Lappeenranta University of Technology
School of Engineering Science
Master's program in Computational Engineering and Technical Physics
Intelligent Computing Major

Master's Thesis

Evgeniya Kosyanenko

CONVOLUTIONAL NEURAL NETWORKS FOR SAIMAA RINGED SEAL SEGMENTATION

Examiners:	Professor Heikki Kälviäinen Docent, Dr. Marina Cherdyntseva
Supervisor:	Professor Heikki Kälviäinen Docent, Dr. Tuomas Eerola

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Master's program in Computational Engineering and Technical Physics

Intelligent Computing Major

Evgeniya Kosyanenko

Convolutional neural networks for Saimaa ringed seal segmentation

Master's Thesis

2017

53 pages, 30 figures, 1 table.

Examiners:	Professor Heikki Kälviäinen Docent, Dr. Marina Cherdyntseva

Keywords: image segmentation, convolutional neural networks, Saimaa ringed seal, computer vision, image preprocessing, animal biometrics

Nowadays, an automated photo-identification of animals is an emerging research topic due to appearance of large labor-intensive image datasets. The Saimaa ringed seal has a distinctive patterning of dark spots surrounded by light gray rings. This pattern is unique to each seal, enabling the identification of individuals in an image. A common approach is to preprocess an image by performing segmentation to separate an object from its background. In this research, convolutional neural networks (CNN) are used for Saimaa ringed seal segmentation. The main objectives of the research were to survey existing CNN based approaches on semantic image segmentation, implement a CNN-based method, and evaluate the method using the Saimaa ringed seal image database. A method based on CNN was implemented. The method was evaluated using a dataset consisting of 168 images of seals. 75% of the images were correctly segmented.

CONTENTS

LIST OF ABBREVIATIONS

AAB	Automated animal biometrics
CNN	Convolutional neural networks
CV	Computer vision
CVS	Computer vision system
FCNN	Fully convolutional neural networks
gPb	Globalized Probability of Boundary
GT	Ground truth
MCG	Multiscale Combinatorial Grouping
MSE	Mean Squared Error
owt	OrientedWatershed Transform
RCNN	Region-based Convolutional Neural Networks
ReLU	Rectified linear unit
SLIC	Simple linear iterative clustering algorithm
SVM	Support vector machine
UEF	University of Eastern Finland
UCM	Ultrametric Contour Map

1 INTRODUCTION

1.1 Background

With the increasing demand for monitoring population demography of various species in recent years, individual wild animal identification has proven increasingly useful. It helps to construct mathematical models, which are used to predict population demographics in future. Any ecosystem consists of nonlinearly interacting subsystems, which can be arranged in a certain hierarchical structure [1]. By means of ecosystem monitoring, the number of animals can be kept track of. Thus, the extinction of certain animal species can be prevented. Various animal species have specific features that make it possible to identify individual animals. Such features are, for example, spots on the body, coat patterns, feather shape [2].

Our research is focused on identifying Saimaa ringed seals, which is one of the most endangered species of seals in the world. Moreover, it is one of the few living freshwater seals. Its population consists of only about 360 individuals nowadays [3]. However, in 1983, the population was even less numerous, with the number of seals lying between 100 and 150.

The Saimaa ringed seal has a distinctive pattern of dark spots surrounded by light gray rings. This pattern is unique to each seal which makes identification of individuals possible. One of the most popular techniques of recording natural markings of an animal is photo-identification. Nowadays, automated photoidentification of animals is an emerging research topic as it implies processing of large labor-intensive image data sets.

A common approach to improve identification performance is to preprocess an image by segmenting it with the goal of separating an object from its background. Segmentation is a process of separating a digital image into different segments.

In this research, we use convolutional neural networks (CNN) [4] for segmentation of images with Saimaa ringed seals. Convolutional Neural network is a type of artificial neural network which design is inspired by the human visual system [5]. They have been shown to achieve state-of-the art performance in various image analysis tasks including semantic image segmentation (see for example [6, 7, 8, 9]).

1.2 Objectives and restrictions

The work is based on the previous studies [10, 11, 12], where the segmentation and identification of the Saimaa ringed seals were considered. The goal of this research is to implement a CNN based approach to Saimaa ringed segmentation. There are several different approaches to segmentation by CNN [4].

The objectives of this research are to:

1. Apply existing CNN based approaches to semantic image segmentation,
2. Implement a method for the segmentation,
3. Evaluate the method using Saimaa ringed seal image database.

In this research, there are several limitations. Firstly, identification of the individual seals is not considered. Secondly, no other animals were considered.

1.3 Structure of the thesis

The rest of the thesis is organized as follows. Chapter 2 consists of information about automated animal biometrics and the description of a computational approach to it. Chapter 3 represents basic knowledge about the segmentation and existing methods. Chapter 4 considers the CNN, basic steps, and field where CNN is used. Chapter 5 describes the idea of the proposed algorithms which were used during the research. Chapter 6 describes the dataset, the experimental arrangement, and the results. In chapter 7 the obtained results and future work is discussed. Finally, chapter 8 summarizes the thesis.

2 AUTOMATED ANIMAL BIOMETRICS

Biometrics refers to methods for identity recognizing based on physiological or behavioral characteristics of a subject. These methods were originally used to identify fingerprints of criminals [13]. This process was paper-based and labor-intensive at the time. Today it is largely digital and highly automated, by means of the image capture technologies and powerful computing [14]. The idea of using biometrics has become increasingly popular over time. Biometric information is used to identify unknown persons in accordance with their unique features or to verify the authenticity of a person's identity by comparing the submitted samples with the existing templates in the database. Moreover, recently biometric recognition methods have been applied also for the identification of individuals animals [15, 16, 17].

Automatic Animal Biometrics (AAB) is one of the new research fields of animal identification based on phenotype. The phenotype is a combination of properties or characteristics of the observation of the body, such as morphological, biochemical and physiological. The phenotype consists of the biometrical data from the subject [17].

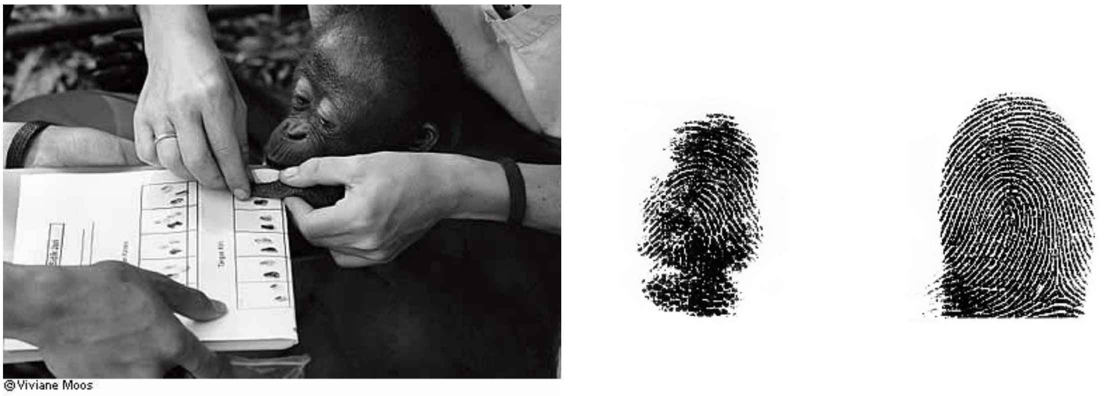
AAB can be applied to both domestic and wild animals, however, the approaches are different. There are several traditional methods for domestic animal identification which are permanent (e.g. ear tags, branding, tattooing, electrical methods) and cannot be applied to wild animals [18] as these methods can hurt animals, affect their appearance, social interaction, other behaviors and therefore can endanger the very survival of these individuals [19].

Automated image-based animal biometrics has been developed as a solution to this problem. This approach is based on analysis of physical characteristics or behavioral signs. Advantages of AAB include low costs and possible automatization of the process. Moreover, this method does not imply any interference into the life of an individual. Thus, recently, image-based biometrics has become a promising trend in various fields of studies, including biometrics, ecology and behavioral research [17].

2.1 Animal biometrics

Animal Biometrics is a field of study, which develops quantified approaches for detecting the phenotypic appearance of species, individuals, their behaviors, and morphological traits [17]. It utilizes different features for the identification, such as fin, fur, and retinal patterns. In some cases, the identification method requires a close contact with an animal, which is often impossible when speaking about wild animals, but it might be applicable to domestic animals [20].

Finding features suitable for identification is a separate problem, but some approaches applied to identify humans might be applicable to animals as well. For instance, fingerprint recognition is among them, although this approach has several disadvantages. First of all, it is not a uniform method as only few animal species (e.g. chimpanzee) have fingerprints (Figure 1) [21, 22]. However, other species can have other features which would be unique for each individual. For example, there is an identification system developed for African penguin which is based on patterns of black and white feathers. This system makes monitoring African penguin populations possible without any human interference.



@Viviane Moos

Figure 1. Chimpanzee's fingerprints. [22]

Another feature which might be suitable for animal identification is the pattern of a footprint. There are rows of dots corresponding to tiny papillae on the penguin's metacarpal pad. For some animal species, nose prints can be used similarly to finger- and footprints. The nasal imprint has a unique special set of dots and lines. The method based on nose printing analysis was proposed in 1975 [23]. It can be used an alternative to ear tags, branding, tattooing and electrical methods. This method is usually applied to cattle. Originally noise prints were made manually without any automatization (Figure 2) [22]. The main tools were ink and paper. This method is not reliable as the procedure should be performed with the same force every time. Moreover, the same type of paper should be used to improve the performance [23].

One more method for animal identification, which is commonly used for people, is face recognition [24]. However, it is not as effective for animals, because it is very difficult to

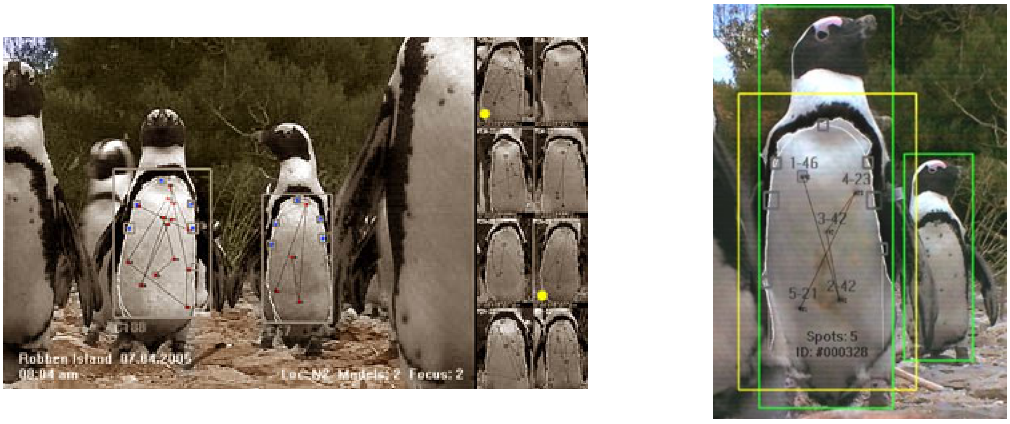


Figure 2. Recognition by fingerprint-like patterns of black and white feathers for penguins. [22]

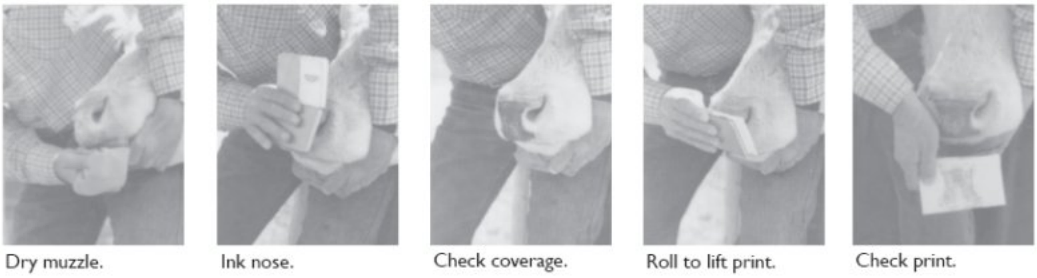


Figure 3. Identification of cattle by noseprint. [22]

get clear images of an animal face. Nevertheless, this method has been used to identify domestic animals, such as dogs (Figure 4) [19, 18, 22].

