Comparative analysis of identification of dynamic objects by scale-invariant feature transform and deep neural networks

E A Ivliev¹ and P S Obukhov¹

¹ Don State Technical University, Rostov-on-Don, Russia

Abstract. The article is devoted to the development and analysis of methods of identifying dynamic objects. A system for identifying information from a luggage tag based on several neural networks with the SSD InceptionV2 architecture has been developed. These neural networks work with sufficiently high accuracy 82-95% and speed 7-10fps. Advantages and disadvantages of application of method of scale-invariant feature transform for identification of luggage tags are considered. The operability of the methods on real images has been tested.

1. Introduction
The task of identifying dynamic objects in the video stream is one of the most popular in the field of technical vision. Based on it, many applications are solved. In this work, baggage tags in the airport sorting area are considered as dynamic objects.

The relevance of this work is due to the fact that employees of the airport sorting facilities are unable to easily identify luggage tags using bar code scanners because the database of luggage tags is inaccessible to airports.

2. Related work and the proposed methods
The task of identifying the baggage tag can be divided into two components:
1. Localization of the luggage tag or its main part - barcode.
2. Recognition of digital-letter information about the IATA airport code.

In the previously described article [1], for the task of localization, we used the algorithm developed by us to search for barcodes on a luggage tag. Later, it was experimentally revealed that this algorithm is inoperable under various conditions, such as: distance from the luggage tag to the video camera, illumination, orientation. Also, this algorithm has a disadvantage - as a result, the algorithm will form an area informing about being in this area of the barcode, but this area in various conditions does not always correspond to the area of the barcode, which nullifies further identification of information about the IATA airport code.

To solve the problem of recognizing IATA airport code, an algorithm was used to search for alphanumeric information. With this approach, you cannot search for all letters and numbers. And to classify these characters, a fully connected neural network was used. This neural network is less suitable for the task of classifying graphic information.

In this work, other methods are considered: the use of deep neural networks and the method of scale-invariant feature transform (SIFT).

To solve the problem of identification the baggage tag using neural networks, several different basic neural network architectures were considered. These neural networks are available in the TensorFlow
Object Detection API, which is an open source platform developed by Google based on TensorFlow [2]. TensorFlow allows you to easily build, train, and deploy object discovery models. For the purposes of this study, the following basic models were considered: faster R-CNN (Regions with CNNs) [3] using Inception V2 [4], SSD (Single Shot Detector) [5] using the InceptionV2 model, and SSD using MobileNetV2 [6]. According to the article [7], the SSD InceptionV2 architecture was chosen as it has a higher identification rate, consumes less memory, and the accuracy of the identification is comparable to the others.

Also, for the identification problem, the SIFT method was considered. SIFT keypoint detection is a powerful method used both for image classification and image correspondence [8]. The procedure of SIFT consists mainly of four steps: 1) scale-space extrema detection, 2) keypoint localisation, 3) orientation assignment, and 4) keypoint descriptor. The intuition behind SIFT relies on finding “keypoints” in an image and then computing a 128-dimensional descriptor around that point to summarize local gradient histograms information in a scale and rotational invariant way. That wealth of local informations can then be aggregated into a compact image representation using bag of visual word-type approaches [9].

3. Results and Discussions

To train the neural network, a training sample was created consisting of 200 images with luggage tags. Data annotation was performed by the LabelImg [10], using which the boundaries of the object of interest are highlighted and the class to which the object belongs is specified.

As a result of training at 240 eras, the accuracy of the neural network for localizing the luggage tag was 95.3%. Results of neural network operation on localization of luggage tag are shown in figure 1. Using this neural network, the search area for flight information is narrowed. However, for more precise narrowing of the search area, it is necessary to search the barcode area as the conceived flight information is located above or below the barcode. Therefore, a neural network was created to solve the problem of localizing the barcode. The same image database was used for training, but barcode regions were highlighted during data annotation. Upon completion of training, the accuracy of localization of the barcode area was 96.7% when using this neural network. The average processing time of one frame is 0.1sec - 10fps. The results of the neural network are shown in figure 2.

![Figure 1. Localization of luggage tag by neural network](image-url)
To solve the second problem of recognizing digital-letter information, the first thing to do is to achieve the correct positioning of the tag image. Article [1] solves this problem with an affine transformation whose input parameter is the barcode rotation angle. To calculate this angle, the algorithm also described in [1] should be used, however, in order to obtain greater accuracy, the input image needs to use the area limited by the neural network to localize the barcode.

To recognize alphanumeric information, a model of the SSD InceptionV2 neural network was used. Neural network training took place on a synthetic training sample consisting of images with randomly placed symbols of different fonts. At the same time, each symbol was rotated, scaled, distorted, noisy and discolored for an acceptable variety of characteristics. The proposed method created 8,000 images with numbers and letters for 36 different classes. When testing the neural network for the task of recognizing alphanumeric information, the accuracy was 82%. The average processing time of one frame is 0.14sec. - 7fps.

