ONE-TO-MANY AND MANY-TO-MANY MATCHING FOR SAIMAA RINGED SEAL RE-IDENTIFICATION

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2022

Antti Vilkman

Examiners: Professor Heikki Kälviäinen
Associate Professor Tuomas Eerola

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Antti Vilkman

One-to-many and many-to-many matching for Saimaa ringed seal re-identification

Master's thesis
2022
50 pages, 29 figures, 2 tables
Examiners: Professor Heikki Kälviäinen and Associate Professor Tuomas Eerola

Keywords: Saimaa ringed seal, computer vision, re-identification, one-to-many, many-to-many

Saimaa ringed seals are among the most endangered seal species. Monitoring Saimaa ringed seals using image data is a useful tool in their conservation efforts. The images can be collected automatically by using camera traps and the amount of data collected can be large. Automatic computer vision methods for identifying individual seals are used to reduce the amount of manual work involved in the identification. Utilising information from multiple images and views simultaneously can give additional information and enhance the features present in the images. In this thesis the one-to-many and many-to-many matching of Saimaa ringed seals is considered. Information from multiple images was utilised by aggregating the features from multiple images to a single Fisher Vector. The experiments were conducted with aggregating only the database features, only the query features, and aggregating both. The achieved results were compared to one-to-one matching where features from a single image are used at a time. Aggregating the only the database features failed to improve results over the baseline one-to-one matching. The other two aggregation methods significantly improved the re-identification accuracy.

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Laskennallinen tekniikka

Antti Vilkman

Yksi-moneen- ja moni-moneen-vastaavuus Saimaannorppien uudelleentunnistamisessa

Diplomityö
2022
50 sivua, 29 kuvaa, 2 taulukkoa
Tarkastajat: Professori Heikki Kälviäinen ja Tutkijaopettaja Tuomas Eerola

Hakusanat: saimaannorppa, tietokonenäkö, uudelleentunnistaminen, yksi-moneen, moni-moneen
Keywords: Saimaa ringed seal, computer vision, re-identification, one-to-many, many-to-many

Saimaannorppa on yksi uhanalaisimmista hyljelajeista. Saimaannorppien tarkkailu kuvien avulla on hyödyllinen työkalu niiden suojelutyössä. Kuvia voidaan kerätä automaattisesti käyttämällä riistakameroita ja kerätyn datan määrä voi olla suuri. Manuaalisen työn vähentämiseksi norppayksilöiden tunnistamisessa hyödynnetään automaattisia tietokonenäköön pohjautuvia metodeja. Useiden kuvien ja kuvakulmien käyttö samanaikaisesti voi antaa lisäinformaatiota ja tarkentaa kuvissa olevia piirteitä. Tässä työssä käsitellään saimaannorppien yksi-moneen- ja moni-moneen-vastaavuutta. Informaatiota useista kuvista hyödennettiin yhdistämällä useista kuvista kerätyt piirteet yhteen Fisher-vektoriin. Suoritetuissa kokeissa yhdistettiin piirteet vain tietokantakuvista, vain kyselykuvista, sekä molemmista. Saavutettuja tuloksia verrattiin yksi-yhteen -vastaavuuteen, jossa käytetään piirteitä vain yhdestä kuvasta kerrallaan. Vain tietokantakuvien piirteiden yhdistäminen ei parantanut tuloksia suhteessa yksi-yhteen-vastaavuuteen. Kaksi muuta kokeiltua yhdistämismetodia paransivat tunnistamistarkkuutta merkittävästi.

ACKNOWLEDGEMENTS

I would like to thank my supervisors M.Sc Ekaterina Nepovinnykh, Associate Professor Tuomas Eerola and Professor Heikki Kälviäinen for their guidance during the process of making this thesis. As my skills in the realms of scientific research and writing are quite limited, their feedback has been invaluable in making this thesis into something that is hopefully not a complete waste of paper and bytes.

I also greatly appreciate all of the people working towards the conservation of Saimaa ringed seals, or any animal for that matter. Taking the time to try and help the animals who can't help themselves is a selfless act and worthy of all the praise.

To all of my friends I have met during my time in Lappeenranta, thank you for making this town feel like home and my time here a memorable experience. To kurssi-1 gang, the peer support and the long nights spent hanging out are a memory I will always cherish.

To all those I hold dear, family, friends, thank you for being here. Sending you my love and light.

Lappeenranta, May 31, 2022

Antti Vilkman

LIST OF ABBREVIATIONS

CBIR	Content Based Image Retrieval
CMS	Cumulative Match Score
CNN	Convolutional Neural Network
DB	Database
FV	Fisher Vector
GMM	Gaussian Mixture Model
MAC	Maximum activations of convolutions
NORPPA	Novel Ringed seal re-identification by Pelage Pattern Aggregation
PCA	Principal Component Analysis
RANSAC	Random Sample Consensus
SFTA	Segmentation Based Fractal Texture Analysis
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
VKD	Views Knowledge Distillation

CONTENTS

1 INTRODUCTION

1.1 Background

Many species of animals are endangered due to climate change, human land usage and other human activities. Monitoring endangered animal populations and tracking their evolution over time is important for conservation efforts of the endangered species [1]. Observing animals from image data is a non-invasive method and can be done by utilising automatic camera traps or images from photographers. Due to large amounts of data generated through these methods, automatic computer vision methods are used in re-identification of known individuals [1].

One endangered species is the Saimaa ringed seal, as shown in Fig. 1 [2]. With a population size of only 400 individuals, it is among the most endangered pinnipeds in the world [3]. Photo identification has been used in conservation efforts to monitor the seal population as it is a non-invasive alternative to methods that require catching the seal, thus reducing the amount of stress that is caused to the animals. Identification from images has been traditionally done manually [4], but it is laborious and time-consuming, creating a need for automatic methods.

Figure 1. Example images of Saimaa ringed seals. [2]

Multiple methods for automatic animal re-identification, such as [5, 6] exist in literature,

many achieving very successful results. Often these methods make use of identifiable features in fur patterns present in the animals. Saimaa ringed seals have a ring pattern in their fur which is unique in each individual, allowing for re-identification based on the pattern [2]. Photographs of Saimaa ringed seals offer a challenging task in automatic re-identification, with large variance in poses, lighting, and low contrast between the ring pattern and rest of the fur making the task more demanding. As a part of the CoExist project [7], various methods [8–13] utilising computer vision have been proposed for automatic re-identification of Saimaa ringed seals, matching features found in the pelage patterns in different images in order to identify the individual animals, as shown in Fig. 2. A general example of steps in an animal re-identification pipeline is shown in Fig. 3. From the raw image, the seal is segmented and the image is cropped to the bounding box of the seal. The pelage pattern is then extracted and taken through a feature extraction process, finding interesting regions from the pelage pattern and encoding them. The extracted features are aggregated to form a decriptor of the seal, and the descriptor is then compared against descriptors from known identified seals from the database (DB) to find the closest matches.

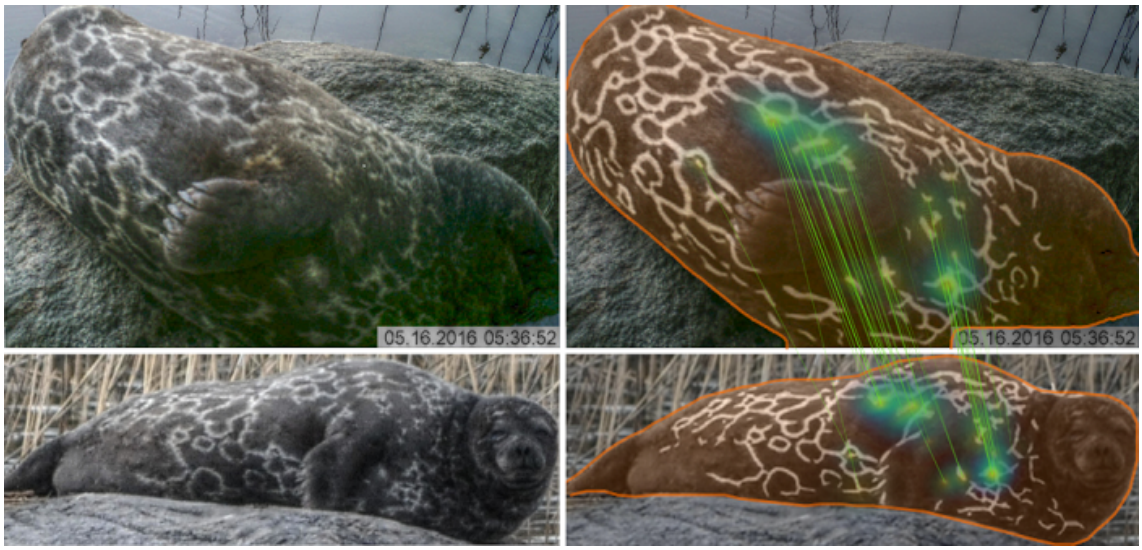

Figure 2. Feature matching between two Saimaa ringed seal images. [13]

The matching of Saimaa ringed seals has been previously done one-to-one, comparing images from the database to the query image one by one. Aggregating information from multiple images of an individual can provide more features or enhance the features if they are seen from multiple angles. Comparing descriptors built from single or multiple query images to descriptors built from multiple database images, the use of one-to-many, many-to-one and many-to-many matching has the ability to improve speed [5] and accuracy [14]

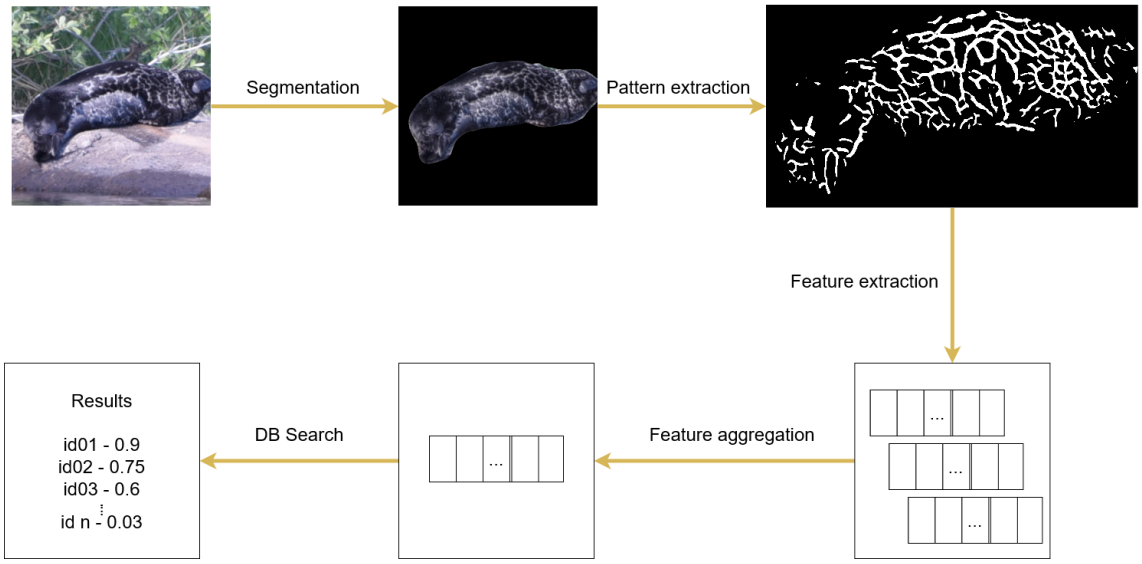

Figure 3. An example of steps taken in automatic re-identification of Saimaa ringed seals.

of re-identification algorithms. While image sequences and multiple views have been used to improve methods in human re-identification [15,16], relatively few works utilising such techniques exist in the realm of automatic animal re-identification.

1.2 Objectives and delimitations

The aim of this thesis is to implement one-to-many, many-to-one and many-to-many matching for Saimaa ringed seal re-identification, and to evaluate their performance for the task, with the delimitation of focusing only on Saimaa ringed seals. The many-to-many matching is visualized in Fig. 4 where information from multiple images is aggregated to a single descriptor. For the query individual and each database individual, all of the features from all of the images of an individual are aggregated to a single descriptor. The distance between the query descriptor and each database descriptor is then calculated to perform the final re-identification. The database individuals with the shortest distances to the query individual are the most likely matches. In one-to-many matching only a single query image would be used instead of many.