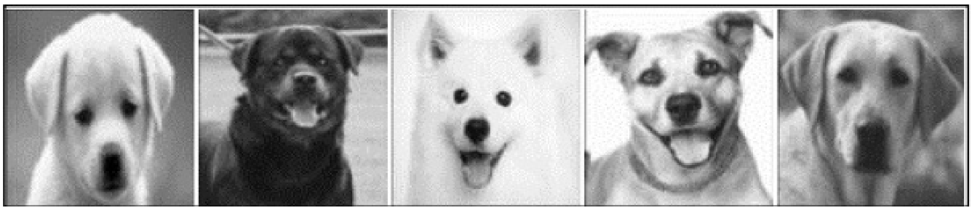


Figure 4. Face recognition for dogs. [18]

Retinal patterns can be used for identification as well. It is a highly accurate method as retinal patterns are unique. Every animal has its unique branching patterns of the retinal vessels and they do not change with the passing of time. The information can be obtained from retinal vessels using a hand-held scanner. This method is relatively cheap. When compared to nose prints, it is more accurate but slower (Figure 5) [25, 26].

There are other methods, which are not used for people but are highly effective for animals. For instance, an approach to identify water animals (e.g. dolphins, whales, sharks) by fins was invented 40 years ago. It examines curves, notches, nicks, and tears on fins

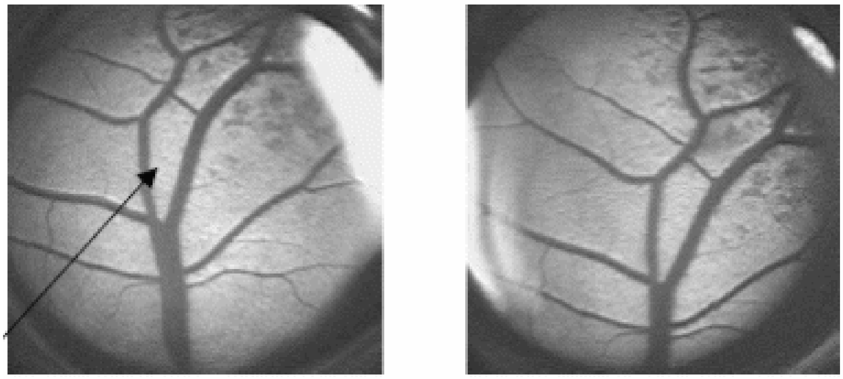


Figure 5. Identification by retinal patterns. [26]

which are unique for each individual (Figure 6) [27, 28].

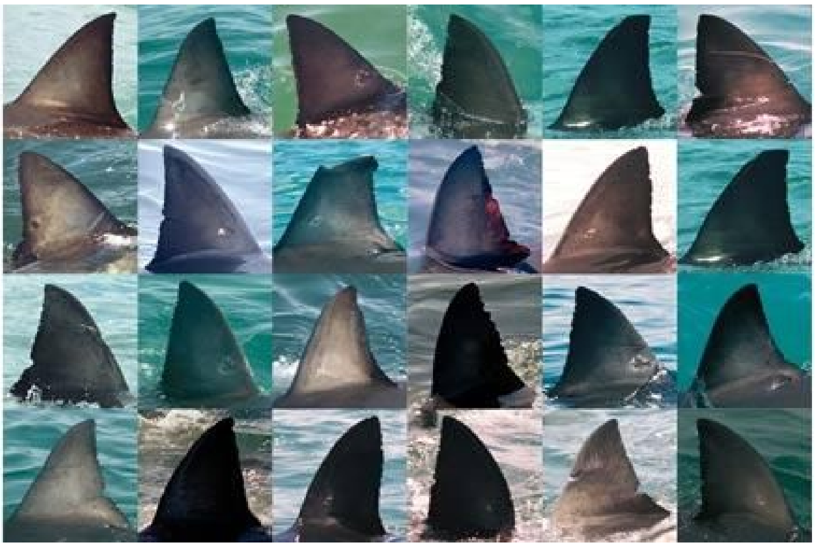


Figure 6. Identification of Great White Shark by fins. [28]

Several animal species can be identified by their ear circulatory system. The main idea is to capture a high-quality image so that blood vessels are visible. They have unique node points which can be used to build an identification system (Figure 7) [2, 29].

The most popular method for animal identification is based on fur, feather or skin patterns. Such patterns can include stripes (zebra) and spots (ringed seal) (Figure 8) [22]. The main advantage of these methods is their non-invasive nature and invariance to the camera angle and animal's pose. Moreover, images can be collected by tourists, researchers, or camera traps. It increases the number of images, and therefore the quality of the resulting identification system [2].

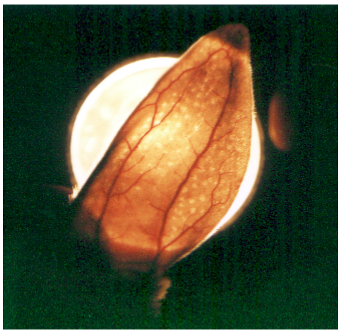


Figure 7. Identification by ear's circulatory. [29]

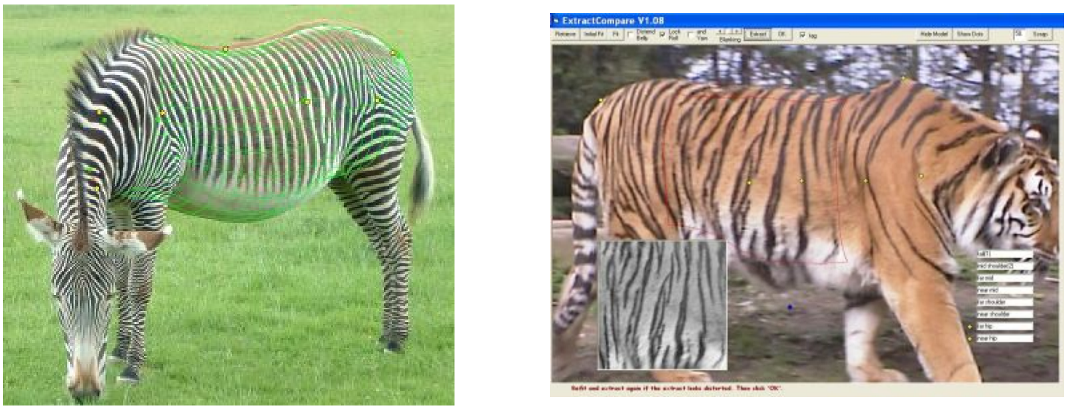


Figure 8. Stripes of zebra and tiger. [22]

Saimaa ringed seals can be identified by fur pattern (Figure 9). They have a distinctive pattern of dark spots surrounded by light gray rings. This pattern is unique to each seal enabling the identification of individuals [2].

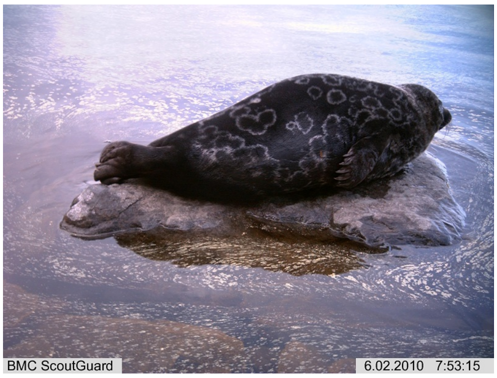


Figure 9. Saimaa ringed seal (from the Saimaa seals database).

2.2 Computer vision based approaches

All the methods listed in the last section can be automated using computer vision (CV). Computer vision is a field of science focused on developing a system, which would be able to interpret images automatically. The main concept is to get information from the image with the help of artificial intelligence systems. The information from a camera can be represented in various ways, such as a video sequence, a set of images from different cameras or 3D data, for example, produced by Kinect device or a medical scanner. Examples of such systems include: process control systems (industrial robots), systems of video observation, information management systems (for example, for indexing image databases), systems for modeling objects or the environment (analysis of images), interaction systems, systems based on augmented reality (Figure 10) [30].

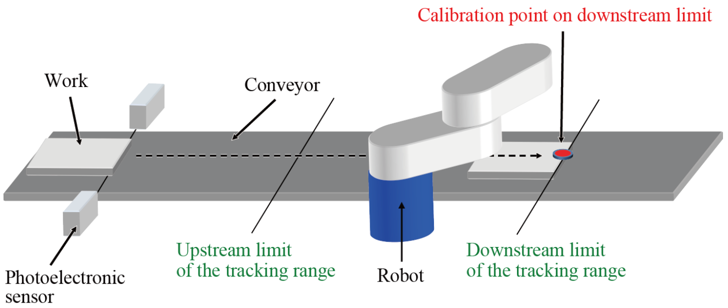

Calibration point on downstream limit
Work
Conveyor
Photoelectronic sensor
Upstream limit of the tracking range
Robot
Downstream limit of the tracking range

Figure 10. Example of the computer vision system: industrial robot [31].

There is no uniform solution to the problem of computer vision. Usually, methods depend on the task and they cannot be generalized for other tasks. Nowadays the number of methods is increasing, because they can be used in different applications [30].

Implementation of a computer vision system is dependent on its applications, hardware platform, and performance requirements. However, there are basic stages that are typical for the majority of computer vision systems [32]:

1. Acquiring images: Digital images are obtained from one or more image sensors, which may include distance sensors, radar, ultrasonic cameras in addition to different types of photosensitive cameras.

2. Preprocessing: Before computer vision methods would be applied to analyze an image, it is necessary to process the image data so that it satisfies certain conditions, depending on the method being used. The most common examples of data

preprocessing include filling in missing values, smoothing noisy data, identifying or removing outliers and noisy data, and resolving inconsistencies.

3. Feature extraction: It builds derived values, i.e. features, which are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations.
4. Detection / Segmentation: Points or areas of the image, which are important for further processing, are detected.
5. Classification: different objects in the image data are classified.
6. Analyzes of the results: the output is compared with the expected results.

There are a lot of different computer vision systems based methods for identifying individual animals [14]. To choose an appropriate method, main features, which can identify an individual, should be selected. Such features can be, for example, fin, fur, and retinal patterns [33].

One of the examples of computer vision system is a software system identifying individual Thornicroft's giraffes [34]. This program uses a Scale Invariant Feature Transform (SIFT) algorithm [35] to extract and match distinctive image features regardless of scale and orientation.

This software is an example of application an identification method called wild-ID, which consists of three stages:

1. Extract the SIFT features [35] for each image in a database,
2. Candidate matched pairs of SIFT features are identified from two images by locating features of one image on the other image and minimizing the Euclidean distance between descriptors of the features,
3. Choose the best matches from all the potential matches.

Another popular method for identification is hot spotter [36]. This method is based on two approaches: one-vs-one and one-vs-many strategies. The former one is similar to the Wild ID, where for every image the key points are recognized and the descriptor corresponding to it is extracted. Then descriptors are compared in order to find matches among the images. In the one-vs-one approach a query image is compared to each image in the database separately, then the database images are sorted by the similarity score so that the final rank of the result can be found.

In the one-vs-many approach, the query image descriptor is matched to the descriptors corresponding to the rest of images in the database. The scores are generated for each image according to these matches, and finally, the scores are aggregated to generate the final similarity scores for each image. In this method, the Local Naive Bayes Nearest Neighbor methods are introduced for scoring matches [36].

3 SEGMENTATION

The typical method to capture image material for the photo-identification is to use static camera traps. Therefore, the same animal individual is often captured with the same background. This increases the risk that a supervised identification algorithm learns to 'identify' the background instead of the actual animal if the full image is used. To avoid this problem, it is useful to segment the animal from the background [37, 9]. Automatic segmentation of animals is, however, often difficult due to the camouflage colors of animals, that is the coloration patterns of that are similar to the visual background (Figure 11)

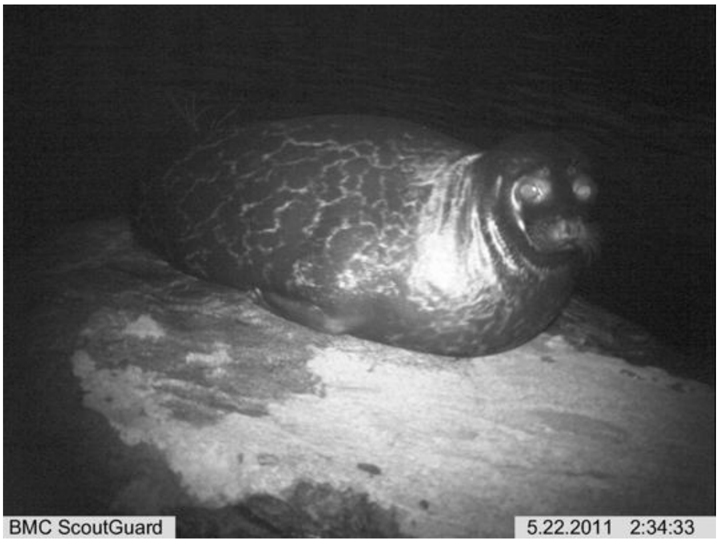


Figure 11. Example of camouflage colors of animals.

