Machine Learning Based Analysis of Finnish World War II Photographers

Kateryna Chumachenko¹, Anssi Männistö², Alexandros Iosifidis³, and Jenni Raitoharju¹,*

¹Unit of Computing Sciences, Tampere University, Tampere, Finland
²Unit of Communication Sciences, Tampere University, Tampere, Finland
³Department of Engineering, Aarhus University, Aarhus, Denmark
*firstname.lastname@tuni.fi

ABSTRACT

In this paper, we demonstrate the benefits of using state-of-the-art machine learning methods in the analysis of historical photo archives. Specifically, we analyze prominent Finnish World War II photographers, who have captured high numbers of photographs in the publicly available SA photo archive, which contains 160,000 photographs from Finnish Winter, Continuation, and Lapland Wars captures in 1939-1945. We were able to find some special characteristics for different photographers in terms of their typical photo content and photo types (e.g., close-ups vs. overview images, number of people). Furthermore, we managed to train a neural network that can successfully recognize the photographer from some of the photos, which shows that such photos are indeed characteristic for certain photographers. We further analyzed the similarities and differences between the photographers using the features extracted from the photographer classifier network. All the extracted information will help historical and societal studies over the photo archive.

Introduction

The Finnish army collected a unique and internationally significant database of photographs during the Winter War, Continuation War, and Lapland War in 1939-1945. This collection is known as the SA (from Suomen armeija = Finnish army) photo archive¹ and it consists of almost 160,000 photographs captured by men who served in TK (Tiedotuskomppania = Information company) troops. The archive has been digitized in at the beginning of 2010s and made publicly available in 2013. In its extent and historical significance, the SA photo archive is comparable to the American Farm Security Administration/Office of War Information Photograph Collection², which contains about 175,000 photos taken during the depression and drought in 1930s and World War II.

When considering the SA photo collection, it is necessary to bear in mind that the photos are not independent journalistic works, but the Finnish army regulated the topics that should or should not be captured. The photographers could not express their own interpretations of the events. The photos had an important task to keep up the spirits in the home front and they were also used for clearly propagandistic purposes. Nevertheless, the SA archive provides a unique view into the every day life of people behind the scenes. One of the official tasks of the TK troops was to collect ethnographic records. The archive provides a unique cross section of the life especially in the Eastern Karelia occupied by Finnish troops during the Continuation War.³

The SA archive provides a valuable source of information for historians, photojournalists, and other researcher searching information of the life and sentiments behind the battles. However, the original photograph labeling typically provides only the date, the place, the photographer, and a brief description of the key content. Much of the content providing insight into the every day life and sentiments of the people has not been described. Therefore, humanistic researchers have invested a considerable amount of time and effort to manually go through the collection and search for the information related to the studies at hand. In this paper, we demonstrate that machine learning algorithms can ease the photo analysis and provide information that would be hard to obtain by manual inspection.

The SA collection has been captured by several hundreds of photographers. However, most of them only took one or few images and just a few dozen photographers captured half of the images. While the photographers did not have the freedom to select their topics freely, each photographer still provides a subjective view of the events. The objects appearing in the photos and the scene setup vary based on the position and personal preferences of the photographers. Some of the photographers can be considered as skillful photo artists, while others simply recorded the events with their cameras. Therefore, a better understanding of the differences of the individual TK photographers can provide deeper insight into the significance of the content and help researchers to find the content they are looking for.
In this paper, we exploit the state-of-the-art machine learning algorithms to analyze the characteristics and differences of 21 active TK photographers. We examine the typical topics and photo setup (i.e., close-ups vs. overview images) for each photographer and we evaluate how how distinguishable different photographers are.

Results

We selected the 21 Finnish war photographers, who were among the most active ones, in terms of the numbers of photos in the SA archive. Due to the missing or unclear original image annotations, we considered only a subset of images which could be easily connected with the certain photographer. The selected photographers along with the number of considered images for each photographer are listed in Table 1. The table also assigns photographer IDs used in later tables and illustrations. The total number of photographs considered in our analysis is 34910.