The objectives of this thesis are as follows:

- to prepare a dataset for the one-to-many, many-to-one and many-to-many Saimaa ringed seals photo-identification,

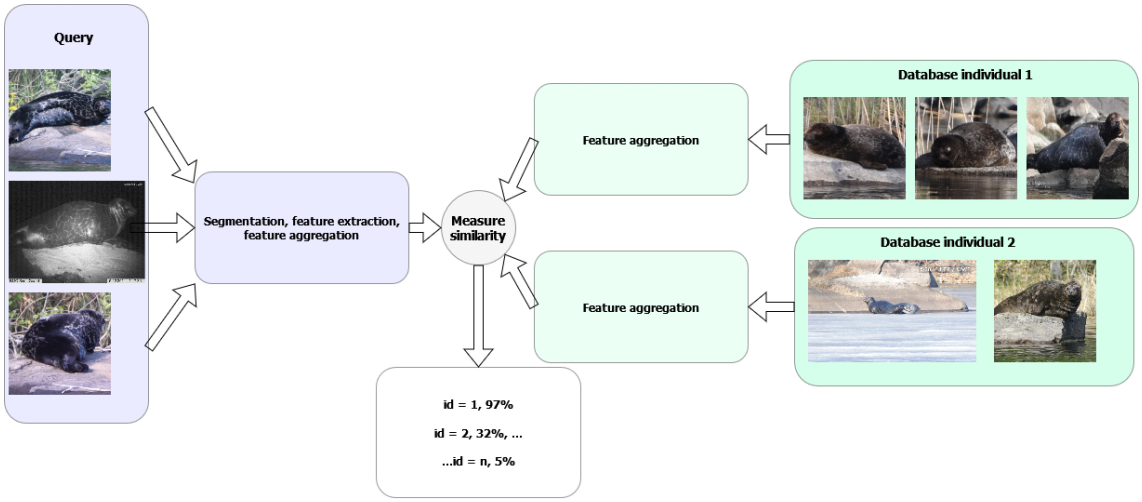

Figure 4. An illustration of many-to-many matching for Saimaa ringed seals.

- to update the Saimaa ringed seal identification method by implementing one-to-many, many-to-one and many-to-many image matching, and
- to evaluate the updated Saimaa ringed seal identification method on the prepared data sets.

A delimitation for thesis is that only the re-identification of Saimaa ringed seals is considered in this thesis. However, it should be possible to apply the method to other patterned species as well.

1.3 Structure of the thesis

The thesis is structured as follows: Chapter 2 discusses the development of methods and different aspects in automatic animal re-identification. Chapter 3 describes use of information from multiple images in re-identification tasks. In Chapter 4 the proposed methods to utilise one-to-many and many-to-many matching in Saimaa ringed seal re-identification are presented. Chapter 5 describes the conducted experiments and the results of the experiments while Chapter 6 discusses the achieved results and possible future work. Chapter 7 concludes the thesis with a summary of the conducted work and achieved results.

2 ANIMAL RE-IDENTIFICATION

2.1 Background

Animal re-identification is an effective tool in environmental conservation efforts, allowing for monitoring of animal populations [1]. The data gathered from animal re-identification can be used for evaluating multiple aspects of animal populations, such as their evolution over time, travel patterns, responses to environmental changes [17]. With many species being endangered and facing threats such as changes in climate, habitat loss due to human land usage, pollution, diseases and poaching [18], being able to monitor their populations and behaviour is a key part of conserving biodiversity.

Traditionally animals have been identified with techniques such as tagging, scarring or DNA analysis [1]. These methods are reliable for accurately identifying the individual animals but they are also intrusive, requiring contact with the animal and causing stress to it. These methods are also laborious to the researchers due to needing field work, catching the animal and performing the re-identification manually [9]. Using image data of the animals, such as images acquired from camera traps, has become a preferred method for performing animal re-identification. Camera traps are inexpensive, easy to maintain [19] and require less work from the researchers [1]. Camera traps are non-invasive, and a large number of them can be installed in the environment that is being monitored [20] allowing for better and continuous monitoring of animals without causing disturbance in their behaviour [1].

While having researchers perform the re-identification from images is a better method than the ones traditionally used in animal re-identification, it still has some issues of its own. Vast amounts of data is produced by the cameras, and it needs to be analyzed by researchers in order to gain the desired information on animal individuals [20]. While not as laborious as performing the work out in the field, the manual identification still takes a large amount of time. In addition, accuracy can become an issue, as human error can lead to a researcher identifying the same individual as another individual [17] or identifying an animal as being of another species altogether [21]. Presence of clear patterns or markings on the animal has traditionally been a needed feature for reliable animal re-identification from images [1].

In order to reduce the amount of manual work that needs to be done, computer vision techniques have been used to perform the animal re-identification automatically. Early

methods for automatic re-identification were based on using feature engineering, where the common features in an animal species were identified and used to identify the individuals. These methods are however lacking in their ability to generalize, as features that have been designed for identifying individuals from one species would not necessarily work on other species [1]. Emergence of deep learning and the availability of large data sets has allowed for creation of algorithms that are capable of learning the needed features from the data, with even the ability to use the same algorithm on multiple species [5].

2.2 Animal detection

Determining whether an image contains an animal or not is a relatively simple and fast task for a human observer [22], but in order to enable automatic re-identification and monitoring of animals, the automatic detection methods are needed. Steps in animal detection can range from determining if an image contains an animal to determining the the location of the animal in the image or segmenting the animal from the image. Problems can be posed by multiple animals being present and the possibility of them being of different species [23]. Further, natural scenes tend to be cluttered and dynamic with movement from trees, water and natural lighting making the task of animal detection more complex and rendering more traditional motion-based object detection methods void for the task [19].

Traditional methods were based on detecting the face of the animal or subtracting the background from the image to locate the animal [24]. Due to the aforementioned challenges in detecting wild animals in natural scenes, Convolutional neural networks (CNN) have risen to be the tool of choice for animal detection due to their ability to learn representations and excellent performance in image classification and object detection [25]. Instead of using hand crafted features, convolutional layers in CNNs are capable of learning features, such as contours and other shapes and patterns from training data, classifying inputs based on the presence of these learned features. A CNN detector, such the R-CNN [26], typically consists of two stages with the first stage generating regions which may contain the object to be detected, and the second stage classifying the proposed objects into true and false objects. Example results from animal detection can be seen in Fig. 5 [27].

CNNs tend to be computationally intensive [27], and for faster performance methods such as YOLO [28] that consist of a single CNN have been presented. While not as accurate as the methods with separate networks for region proposals and classification, they do have

Figure 5. Results from animal detection, displaying the bounding boxes. [27]

their uses when low latency is needed.

Animals with high site fidelity, such as Saimaa ringed seals, can end up being captured with a fairly static background when using camera traps, leading to supervised identification algorithms running the risk of learning features from the background instead of learning only the features from the animal [8]. In such cases segmentation of the animal at pixel level is important to separate it from the background. Superpixel based methods [29, 30] and the Mask R-CNN [31] are examples of methods that have been shown to accurately achieve the separation of animals from the image background. Mask R-CNN has also been successfully used in segmenting multiple animals from single images [32] as can be seen in Fig. 6 for cattle and in Fig. 7 [14] Ladoga ringed seals, a sister species of the Saimaa ringed seal.

Figure 6. Segmentation masks produced by the Mask R-CNN for cattle. [32]

Figure 7. Segmentation masks produced by the Mask R-CNN for Ladoga ringed seals. [14]

2.3 Re-identification

Early works utilising computer assistance in animal re-identification were based on feature engineering [1] such as [33] where whales were identified based on the features in their tail flukes. The predetermined features, such as notch shape and pigmentation, were manually coded from each image and a matching algorithm was then used to find the best matches. While aiding in the re-identification process, the images still needed to be manually tagged and the system could fail to account for characteristics that have not been previously encountered in the tails.

In [34], the Finscan algorithm was presented, where the edge pattern of the dorsal fin in multiple delphinid species was used to identify the animals. The system utilises automatic recognition of the boundaries of the dorsal fin, and the shape of the fin is automatically computed using curve matching and string matching [35]. The automatically computed features were used for making database query, finding the images with the highest similarity to the features in the query image. Using top-1 accuracy to measure, as in how often the closest match provided by the algorithm was the correct individual, Finscan achieved promising results and was a step towards autonomous re-identification.

More complex features have been introduced to use visual patterns in the animal for identification, such as using local colour and brightness information as well as the shape for salamanders [36], spot patterns for whale sharks [37] or whisker spot patterns in polar bears [38]. In [39] feature patterns were extracted by 3D modelling the animal, with the ability to use the same method on different patterned species, only needing to train a new classifier.

The use of Scale Invariant Feature Transform (SIFT) features to find keypoints in images [40] has been a popular method in animal re-identification due to their invariance to scale and orientation. For example in the Wild-ID algorithm presented in [41], the SIFT features are extracted from images of giraffes and identification is performed by finding the database image with the most similar features, with a modified Random Sample Consensus (RANSAC) algorithm [42] being used to ensure the geometric consistency of the matched features. An example of SIFT features and matching them can be seen in Fig. 8.

Figure 8. Feature matching using SIFT features. White dots in each image are the locations of the matching SIFT features. [41]

The HotSpotter [5] built upon similar methods as the Wild-ID, using the Root-SIFT [43] for feature descriptors and finding the database images with similar descriptors. HotSpot-

ter uses RANSAC to ensure the consistency of the descriptors in the found matches and computes the combined scores for each label in the database to classify the query image instead of simply using the highest scoring image. Achieving high re-identification accuracy on jaguars, giraffes, zebras and lionfish, HotSpotter has been shown to be accurate and work on multiple patterned species [5].

Deep learning methods, particularly Convolutional Neural Networks have shown very strong results in recognition tasks. With their ability to learn the needed features from training data, removing the problem of crafting the needed features by hand. In human re-identification CNNs have been used to great effect, but traditional CNN architectures for image classification face problems with needing sufficient data for each individual and needs to re-trained for every new individual [1].

The Siamese network [44] is an architecture capable of tackling the above problems. A Siamese Network consists of a pair of networks with identical structure and parameters, with the outputs from each network connected at the final level by a function that compares the similarity of the two outputs [45]. With two similar input images, both networks should give similar outputs, and each network produces the same output if it is given the same input as the other [46]. The twin network structure thus allows for evaluating whether the two inputs are similar to each other. For example in [47], the Siamese architecture is utilised in re-identification of lemurs, chimpanzees and golden monkeys. Using CNNs to create the representations of facial images of the animals, the cosine similarity between the feature vectors is used as the metric for determining similarity.

The Triplet Loss network [48] is a developed version of the Siamese network, but uses a triplet of networks as shown in Fig. 9 [11], instead of twins. A Triplet Loss network is trained by inputting a reference image, an image of the same class and an image from a different class to each of the respective networks, and the feature representations can then be used to determine which image better matches the reference [48]. Use of the triplet loss has shown improved results in digit classification over a Siamese network with twin structure [48], as well as better re-identification results in multiple species of animals [6]. Triplet loss does come with the triplet mining problem, meaning that during training the triplets need to be chosen in such a way that the samples positive and negative images are close to the reference, to make the network learn subtler differences. Choosing a suitable strategy for triplet mining can be difficult [14].

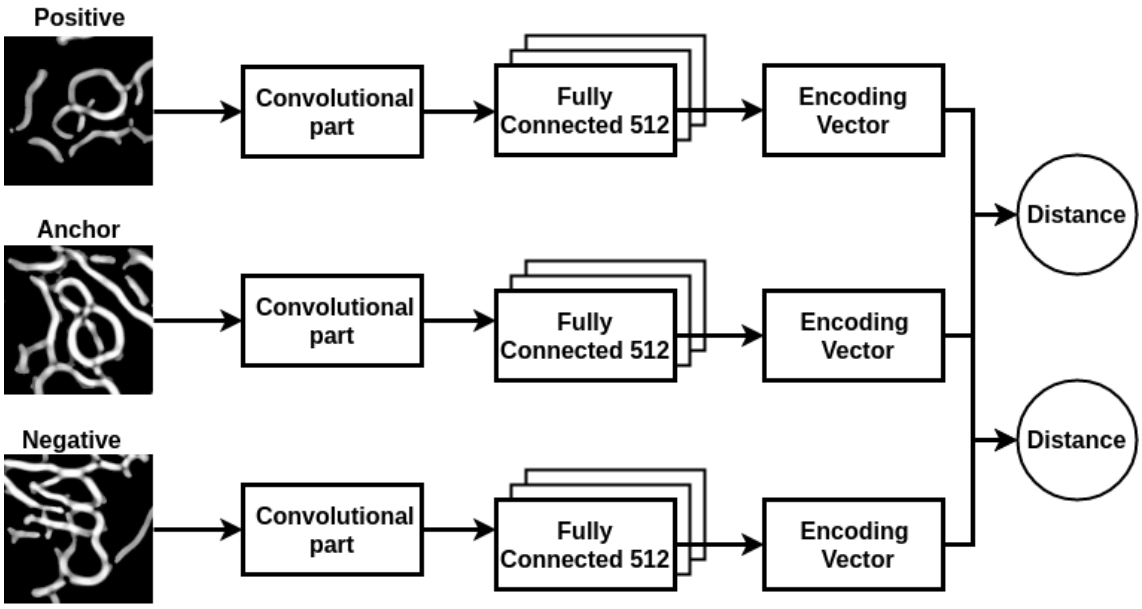

Figure 9. An example of the Triplet loss network training architecture. [11]

2.4 Ringed seal re-identification

Multiple methods and subsequent improvements have been presented for automatic re-identification of ringed seals. In [8], the segmentation of Saimaa ringed seals from images is considered, using a superpixel based method [29,30] to extract the seals from the image background. The results were evaluated using the Cumulative Match Score (CMS) histogram, in essence measuring the top-N accuracy of the identification at different numbers of N closest matches.