The segmentation is the process of separation of a digital image into different segments, which satisfy certain criteria of homogeneity, for example, the selection in the image areas of approximately equal brightness. Image parsing is the decomposition (segmentation) of an image to natural parts (Figure 12) [8, 38, 39].

3.1 Categorization of the methods

Segmentation methods can be divided into two types: automatic methods that do not require interaction with the user and interactive methods that use user input directly in the process. The tasks of automatic segmentation can be further divided into two classes [38]:

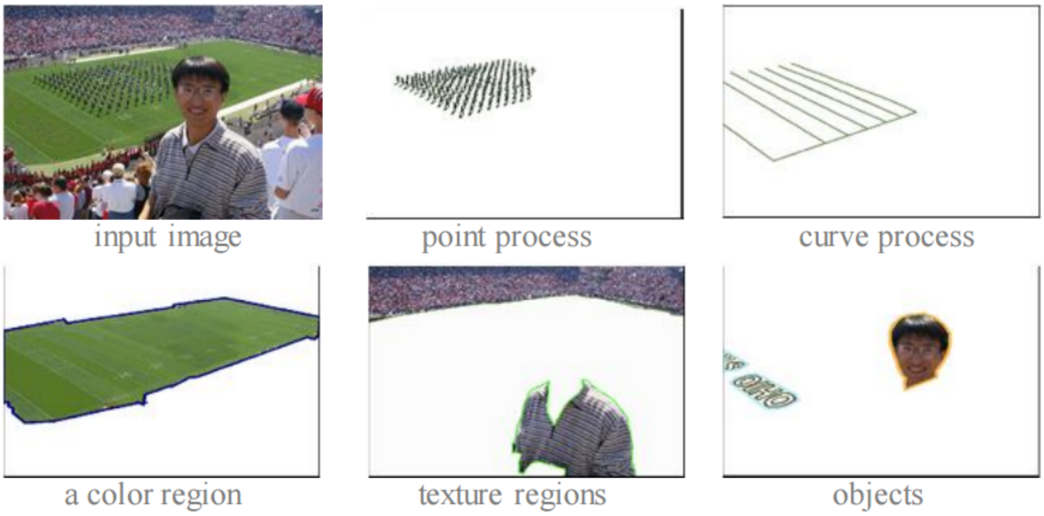


Figure 12. Example of decomposition(segmentation) of an image to natural parts. [40]

1. Selection of image areas with known properties (semantic segmentation),
2. Splitting an image into homogeneous areas (unsupervised segmentation).

There is a fundamental difference between these two statements of the problem. In the first case, the task of segmentation is finding certain areas for which there is a priori information known, for example, the color, the shape of regions, or the regions of interest which are the parts of a known object. The methods of this group are narrowly specialized for each specific task. Segmentation in this formulation is used mainly in problems of computer vision (analysis of scenes, search for objects on the image) [38, 41].

In the second case, no a priori information about the properties of the regions is used, but certain conditions are imposed on the image itself, for instance, all the areas should be uniform in color and texture. Since this setting of the segmentation task does not use a priori information about the depicted objects, the methods of this group are universal and applicable to any images. Basically, segmentation in this setting is applied at the initial stage of solving the problem to get an image representation in a more convenient form for the further work [38, 41].

There are a lot of different methods for segmentation such as clustering methods, compressive methods, histogram based methods, edge detection, dual clustering methods, region-growing methods, partial differential equation-based methods, variational methods, graph partitioning methods, watershed transformation, model based segmentation, multi-scale segmentation, semi-automatic segmentation and trainable segmentation [42, 7, 8, 9]. All of them could be categorized based on the use of information about the connectivity areas: region growing, merging, segmentation by morphological watershed, graph theory methods and thresholding [38].

3.1.1 Thresholding

One of the simplest ways to perform segmentation is the thresholding. The threshold is an attribute (property) which helps to divide the desired signal into classes (object and background). It is done by comparing their brightness with a certain threshold value [7, 8].

A single thresholding is given as a result function $g(x, y)$, where T is a threshold value as:

$$g(x,y) = \begin{cases} 1, iff(x,y) > T \\ 0, iff(x,y) \leq T \end{cases} \tag{1}$$

Multiple thresholding is defined as a result function $g(x, y)$ where a, b, c are three distinct intensity threshold values:

$$g(x,y) = \begin{cases} a, iff(x,y) > T2 \\ b, ifT1 < f(x,y) \leq T2 \\ c, iff(x,y) \leq T1 \end{cases} \tag{2}$$

3.1.2 Region growing

Region growing is a simple region-based image segmentation method. This approach to segmentation examines neighboring pixels of initial seed points and determines whether the pixel should be added to the region. The process is iterative, in the same manner as general data clustering algorithms. The method includes the following steps:

1. Points (nucleation sites), which presumably belong to the allocated areas, are selected on the original image. For example, these may be the points of maximum brightness level.

2. Growth of the areas begins from these points. It means that points adhere to the

existing points of the neighboring region, if they satisfy certain proximity criteria.

3. Growth of areas stops after a predefined condition is reached.

The disadvantages of this method are a long computational time and sensitivity to noise [9, 43, 44].

3.1.3 Merging

Another way to perform segmentation is to apply the area merging algorithm. It is based on the idea that pixels of the original image compose homogeneous areas. In this case, the segmentation method performs the union of the closest neighboring areas based on a predefined distance function. Determining the completion condition for this process is a separate problem. [7, 8].

3.1.4 Segmentation by morphological watershed

The algorithm works with the image as a function of two variables f = I (x, y), where x, y are the pixel coordinates. The value of the function can be, for example, intensity or the modulus of the gradient. If the absolute value of the gradient is plotted along the z-axis, then in the places of the intensity drop ridges are formed. Homogeneous regions are refered to as plains. After finding the minima of the function f, the process of filling with "water" starts, which begins with a global minimum. As soon as the water level reaches the next local minimum, it begins to fill with water. When two regions begin to merge, a partition is constructed to prevent the merging of areas [45]. The water will continue to rise until the regions are separated only by artificially constructed partitions (Figure 13).

The result of the algorithm is a mask with a segmented image, where the pixels, which are labeled with the same label, form a connected area. The main drawback of this algorithm is the use of a preprocessing procedure for pictures with a large number of local minima (images with a complex texture and an abundance of different colors) [45].

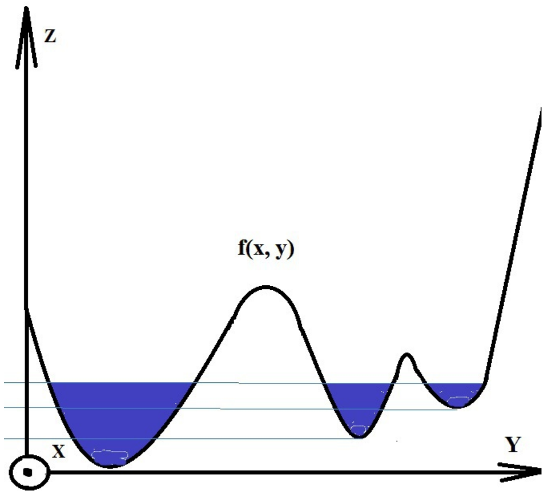


Figure 13. Illustration of the process of filling with water. [46]

3.1.5 Graph theory methods

The basic idea is to match the pixels of an image with the vertices of a graph. All vertices are connected with edges, with their weights showing similarity measure based on the distance between characteristic values for the two vertices. Further, the graph in the subgraph, the weight of the edges is much larger weights of edges connecting the subgraphs. The main idea of segmentation is to compare the image graph with subsequent clustering based on the construction minimal spanning tree, where minimal spanning tree is a subset of the edges of a connected, edge-weighted undirected graph that connects all the vertices together, without any cycles and with the minimum possible total edge weight [47].

3.2 Saimaa ringed seal segmentation

Saimaa ringed seal segmentation is a topic that has been studied for some time. This research is based on several works which were dedicated to this topic [10, 12].

The first work was focused on the automatic image-based identification of Saimaa ringed seals [10]. It consists of the detection and segmentation of a seal in an image, analysis of its ring patterns, and identification of the detected seal based on the features of the ring patterns. The proposed segmentation algorithm from the previous research consists of the following steps [10]:

1. Unsupervised segmentation,

2. Training of the classifier,
3. Classification of the segments.

In the first step, the image is divided into small segments (superpixels) for further classification. For this purpose, the Globalized Probability of Boundary, Oriented Watershed Transform, Ultrametric Contour Map were used. In the first step, the gPb contour detector is applied to produce a probability map E(x; y; Θ) which describes the probability of an image boundary at location (x; y) and orientation Θ. Then, hierarchical regions are built by exploiting the information in the contour probabilities using a sequence of two transformations: Oriented Watershed Transform (OWT) and Ultrametric Contour Map (UCM) [10]. These methods are given in detail in the previous research [10].

In the second step, a classifier is trained using features from manually labeled segments to classify segments as either seal or background. The main purpose of this research was to create automatic supervised segmentation system [10]. The system should detect image segments containing parts of a seal and combine them into one large segment that contains all the pixels that belong to the seal. There were several features for describing segments: RGB colors, center distance, area, Segmentation-based Fractal Texture Analysis descriptor, Local Binary Pattern Histogram Fourier descriptor, Local Phase Quantization descriptor [10]. These methods are described in detail in the previous research [10].

In the third step, classification of the segments is performed. The segments containing the parts of the detected seal form the output of the method. During the research, several classifiers were considered. There are Naive Bayes Classifier, k-nearest neighbor classifier and support vector machine classifier (SVM). Descriptions of these classifiers are given in the previous research [10].

Testing performed in the previous research shows that the most informative feature is LPQ and the classifier that has shown the best results is SVM [10].

The second work, which we considered, was about enhanced methods for Saimaa ringed seal identification [12]. In this work the identification algorithm consists of three main steps: image segmentation, image enhancement, and identification. The image segmentation method in the previous research contains two steps [12]:

1. Unsupervised segmentation (the image of seal is partitioned into several segments called superpixels),

2. Classification of the superpixels (all superpixels are classified into two classes of the seal and the background based on the information extracted by feature descriptors).

The first step is unsupervised segmentation. For this purpose, the same methods as in the first work were used [10]. However, the Globalized Probability of Boundary algorithm was replaced by Multiscale Combinatorial Grouping (MCG). The descriptions of these algorithms are given in the previous research [12] and in the chapter 5.1.1. MCG was chosen as it requires less computational effort in comparison with the gPb edge detector.

In next step of the segmentation, features describing the image segments are utilized. These features are used to train the superpixel classifier and to classify the segments in the new images. To describe the superpixels, various features were tested: mean RGB colors, mean variance of local range, local standard deviation, and local entropy, Local Binary Pattern Histogram Fourier descriptor, Local Phase Quantization descriptor. Descriptions of these features are given in the previous research [12].

The second step is classification of the superpixels. Superpixels are classified into two classes — the background and the seal body — to construct the segmented image. For the classification of the superpixels, three classifiers were selected: Naive Bayes, k-nearest neighbor, and support vector machine. Descriptions of these classifiers are given in the previous research [12].

The best result was achieved using the SVM classifier based on the combination of features: LBP, LPQ, mean color and mean variance. [12].

4 CONVOLUTIONAL NEURAL NETWORKS

Artificial neural network is a mathematical model, as well as its software or hardware implementation, built on the principle of the organization and functioning of biological neural networks. The convolutional neural network is a special architecture of artificial neural network which can effectively analyze images (Figure 14). It is a part of the deep learning technologies, which consists of the features of the visual cortex. The idea of CNN is the alternation of convolutional layers and down sampled layers. The details about the layers are described later in Section 4.2 [4, 48, 49].