| ID  | Photographer      | Total |
|-----|-------------------|-------|
| 1   | Pauli Jänis       | 2280  |
| 2   | Oswald Hedenström | 2234  |
| 3   | Ovaskainen        | 2069  |
| 4   | Tuovi Nousiainen  | 2066  |
| 5   | Kyytinen          | 1914  |
| 6   | Kim Borg          | 1768  |
| 7   | Taube             | 1736  |
| 8   | Väinö Hollming    | 1731  |
| 9   | Uomala            | 1687  |
| 10  | Tauno Norjavirta  | 1690  |
| 11  | Kauko Kivi        | 1696  |
| 12  | Viitasalo         | 1685  |
| 13  | Sundström         | 1674  |
| 14  | Viljo Pietinen    | 1456  |
| 15  | Esko Suomela      | 1426  |
| 16  | Laukka            | 1412  |
| 17  | Esko Manninen     | 1383  |
| 18  | Nurmi             | 1313  |
| 19  | Sjöblom           | 1285  |
| 20  | Niilo Helander    | 1250  |
| 21  | Feldhoff          | 1155  |

Table 1. Selected photographers and the total number of taken photographs

Object detection

We applied pretrained object detection algorithms to detect the objects appearing in images. Out of the 80 object classes, we manually selected 11 relevant classes (people, airplanes, boats, trains, cars, bicycles, skis, dogs, horses, chairs, and ties). We also empirically checked that the detection quality for these classes was high. Some of the potentially interesting classes (e.g., cow) we discarded, because many cow detection were actually horses, reindeer, or other objects. Also for the selected classes, the results should be considered only as indicative. When objects are clearly visible, they are typically well detected. However, there are cases where objects are missed or misidentified. Few examples of object detections are shown in Fig. 1.

It is evident that the results do not provide exact object numbers. Instead, we exploit the results to evaluate relative numbers of occurrences of different objects in the photographs of each photographer. The object detection results for each photographer are given in Table 2, where we report the ratio of images with people and the average number of persons in these images as well as the average number of occurrences of other objects per 100 images for each photographer. For each object class, we highlight the photographers with the most frequent (bolded) and infrequent (italic) occurrences.

As expected, we observe from Table 2 that different photographers concentrated on different content: 19-Sjöblom has people in 95% of his images, while 9-Uomala and 10-Norjavirta have people in less than two thirds of their images. 2-Hedenström and 17-Manninen have the highest average number of people in these images (i.e., only images with persons counted), while 4-Nousiainen and 20-Helander captured images with fewer people.

18-Nurmi and 20-Helander have captured high numbers of airplanes, while 5-Kyytinen and 15-Suomela concentrated on boats. In 11-Kivi’s photos, there are many animals (horses, dogs). Based on our manual inspection, chair pictures are
Figure 1. Examples successful and erroneous object detection results
Table 2. Ratio of photos with people, persons per such an images, and occurrences of other object classes per 100 images for different photographers.

typically taken indoors, while ties are worn by high ranking soldiers or wealthy people in urban conditions. 2-Hedenström and 17-Manninen, who have the highest average number of humans in their pictures, also have the most chairs. 8-Hollming and 9-Uomala have the lowest chair incidence. 9-Uomala has also a low ration of person images, while 8-Hollming pictured several skiing photos and dogs, which are clearly outdoor topics. 19-Sjöblom seems to profile as an urban photographer with a high number of ties and cars but only few animals or skis.

Target distance evaluation

The distance to the main target is one of the main stylistic decisions in photographs. The famous war photographs are typically close-ups showing emotions and capturing a specific moment in a skillful manner. Overview photographs showing crowds of people typically just record the events. Therefore, it is interesting to estimate if specific photographers were more commonly capturing close-ups or overview pictures from far away especially when it comes to photographs depicting people.

We examined the photographs with detected people and considered the relative size of the bounding boxes with respect to the image size. We manually defined two thresholds to divide such photographs into three classes: close-ups, mid-range photos, and overviews. Fig. 2 shows an example photograph belonging to each of these classes.

Fig. 3 shows how the person pictures are divided into different distance ranges for different photographers (the percentages of close-ups and overview photographs are shown, the remaining percentage is mid-range photos). The figure shows that 19-Sjöblom took relatively most close-ups and mid-range images and fewest overview images. From the previous subsection, we know that he had also the highest ratio of photos with people and he covered topics that profiled him as an urban photographer. 13-Sundström took fewest close-ups. 4-Nousiainen and 9-Uomala captured relatively most overview images. For 4-Nousiainen, this is somewhat surprising considering that he also has the fewest number of people per image. 9-Uomala, on the other hand, had only few chairs in his images, which we concluded to mean that he did mainly outdoor photography. These observations support each other as overview images are mainly outdoor images.