Using Segmentation based Fractal Texture Analysis (SFTA) features [49] and a Naive Bayesian classifier to perform re-identification, the top-15 accuracy was decent [8], meaning that often the correct individual was found within the be closest 15 matches given by the algorithm.

In [9] a more elaborate pipeline is presented, featuring colour normalisation and contrast enhancement to better extract the features in the pelage patterns. Trying out Wild-ID and HotSpotter, HotSpotter was the better performing of the two when colour normalisation and contrast enhancement were applied, markedly increasing the accuracy from [8]. The pipeline is visualised in Fig. 10.

In [10], transfer learning is utilised to retrain a pretrained CNN for the Saimaa ringed seal identification task. A CNN is also used to extract features from the images that were then

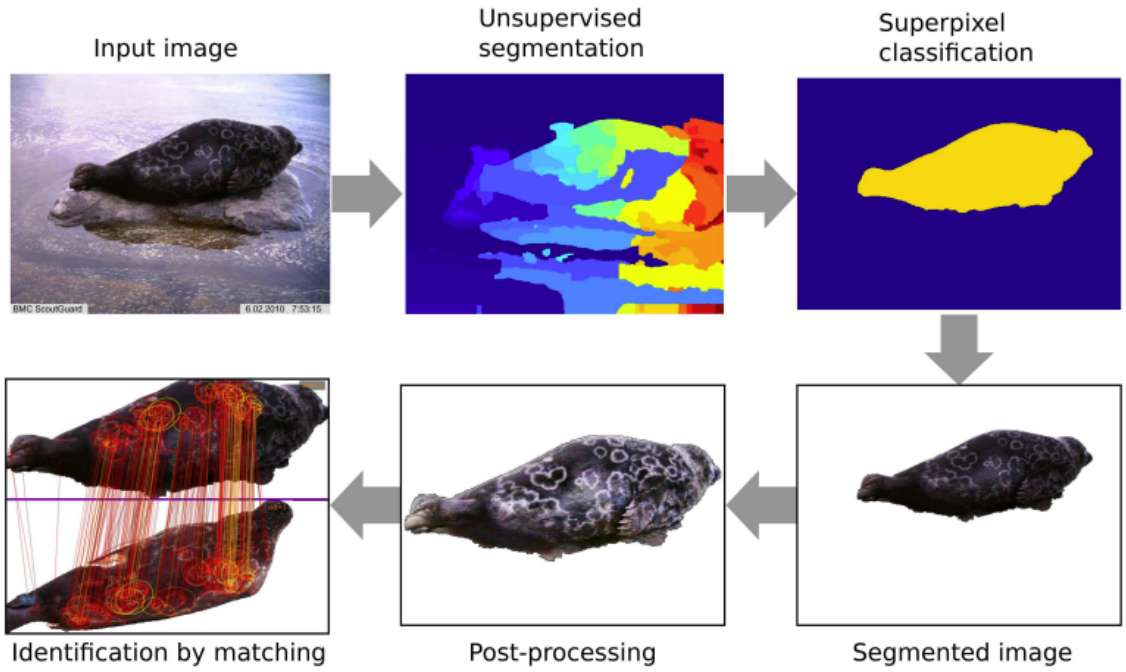

Figure 10. Segmentation of seals using superpixel based methods and classification with Wild-ID and HotSpotter. [9]

used for classification with a Support Vector Machine (SVM). Both the CNN and SVM method achieved very high accuracies. However, the use of CNN faced the issues of needing lots of images of each individual and being unable to deal with new individuals, needing retraining when a new individual is introduced.

A Triplet Loss Network is used in [11] to better generalize when presented with individuals from a new class. In addition multiple other improvements to the Saimaa ringed seal identification pipeline are made as follows: the DeepLab model [50] is used for seal segmentation, Sato tubeness filter [51] is used for extracting the pelage pattern from the segmented seals, and the extracted patterns are divided into patches to increase robustness towards the pose of the seal and the angle it is viewed from. A Triplet Neural Network with a rotation-invariance pass is trained for computing the similarities between patches. The re-identification is then performed by calculating the similarities between images, filtering the possible matches based on topological consistency, and finding the closest match. Quite high top-1 accuracy was achieved, with the extracted pattern patches outperforming the use of raw patches from the segmented images.

In [12] the EDEN pooling method and a further improved pipeline is presented, using Mask R-CNN [31] for the segmentation step and a CNN for the pelage pattern extraction [52]. The ArcFace [53] loss is used to deal with the problem of triplet mining. The pattern

patches are run through a ResNet-18 [54] based CNN, with a custom global pooling layer at the end replacing the global average pooling of ResNet, and after the pooling step, the patch descriptors are aggregated to a Fisher vector [55] to generate the descriptor for an entire image. The EDEN pooling outperformed other methods, with the combination of the EDEN and Maximum activations of convolutions (MAC) [56] poolings achieving the highest accuracy.

A similar pipeline is utilised in [14] for re-identification of Ladoga ringed seals from image sequences where the SphereFace [57] loss is used to avoid triplet mining, Generalized-Mean pooling is used in the global pooling layer, and a grouping step is used for multiple images of the same individuals, aggregating their descriptors from multiple images into a single Fisher Vector. On the 50 images of the test set, very high top-1 accuracy was achieved.

The Novel Ringed seal re-identification by Pelage Pattern Aggregation (NORPPA) pipeline is presented in [13] again improving the accuracy from previous methods. The pipeline is visualised in Fig. 11. The HesAffNet [58] is used to extract affine covariant regions from the pelage pattern and the extracted regions are embedded using the HardNet [59] instead of the pooling step in [12]. Principal Component Analysis (PCA) is then applied to the embedding before aggregating the features into Fisher Vectors. Kernel PCA [60] is also utilised to reduce the dimensionality of the descriptors. For the re-identification step, the distance between the descriptors of query images and database images is computed, with lowest distance being the nearest match.

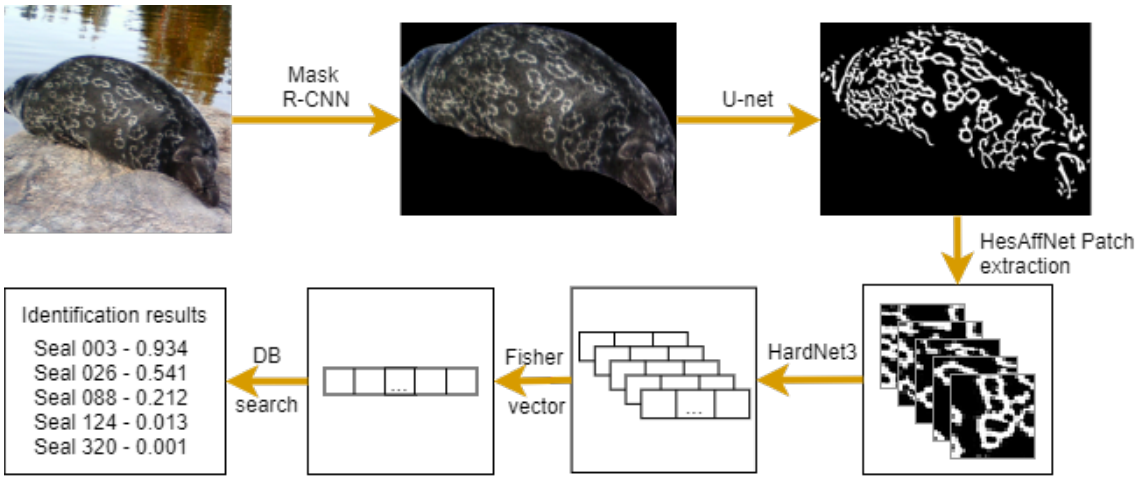

Figure 11. The Novel Ringed seal re-identification by Pelage Pattern Aggregation (NORPPA) pipeline. [13]

A new, more extensive and challenging dataset is presented in [2] consisting of 57 seal individuals and 2080 total images with more variability in appearance compared to previous sets. On the newer more difficult dataset, the NORPPA HardNet version achieved much higher top-1 accuracy than the ArcFace method from [12]. Combining EDEN with HardNet was also attempted but did not improve results.

3 RE-IDENTIFICATION USING MULTIPLE IMAGES

3.1 Background

Traditional re-identification methods search for matches by brute force, that is by computing the similarity between the query image and each database image one at a time. Going through each image is a slow process, and aggregating information from multiple images of an individual to a is a possible way to increase the efficiency of the re-identification [5, 61]. In addition, using multiple images to construct the feature descriptor for an individual can lead to obtaining more reliable features, better representing its appearance [62]. The concept is visualized in Fig. 12 where many-to-many matching is illustrated. In one-to-many matching a single query image would be used instead and in many-to-one matching a single database image would be used instead.

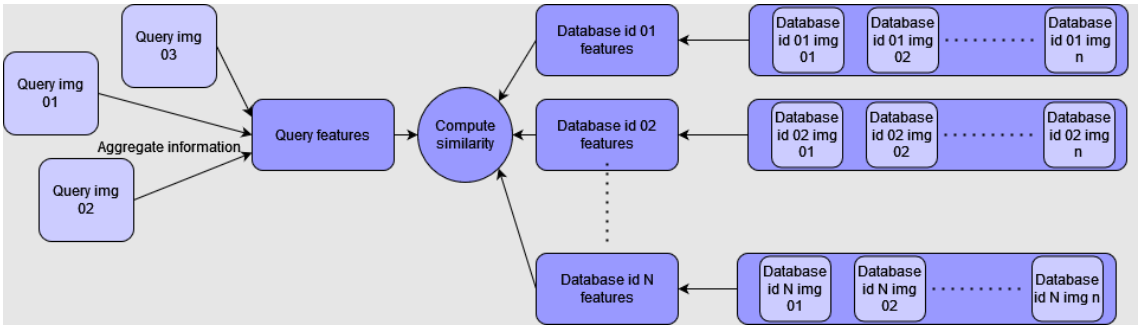

Figure 12. Concept of aggregating information from several images for image matching.

In case of having a single image for the query individual and multiple images for the database individual the matching would be called one-to-many matching. In many-to-one matching multiple images for the query individual are used and a single image for the database individual. In many-to-many matching multiple images are used for both the query and database individual. Use of these matching techniques has been shown to have the ability increase the speed [5] and accuracy [14] of animal re-identification. Thus, finding methods of utilising multiple images in animal re-identification is an alluring prospect.

3.2 Multi-view image features

Multiple views may be utilised to acquire more features of an object or to gain better representation of a feature that is present in multiple views. Features may need to be matched between the views to allow for forming a single feature descriptor from the multiple views. Re-identification tasks and matching features between images could also be thought of as Content Based Image Retrieval (CBIR) tasks [63], where similar methods often apply. Various methods exist for matching features from multiple views.

In [64] matching features between multiple images is explored, using SIFT for feature representation and distances between features to determine matches and distinctiveness of features. Bag-of-Visual words [65] models establish "vocabularies" of visual features, assigning the same encoding for similar features, effectively matching similar features when obtaining encodings from multiple images, as illustrated in Fig. 13.