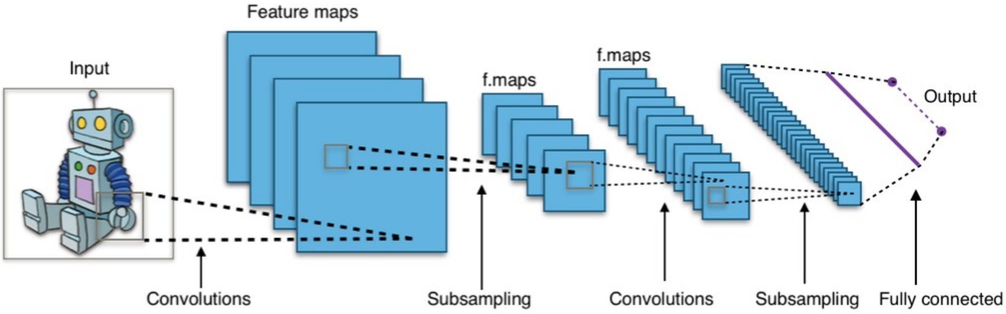


Figure 14. Typical CNN architecture. [50]

The network architecture was named as convolutional because it includes convolution operation. The idea is that each portion of the image is multiplied by the matrix (core) convolution element by element, and the result is summed up and recorded in the same position of the output image [4].

The main idea of the CNN is as follows: an image is taken, it is then passed through a series of convolutional, non-linear layers, layers of unification and fully connected layers, and an output is generated. The output can be the class or probability of the classes that best describe the image [4, 48].

CNN has the several advantages [49]:

1. It is one of the best algorithms for image recognition and classification
2. In comparison with the fully-connected neural network, there are far fewer variables as weights of one core are used for the entire image.
3. Convenient parallelization of computations, and therefore, the possibility of implementing the algorithms and learning networks on GPU.

4. Training with the classical method of backpropagation.

4.1 Typical tasks

CNNs have been used for different tasks. These include object identification [51], semantic segmentation [33], face detection [51], recognition of human body parts [51], a semantic definition of boundaries, highlighting of objects of attention on the image and allocation of normals to the surface. These tasks are presented in Figure 15 [52]:

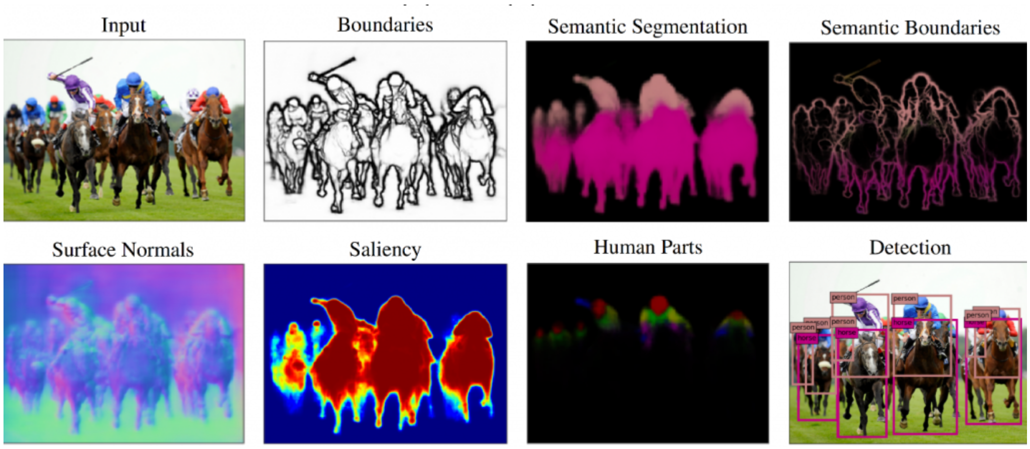


Figure 15. Typical tasks solved with CNN [52].

1. Defining Borders is the lowest-level task for which convolutional neural networks are already classically applied,

2. The definition of the vector to the normal allows reconstructing a three-dimensional image from a two-dimensional one,

3. Saliency means the definition of objects of attention. This is what the person would pay attention to when viewing the image,

4. Semantic Segmentation attempts to partition the image into semantically meaningful parts and to classify each part into one of the pre-determined classes,

5. The semantic delimitation of boundaries is the allocation of boundaries divided into classes,

6. Detection of parts of the human body,

7. The highest-level task is the recognition of objects themselves, as an example of it is face recognition.

Neural networks are becoming more popular. Many scientists compete in the invention of a more accurate neural network. One of the most popular championships is ILSVRC [53]. It is as "Olympic Games" in the field of computer vision. As part of this event, teams from around the world are competing to create a better model for solving tasks such as classification, localization, detection.

AlexNet won ILSVRC in 2012. The convolutional neural network for the first time reached an error rate of 15.4% in the classification task when rated in the top-5 format. When using the top-5 format for each test image, the model predicts the 5 most likely classes. If there is no correct answer among them, an error is counted. Since then, the CNN has been used extensively in such competitions [53]. The architecture of AlexNet is quite simple compared to modern options. The network contains 5 convolutional layers using max-pooling and dropout, as well as 3 fully connected layers. The network allows you to classify images from 1000 different categories [48]. Basic provisions [48]:

1. The network was trained on ImageNet data, which contains more than 15 million annotated images belonging to more than 22,000 categories,
2. ReLU was used as the activation function. This function allows to reduce the learning time several times compared to the traditional hyperbolic tangent,
3. Augmentation of data (creation of additional data) has been realized by means of various transformations of the original images, such as displacement, horizontal reflection, and selection of fragments,
4. Dropout was implemented to prevent retraining.

Further in these competitions in 2013, ZF Net won. The error ratio was 11.2% in the classification task [53]. The solution, for the most part, was based on the fine-tuning of the architecture of AlexNet, and some new approaches were also developed that improved the result. The main contribution of this publication is a slightly improved model of AlexNet and a very interesting way of visualization of feature maps. Basic provisions [54]:

1. The architecture is very similar to AlexNet, with the exception of a few minor modifications,

2. To train ZF Net, only 1.3 million images were used, while AlexNet was trained on 15 million images,

3. Instead of filters 11×11 in the first layer (used in AlexNet), 7×7 filters and a small step (stride) were applied in ZF Net. The meaning of this modification is that a smaller filter in the first convolutional layer allows you to store more information contained in the original pixels. As it turned out, the 11×11 filter misses a lot of useful information, especially in the first convolutional layer,

4. As the network grows, the number of filters used increases,

5. Activation function - ReLU, objective function - cross-entropy (cross-entropy), training method - batch stochastic gradient descent.

Within the framework of the publication, the visualization technology developed by the authors was developed called the "deconvolutional network", which allows analyzing the activation of various characteristics and their interrelation with the input space. The technology received such a name because it displays the attributes per pixels, that is, it performs the reverse of the convolution operation.

In 2014 the error ratio was 7.3% [53] in the classification task. VGG Net was a new neural network, that got this result. The model was a 19-layer CNN, in which 3×3 filters were used (step 1, padding -1, and 2-layer 2 push-ups). Basic provisions [54]:

1. The size of the 3×3 filters differs significantly from the size of the AlexNet (11×11) and ZF Net (7×7) filters. The rationale for this approach is that a combination of two convolutional layers of 3×3 size has a 5×5 receptive field. Thus, it is possible to simulate a larger filter while retaining the advantages of smaller filters. One of the significant advantages is the reduction in the number of parameters. In addition, having two convolutional layers, you can use two ReLU-layers instead of one, which allows increasing the efficiency of the model,

2. The combination of three convolutional layers of size 3×3 has a receptive field of 7×7, While the spatial size of the input data on each layer decreases due to convolution and pulling, the data depth increases as a result of a larger number of filters,

3. It should be noted that the number of filters is doubled after each layer. This reinforces the idea of reducing the spatial dimension and increasing the depth,

4. The network coped well with the classification problem and with the localization problem. When solving the problem of localization, the authors applied regression (see page 10 of the publication),

5. The model was created using the Caffe framework,

6. As one of the methods of data augmentation, the scale jittering method was applied,

7. After each convolutional layer, a ReLU layer was present. The training was done through a package gradient descent.

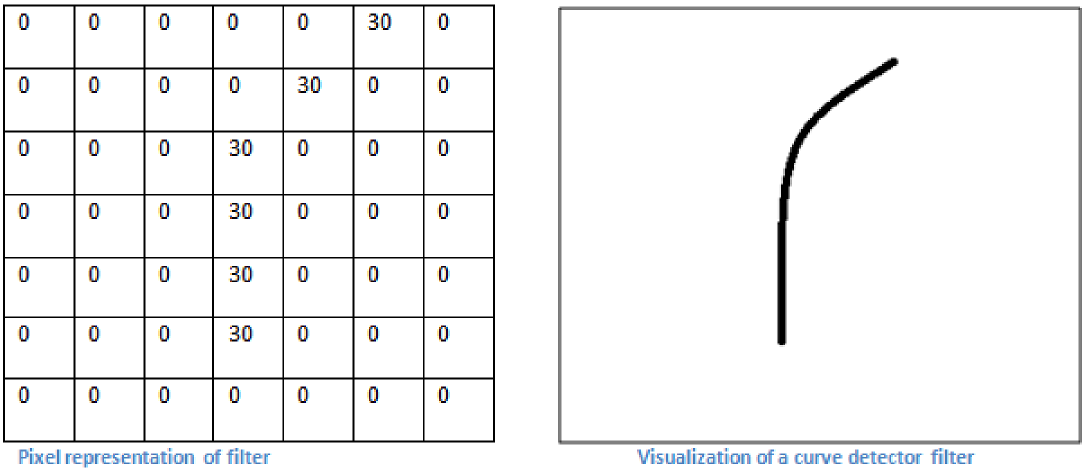
4.2 Structure

Next, the typical structure of a convolutional neural network is considered in more detail. The network consists of the many layers. After the initial layer (the input image), the signal goes through a series of convolutional layers, which alternates the actual convolution and subsampling. Alternating layers allows making "features maps". Each map is the next layer, which is reduced in size, but the number of channels increases. The output of convolutional network layers further establishes multiple layers fully connected neural network (perceptron) [48].

The first layer in the CNN is always convolution layer. Each filter can be considered as a property identifier (direct borders, simple colors, and curves). Usually, the filter is a matrix (such a matrix is also called a matrix of weights or a matrix of parameters). For example, the first filter is a detector of curves. The filter has a pixel structure in which the numerical values are higher along the region defining the shape of the curve. From the mathematical visualization: when a filter is situated in the upper left corner of the input image, it multiplies the filter values by the pixel values of this region (Figure 16) [4, 48].

In fact, if there is a form of the introductory image that is roughly like the curve that this filter represents, and all the multiplied values are summed. When the filter is moved, the value is much lower, if in the new area of the image there is nothing that the curve definition filter could detect. The output of this convolutional layer is the property map [4, 48].

In the simplest case, if there is one convolution filter (and if this filter is a curve detector), the property map will show areas in which there is more probability of having curves. If the value of the map is large, then it shows that, perhaps, something like the curve is present in the image, and such a probability activated the filter. If the filter's value is 0,



0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter

Visualization of a curve detector filter

Figure 16. Example of the filter. [55]

then it means that there was nothing in the picture that could activate the filter (there was no curve in this area). The more filters, the greater the depth of the map, and the more information there is about the introductory picture [4, 5, 48].

In a traditional convolutional neural network architecture, there are other layers that are interspersed with the convolutional ones. The classical architecture of the CNN looks like is shown in Figure 17, where input is the input image, Conv is a convolutional layer, ReLU (rectified linear unit) is a neuron with activation function, which is defined, usually, as f(x) = max(0,x), where x is is the input to a neuron, Poll is the pulling layer (otherwise sub-sampling, subsampling) [55].

Input -> Conv -> ReLU -> Conv -> ReLU -> Pool -> ReLU -> Conv -> ReLU -> Pool ->Fully Connected

Figure 17. A typical architecture of a CNN. [55]

When an image passes through one convolution layer, the output of the first layer becomes the input value of the 2nd layer. In the first layer, only the original image data is input data. For the 2nd layer, the input value is one or more property maps. Each set of input data describes the positions where certain basic characteristics occur on the source image.

When a set of additional filters on top of this will apply, the output will activate filters that represent properties of a higher level. The more convolutional layers the image passes and the further it moves across the network [4, 5, 48, 56].

The way the fully-connected layer works is to access the output of the previous layer and define properties that are more related to a particular class. For example, if a program predicts that some image contains a dog, map of features that reflect high-level characteristics, such as paws or 4 legs, should have high values. Similarly, if the program recognizes that the image contains a bird, it has high values in the map of features represented by high-level characteristics such as wings or beaks. A fully-connected layer explains the fact that high-level functions are strongly associated with a particular class and have certain weights. When the products of the weights with the previous layer are calculated, the correct probabilities for different classes are obtained (Figure 18) [48, 56].