Photographer recognition

To evaluate how distinguishable different photographers are, we used some of the photographs to train a neural network to recognize the photographer from a photograph and tested if the network can be used to recognize the photographer for the unseen photographs not used in training. Here we split the photographs into train and test splits according to the capturing times to ensure that photographs depicting the same event are not be used for both training and testing.
Figure 2. Examples of photos taken from different distance ranges and the corresponding bounding boxes

Figure 3. Percentage of people images in different distance ranges
Overall, the network achieved 16.7% classification accuracy on the test set. The confusion matrix of the classification results is shown in Fig. 4, where all the diagonal elements represent correctly classified samples. We see that some of the photographers cannot be recognized and the network assigns them for 1-Janis as he has the highest number of photographs and, thus, the highest probability of being the correct choice for a random photo. However, there are some photographers (especially 4-Nousiainen, 5-Kyytinen, 8-Hollming, 10-Norjavirta, 13-Sundström, 17-Manninen, and 19-Sjöblom), where the recognition rate shows that some of the photographs are clearly distinguishable as theirs.

![Figure 4. Confusion matrix for photographer recognition](image)

Comparison of the recognition results with the earlier analysis on topics reveals that some of the easily recognized photographers have also specific topics. 4-Nousiainen took photos with lowest average number of persons and he also captured many trains and bicycles. 5-Kyytinen has many airplane pictures. 8-Hollming has the highest number of skiing pictures and only few chairs (i.e., many outdoor photos). 10-Norjavirta took many pictures with no people in them, and 13-Sundström took fewest close-ups. 17-Manninen had highest average number of people in his people photos and the highest occurrence of chairs (i.e., indoor photos). 19-Sjöblom captured urban topics. The confusion matrix also reveals some similarities between photographers who are often confused with each other. 5-Kyytinen and 6-Borg form such a pair. Photographs taken be 11-Kivi are often misclassified to be 2-Hedenström’s or 5-Kyytinen’s and photographs taken by 16-Laukka also as 5-Kyytinen’s.

We further examine the similarities and differences between the photographers by extracting the features learned by the classifier network for the test images and visualize them using the t-SNE algorithm in Fig. 5. In the figure, the dots denote photographs and different colors correspond to different photographers. Some of the colors are clearly concentrated on certain spots further confirming that different images are characteristic for different photographers.

**Discussion**

We showed that the modern machine learning algorithms can help in societal research on historical photo archives in many ways. In this paper, we applied state-of-the-art object detection models and neural network architectures to obtain statistics
and characteristics of prominent Finnish World War II photographers. We examined the typical topics in the photos of each photographer and analyzed the differences in their ways of capturing people. Furthermore, we showed that a convolutional neural network was able to some extent recognize photographers from the photos leading to the conclusion that certain photos can be considered typical for a specific photographer. The confusion matrix of the photographer classifier revealed some similarities between the photographers. All this information will help the historians and photo journalists in their work when analyzing the works of a certain photographer and their meaning in the photo collection.

This paper demonstrated the benefits of the publicly available pretrained machine learning models along with straightforward application of the existing labeling (photographer info) for training a photographer recognizer. The algorithms showed good performance on the historical black-and-white photographs even though pretrained with modern color photos. Thus, it can be concluded that the same methods can be easily applied on other historical photo archives.

In the future, we will concentrate on issues requiring more specialized methods such as recognizing object classes only appearing in Finnish historical photos or during World War II. We aim at exploiting the original photo descriptions to produce more complete object labeling and event recognition. We aim at eventually publishing our object detections and photo classifications in the archive to assist different types of societal studies on the archive.
Methods

Object Detection

For the detection of various objects in the photographs, we applied three state-of-the-art object detectors, namely Single-Shot Detector (SSD)\(^7\), You Only Look Once v3 (YOLOv3)\(^8\), and RetinaNet\(^7\). All models were pretrained on MS-COCO dataset\(^8\) that contains 80 classes. Among those, we considered people, airplanes, boats, trains, cars, bicycles, skis, dogs, horses, chairs, and ties as shown in\(^7\). At the end, we aggregated the information obtained from each object detector.

SSD

The first object detector applied was SSD that is one of the most well-known single-shot detectors. The detector is based on the VGG-16\(^9\) model pretrained on ImageNet dataset\(^10\) that is used as a backbone feature extractor, followed by several convolutional layers that downsample the image and result in multiple feature maps. Using these feature maps from different layers, the detection can be done on multiple scales, while preserving the parameters across all scales, ensuring that both large and small objects are detected equally well. In addition to that, the single-shot approach results in high inference speed of this detector.