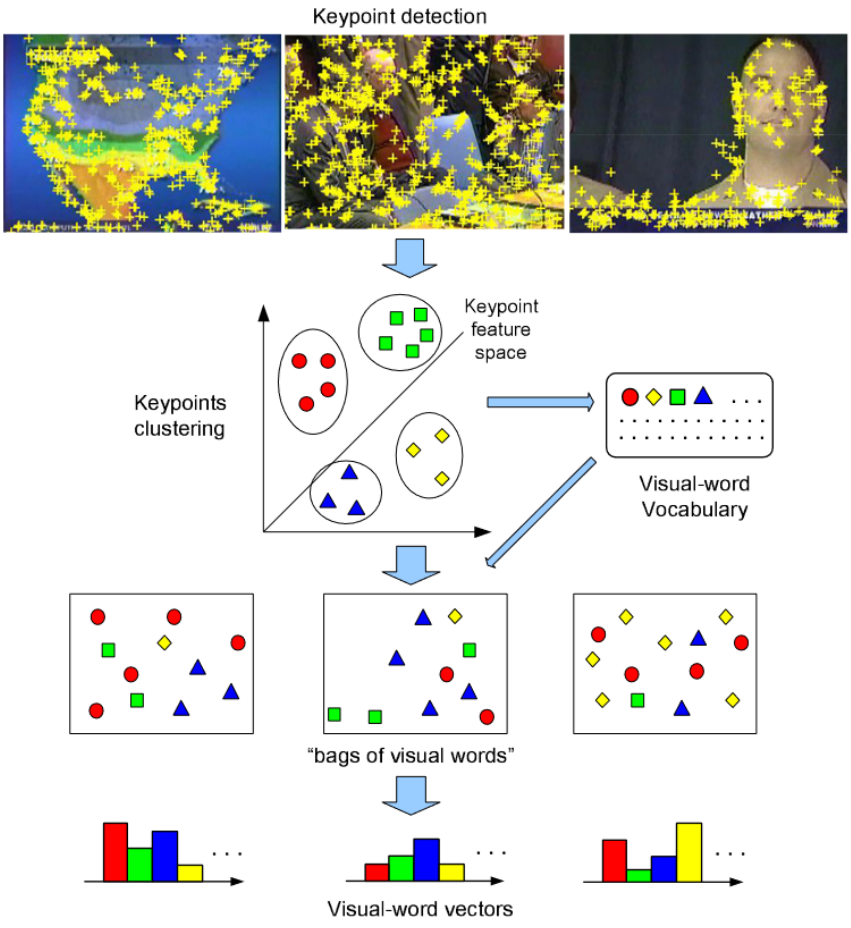

Figure 13. An example of a Bag-of-Visual words model. [65]

Fisher Vectors (FV) are a similar way of representing the features in an image and can be understood as a probabilistic visual vocabulary [66]. Using a generative model to create a vocabulary of features from a training set of features, the obtained parameters are then used to describe query features by their deviation from the model. In [66] Fisher Vectors were created by using a Gaussian Mixture Model (GMM) to create the vocabulary to achieve competitive results with state-of-the-art techniques in standard datasets.

In some cases use of multiple views has been shown to outperform methods that utilize a single view. In [15], videos of human faces were considered. Pooling features from multiple frames is performed using two different methods: 1) computing Fisher Vectors for each frame and averaging the vectors and 2) pooling all SIFT features from each frame and computing one Fisher Vector from the pooled features. The Video Fisher Vector Faces representation achieved state-of-the-art performance across multiple datasets.

CNN based techniques have also been used for multi-view tasks. A multi-view CNN for object recognition was presented in [67]. Multiple views of an object are fed to the network and a pooling layer is used to aggregate the views before classification. The concept is illustrated in Fig. 14.

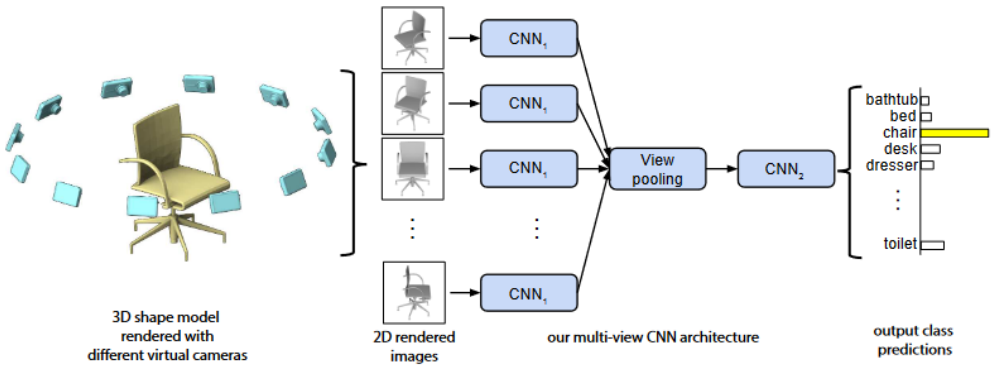

Figure 14. Concept of the Multi-view CNN. [67]

In [68], the Views Knowledge Distillation (VKD) is presented, in which a teacher CNN is used to optimise a student CNN with the same architecture. During training, the teacher receives more views of of the targets and the student receives fewer views. The Knowledge Distillation loss [69] is used to make the student network match the representations of the teacher network while having access to fewer views. This training method is illustrated in Fig. 15.

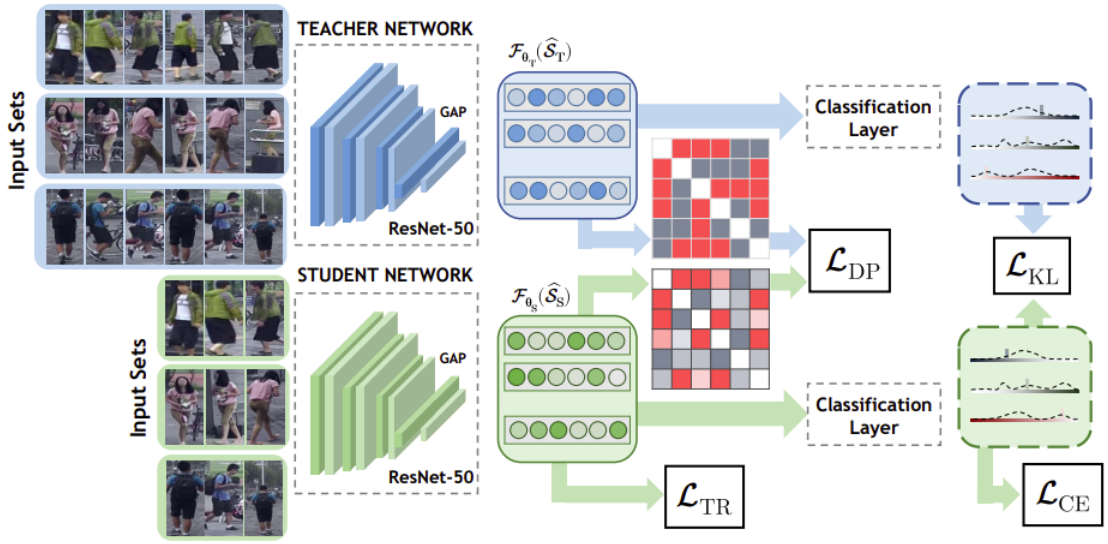

Figure 15. Training a CNN using the Views Knowledge Distillation. [68]

3.3 Re-identification from image sequences

Various works exist exploring the use of image sequences or video data in human re-identification. In addition to using spatial appearance, the sequential nature of the data enables use of temporal features for the re-identification. In [70], a model for selecting the most discriminative video fragments by using motion intensity is presented. These fragments are then used for person re-identification by matching the features between the most discriminative fragments of two sequences. In [71], the Adaptive Fisher Discriminant analysis algorithm is presented for multi-shot re-identification, aiming to create a feature space in which images of different people are separated and images of the same person are clustered near each other. This is achieved by adapting the Local Fisher Discriminant Analysis and iteratively updating the feature space, preserving diversity in the samples. The method has been shown to achieve better performance than the earlier state-of-the-art methods.

In [16], a Recurrent Convolutional network with temporal pooling was utilised to make use of the temporal information as shown in Fig. 16. Both optical flow and colour channels are used as the inputs to the network and a convolutional layer is used for feature extraction. The recurrent layers use information from both current and previous time-steps to produce their output. The outputs from different time-steps are then pooled and the Siamese architecture is used to train the network for extracting the features that are used in classification. The method was shown to outperform other state-of-the-art video

re-identification methods.

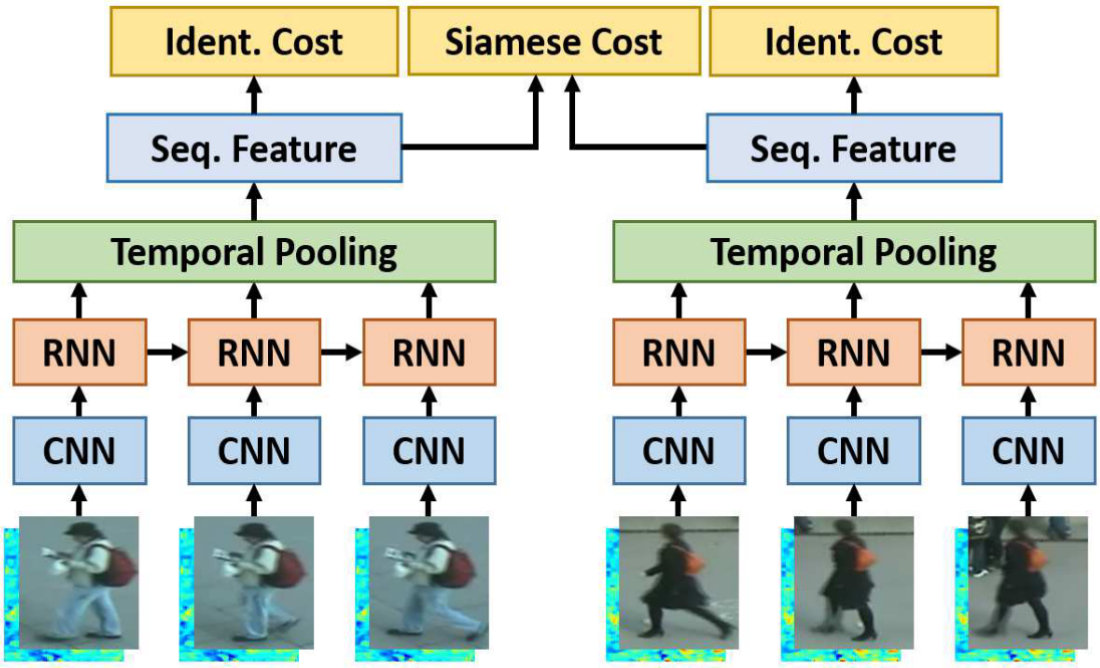

Figure 16. The Recurrent CNN architecture with temporal pooling. [16]

3.4 Animal re-identification using multiple images

For animal re-identification, features aggregated from multiple images have been occasionally used to enhance the performance of the algorithms. In [5], a one-to-many variant the HotSpotter animal individual matching algorithm is presented in addition to a one-to-one version. Instead of comparing the query features to each database individual, all database descriptors are aggregated to a forest of kd-trees [72] from which the closest matches for query descriptors were sought. Tested on datasets of jaguars, giraffes, zebras and lionfish, the one-to-many version outperformed the one-to-one algorithm in accuracy and was vastly faster [5].

In [14], instances of individual Ladoga ringed seals are grouped by utilising distance between their descriptors, grouping individuals with similar descriptors to a single group, as shown in Fig. 17. The grouping is done by using the ResNet-101 to obtain descriptor vectors for each instance, forming initial groups from the image with the largest amount of seals, and then adding the instances from each image to groups with minimum mean distance to their descriptor. For the re-identification process, the pattern descriptors from

each instance in a group are aggregated using a Fisher Vector to create the final descriptor of each seal, gaining more information for matching the seals. The top-1 re-identification accuracy was improved by 30% when utilising the grouping method.

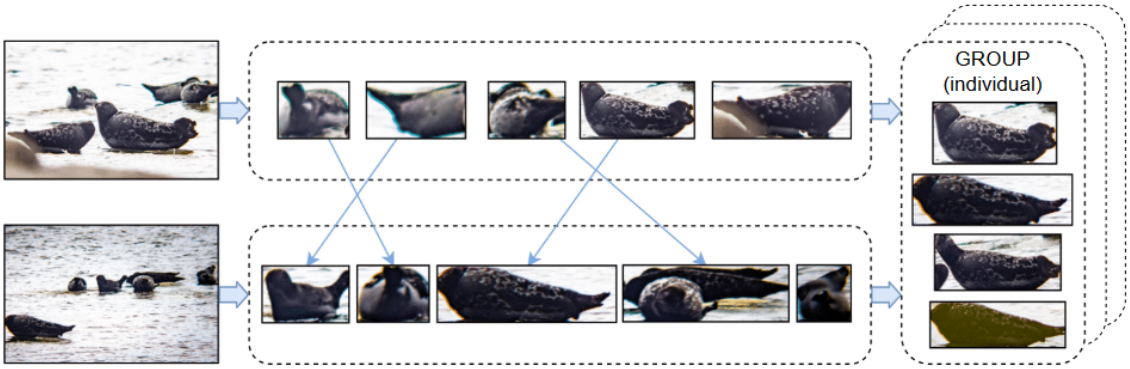

Figure 17. Grouping instances of the same Ladoga ringed seal individuals. [14]

4 ONE-TO-MANY AND MANY-TO-MANY MATCHING METHOD

4.1 Pipeline

In this chapter the proposed implementation of one-to-many and many-to-many matching for the Saimaa ringed seal re-identification is presented. A Fisher Vector based approach was chosen for the aggregation of the features across multiple images since using the Fisher Vector based grouping method gave good results in [14].

