Normalization of data for training and analysis by the MaskRCNN model using the k-means method for a smart refrigerator's computer vision

M G Dorrer¹,³ and A E Alekhina²

¹ Reshetnev Siberian State University of Science and Technology, Krasnoyarsky Rabochy Av. 31, Krasnoyarsk, 660037, Russian Federation
² Solution Factory, Krasnoy Armii, 10c3, Krasnoyarsk, 660001, Russian Federation
³ E-mail: mdorrer@mail.ru

Abstract. The authors propose to use the k-means method for semantic image segmentation in the artificial vision of a smart refrigerator. The authors have developed a new two-tier architecture for the semantic segmentation system. In the proposed architecture, various sets of image contrast optimization settings are used to classify pixels belonging to fragments of the studied image. Experiments have shown that the proposed method has decisive advantages over existing works. The k-means method was used to cluster pixels directly into semantic groups. The clustered data is used to solve the semantic segmentation problem. For the correct selection of the number of centroids of the k-means method, the mean shift method was used. Extensive experiments with the traditional k-means algorithm and mean shift method have shown the advantage of the proposed approach in accuracy and performance.

1. Introduction
Nowadays, the use of artificial intelligence (AI) in retail is an urgent problem. The management of retail chains realizes that their companies cannot be successful in the 6th technological order without the use of artificial intelligence systems. Analysis by McKinsey showed that the retail industry is one of the leaders in AI use [1]. The high level of demand for AI solutions by enterprises in retail is also confirmed by the data of a survey conducted in 2018 by the IBM Institute for Business Value (IBV) in conjunction with Oxford Economics [2]. A study by Juniper Research shows that by 2023, 325 thousand retail industry companies are expected to use AI technologies [3]. The same trends are observed in the vending device niche. According to a US Grand View Research report, the market for vending machines will reach $11.84 billion by 2025 [4]. According to Berg Insight analysts, the number of connected (IoT / M2M) smart vending machines in the world will reach 5.4 million units by 2022 [5].

One of the vending machines’ essential functions is identifying the product and the buyer’s actions. In the ”smart refrigerator” considered in this article, the identification of the types of goods is used to issue an invoice to the client and reflect transactions in the goods accounting system [6, 7].

At the current stage of development of the smart refrigerator project, product identification is solved by typed labels marked with easily recognizable figures - circles, triangles, rectangles, etc. However, even with this simplification, the task of identifying classified objects in a group of similar ones is urgent. With this approach, the precise identification of unique areas of labels is critical to reliable
product detection. One technique for increasing object recognition accuracy is image preprocessing, particularly clustering, to increase contrast and exclude minor details [8].

The k-means method is in demand in image processing in various fields - analysis of the results of remote scanning of the earth [9, 10] in medicine [11], etc.

However, its use in processing a video stream in the smart refrigerators has led to specific problems. In particular, difficulties arose with the selection of the cardinality of the set of clusters. At particular values of these indicators, the clustered image was distorted by artifacts that complicate image identification and sometimes require manual intervention in the system setup and the recognition process.

Thus, this article is devoted to solving the problem of tuning the k-means algorithm for practical use in the smart refrigerator's computer vision.

2. Methods and experimental materials

2.1. Methods

2.1.1. The k-means method. The k-means algorithm is one of the machine learning algorithms that solve the clustering problem. This algorithm is a non-hierarchical, iterative clustering method; it has gained tremendous popularity due to its simplicity, clarity of implementation, and reasonably high work quality. It was invented in the 1950s by the mathematician Hugo Steinhaus and almost simultaneously by Stuart Lloyd. It gained particular popularity after the publication of McQueen's work in 1967 [12].

The main idea of the k-means algorithm is that the data is randomly divided into clusters, after which the centers of the clusters obtained in the previous step are iteratively recalculated, then the vectors are divided into groups again by which of the new centers is closer according to the chosen metric.

Researchers often use k-means to pre-partition a large dataset, followed by a more powerful cluster analysis of subclusters. The k-means method can also preliminary estimate the number of clusters and check for unaccounted-for data and relationships in the sets [13].

2.1.2. OpenCV 4 Computer Vision Library. OpenCV is an open source library of computer vision, image processing and general purpose numerical algorithms [14]. Originally developed for the C / C ++ programming language but also ported to programming languages such as Python.

In this library, the k-means algorithm is implemented in the cv.kmeans function.

2.1.3. MaskRCNN neural network. The Mask-R CNN concept is outlined in the works [15].

![Figure 1. An example of the architecture of the MaskRCNN model.](image)
Mean average precision and Intersection over Union metrics. To measure object detection accuracy, the mAP ranking quality metric (mean average precision) is used. This is an integral parameter: the average AP (average precision) \[ AP = \frac{1}{11} \sum_{r \in (0.01-1)} P(r) \] for each class from the training sample.