SSD relies on the idea of default bounding boxes, meaning that prior to training, several default bounding boxes are determined based on the amount of feature maps to be used and the size of the feature maps. Bounding boxes are created for the aspect ratios of \(\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}\). During training, each groundtruth bounding box is associated with one of the default bounding boxes, determined by the highest Jaccard similarity, also referred to as Intersection over Union\(^11\). Intersection over Union is defined by the area of the intersection of two boxes divided by the area of the union of these boxes. This default bounding box becomes a positive example for the groundtruth box, while the others become negative examples.

At each scale, a feature map of different size is created and divided into a grid cell. During inference, a set of default bounding boxes is evaluated for each cell of the feature map and for each default bounding box, a shape offset is predicted along with the class probabilities for each class. Training is done with the combination of localization loss that is a Smooth L1 loss\(^12\) between the predicted box and the groundtruth box; and the confidence loss that is the cross-entropy loss over multiple class confidences. In our experiments, we use images rescaled to the size of 512 × 512 pixels as an input to SSD detector.

YOLOv3

The second object detector used was YOLOv3 that is in many ways similar to SSD: YOLO is a single-shot detector that makes predictions on multiple scales by performing detection on feature maps from different parts of the network. Prediction is done across three different scales obtained by dividing the image size into 32, 16, and 8.

YOLO relies on the ImageNet-pretrained Darknet-53 architecture that is used as a feature extractor backbone and multiple convolutional layers are added on top of it. Similarly to SSD, an image is divided into a grid cell and each cell is responsible for detecting the object, the center of which is located within its boundaries. Each grid cell predicts several bounding boxes along with the corresponding class label and confidence score.

Rather than predicting bounding box coordinates directly, YOLO predicts the offsets from the predetermined set of boxes, referred to as anchors boxes or prior boxes, and each box is represented by the width and height dimensions\(^13\). These anchor boxes are obtained by applying k-means clustering\(^14\) of the boxes in the training set with the distance defined as

\[
d(box, centroid) = 1 - IoU(box, centroid),
\]

where both box and centroid are the width and height of the corresponding box, IoU stands for Intersection over Union and \(k = 9\) is chosen. For calculation of IoU we assume that the centers of the boxes are located at the same point. More specifically, for the model trained on COCO dataset and 416 × 416 images, the anchor boxes are \(10 \times 13, 16 \times 30, 33 \times 23, 30 \times 61, 62 \times 45, 59 \times 119, 116 \times 90, 156 \times 198, 373 \times 326\).

For each detected bounding box, class prediction is obtained by multi-label classification with separate logistic classifiers. During training, the loss comprised of binary cross-entropy loss for object classification, and sum of squared error loss for bounding box prediction is used. YOLO operates on images of fixed size, and for our experiments all images were rescaled to 416 × 416 pixels size.

RetinaNet

The RetinaNet object detector is the third state-of-the-art object detector used in this work. Overall architecture of RetinaNet consists of the backbone network for feature extraction, namely, Feature Pyramid Network\(^15\) built on top of ResNet\(^16\), and two subnetworks, one of which is responsible for object classification, and the other one - for the bounding box regression. Similarly to previous detectors, the backbone network in pretrained on ImageNet dataset.

In a similar way to other detectors discussed so far, RetinaNet performs detection on multiple scales and relies on a predefined set of anchor boxes. Here, for each scale, anchors of 3 aspect ratios \(\{1 : 2, 1 : 1, 2 : 1\}\) and 3 sizes \(\{2^0, 2^4, 2^5\}\) are used, resulting in 9 anchor boxes per scale level.
The subnet for object classification is a small fully-connected network, where the parameters are shared between different scale levels. The network is comprised of $3 \times 3$ convolutional layers. For each spatial position and anchor box, a sigmoid activation function predicts the probability of presence of the object of a certain class, therefore making $A \times K$ binary predictions at each location, where $A$ is the number of anchor boxes and $K$ is the number of classes. The bounding box regression subnet is a fully-connected network that predicts four coordinates for each anchor box at each spatial location. The predicted coordinates correspond to the offset relative to the anchor.