The proposed re-identification pipeline utilising one-to-many and many-to-many matching is largely based on the seal identification pipeline from [13] which is illustrated in Fig. 11. In [13] features from each image were aggregated to a single Fisher Vector. With the proposed one-to-many and many-to-many matching methods, the Fisher Vectors are instead formed from all of the features from all images of a single seal. The updated pipeline is presented in Fig. 18.

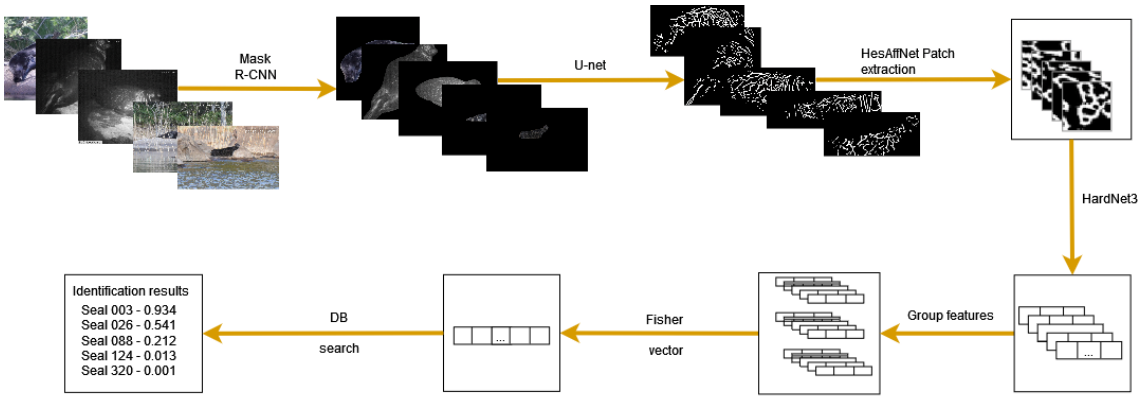

Figure 18. The proposed Saimaa ringed seal re-identification pipeline.

4.2 Segmentation and feature extraction

Instance segmentation of seals and feature extraction follow the same process that is presented in [13]. Mask R-CNN [31] is used to remove the background from images, leaving only the seal visible. U-net CNN is then used to extract the pelage pattern of the seal into a binary image. As the images vary in resolution, the pattern images are resized to make

the pattern lines have similar width in each image. Examples of pattern images are shown in Fig. 19.

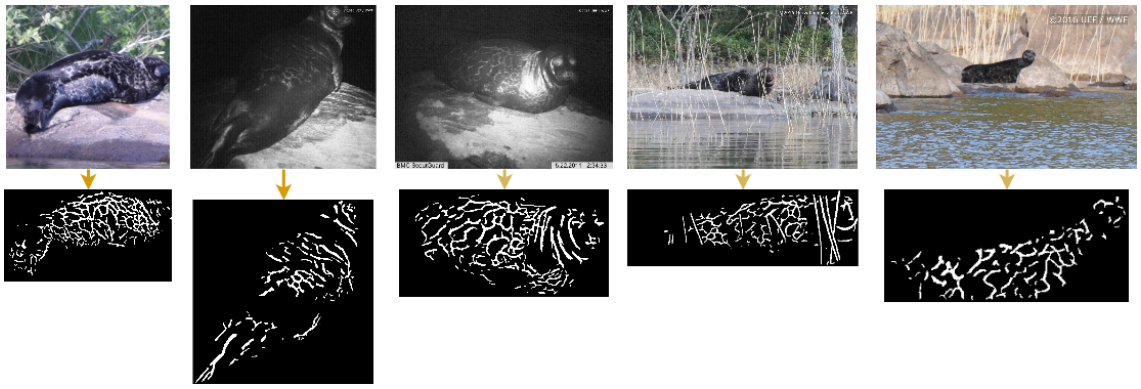

Figure 19. Examples of pattern images extracted by U-net CNN.

With the pattern extracted, affine covariant regions are found and extracted from the pattern image by using HesAffNet [58]. HesAffNet extracts local descriptors from images and transforms them in such a way that similar patches from different images are geometrically corresponding. The HesAffNet is trained by using the HardNegC loss function [58], which is similar to the triplet loss but within the training batches the distance to the closest negative sample is constant. The HardNegC loss is defined as [58]

$$L = \frac{1}{n} \sum_{i=1,n} max(0, 1 + d(s_i, \dot{s}_i) - d(s_i, N)), \frac{\partial L}{\partial N} := 0 \tag{1}$$

where $d(s_i, \dot{s}_i)$ is the distance between matching patches and N is the hardest negative sample in the training batch, making $d(s_i, N)$ the distance to the that sample. The derivative of L with respect to N is set to zero. Examples of extracted patches are shown in Fig. 20.

The extracted patches are then embedded into vectors of size 1×128 by using HardNet [59]. HardNet is trained to correctly embed and match descriptors while avoiding false positives from similar appearing descriptors by using triplet margin loss which is defined as [59]

$$L = \frac{1}{n} \sum_{i=1,n} max(0, 1 + d(a_i, p_i) - min(d(a_i, p_{j_{min}}), d(a_{k_{min}}, p_i))) \tag{2}$$

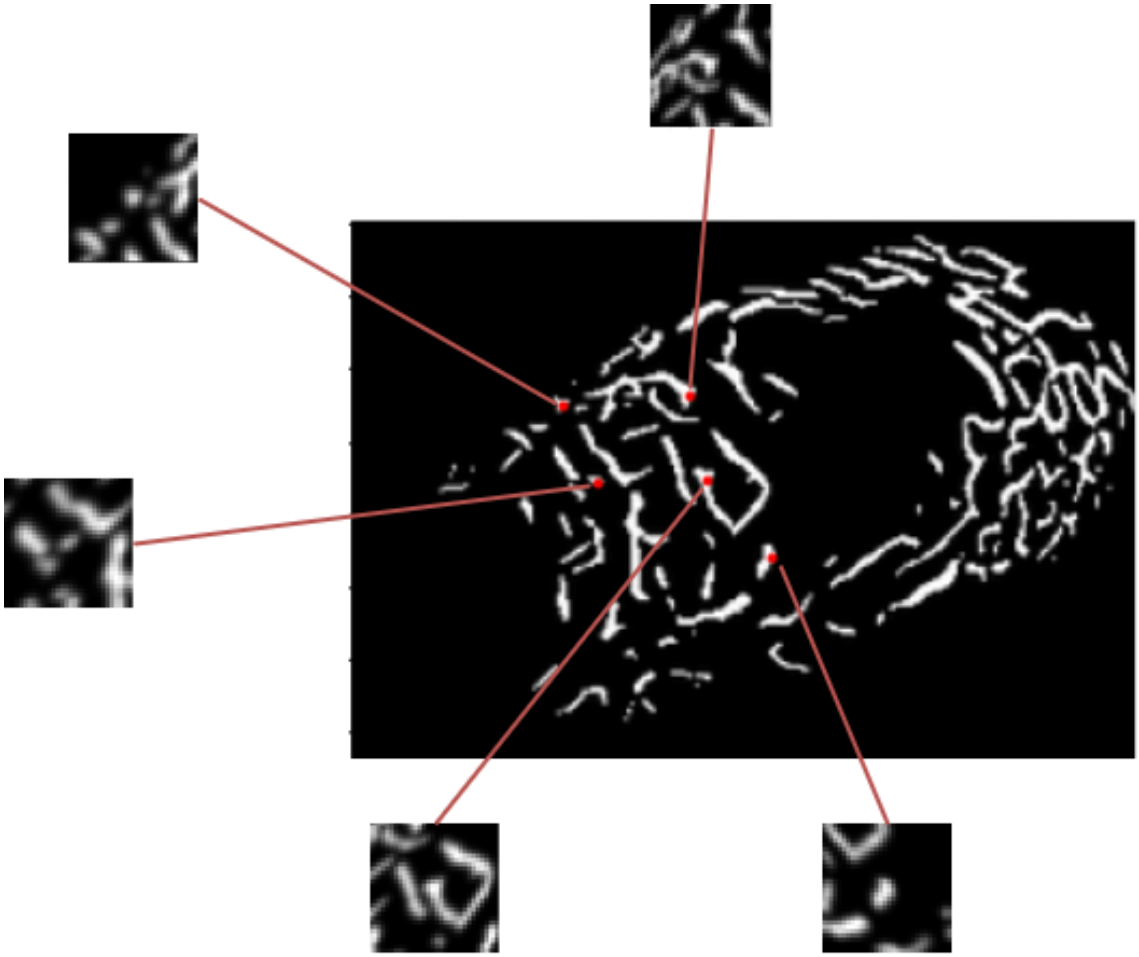

Figure 20. Patches extracted by HesAffNet and their locations in the pattern image.

where d is the chosen distance function for measuring the distance between descriptors, a_i is the reference descriptor from group of descriptors A, p_i is from another group of descriptors P and is a positive match to a_i, $p_{j_{min}}$ is the closest negative match to a_i from P and $a_{k_{min}}$ is the closest negative match to p_i from A. Choosing the minimum distance between the positive and negative matches in the loss function effectively always picks the most difficult sample into the triplet, improving the net's ability to avoid false positives. After the HardNet embedding, PCA is applied to the features to achieve decorrelation and dimensionality reduction.

4.3 Feature aggregation

The main difference to the method presented in [13] is in the aggregation of features from multiple images to a single descriptor. The features embedded by HardNet are grouped

for each seal individual. The grouping is done based on the metadata of the images which contains information such as the identity of the individual and the time and place where the image was captured.

The Fisher Vectors for the extracted features are created by using parameters from a Gaussian Mixture Model and the feature vectors $I = (x_1, \ldots, x_N)$ [73]. A GMM is a mixture of Gaussian distributions with K components (distributions). The parameters of a GMM are

$$\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K) \tag{3}$$

where μ_k are the means, Σ_k the covariances and π_k the priors of the components.

In order to create the Fisher Vectors, each feature vector x_i is connected to each of the K components in the GMM with a strength that is the posterior probability of the component defined as [73]

$$q_{ik} = \frac{\exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right]}{\sum_{t=1}^{K} \exp\left[-\frac{1}{2}(x_i - \mu_t)^T \Sigma_k^{-1}(x_i - \mu_t)\right]} \tag{4}$$

The mean and covariance deviation vectors for the K components are [73]

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}, \tag{5}$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{6}$$

where $j = 1, 2, \ldots, D$ spans the vector dimensions.

The Fisher Vector is formed by stacking each of the vectors u_k and the vectors v_k for each of K components in the GMM, defined as [73]

$$\Phi(I) = \begin{bmatrix} \vdots \\ u_k \\ \vdots \\ v_k \\ \vdots \end{bmatrix} \tag{7}$$

All of the grouped features are encoded to a single Fisher Vector to form a descriptor for the seal, utilising information from multiple images. The vocabulary of features for the Fisher Vectors is created by applying a Gaussian Mixture Model to the database patch features. The grouping step is illustrated in Fig. 21.