The assessment's main metrics are precision and recall, which are based on an adjacency matrix (table 1).

**Table 1.** Adjacency matrix as an example of parsing mAP.

| Class A predicted: Yes | Class A predicted: No |
|-----------------------|-----------------------|
| Class A: Yes          | True positive         |
| Class A: No           | False positive        |
|                       | False positive        |
|                       | True positive         |

Thus, based on the adjacency matrix, it is possible to calculate the accuracy, which shows how accurate the predictions are, that is, the percentage of correctly predicted classes (2) and completeness, which is the proportion of objects belonging to the class found by the network relative to all objects of this class in the sample (3):

\[ Precision = \frac{TP}{TP + FP} \] (2)
\[ Recall = \frac{TP}{TP + FN} \] (3)

Another indicator is IoU (Intersection over Union), which allows you to measure the coverage of two boundaries: predicted by the classifier and the real one that frames the object in the image (4):

\[ IoU = \frac{overlap}{union} \] (4)

**2.2. Materials**

2.2.1. Refrigerator test area. The kiosk's test stand is assembled based on a brix computer; the neural module Neural Computer Stick 2 is used for computing neural networks [17]. The stand has five shelves with side white matte lighting and a 5 MP camera with a fisheye lens in the middle of the frame.

Figure 2 shows one of the shelves of the test stand of the kiosk used in the work.

![Figure 2](image-url)
Figure 3 shows a view of the location of the camera on the shelf of the kiosk test bench.

![Camera on shelf](image)

**Figure 3.** An example of installing a camera on a shelf of a kiosk test bench.

2.2.2. *Training dataset.* During testing, the original dataset was created without any preprocessing. Based on its basis, two additional datasets were created using the k-means method parameters with parameter values 16 and 32. All datasets were completely independent of each other.

3. **Results**

3.1. *Model training results*

The neural network was trained on the pre-trained scales of the Resnet-101 architecture on an Nvidia GeForce 1060 video card, and each model was trained for 12 hours.

The model trained on original images has an error of 0.21 (see figure 4)

![Error dynamics](image)

**Figure 4.** Dynamics of the model's error in the learning process.

It was impossible to complete the training process on training datasets with the parameters $k = 16$ and $k = 32$. 
3.2. mAP accuracy metric
Table 2 shows the accuracy results of the trained MaskRCNN model on the original images.

| Category | Average Precision Category | BoxClassifierLoss | MAP |
|----------|---------------------------|-------------------|-----|
| circle   | 0.99                      |                   |     |
| icon     | 1                         |                   |     |
| square   | 0.99                      | 0.007             | 0.32| 0.99 |
| star     | 0.99                      | 0.002             |     |
| triangle | 0.99                      |                   |     |

Based on the results of calculating the mAP metric, it can be seen that the accuracy of the model trained on the original images on the test sample shows high results, which allows it to be used for testing in a production system.

3.3. Normalization of the training sample
In the training dataset, stickers of different image quality were used, as well as in motion.

When applying normalization with parameter 16, 32, the following results were obtained. As can be seen, at k = 16, the sticker's unique areas were distorted to an inadequate state. The unique mark is half distorted at k = 32, but in this sample, you can recognize the pattern's unique features.

Figure 5 shows an example of changing the frame of an object in motion for different parameters of k-means.

![Figure 5. Applying different k-means parameter values (in dynamic).](image)

Figure 6 shows an example of changing an object's frame in statics with different parameters to the average neighbors.
Figure 6. Applying different k-means parameter values (in static).

As you can see, unique regions are visible everywhere, but this is most clearly manifested at k = 16. Arbitrary assignment of cluster centers in the k-means method led to the appearance of anomalies in normalized images. Figure 7 shows the result of applying the method with the parameter k = 16.

Figure 7. Result of the k-means method on one frame.

When using this method as image normalization, it is unprofitable for replacing the old training sample since it is necessary to look at it for defects so that the quality of the trained models does not decrease.

3.4. Testing the MaskRCNN neural network models on images

The original model's operation on unnormalized images, normalized by the parameters k 16 and 32, is shown in figure 8.

Figure 8. The results of the original model on different images.
Figure 9 illustrates the accuracy of image extraction and classification when working with an unnormalized image and an image clustered using the k-means method with parameters 16 and 32.

![Figure 9](image_url)

**Figure 9.** Classification accuracy of the original model.

As you can see, for k=32, the probability value is higher, and the boundaries of the mask are more precise than the original and k = 16. At the moment of incorrect operation at parameter k = 16, the actual network did not detect anything.