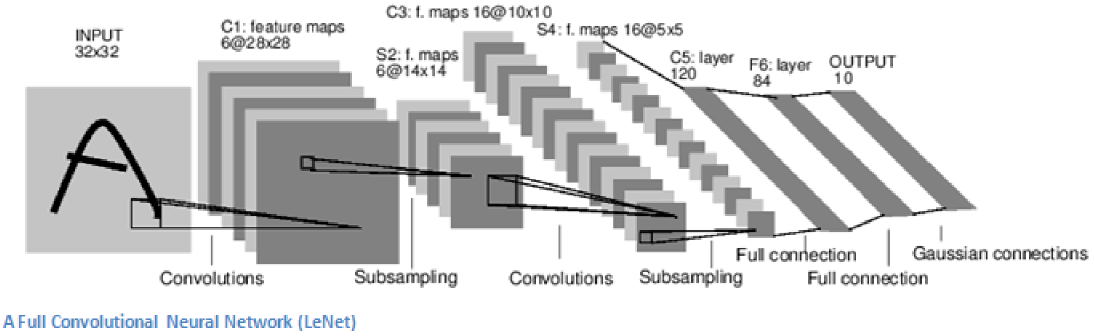


Figure 18. Classical architecture of the CNN. [55]

4.3 Learning algorithms

After the structure of the CNN is selected the weights of the network need to be learned. It is a difficult process, which could be compared to how people learn to recognize objects. The weight of the network are typically initialized as random numbers. The CNN is trained by showing labeled images to it. For example, a set of training images, in which there are thousands of images of dogs, cats, and birds. Each image has a label with the name of the animal [48, 57].

The most common method to train CNN's is backpropagation. The method of backpropagation of errors can be divided into 4 separate blocks: direct distribution, loss function, back distribution, and weight update. During direct propagation, a training image is taken and passed through the entire network. In the first training example, since all weights or filter values were initialized randomly, the output value does not give preference to a particular class. A network with such weights cannot find a base-level property and cannot reasonably determine an image class. This leads to a loss function. The loss function can be expressed in different ways, but the Mean Squared Error (MSE) is often used, where the target is real data and output is a predicted value:

$$E = \sum 1/2(target - output)^2 \tag{3}$$

The main idea is to ensure that the predicted label (the output of the convolution layer) is the same as the label of the training image (this means that the network has made the right assumption). To achieve this, the number of losses should be minimized. Visualizing this as an optimization problem from mathematical analysis, inputs (weights) most directly contributed to the losses (or errors) of the network should be found.

One way to visualize the idea of minimizing loss is a three-dimensional graph, where the weights of the neural network (obviously more than 2, but the example is simplified) are independent variables, and the dependent variable is a loss. The task of minimizing losses is to adjust the weights so that the loss is minimized, visually. The lowest point of the cup-like object needs to be reached. To achieve this, the derivative loss considering the weights needs to be found.

This is the mathematical equivalent of dL / dW, where W is the weight of a certain layer, L is a loss function. Now network should determines which weights have had greater impact on the loss and makes it possible to adjust them to reduce losses. After derivative are calculated. All the filter weights are updated, so that they change in the direction of the gradient as [4, 5, 48, 56]:

$$w = w_i - \eta \frac{dL}{dW} \tag{4}$$

Where w_i is initial weight, η - learning rate. Learning rate is an additional parameter. A high learning rate means that the weight updates take larger steps, so the sample may take less time to gain the optimal set of weights. But too high a speed of training can lead to very large and insufficiently precise jumps.

The process of direct propagation, loss function, back propagation and updating of weights is usually called one sampling period (or epoch-era). The program will repeat this process a fixed number of periods for each training image. After the parameters update is completed on the last training sample, the network, in theory, should be sufficiently well-trained and the weights of the layers are set correctly. [4, 5, 48, 56]

4.4 CNN for segmentation

CNNs can also be used for the image segmentation, for example SegNet (Figure 19) [58].

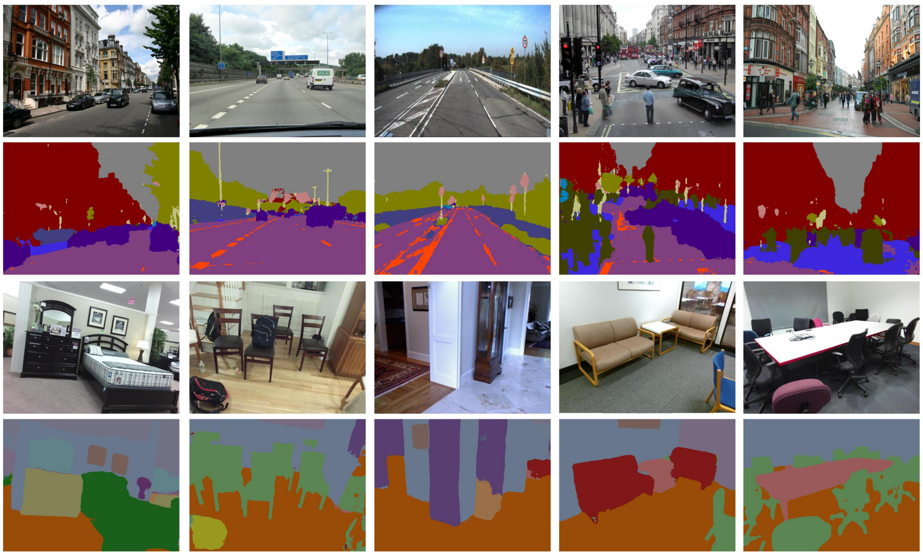


Figure 19. Example of the results by SegNet [59].

SegNet is a deep encoder-decoder architecture for multi-class pixelwise segmentation researched and developed by members of the Computer Vision and Robotics Group at the University of Cambridge, UK [59]. Encoder-decoder architecture works as follow: encoder receives the input data and generates some condensed representation of them. This condensed representation is fed to the input of the decoder, which must already generate the output data [59]. The main idea of the SegNet is that instead of the label, not a number but an image is supplied, a new layer "Upsample" is added to increase the dimension of the layer.

The architecture of the SegNet is presented on the figure 20. It consists of a sequence of non-linear processing layers (coders) and a corresponding set of decoders, followed by a pixel classifier. The encoder consists of convolutional layers with batch normalization and nonlinearity of ReLU, followed by nonoverlapping max pooling and sub-sampling. The most important part of the SegNet is the use of maximum volume indices in decoders for a low-resolution sampling of cards. Thereby it helps to serve high-frequency parts in segmented images and reduce the total number of learning parameters in decoders. The whole architecture can be trained from end to end, using stochastic gradient descent [59].

One more neural network for the segmentation is a new form that combines Conditional

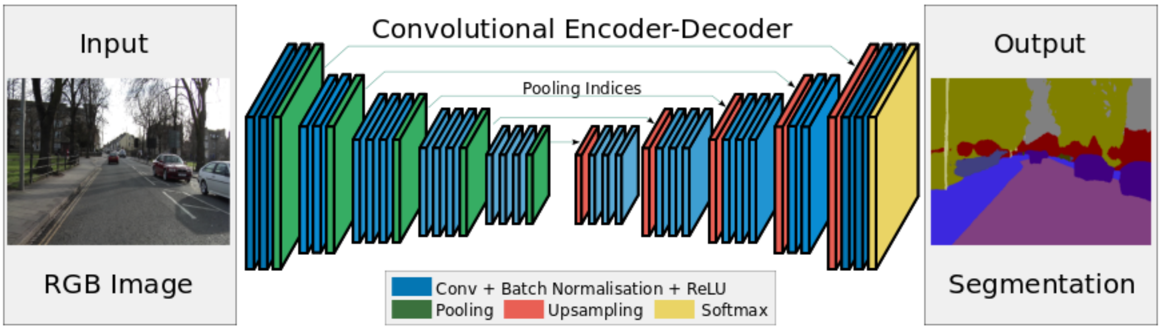
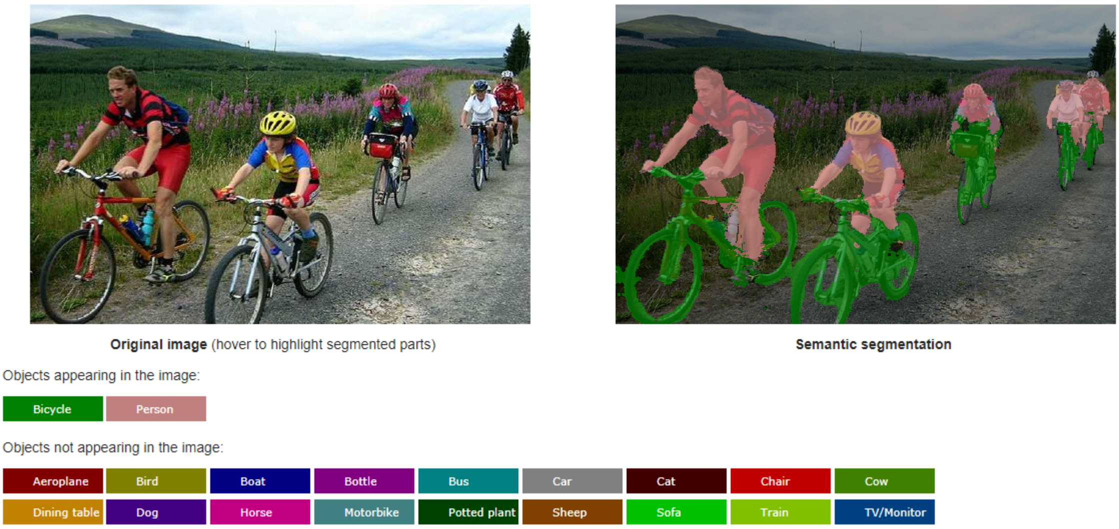


Figure 20. Example of the results by SegNet [59].

Random Fields and Recurrent Neural Networks (CRF-RNN) [60]. Conditional Random Field, CRF is a statistical method of classification, the characteristic difference of which is the ability to take into account the "context" of the object being classified. One of the main advantages of this model is that it does not need to model the probabilistic relationships between the so-called observable variables. CRF-RNN takes the advantages from the CRF and CNN and combines them. This system makes possible to train the whole deep network end-to-end with the usual back-propagation algorithm, avoiding offline post-processing methods for object delineation [60].

Experimental results were presented with the proposed CRFRNN framework. The Pascal VOC 2012 dataset and the Pascal Context dataset were used for the training. On the figure 21 presents the example of the framework, where the system segments and recognizes objects on the image.



Original image (hover to highlight segmented parts)

Semantic segmentation

Objects appearing in the image:

Bicycle Person

Objects not appearing in the image:

Aeroplane Bird Boat Bottle Bus Car Cat Chair Cow

Dining table Dog Horse Motorbike Potted plant Sheep Sofa Train TV/Monitor

Figure 21. Example of the results by CRFRNN [61].

5 PROPOSED METHODS

The main task of this research is to implement a CNN based approach for the segmentation. For this purpose the method was evaluated, which contains the following steps:

1. Unsupervised segmentation,
2. Superpixel feature extraction,
3. Superpixel classification.

Unsupervised segmentation means that image is split into several parts. They are called superpixels. The input parameters for the first step are ground truth images. The output parameters are superpixels. For this purposes, two methods are considered:

1. Multiscale Combinatorial Grouping (MCG) algorithm,
2. Simple linear iterative clustering (SLIC) algorithm.

In the second step, CNN was implemented for superpixel feature extraction. Input parameters for this step are superpixels, the output parameters are feature vectors.

The results from the second step are applied for the further analyses - superpixels classification. Support vector machine was used for these purposes. Where input parameter is feature vectors. Classification means to define what is represented on the image: seals (this is first class) or background (this is the second class).

The main steps of the segmentation method are presented on the Figure 22.