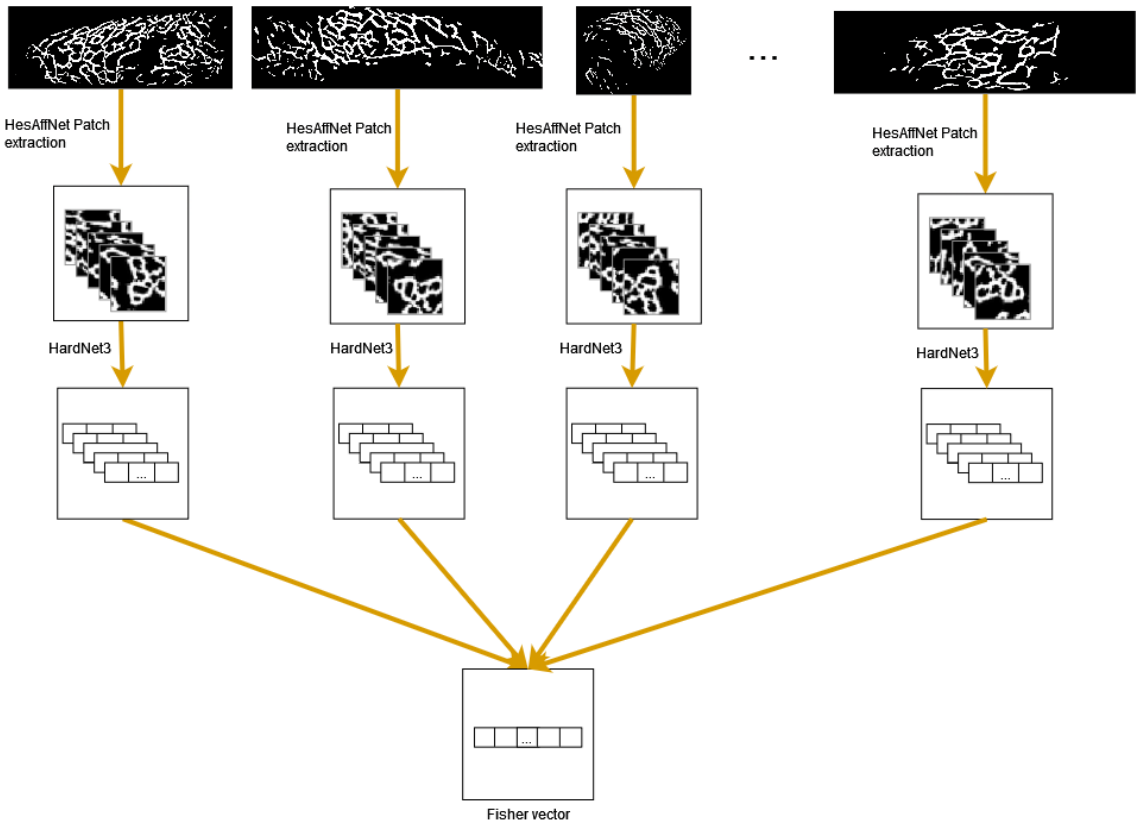

Figure 21. Illustration of the grouping step.

4.4 Re-identification

The final re-identification is performed by measuring the distance between the query descriptor and each of the database descriptors. This is achieved by computing the cosine distance between the descriptors. The cosine distance is calculated as [74]

$$D_{cos} = 1 - \frac{u \cdot v}{||u||_2||v||_2} \tag{8}$$

where u and v are the seal descriptor Fisher Vectors.

Once the distances between the query descriptor and all database descriptors have been computed, the class of the database descriptor with the shortest distance to the query descriptor is chosen as the predicted class. The database descriptors can be stored in the database to remove the need to compute them for each query, reducing the amount of computation needed when a query is made.

Three different methods of utilising the aggregated features were implemented with respect to whether the query features, database features or both were aggregated from multiple images. The implemented methods were one-to-many, where the database features are aggregated, many-to-one, where the the query features are aggregated and many-to-many, where both the query and database features from multiple images are aggregated to a single descriptor. The different methods are illustrated in Fig. 22, visualizing what is grouped and when.

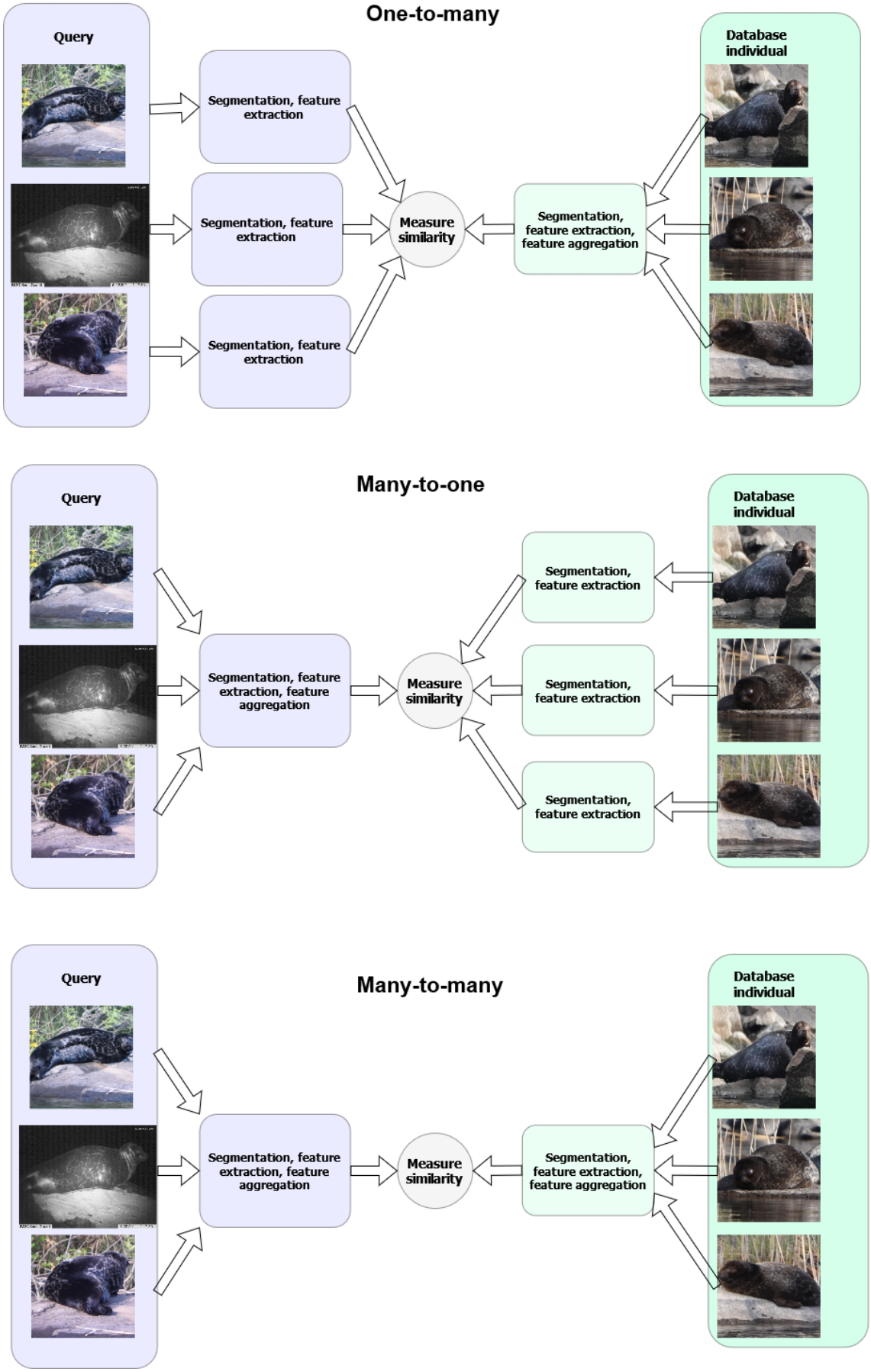

Figure 22. The different manners of grouping visualized, showing when the features from the query, database or both are aggregated.

5 EXPERIMENTS

5.1 Data

Two different datasets were used for evaluation of the re-identification performance. The first dataset is the SealID dataset [2] which was used for the re-identification experiments in [13] as well. The SealID dataset consists of 2080 segmented seal images with 57 different individual seals. The dataset is divided into database and query sets with the database set containing 430 images and the query set containing 1650 images. Example images from SealID dataset are shown in Fig. 23.

Figure 23. Example images of various individuals from SealID dataset.

The second dataset, SealID_seq, is a larger, more difficult set of images consisting of seal images from camera traps. The images in the set are generally lower quality with worse pelage pattern extraction performance than those in the SealID set. SealID_seq contains

multiple image sequences for each individual, with the sequences separated by the date they were captured. After segmenting and extracting patterns from the raw images, images with less than 10% of the area containing pattern or images with no patches found were discarded. After this preprocessing step, the dataset used for the re-identification experiments contains images of 29 individuals, consisting of 47171 pattern images in 183 separate sequences. The mean sequence length is 258 images, with the shortest sequence containing 1 image while the longest sequence contains 1565 images.

Especially in larger sequences the pose of the seal can vary greatly, allowing for a better representation of the entire pelage pattern. The lighting conditions can also change within a sequence with some images for example having the seal lit up by the sun while other images are darker or night vision as the day goes on. An example sequence from SealID_seq is shown in Fig. 24 where a gradual change in the pose and illumination can be seen throughout the sequence.

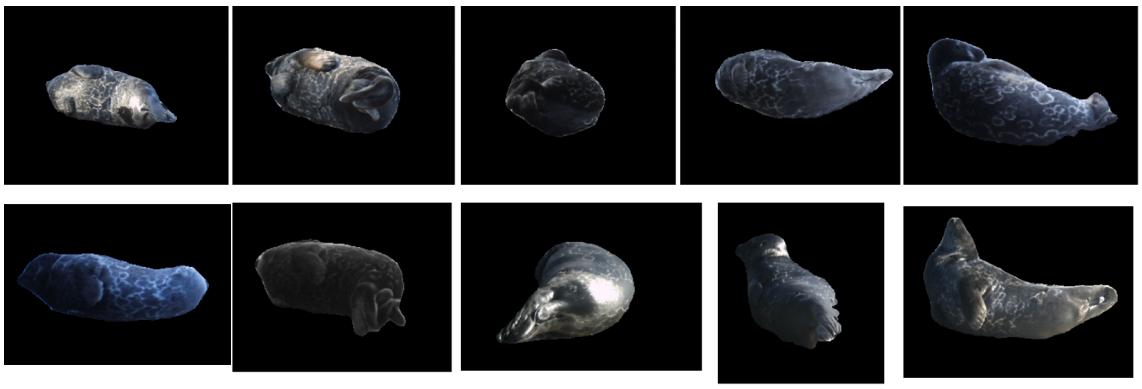

Figure 24. Example sequence from SealID_seq dataset.

5.2 Description of experiments

In Experiment 1 the query and database images from SealID dataset were used to evaluate the performance of the one-to-many and many-to-many matching methods in relation to the existing one-to-one method. When using the grouping method, all features from each image of each individual were aggregated to a single descriptor.

The following methods utilising the grouping were tested: one-to-many, many-to-one and many-to-many. In the one-to-many matching the database features for each individual are grouped to single descriptor and every query image is matched to these descriptors. In

the many-to-one matching the query features for each individual are instead grouped and matched to the individual database images. In the many-to-many matching both the query and database features are grouped to single descriptors for the matching process.

In addition to the proposed methods, a voting method was tested as a more naive way of utilising information from multiple images in the re-identification process. The distances between the descriptors were computed using the one-on-one method as in [13], and a majority voting was done with the ten nearest database images. If a class had more images among these ten nearest images than other classes, thus a majority, it was chosen as the prediction. In case of ties the classes were ranked by the lowest minimum distance.

In Experiment 2 the database images from SealID dataset were still used as the database set, but the entirety of SealID_seq dataset was used as the query set. As SealID_seq contained multiple sequences of each individual, images from each sequence were grouped when using the grouping method, instead of grouping all images of an individual to a single descriptor.

Some of the query sequences in SealID_seq contain very large amounts of images with only a small variation in pose between consequent images. The impact of discarding the overlapping information from similar images was tested by implementing a sparse version of the algorithm, where only every 10th image from a sequence is used if the sequence contains more than 20 images.

5.3 Evaluation criteria

To evaluate the re-identification performance of the proposed methods, Top-1-accuracy and Top-5-accuracy were used as the metrics to determine the accuracy of the predictions. When using Top-1-accuracy, a correct prediction is one where the nearest matching image predicted by the model is of the correct class. When using Top-5-accuracy a correct prediction is such that an image of the correct class is found within the five nearest images predicted by the model. The accuracy is then calculated as

$$accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ query\ samples} \tag{9}$$

5.4 Results

Results from Experiment 1 are presented in Table 1 and Fig. 25. The one-to-one accuracy was calculated by using the version of the pipeline from [13].

Table 1. Experiment 1 results.