When using this SIFT algorithm to solve the problem of localizing the luggage tag, there were advantages and disadvantages.

The advantages of this algorithm include the speed of operation - the average processing time of one frame is 0.27sec.

The disadvantages of this algorithm include:
1. Algorithm inoperability in non-uniform lighting [11].
2. Its inoperability when moving an object, leading to blurring the image and as a result of the impossibility of searching for key points, and further constructing descriptors.
3. It cannot be applied to barcode localization. The calculation produces a different number of key points and different distances when calculating descriptors because all barcodes are unique [12].

The work of the SIFT method showed (Figures 3 and 4) that even when using images of the same tag, there were few matches, and when using images of different tags, there were even fewer matches.
The experiment shows that the simple SIFT algorithm does not cope with this task and requires improvements. This task can be solved with application of more difficult algorithm SIFT flow [14]. One popular method for computing the dense features is the Dense SIFT (DSIFT) descriptor which extracts SIFT histogram at a single scale for all pixels with overlapping patches. Using DSIFT, SIFT flow [14] aligns two images by minimizing matching cost and keeping the flow field smooth. Since the DSIFT feature is only computed at one scale in SIFT flow, it requires that objects in two images share the same/similar scales. This makes SIFT flow problematic in dealing with images of large scale change, which was observed in [13]. To overcome this problem, a recent method called Scale-Less SIFT (SLS) was proposed in [13]. When performing the matching, SLS extracts a set of DSIFT features at multiple scales for every pixel and uses set-to-set distance to measure the matching cost between the corresponding pixels. More specifically, SLS uses Projection Frobenius Norm to compute the set-to-set distance. In practice, if there are 100,000 pixels in each image then a locality constraint is needed; running SLS with the Projection Frobenius Norm is both time- and memory consuming. For an image of standard size $640 \times 480$, it consumes more than 10G memory and takes hours to perform one matching, which makes the SLS algorithm nearly impractical to scale up.

4. Conclusions
In the course of this work, neural networks with the SSD InceptionV2 architecture were developed to solve the problem of detecting luggage marks, barcodes, and alphanumeric information. These neural networks work with sufficiently high accuracy 82–95% and speed 7–10fps. As a result of the experiments, it was revealed that the SIFT algorithm has many disadvantages and the algorithm speed is significantly lower than the speed of the neural networks. Based on the above, we can conclude that the solution of our application problem is possible using deep neural networks and absolutely impossible using the SIFT method.

References
[1] Obukhov P S, Ivliev E A and Ivliev V A 2019 Identification of alphanumeric information from baggage tag based on neural network a baggage Proc. Int. Conf. on Dynamics of Technical Systems (Ros-tov-on-Don) pp 65-68 (In Russian)
[2] Tesseract wiki, https://github.com/tesseract-ocr/tesseract/wiki, last accessed 2020/08/18.
[3] Ren S, He K, Girshick R and Sun J 2015 Faster r-cnn: Towards real-time object detection with region proposal networks Advances in neural information processing systems vol 39 pp 91–99
[4] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z 2015 Rethinking the inception architecture for computer vision. Preprint gr-qc/1512.00567
[5] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C 2016 SSD: Single shot multibox detector Proc. European. Conf. on Computer Vision vol 1 pp 21–37
[6] Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M and Adam H 2017 Mobilenets: Efficient convolutional neural networks for mobile vision applications Preprint gr-qc/1704.04861

[7] Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y and Guadarrama S 2017 Speed/accuracy trade-offs for modern convolutional object detectors Preprint gr-qc/1611.10012v3

[8] Lowe D G 2004 Distinctive Image Features from Scale-Invariant Keypoints International Journal of Computer Vision vol 60 pp 91-110

[9] Sivic J, Zisserman A, Kingdom U 2003 Video Google: A Text Retrieval Approach to Object Matching in Videos Ninth IEEE International Conference on Computer Vision (ICCV’03), (Nice, France)

[10] Tzutalin. 2015. LabelImg (version 1.8.3). Gitcode.

[11] Krittachai Boonsivanon; Anupap Meesomboon 2016 IKDSIFT: An Improved Keypoint Detection Algorithm Based-on SIFT Approach for Non-uniform Illumination Procedia Computer Science vol 86 pp 69-272

[12] Zhang Y, Lu B and Su L A 2012 Barcode Recognition Based on Multiform Algorithms 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, (Nanchang, Jiangxi) pp 17-20

[13] Hassner T, Mayzels V, and Zelnik-Manor L 2012 On sifts and their scales IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) on Intelligent Human-Machine Systems and Cybernetics, (Nanchang, Jiangxi) pp 17-20

[14] C. Liu, J.Yuen, and A.Torralba. 2011 SIFT Flow: Dense correspondence across Scenes and Its Application IEEE Transactions on Pattern Analysis and Machine Intelligence vol 33 pp 978-994