5.1 Unsupervised Segmentation

The first step of the proposed method is unsupervised segmentation. The aim of the unsupervised segmentation has divided an image into different segments. Segments are the pixels adjacent to each other and having similar spectral characteristics. Usually, these segments are called superpixels.The major difference in segmentation method used in this

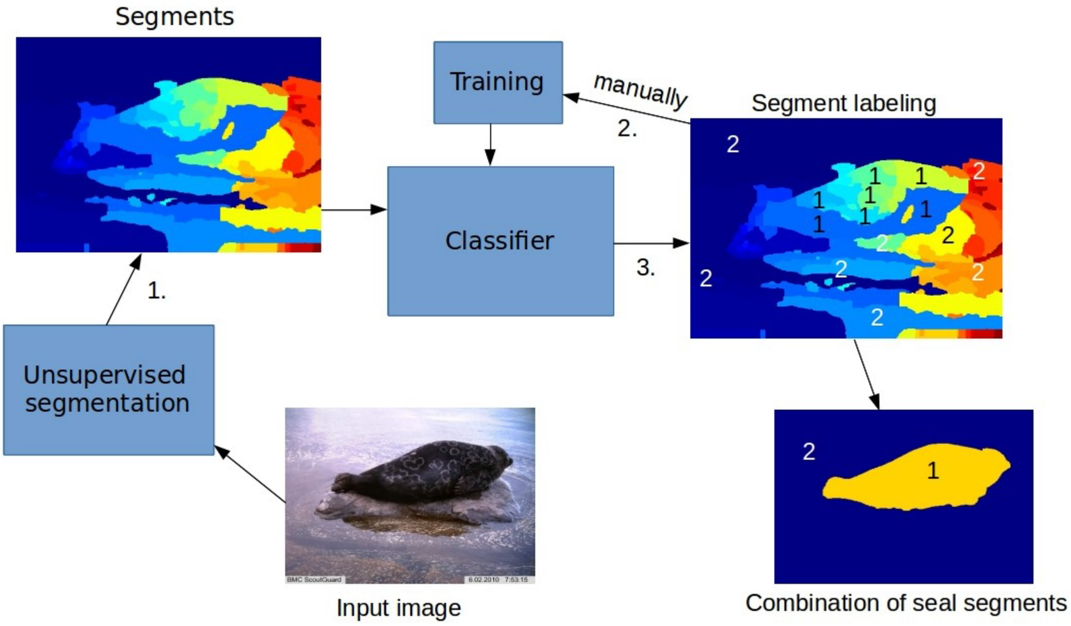


Figure 22. The main steps of the segmentation method based on Saimaa ringed seal segmentation proposed [10, 11].

study and previous study of ringed seal segmentation is in unsupervised segmentation part.

In this part, two algorithms for the unsupervised segmentation were used. There are Multiscale Combinatorial grouping (MCG) and Simple Linear Iterative Clustering (SLIC).

5.1.1 Multiscale combinatorial grouping

The idea of the Multiscale Combinatorial Grouping is to exploit multiscale hierarchal information for image segmentation. Then using the smart combining technique, which would combine regions from different scales into possible object candidates [62].

MCG is a unified approach that generates and then group together high-quality multiscale regions. It does not depend upon pre-computed hierarchies and superpixels.

The image is segmented independently into multiple resolutions. Each image now represents a family of super pixels, from a fine set of superpixel to the complete domain. Hierarchical representation of the image boundaries are called Ultrametric Contour Map UCM, further, represents these sets by assigning the weight to the boundary of adjacent regions in the hierarchy by the index at which they were merged. Now simply, thresholding at a particular level in UCM produces segmentation [62].

In the second phase, all hierarchical boundaries are aligned and combined in a multiscale hierarchy. Eventually, a grouping component scans efficiently through the combinatorial spaces and produces a ranked list of object candidates. The ranking strategy utilizes the information about size, location, shape, and contours [62].

5.1.2 Simple linear iterative clustering

SLIC algorithm is a modified algorithm for clustering k-means, in which the clustering error function is minimized. The main difference between SLIC and classic k-means is the restriction of the search area: the pixels for each segment are searched not in the whole image, but in a small area proportional to the average size of the segment. As a measure of proximity, the weighted sum of Euclidean distances by coordinates and three color components is used [63].

Each point of the image is characterized by a five-dimensional vector p = (c1, c2, c3, x, y), where c1, c2, and c3 are the coordinates of the point in the selected color space; x and y are the spatial coordinates of the image point [63].

The SLIC algorithm includes the following steps [63]:

1. The image is divided into K fragments of size $a \times a$ that specify the initial approximation of superpixel clusters. As the initial centers of superpixel fragments, their geometric centers C_k are chosen.

2. The coordinates of the fragment centers are corrected from the condition of the minimum value color gradient in a 3×3 neighborhood of the geometric center,

3. The formation of local clusters in $2a \times 2a$ neighborhoods of the centers C_k is similar to the method of k-means. The distance D between the center and the points of the fragment is computed as a combination of Euclidean distances along the color d_c and the spatial d_s component of the description of the point p:

   $d_c = \sqrt{(c_j1 + c_i1)^2 + (c_j2 + c_i2) + (c_j3 + c_i3)}$

   $d_s = \sqrt{(x_j + x_i)^2 + (y_j + y_i)}$

   $D = \sqrt{d_c^2 + (\frac{d_s}{a})^2 * m^2},$

where m is the parameter specifying the ratio of the contributions of the two components of the image description to the distance D. I and j are the numbers of the points between which the distance is calculated,

4. Determination of new cluster centers and calculation of center displacements,

5. Repeat steps 3 and 4 until the center offset between iterations is less than the specified value.

5.2 CNN for the superpixel feature extraction

For superpixel feature extraction CNN was used. The input parameters are superpixels, which were received from the previous step (unsupervised segmentation). Neural networks are trained to retrieve information from images, the computational complexity grows with each level. For example, the first layer memorizes strokes, color changes, and brightness, The second layer examines the combinations of the outputs of the first layer and so on.

5.2.1 AlexNet

For this research network, AlexNet was chosen. The architecture of AlexNet is quite simple compared to other modern CNNs. The network contains 5 convolutional layers using max-pooling and dropout, as well as 3 fully connected layers. The network allows classification of images of 1000 different categories. AlexNet has 25 layers, including input and output layers. Hidden layers are divided into 8 subcategories. The first and the second layers perform convolution, ReLU, cross-channel normalization. The fifth performs convolution, ReLU, and max pooling. The six and the seventh are fully connected layers with ReLU and drop-out which helps reduce overfitting during training. The eighth is the last fully-connected layer with 1000 outputs which correspond to the number of classes. Logical layers 1-8 correspond to actual layers 2-23. Layer 24 is softmax and layer 25 is the classification output. It is a fully connected layer. The structure of the AlexNet is presented on Figure 23.

For our purposes, layers that have been trained to extract features from the image have to be selected. The layers that use these features for classification are not suitable for this task. Based on the structure of the AlexNet, the classification layers start with layer 23, that is the last fully-connected layer with 1000 outputs. Layers 21 and 22, 'relu7'

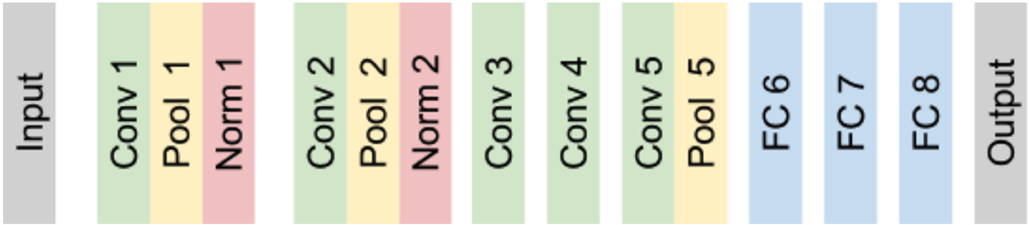


Figure 23. The architecture of layers of the neural network AlexNet. Layer designations: Conv - convolutional, Pool - max-pooling, Norm - normalization [64].

and 'drop7' are rectifier and dropout which were just used during training and do not hold any learned weights themselves, that is why the last feature extraction layer in this research is 20, 'fc7'. It is a fully-connected layer with 4096 outputs. The size of the image changes dimensionality reduction from 227x227x3 to 4096-dimensional vector. To construct classifier after 20 layers of AlexNet, it is only needed to train any sort of classifier over 4096-dimensional vectors representing our images.

5.3 Superpixel classification

The third step of the proposed method is superpixel classification. The task of classification consists of determining which class (of at least two originally known) the given object belongs to. Usually, such an object is a vector in an n-dimensional real space. The vector coordinates describe individual attributes of an object [65].

For this purpose, Support Vector Machine (SVM) was used. It classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin is the maximal width of the slab parallel to the hyperplane that has no interior data points [65, 66].

An example of the SVM are represented at the (Figure 24), here "+" indicating data points of the first class, and "-" indicating data points of the second class.

For this research the first class represents seals, the second class is for background.

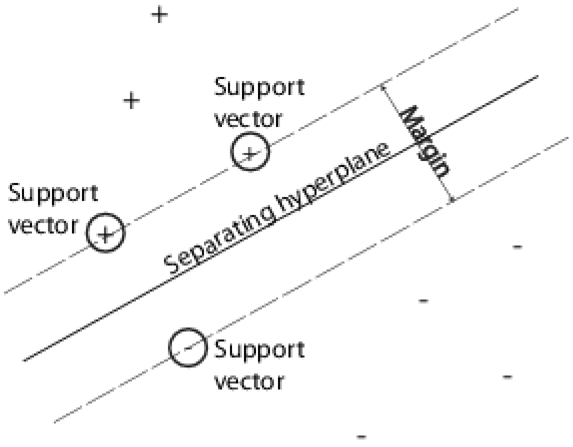


Figure 24. Example of the svm [67]

6 EXPERIMENTS

6.1 Data

The data for this research was collected by UEF. The dataset consists of 168 images, with a total of 47 individual seals. All of the images were taken in the natural habitat of seals (Figure 25).

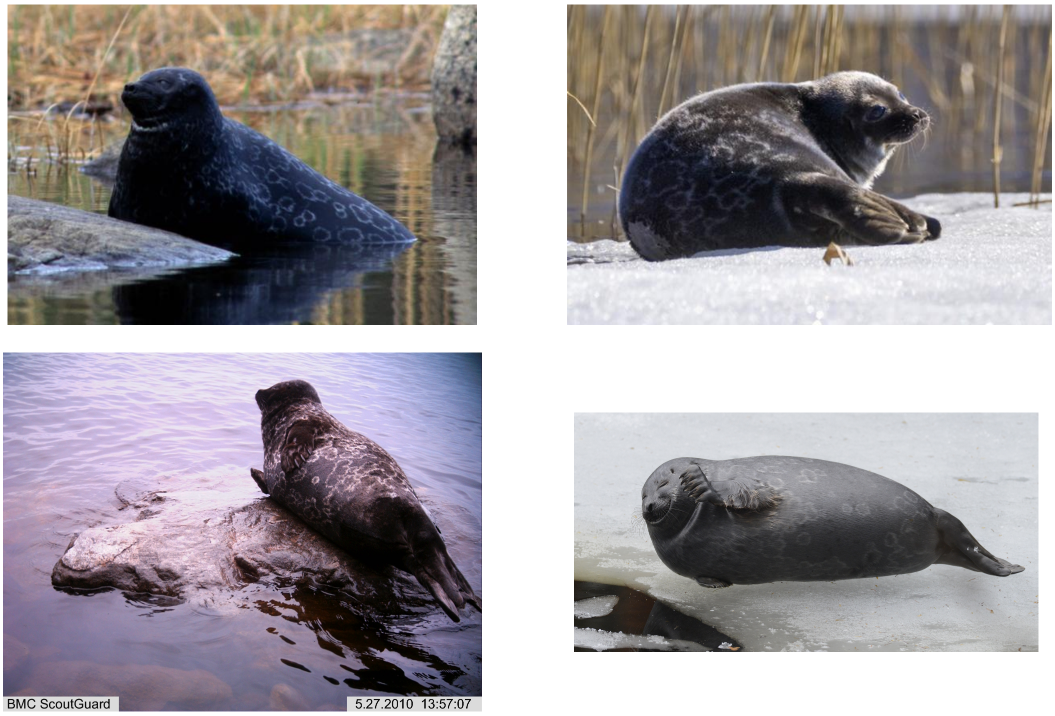


Figure 25. Examples from the Saimaa ringed seals database

Images in the database do not contain just seals. Apart from seals, there can be other objects in the images, such as some parts of human body, a sensor on a seal's body. Some images contain more than one seal. Other problems are connected with varying image quality. On the one hand, there are some images that were captured at night time, which complicates detection of a seal, as it is dark and does not stand out against the background. On the other hand, there are images captured in the daytime which is overexposed. Some images contain seals only partially (Figure 26).