Method	Top-1 accuracy	Top-5 accuracy
One-to-one	77.33%	84.79%
One-to-many	69.27%	81.15%
Many-to-one	100.00%	100.00%
Many-to-many	100.00%	100.00%
Voting	59.45%	84.85%

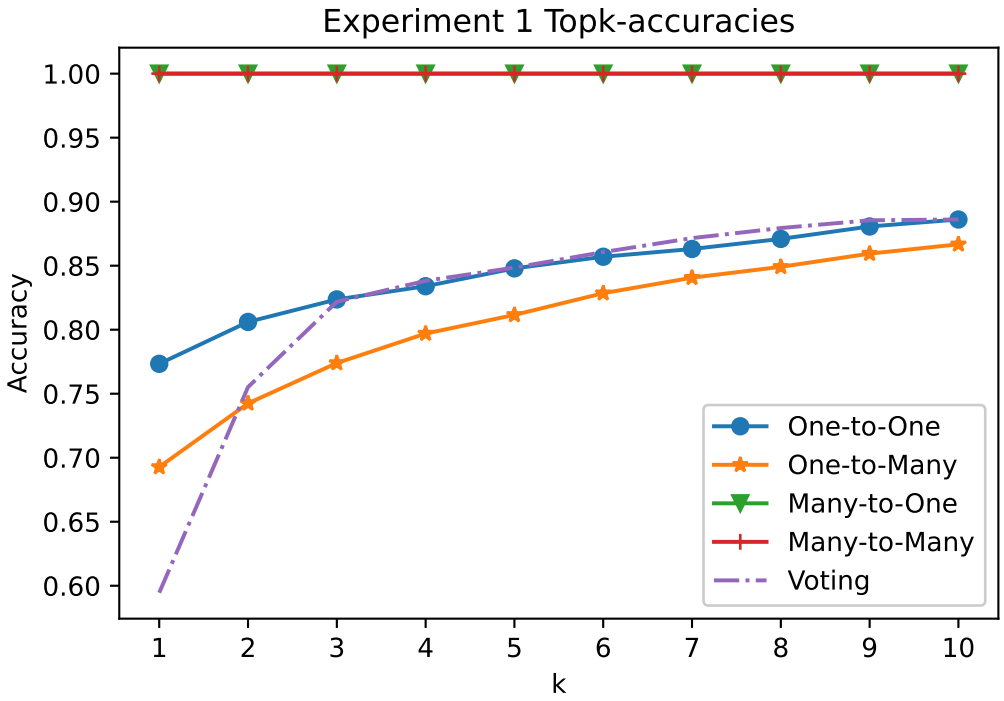

Figure 25. Top-k-accuracies from Experiment 1 for values of k from 1 to 10.

Results from Experiment 2 are presented in Table 2. The sparse parameter indicates whether sequences over 20 images long were truncated by only using every tenth image in the sequence. The achieved accuracies are significantly lower than those achieved for

the SealID dataset, showcasing the more difficult nature of SealID_seq. Figures 26 and 27 show the topk-accuracies for values of k from 1 to 10.

Table 2. Experiment 2 results.

Method	Sparse	Top-1 accuracy	Top-5 accuracy
One-to-one	No	40.40%	55.93%
	Yes	40.68%	55.92%
One-to-many	No	34.72%	52.21%
	Yes	34.87%	52.11%
Many-to-one	No	50.27%	75.41%
	Yes	50.82%	75.96%
Many-to-many	No	55.19%	71.04%
	Yes	54.64%	71.58%
Voting	No	31.87%	56.81%
	Yes	31.98%	56.92%

In Fig. 28 example images from a correct match produced by many-to-many matching are shown. The query images have been picked with intervals of ten starting from the beginning of the sequence to illustrate the change in the pose of the seal over the course of the sequence. With the database set containing enough images to create a representation of the entire seal a successful match has been made.

In Fig. 29 example images from an incorrect match produced by many-to-many matching are shown. Over the course of the query sequence the seal is largely captured from a single angle. With fewer angles to work with, gaining information of the entire pelage pattern can be difficult, offering a possible explanation for the misclassification.

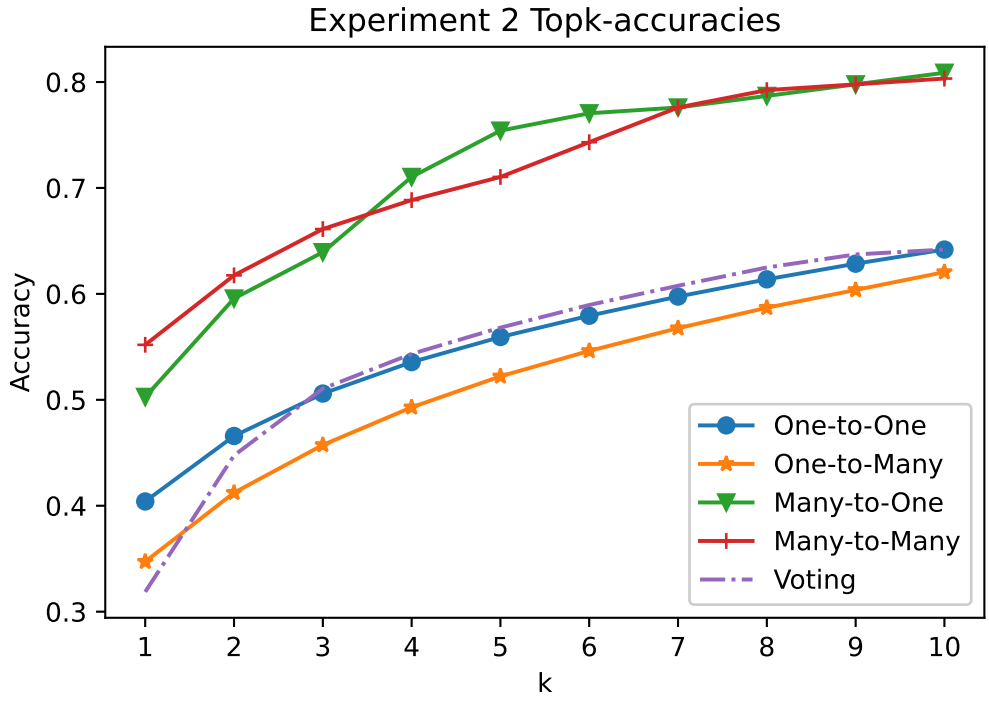

Figure 26. Top-k-accuracies from Experiment 2 for values of k from 1 to 10.

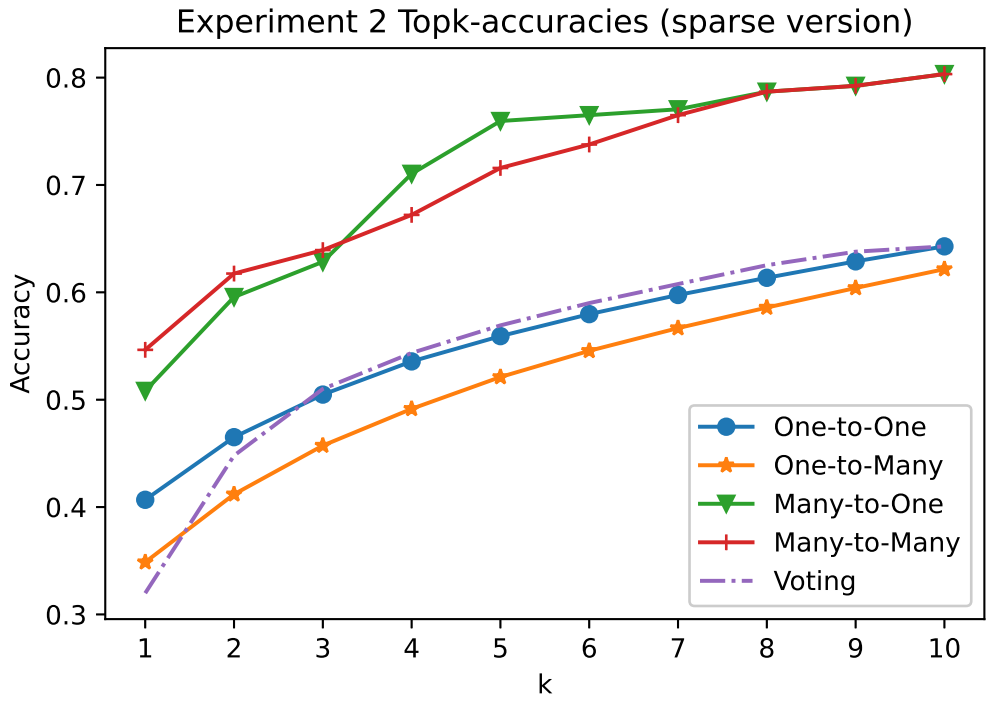

Figure 27. Top-k-accuracies from Experiment 2 for values of k from 1 to 10 (sparse version).

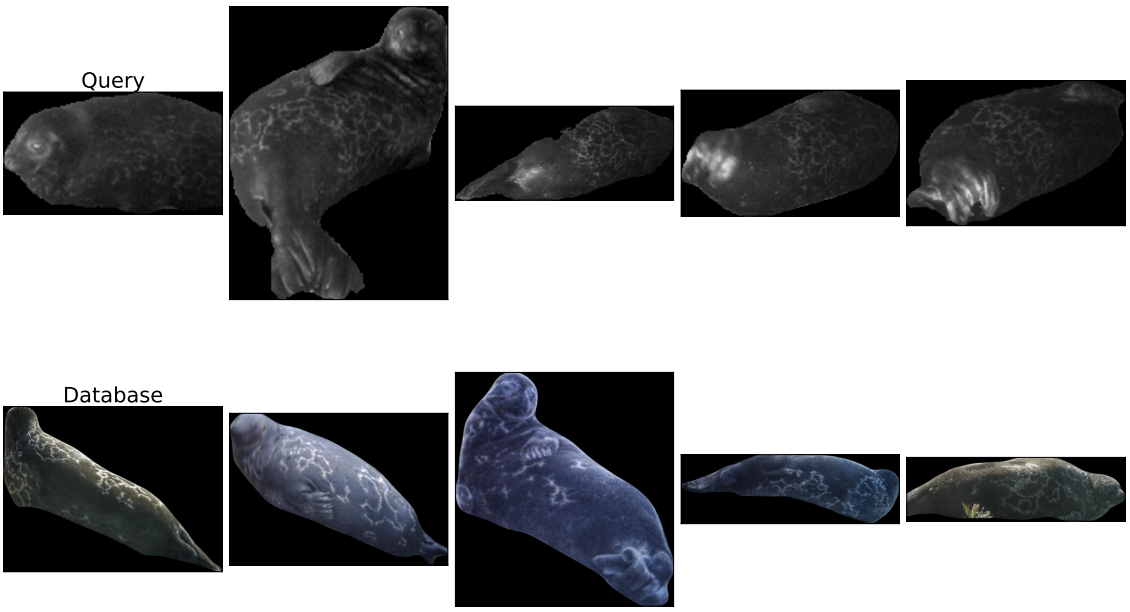

Figure 28. Example images from correctly matched query and database sequences.

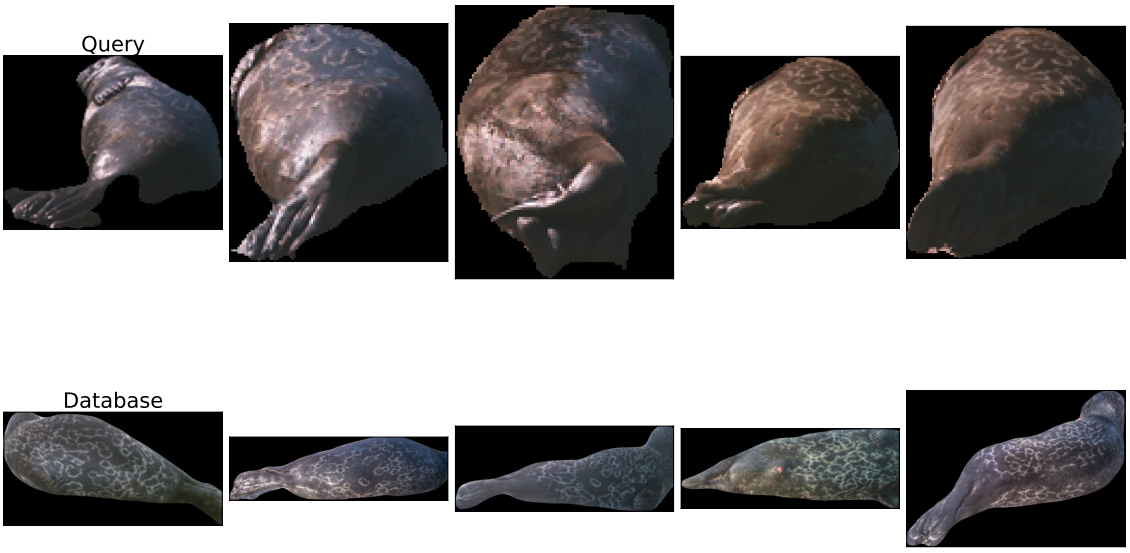

Figure 29. Example images from incorrectly matched query and database sequences.