The k = 16 model did not show any results. Problems arose in the training sample k = 16.

During testing, the reason for the inadequate performance of models trained on normalized images was found out. In this case, it is necessary to review the received training samples and retrain the models.

To check the model's results' robustness, 300 frames were taken from a shelf on which nothing was happening.

Table 3 shows the results of the video stream analysis on different parameters to the average.

**Table 3.** The results of the model's operation on different k-means parameters.

| k-mean parameter | Object | Skip | False detection | False positive | Note |
|------------------|--------|------|-----------------|----------------|------|
| No               | triangle | 3    | 0               | 0              | * Defined as a triangle class, probability percentages 85 and 97 |
|                  | square  | 0    | 2*              | 0              |      |
4. Discussion

When applying k-means on the original images, the masks are fuzzy, increasing the probability percentage by hundredths of a percent. There is blinking; the boundaries of localization of objects are distorted when objects are static on the shelf. According to the test results, the best work was given by the original model - the analyzed image with k = 64, k 84.

It was decided to abandon the parameter k = 16, since anomalies in the form of blots occurred.

5. Conclusion

Suppose we use normalization by using k-means neighbors with a small parameter (16, 32). In that case, it is necessary to view the obtained normalized images for the fact of anomalies of areas since this effect can affect the quality of network training. This normalization method can solve the problem of augmentation of the original dataset with image quality degradation.

References

[1] McKinsey Artificial intelligence: The time to act is now https://www.mckinsey.com/industries/advanced-electronics/our-insights/artificial-intelligence-the-time-to-act-is-now

[2] The coming AI revolution in retail and consumer products, IBM Institute for Business Value (IBV) & Oxford Economics. https://www.ibm.com/downloads/cas/NDE0G4LA

[3] AI in Retail: Segment Analysis, Vendor Positioning & Market Forecasts 2019-2023, Juniper Research. https://www.chainstoreage.com/technology/study-retailer-spending-on-ai-will-reach/

[4] Vending Machine Market Size, Grand View Research. https://www.grandviewresearch.com/industry-analysis/global-vending-machine-market

[5] Connected Vending Machines, Berg Insight. http://www.berginsight.com/ReportPDF/ProductSheet/bi-vending3-ps.pdf

[6] Dorrer M G and Tolmacheva A E 2020 Application of the Canny’s operation and the template matching method to improve the training set in computer vision training’s issue system J. Phys. Conf. Ser. 1679 p 032089 doi: 10.1088/1742-6596/1679/3/032089

[7] Dorrer M G and Tolmacheva A E 2020 Comparison of the YOLOv3 and Mask R-CNN architectures’ efficiency in the smart refrigerator’s computer vision J. Phys. Conf. Ser. 1679 042022 doi: 10.1088/1742-6596/1679/4/042022

[8] Russ J C and Russ J C 2002 The Image Processing Handbook CRC Press

[9] Filho A G S et al. 2003 Hyperspectral images clustering on reconfigurable hardware using the k-
means algorithm 16th Symposium on Integrated Circuits and Systems Design SBCCI pp 99–104 doi: 10.1109/SBCCI.2003.1232813

[10] Kumar G, Sarth P P, Ranjan P and Kumar S 2016 Satellite image clustering and optimization using K-means and PSO IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) pp 1-4 doi: 10.1109/ICPEICES.2016.7853627

[11] Anushya D A 2020 Herbal Leaves Image Clustering via K-Means Int. J. Res. Appl. Sci. Eng. Technol. 8(1) pp 282-5 doi: 10.22214/ijraset.2020.1051

[12] MacQueen J 1967 Some methods for classification and analysis of multivariate observations,” Proceedings fifth Berkeley Symp. Math. Stat. Probab. 1(14) Retrieved from: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf

[13] Prasad R K, Sarmah R and Chakraborty S 2019 Incremental k-Means Method pp 38-46

[14] OpenCV Retrieved from: https://opencv.org/

[15] Van Gansbeke W, Vandenhende S, Georgoulis S and Van Gool L 2021 Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals Retrieved from: http://arxiv.org/abs/2102.06191

[16] Klamanos I A 2009 Manning Christopher, Prabhakar Raghavan, Hinrich Schütze: Introduction to information retrieval Inf. Retr. Boston. 12(5) pp 609-12 doi: 10.1007/s10791-009-9096-x

[17] Tolmacheva A E, Ogurtsov D A and Dorrer M G 2020 Justification for choosing a single-board hardware computing platform for a neural network performing image processing IOP Conference Series: Materials Science and Engineering 734(1)