The ground truth for the seal images was formed manually. In these images, seals were segmented from the background. For this purpose binary masks from the previous research were used [12]. A total 168 images were segmented manually. Ground truth (GT)

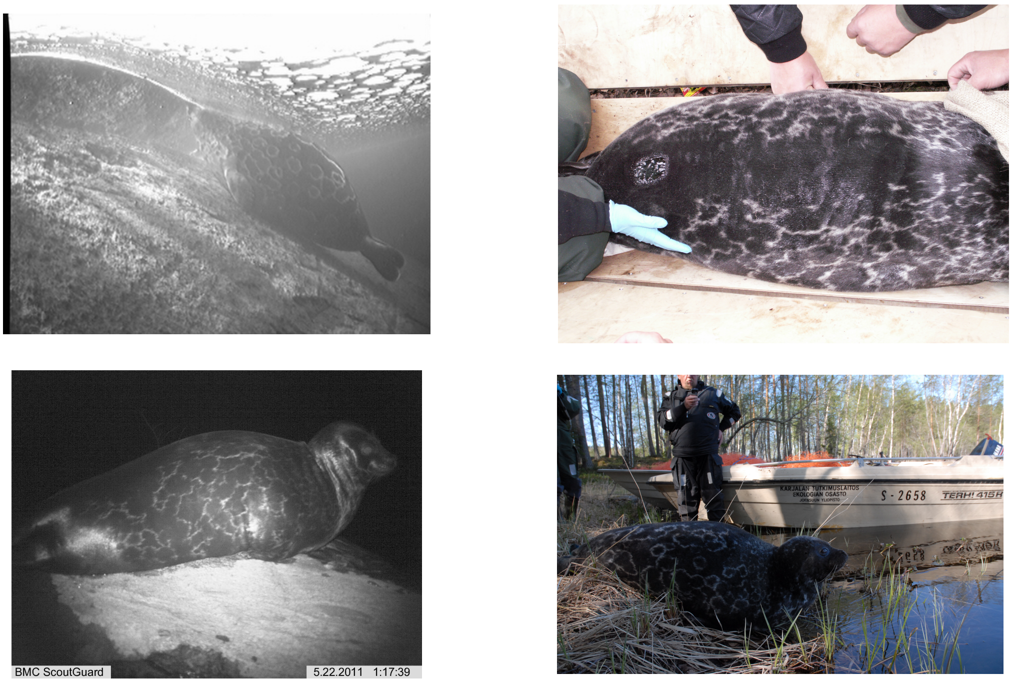


Figure 26. Examples of challenging images from the Saimaa ringed seals database

images were used to evaluate segmentation results. Examples of the GT images are shown in Figure 27.

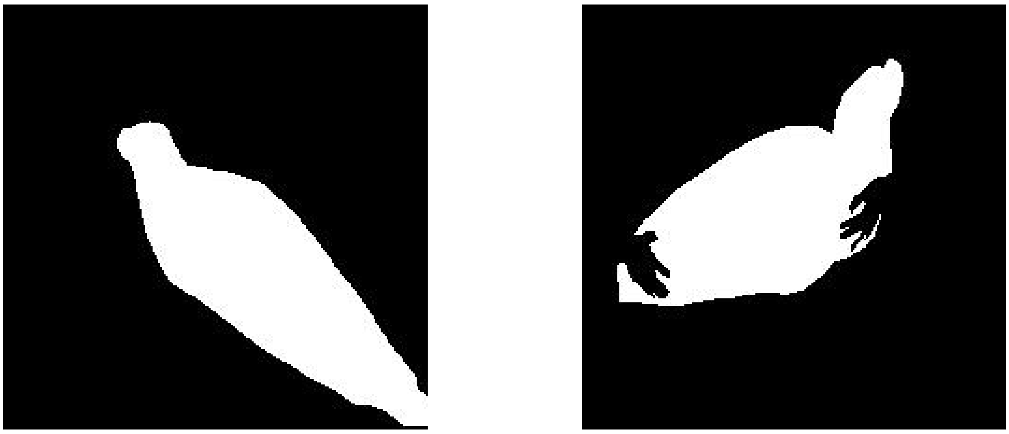


Figure 27. Examples of the GT images

6.2 Evaluation criteria

During this research, two evaluation criteria were applied: the confusion matrix for the superpixel classification [68] and the Jaccard similarity for the final segmentation results [69].

The confusion matrix is a matrix of size N on N where N is the number of classes. Columns of this matrix are reserved for expert decisions, and rows for the classifier solutions. When a superpixel from the test set is classified, the number of the class line is incremented. The class line is defined by the output of the classificator. Also, the class column to which the document really belongs is incremented to [68].

In this case, there are only two classes: the seals and the background. The accuracy of classifying is equal to the ratio of the corresponding diagonal matrix element and the sum of the entire class line. Completeness is the ratio of the diagonal element of the matrix and the sum of the entire column of the class. The Matlab function confusion at was used during this research to calculate confusion matrix [68].

The other evaluation criterion is Jaccard similarity. For this research, each binary matrix of the result images was compered with the corresponding ground truth. The range of similarity is from 0% to 100%. The higher the percentage, the more similar result was got. The Jaccard coefficient is defined as the size of the intersection of the image's ground truth (I_{GT}) and the segmentation result of the image (I_S) divided by the size of the union the image's ground truth and the segmentation result [69]:

$$S_{Jaccard} = \frac{|I_S \cap I_{GT}|}{|I_S \cup I_{GT}|}. \tag{5}$$

6.3 Implementation

The research consists of two experiments. They are presented in more detail in the 6.4.

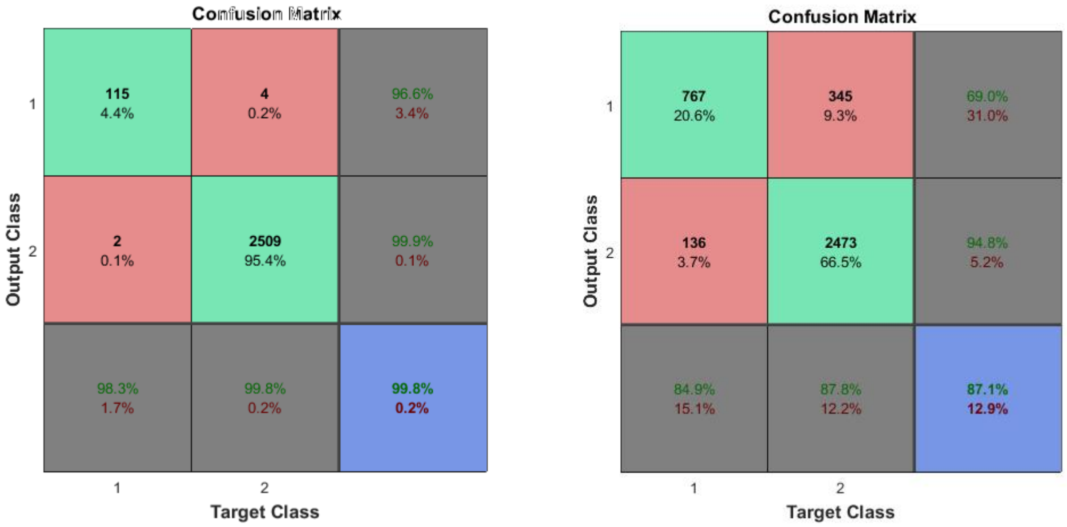
Information about superpixels image labels was prepared in an earlier research [12]. For MCG, the number of superpixels varies between images and for SLIC the number of superpixels is a constant variable which is equal 50 in this research.

The next step is an implementation of the CNN. For this task the MATLAB function, AlexNet was used. The size of the images for this function should be 227x227x3. The network in Matlab has 25 layers, including input and output. The 20th layer ('fc7') was used fo feature extraction in this research. The CNN transforms the image of size 227x227x3 to a 4096-dimensional feature vector.

SVM was used for classifying the superpixels. For this task Matlab function, 'fitcecoc' was used. The parameters for SVM were defined automatically. The kernel is linear. Kernel scale is selected automatically using a heuristic procedure. Kernel offset is set to 0.1.

6.4 Results

The proposed method was tested on a dataset comprising 168 images. The images were divided into training and test data as follows: 52 images formed the training set and 117 images were used for testing. The number of MCG superpixels was 1129 for the training set and 2630 for the test set. The result of superpixel classification is shown in Figure 28, where the output class is a class predicted by the classifier, and target class is the actual class of a superpixel. The first class stands for an object (a seal), and the second class is a background. Based on the results for MCG superpixels, the number of incorrectly classified superpixels for the first class was two and it was equal to four for the second class, while for the SLIC superpixels there were 136 and 345 incorrectly classified superpixels, respectively. The classification accuracy of the method using MCG superpixels was 99,8%, where it reached only 87,1% when SLIC superpixels were used.



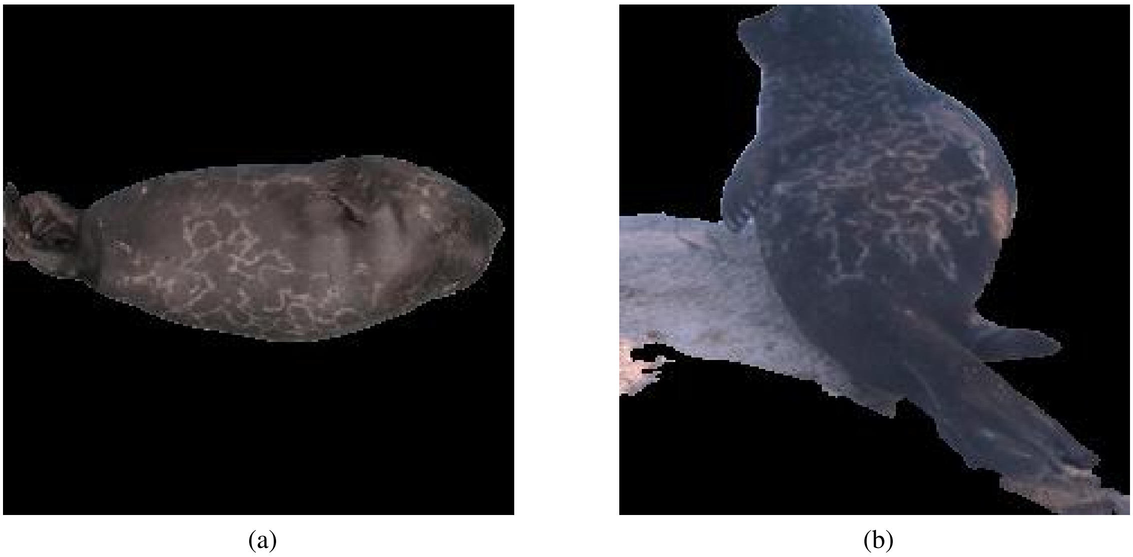
(a)

(b)

Figure 28. Confusion matrix for the superpixel calssification: a) superpixels obtained using MCG; b) superpixels obtained using SLIC.

The result, which was obtained by MCG superpixels, is more promising. For the classification, the size of the dataset is the same, but the amount of superpixels are different. Results of superpixels obtained using MCG is better (99.8% and 87.1%) then results of the superpixels obtained using SLIC. There can be several reasons for this: not enough superpixels, when superpixels were obtained using SLIC, superpixels from the Matlab function did not obtain the border of the seals.

Example results for the segmentation are shown in Figure 29. Some images were segmented incorrectly. There are various reasons for it. Firstly, the superpixels were not ideal, therefore some images contain seals with a background. In addition, the illumination varies between images, which also makes segmentation difficult.



(a) (b)

Figure 29. Examples of segmentation results: a) Correctly segmented images; b) Incorrectly segmented image.

The percentage of correctly segmented images with different Jaccard similarity thresholding are presented in Figure 30. The mean value of the Jaccard similarity over all images was 75.81%.

To compare results of the segmentation, which are presented in the Table 1, MCG shows the better results. The amount of the better similar images is bigger than in the method, where superpixels were getting by the Matlab function.