6 DISCUSSION

6.1 Current study

The objectives of this thesis were to implement the one-to-many and many-to-many matching for Saimaa ringed seal re-identification and to evaluate the performance of the implemented methods. The results of the experiments show that aggregating over the query images gives a substantial improvement in re-identification accuracy, with many-to-one and many-to-many matching achieving perfect accuracy with the SealID dataset, and outperforming the use of a single query image with the more challenging SealID_seq dataset as well. The achieved perfect accuracies are very high and a possible explanation is offered by the fact that combining features from all images reduces the amount of descriptors, reducing the chances of getting the re-identification wrong. In Experiment 1 in both many-to-one and many-to-many matching the amount of query descriptors is the amount of individuals in the dataset (57) which is not very many. The simple voting method did not perform as well as the other grouping methods, only achieving better Top-5 accuracy than one-to-many matching and being outclassed in all other cases. The performances of one-to-one matching and the voting method indicate that aggregating features from multiple images to a single descriptor does offer better re-identification accuracy.

Surprisingly, one-to-many matching fails to improve the accuracy over one-to-one matching. A possible explanation could arise from the fact the database set from the SealID dataset contains just enough images for each seal that their full body pattern is covered. While aggregating the features from each image does increase the overall information of the individual's pattern, it effectively increases the amount of noise for query images where only a certain angle of the pattern is visible. The poor performance of the voting method could be explained in a similar fashion. While one of the angles from database could match decently to the query image, the rest would be just noise and this time without the multi-view information from one-to-many.

For similar reasons one could argue that many-to-one matching should also perform worse than one-to-one matching. However, on the query side the amount of images is larger, with more overlap in views between the images. Thus, aggregating features from the query images can lead to better representation of features that are matched to the database features. Many-to-one matching also performed nearly as well as many-to-many matching indicating that aggregating features from the query images is more important than

aggregating over database images. The query images are generally of a lower quality with less of the pelage pattern showing while the database images are much better in terms of how much pattern can be extracted from an image. Thus the query descriptors would gain more from the aggregation from multiple images.

Based on the performance of the sparse version more images in a sequence does not seem to make a large difference in the accuracy. Long sequences tend to have little variation in pose between consequent images and leaving out some of the images can greatly reduce the required computation time while not impacting the re-identification accuracy. Gaining a complete representation from multiple angles of the seal appears to be a more important factor based on the observations from Figures 28 and 29.

6.2 Future work

As discussed above, one-to-many matching failed to improve upon the baseline of one-to-one matching. A larger database set of images with more overlap in views between the images could be experimented with to further test the performance of one-to-many matching, and seeing whether adding more information to the database side would improve its accuracy. A new aggregation method could also be developed to gain representation of different sides of a database seal. This could be achieved by using for example clustering to group the features and form the Fisher Vectors from the clusters, effectively creating representations of different sides. Since a single query image has a limited view of the seal, this could improve the chances of a correct match.

As was observed from the performance of the sparse version, the number of images in a sequence did not seem to have a large impact on the re-identification accuracy, while acquiring a complete representation of the seal seems to carry more importance. Finding a method to select the minimum required images from a sequence to gain as much information from different views while discarding overly overlapping images could be a way to further increase the efficiency.

7 CONCLUSION

In this thesis, the one-to-many, many-to-one and many-to-many matching of Saimaa ringed seals were considered to utilise information from multiple images or multiple views in the re-identification of the seals. An approach using Fisher Vectors to aggregate features from multiple images to a single descriptor was implemented and experimented with.

In the experiments three different ways of aggregating the features were considered: 1) one-to-many, where database features for a seal were aggregated from multiple images, 2) many-to-one, where query features for a seal were aggregated from multiple images, and 3) many-to-many, where both query and database features for a seal were aggregated from multiple images. Of these tested methods, the many-to-many matching outperformed both the other grouping methods and the existing one-to-one matching where the descriptor for a seal was formed from a single image.

REFERENCES

[1] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

[2] Ekaterina Nepovinnykh, Tuomas Eerola, Vincent Biard, Piia Mutka, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. SealID: Saimaa ringed seal re-identification database. 2022. To be submitted.

[3] Mervi Kunnasranta, Marja Niemi, Miina Auttila, Mia Valtonen, Juhana Kammonen, and Tommi Nyman. Sealed in a lake—biology and conservation of the endangered saimaa ringed seal: A review. *Biological Conservation*, 253:108908, 2021.

[4] 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.

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

[6] Stefan Schneider, Graham W Taylor, and Stefan C Kremer. Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 44–52, 2020.

[7] CoExist Towards Sustainable coexistence of seals and humans. LUT School of Engineering Science, Computer Vision and Pattern Recognition Laboratory. `http://www2.it.lut.fi/project/coexist/index.shtml`, 2020. Accessed: 2022-01-31.

[8] Artem Zhelezniakov, Tuomas Eerola, Meeri Koivuniemi, Miina Auttila, Riikka Levänen, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. Segmentation of saimaa ringed seals for identification purposes. In *International Symposium on Visual Computing*, pages 227–236, 2015.

[9] Tina Chehrsimin, Tuomas Eerola, Meeri Koivuniemi, Miina Auttila, Riikka Levänen, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. Automatic individual identification of saimaa ringed seals. *IET Computer Vision*, 12(2):146–152, 2018.

[10] Ekaterina Nepovinnykh, Tuomas Eerola, Heikki Kälviäinen, and Gleb Radchenko. Identification of saimaa ringed seal individuals using transfer learning. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 211–222, 2018.

[11] Ekaterina Nepovinnykh, Tuomas Eerola, and Heikki Kalviainen. Siamese network based pelage pattern matching for ringed seal re-identification. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 25–34, 2020.

[12] Ilja Chelak, Ekaterina Nepovinnykh, Tuomas Eerola, Heikki Kalviainen, and Igor Belykh. Eden: Deep feature distribution pooling for saimaa ringed seals pattern matching. *arXiv preprint arXiv:2105.13979*, 2021.

[13] Ekaterina Nepovinnykh, Ilja Chelak, Tuomas Eerola, and Heikki Kälviäinen. NORPPA: Novel ringed seal re-identification by pelage pattern aggregation. 2022. To be submitted.

[14] Ekaterina Nepovinnykh, Ilja Chelak, Andrei Lushpanov, Tuomas Eerola, Heikki Kälviäinen, and Olga Chirkova. Matching individual ladoga ringed seals across short-term image sequences. *Mammalian Biology*, 2022.

[15] Omkar M Parkhi, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. A compact and discriminative face track descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1693–1700, 2014.

[16] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016.

[17] Tanya Y Berger-Wolf, Daniel I Rubenstein, Charles V Stewart, Jason A Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880*, 2017.

[18] David Tilman, Michael Clark, David R Williams, Kaitlin Kimmel, Stephen Polasky, and Craig Packer. Future threats to biodiversity and pathways to their prevention. *Nature*, 546(7656):73–81, 2017.

[19] Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao. Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092, 2016.

[20] Roland Kays, Sameer Tilak, Bart Kranstauber, Patrick A Jansen, Chris Carbone, Marcus J Rowcliffe, Tony Fountain, Jay Eggert, and Zhihai He. Monitoring wild animal communities with arrays of motion sensitive camera traps. *arXiv preprint arXiv:1009.5718*, 2010.

[21] Paul Douglas Meek, Karl Vernes, and Greg Falzon. On the reliability of expert identification of small-medium sized mammals from camera trap photos. *Wildlife Biology in Practice*, 9(2):1–19, 2013.

[22] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[23] Jason Parham, Charles Stewart, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, and Tanya Berger-Wolf. An animal detection pipeline for identification. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1075–1083, 2018.

[24] Boon Tatt Koik and Haidi Ibrahim. A literature survey on animal detection methods in digital images. *International Journal of Future Computer and Communication*, 1(1):24, 2012.

[25] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[26] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[27] Hayder Yousif, Jianhe Yuan, Roland Kays, and Zhihai He. Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. In *IEEE International Symposium on Circuits and Systems*, pages 1–4, 2017.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[29] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. From contours to regions: An empirical evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2294–2301, 2009.

[30] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2010.

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[32] Yongliang Qiao, Matthew Truman, and Salah Sukkarieh. Cattle segmentation and contour extraction based on mask r-cnn for precision livestock farming. *Computers and Electronics in Agriculture*, 165:104958, 2019.

[33] Sally A Mizroch, Judith A Beard, and Macgill Lynde. Computer assisted photo-identification of humpback whales. *Report of the International Whaling Commission*, 12:63–70, 1990.

[34] GR Hillman, B Wursig, GA Gailey, N Kehtarnavaz, A Drobyshevsky, BN Araabi, HD Tagare, and DW Weller. Computer-assisted photo-identification of individual marine vertebrates: a multi-species system. *Aquatic Mammals*, 29(1):117–123, 2003.

[35] BN Araabi, N Kehtarnavaz, T McKinney, G Hillman, and B Würsig. A string matching computer-assisted system for dolphin photoidentification. *Annals of Biomedical Engineering*, 28(10):1269–1279, 2000.

[36] S Ravela and L Gamble. On recognizing individual salamanders. In *Asian Conference on Computer Vision*, pages 742–747, 2004.

[37] Z Arzoumanian, Jason Holmberg, and Brad Norman. An astronomical pattern-matching algorithm for computer-aided identification of whale sharks rhincodon typus. *Journal of Applied Ecology*, 42(6):999–1011, 2005.

[38] Carlos Anderson. *Individual identification of polar bears by whisker spot patterns*. PhD thesis, University of Central Florida, Orlando, Florida, 2007.

[39] Tilo Burghardt and Neill Campbell. Individual animal identification using visual biometrics on deformable coat patterns. In *International Conference on Computer Vision Systems*, 2007.

[40] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[41] Douglas T Bolger, Thomas A Morrison, Bennet Vance, Derek Lee, and Hany Farid. A computer-assisted system for photographic mark–recapture analysis. *Methods in Ecology and Evolution*, 3(5):813–822, 2012.

[42] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[43] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012.

[44] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[45] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. Are you convinced? choosing the more convincing evidence with a siamese network. *arXiv preprint arXiv:1907.08971*, 2019.

[46] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Deep Learning Workshop*, 2015.

[47] Debayan Deb, Susan Wiper, Sixue Gong, Yichun Shi, Cori Tymoszek, Alison Fletcher, and Anil K Jain. Face recognition: Primates in the wild. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–10, 2018.

[48] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-based Pattern Recognition*, pages 84–92, 2015.

[49] Alceu Ferraz Costa, Gabriel Humpire-Mamani, and Agma Juci Machado Traina. An efficient algorithm for fractal analysis of textures. In *SIBGRAPI Conference on Graphics, Patterns and Images*, pages 39–46, 2012.

[50] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018.

[51] Yoshinobu Sato, Shin Nakajima, Nobuyuki Shiraga, Hideki Atsumi, Shigeyuki Yoshida, Thomas Koller, Guido Gerig, and Ron Kikinis. Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. *Medical Image Analysis*, 2(2):143–168, 1998.

[52] Denis Zavialkin. Cnn-based ringed seal pelage pattern extraction. Master's thesis, Lappeenranta-Lahti University of Technology LUT, 2020.

[53] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[55] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010.

[56] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[57] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[58] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *European Conference on Computer Vision*, pages 284–300, 2018.

[59] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *arXiv preprint arXiv:1705.10872*, 2017.

[60] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[61] Yang Li, Ziyan Wu, and Richard J Radke. Multi-shot re-identification with random-projection-based random forests. In *Winter Conference on Applications of Computer Vision*, pages 373–380, 2015.

[62] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, pages 68.1–68.11, 2011.

[63] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Recent developments of content-based image retrieval (CBIR). *Neurocomputing (Amsterdam)*, 452:675–689, 2021.

[64] Zachary Serlin, Guang Yang, Brandon Sookraj, Calin Belta, and Roberto Tron. Distributed and consistent multi-image feature matching via quickmatch. *The International Journal of Robotics Research*, 39(10-11):1222–1238, 2020.

[65] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *International Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.

[66] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[67] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *IEEE International Conference on Computer Vision*, pages 945–953, 2015.

[68] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *European Conference on Computer Vision*, pages 93–110, 2020.

[69] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[70] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703, 2014.

[71] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *British Machine Vision Conference*, page 2, 2015.

[72] Chanop Silpa-Anan and Richard Hartley. Optimised kd-trees for fast image descriptor matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[73] Fisher vector fundamentals. The VLFeat Authors. `https://www.vlfeat.org/api/fisher-fundamentals.html`, 2007. Accessed: 2022-05-23.

[74] SciPy API reference, distance computations, cosine. The SciPy community. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html`, 2022. Accessed: 2022-05-30.