The main difference from other detectors lies in the utilization of the new loss function, referred to as Focal Loss, designed to address the issue of imbalanced classes in the object classification subnet:

$$FL(p_t) = -\alpha (1 - p_t)^\gamma \log(p_t); \quad p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise} \end{cases}$$

(2)

where $y = \pm 1$ is the binary class label for the evaluated class, $p$ is the class probability, $\gamma$ is a focusing parameter, and $\alpha$ is a balancing parameter. For the input to this detector, we rescale the images preserving the aspect ratio and setting the size of the smaller side to 800 pixels, while keeping the size of a larger side at 1333 pixels maximum.

**Detection aggregation and further analysis**

From each detector, we obtain a set of bounding boxes that are given as 4 coordinates and a class label with a corresponding confidence score. In order to determine the final bounding boxes, the aggregation of the results from multiple detectors should be performed, and it can be achieved by multiple approaches. In our approach, we first identify which bounding boxes correspond to the same object by grouping together the bounding boxes with Intersection over Union above certain threshold, which we manually set to 0.1. Then, either the bounding box with the highest confidence score can be selected or the mean of each coordinate of all bounding boxes corresponding to the same object can be taken.

Following the first approach, issues related to different scoring systems of different detectors can arise, i.e., some detector might produce higher scores for all of its detections in general, while its bounding boxes might be less accurate. In our experiments, we follow the second approach of taking the mean value of the coordinate produced by all the detectors and we observe that generally this results in more accurate positioning of the bounding box, although this cannot be evaluated quantitatively without the groundtruth information. This process is applied to bounding boxes of each class separately.

After combining the predictions of each detector, based on the area occupied by the bounding box of person class, the prediction of the distance from the photographer to the scene was made - if the bounding box occupies more than 65% of the area, the photo was classified as a close-up, 10-65% - mid-range, and <10% - as an overview photo.

**Photographer recognition**

For recognizing the photographer from the photos, we applied a pretrained and finetuned convolutional network. We used VGG-19 architecture\(^3\), pretrained on ImageNet dataset as a backbone, with Dropout layers added after each pooling layer with keeping 20% of connections after the first 4 pooling layer, and 50% after the last one. Stochastic gradient descent was used for training with the learning rate of $10^{-5}$.

The training, validation, and test splits were selected in a way to exclude the duplicate photographs, where 60% of the earliest photos were selected as training set, the following 20% - as a validation set, and the rest - as the test set. For our experiments, histogram equalization was performed on each photo on the value component in the HSV space in order to improve the contrast of each photo. Then, each image was resized into $224 \times 224$ pixels size. Training was done for 100 epochs with batch size of 32 and categorical cross-entropy as a loss function.

**Photographer clustering**

In order to visualize the relationships between the photos of different photographers, we extract the feature map of the second last layer of the network trained for photographer recognition. The resulting feature map has high dimensionality and for the visualization purposes we exploit the t-Stochastic Neighbour Embedding algorithm (t-SNE)\(^4\). t-SNE is a data visualization method for high-dimensional data, that aims at mapping the data instances in the high-dimensional space to some low-dimensional space, where the similarity between instances is preserved. This is achieved by modelling the similarities between instances as conditional probabilities. In the high-dimensional space, the similarity between data instances $x_i$ and $x_j$ is represented by the probability of $x_j$ to be selected as the nearest neighbor of $x_i$ if neighbors were selected proportionally to their probability density under a Gaussian distribution centered at $x_i$. In the low-dimensional space, instead of using the Gaussian distribution, the Student’s t-distribution with one degree of freedom is used. Using a heavy-tailed distribution helps to model moderate distances in the high-dimensional space with a much larger distances in the low-dimensional space, resulting in better results compared to other methods. The Kullback-Leibler divergence of these probability distributions is then minimized with a gradient descent. The result of the visualization can be seen in Fig. 5.
References
1. The Finnish Defence Forces. SA photo archive. sa-kuvat.fi (2013).
2. Library of Congress. Farm security administration/office of war information black-and-white negatives. http://www.loc.gov/pictures/collection/fsa/ (2009).
3. Männistö, A. & Kivimäki, V. Sodan särkemä arki (WSOY, 2016).
4. van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605 (2009).
5. Liu, W. *et al.* Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37 (Springer, 2016).
6. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).
7. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2018).
8. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).
9. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
10. Deng, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*.
11. Levandowsky, M. & Winter, D. Distance between sets. *Nature* 234, 34 (1971).
12. Girshick, R. Fast r-cnn. In *IEEE international conference on computer vision*, 1440–1448 (2015).
13. Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).
14. Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory* 28, 129–137 (1982).
15. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125 (2017).
16. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 770–778 (2016).