createCLAHE* from the openCV package. The used tile size was the default of 8x8 pixels and the cliplimit was set to 3.