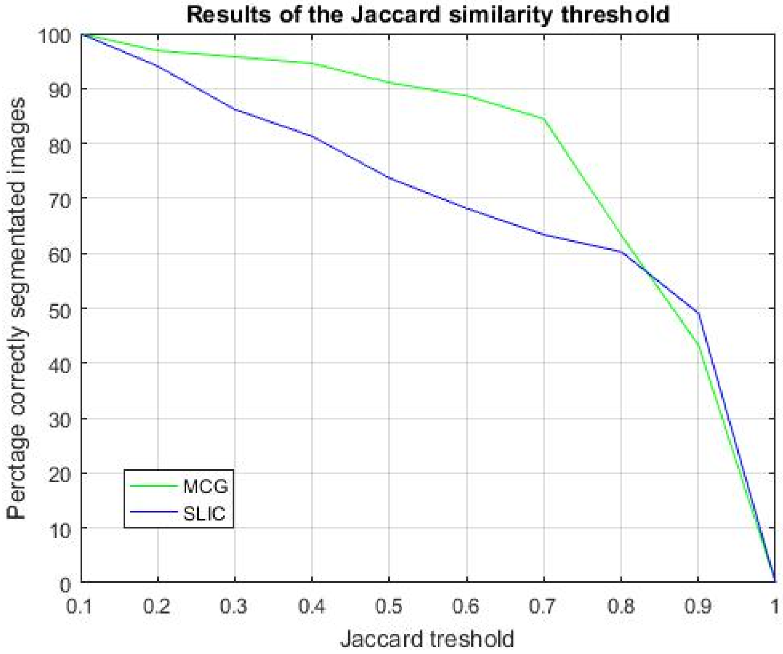


Figure 30. Correctly segmenting images by the Jaccard similarity

Table 1. Correctly segmented images with different thresholds by the Jaccard similarity

	MCG	SLIC
Correctly segmented images with Jaccard threshold 0.9	43.4%	49.1%
Correctly segmented images with Jaccard threshold 0.7	84.5%	63.4%
Correctly segmented images with Jaccard threshold 0.5	91.1%	73.7%
Correctly segmented images with Jaccard threshold 0.3	95.6%	86.2%
Mean Jaccard similarity	75.81%	67.63%

7 DISCUSSION

7.1 Current results

In this work, Saimaa ringed seal was an object of research. It is one of the most endangered seals in the world. They have a distinctive patterning of dark spots surrounded by light gray rings. This pattern is unique to each seal enabling the identification of individuals [10, 11, 12].

At the first step unsupervised segmentation was applied. Two methods for unsupervised segmentation (i.e. MCG and SLIC) were considered. As the result superpixels, which were obtained by MCG method, more accurately determine the boundary of objects. Therefore, the results of classification and segmentation of the method where MCG algorithm was used, is better.

AlexNet was used to extract features from the images. SVM was used for superpixel classification. The proposed method was evaluated using the Saimaa ringed seal image database. The Jaccard similarity and the confusion matrix were used as evaluation criteria. The method was evaluated using a dataset consisting of 168 images of seals. The accuracy of the seal classification was 99.7%. About 75% of the images were correctly segmented with a Jaccard similarity threshold.

7.2 Future work

The main task for the future work is the implementation and evaluation of more sophisticated CNN based segmentation such as SegNet, CRF-RCNN, ZF Net and VCG Net and other new neural networks, which will give better results.

Also the future work can consist of the following points:

- Apply another method for classification.

  During this research, only SVM was used for classification. For example, CNN can be used for classification.

- Use GPU for the algorithm.

The implementation of the CNN can be improved by process parallelization using GPU for CNN, and thus processing the time can be decreasing.

8 CONCLUSION

The Saimaa ringed seal is among the most endangered species of seals in the world. Its population consists of only about 360 individuals nowadays. Therefore, constant monitoring of the seal population demographics is necessary to save this species from extinction. The Saimaa ringed seal has a distinctive patterning of dark spots surrounded by light gray rings, which is unique to each seal. This pattern makes identification of individuals possible. To capture images of seals camera traps are used. In this research, a method was proposed to segment an image in order to separate a seal from the image background. This approach helps improve the identification process performance.

During the research, existing CNN-based segmentation methods were applied. We propose an approach consisting in unsupervised segmentation and superpixel classification. Two methods for unsupervised segmentation (i.e. MCG and SLIC) were considered. The produced superpixels were classified using SVM. We used a trained AlexNet to extract features from the superpixels. We assessed the effectiveness of the proposed method on the Saimaa ringed seal image dataset consisting of 168 images. The Jaccard similarity and the confusion matrix were used as evaluation criteria. Out of the two unsupervised segmentation methods considered, MCG produced better overall results. The accuracy of the superpixel classification is 99.7%. About 75% of images were correctly segmented.

In future, we would like to base our approach on more sophisticated CNN-based segmentation methods such as SegNet, CRF as RNN.

REFERENCES

[1] Fred Brauer and Carlos Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 40. Springer, 2001.

[2] Cecilie E Bugge, John Burkhardt, Kari S Dugstad, Tone Berge Enger, Monika Kasprzycka, Andrius Kleinauskas, Marit Myhre, Katja Scheffler, Susanne Ström, and Susanne Vetlesen. Biometric methods of animal identification. *Course notes, Laboratory Animal Science at the Norwegian School of Veterinary Science*, pages 1–6, 2011.

[3] Suomen luonnonsuojeluliitto ry. website: http://www.sll.fi.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[5] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[6] Hongyuan , Fanman Meng, Jianfei Cai, and Shijian Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016.

[7] Wilburn E Reddick, John O Glass, Edwin N Cook, T David Elkin, and Russell J Deaton. Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. *IEEE Transactions on Medical Imaging*, 16(6):911–918, 1997.

[8] Paulo Correia and Fernando Pereira. Objective evaluation of relative segmentation quality. In *Proceedings of International Conference on Image Processing*, volume 1, pages 308–311. IEEE, 2000.

[9] Eitan Sharon, Achi Brandt, and Ronen Basri. Segmentation and boundary detection using multiscale intensity measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001.

[10] Eerola T. Koivuniemi M. Auttila M. Levänen R. Niemi M. Kunnasranta M. Kälviäinen H. Zhelezniakov, A. Segmentation of saimaa ringed seals for identification purposes. In *Advances in Visual Computing*, pages 227–236. Springer Lecture Notes in Computer Science, LNCS Vol. 9475., 2015.

[11] Artem Zhelezniakov. Automatic image-based identification of saimaa ringed seals. Master's thesis, Lappeenranta University of Technology, Finland, 2015.

[12] Tina Chehrsimin. Enhanced methods for saimaa ringed seal identification. Master's thesis, Lappeenranta University of Technology, Finland, 2016.

[13] Simon A Cole. *Suspect identities: A history of fingerprinting and criminal identification*. Harvard University Press, 2009.

[14] Wayne C Boncyck and Ronald H Cohen. Image capture and identification system and process, January 3 2017. US Patent 9,536,168. Accessed: 2017-07-31.

[15] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

[16] Anil Jain, Lin Hong, and Sharath Pankanti. Biometric identification. *Communications of the ACM*, 43(2):90–98, 2000.

[17] Hjalmar S Kühl and Tilo Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in ecology & evolution*, 28(7):432–441, 2013.

[18] Santosh Kumar and Sanjay Kumar Singh. Biometric recognition for pet animal. *Journal of Software Engineering and Applications*, 7(5):470, 2014.

[19] Louis B Meadows. Pet identification system and method, January 18 2005. US Patent 6,845,382. Accessed: 2017-07-31.

[20] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013(1):52, 2013.

[21] Rahul Sharma, Nidhi Mishra, and Sanjeev Kumar Yadav. Fingerprint recognition system and techniques: A survey. *International Journal of Scientific & Engineering Research*, 4(6):1670, 2013.

[22] Jean-Francois Mainguet. Biometrics for animals. website: http:// biometrics. mainguet.org/types/animals.htm, 2017. Accessed: 2017-07-31.

[23] Richard B Sherley, Tilo Burghardt, Peter J Barham, Neill Campbell, and Innes C Cuthill. Spotting the difference: towards fully-automated population monitoring of african penguins spheniscus demersus. *Endangered Species Research*, 11(2):101–111, 2010.

[24] KM Kramer, DS Hedin, and DJ Rolkosky. Smartphone based face recognition tool for the blind. In *Annual International Conference of the IEEE*, pages 4538–4541. IEEE, 2010.

[25] Yong Zhu, Tieniu Tan, and Yunhong Wang. Biometric personal identification based on iris patterns. In *Proceedings of the 15th International Conference of Pattern Recognition*, volume 2, pages 801–804. IEEE, 2000.

[26] Clinton P Rusk, Christine R Blomeke, Mark A Balschweid, SJ Elliot, and Dan Baker. An evaluation of retinal imaging technology for 4-h beef and sheep identification. *Journal of Extension*, 44(5):1–33, 2006.

[27] Benjamin Hughes and Tilo Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, pages 1–16, 2016.

[28] Dr Tilo Burghardt. Plan for fingerprinting great white sharks. http:// www.bristol.ac.uk / news /2010/7177.html, 2010. Accessed: 2017-07-31.

[29] Arid Recovery News. Keeping cool in the desert. website: http://aridrecovery. org.au/ blog/AridRecoveryNews/post, 2015. Accessed: 2017-07-31.

[30] Robert J Schalkoff. *Digital image processing and computer vision*, volume 286. Wiley New York, 1989.

[31] DENSO WAVE INCORPORATED. Terminology and definitions. http: //densorobotics .com, 2012.

[32] Rafael C Gonzalez and Richard E Woods. Image processing. *Digital image processing*, 2, 2007.

[33] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3378–3385. IEEE, 2012.

[34] Kelly M Halloran, James D Murdoch, and Matthew S Becker. Applying computer-aided photo-identification to messy datasets: a case study of thornicroft's giraffe (giraffa camelopardalis thornicrofti). *African Journal of Ecology*, 53(2):147–155, 2015.

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[36] Jonathan P Crall, Charles V Stewart, Tanya Y Berger-Wolf, Daniel I Rubenstein, and Siva R Sundaresan. Hotspotter-patterned species instance recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 230–237. IEEE, 2013.

[37] Meeri Koivuniemi, Miina Auttila, Marja Niemi, Riikka Levänen, and Mervi Kunnasranta. Photo-id as a tool for studying and monitoring the endangered saimaa ringed seal. *Endangered Species Research*, 30:29–36, 2016.

[38] Robert M Haralick and Linda G Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985.

[39] Liming Wang, Jianbo Shi, Gang Song, and I-Fan Shen. Object detection combining recognition and segmentation. In *Proceedings of Asian Conference on Computer Vision*, pages 189–199. Springer, 2007.

[40] Alexey Efros by Microsoft research. Image parsing. website: http:// courses. graphicon .ru /main/vision2, 2011. Accessed: 2017-07-31.

[41] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[42] Mehmet Ozkan, Benoit M Dawant, and Robert J Maciunas. Neural-network-based segmentation of multi-modal medical images: a comparative and prospective study. *IEEE Transactions on Medical Imaging*, 12(3):534–544, 1993.

[43] Alain Tremeau and Nathalie Borel. A region growing and merging algorithm to color segmentation. *Pattern Recognition*, 30(7):1191–1203, 1997.

[44] Frank Y Shih and Shouxian Cheng. Automatic seeded region growing for color image segmentation. *Image and Vision Computing*, 23(10):877–886, 2005.

[45] André Bleau and L Joshua Leon. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, 77(3):317–370, 2000.

[46] Intel. website: https:// habrahabr.ru/ company /intel/blog/266347/. Accessed: 2017-07-31.

[47] Bo Peng, Lei Zhang, and David Zhang. A survey of graph theoretical approaches to image segmentation. *Pattern Recognition*, 46(3):1020–1038, 2013.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[49] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[50] Doris Y Kim. Computing and data handling. In *Proceedings of the 38th International Conference on High Energy Physics*, 2016.

[51] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665, 2014.

[52] HighLoad. Classical tasks for cnn. website: https:// pinme.club /pin/nejronnye-sety/, 2016. Accessed: 2017-07-31.

[53] Stanford Vision Lab. Large scale visual recognition challenge (ilsvrc). website: http://www.image-net.org/challenges/LSVRC/, 2015. Accessed: 2017-07-31.

[54] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision (ECCV)*, pages 818–833. Springer, 2014.

[55] Adit Deshpande. Convolutional neural network. website: https:// habrahabr.ru / post /309508/.google.com, 2016. Accessed: 2017-07-31.

[56] M.C. Munteanu, A. CALIMAN, and C. Zaharia. Convolutional neural network. Google Patents, https://www.google.com/patents/US9665799, 2017. US Patent 9,665,799. Accessed: 2017-07-31.

[57] Thibaut Durand, Nicolas Thome, and Matthieu Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752, 2016.

[58] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[59] Alex Kendall. Segnet. website: http://mi.eng. cam.ac.uk/projects/segnet/, 2015. Accessed: 2017-07-31.

[60] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[61] Bernardino Romera-Paredes Philip Torr Shuai Zheng, Sadeep Jayasumana. Crf as rnn. website: http://www.robots.ox.ac.uk/ szheng/crfasrnndemo, 2015. Accessed: 2017-07-31.

[62] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335, 2014.

[63] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[64] Aleksei Paevskiy. Neuronovosti. website: http://neuronovosti.ru/convolutional/, 2016. Accessed: 2017-07-31.

[65] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14:85–86, 1998.

[66] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874, 2008.

[67] Inc. The MathWorks. Support vector machines for binary classification. website: https://se.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html, 1994-2017. Accessed: 2017-07-31.

[68] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[69] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.