Extraction of Human Body Parts in Image Using Convolutional Neural Network and Attention Model

Viktoria Sorokina
Belarusian State University
Minsk, Belarus
viktoria.sorokina.96@gmail.com

Sergey Ablameyko
Belarusian State University
United Institute of Informatics Problems
of NAS of Belarus
Minsk, Belarus
ablameyko@bsu.by

Abstract. **In computer vision, human body parts extraction is a challenging task for many applications. In this work, we propose the algorithm to extract human body parts in images using the OpenPose system and attention model. The novelty of the proposed work is that algorithm is based on a convolutional neural network that uses a nonparametric representation to associate body parts with people in an image in combination with a new attention model that learns to focus on specific regions of different input features. The algorithm is a part of Smart Cropping system developed by us which aim is to cut the images and prepare e-commerce catalog.**

Keywords: **human body parts extraction, attention model, convolutional neural network, smart cropping**

I. INTRODUCTION

Task of human body parts extraction is very important for many applications, particularly in e-commerce. It is well known that deep learning can be used to identify parts of a person's body [1]. Extraction of human body parts is the foundation of another task of computer vision – pose estimation, which can also be considered as the problem of determining the position and orientation of the camera relative to a given person or object.

Solving the problem of a human body part extraction, an object or person (or several people) could be tracked in real world at an incredibly detailed level. This powerful capability opens up a wide range of possible applications.

In this article, we will consider the application of the human body parts extraction in e-commerce tasks, namely, when creating an e-commerce catalog.

Creating a catalog for an e-commerce store includes preparing images and content for them [2]. When preparing images of clothes, a full-length photograph of a person presenting several items of clothing at the same time is usually used. Such an image is cut into pieces in accordance with certain rules. For example, for a skirt it is necessary to show the part from the waist to the feet, for a shirt - from the crown of the head to the hips. Currently, the slicing process is done manually. To automate the process, the Smart Cropping system has been developed, which allows, by solving the problem of human body parts detection, to cut images.

In this paper, we propose an algorithm for human body parts extraction using a neural network and attention model. Having received the key points of the human body, the positional relationship between them is calculated, after which it is used to crop the original image and create a set of images representing the goods. The algorithm is capable of preparing images of shoulder clothing (clothing resting on the upper supporting surface of the body, bounded from above by the articulation lines of the body with the neck and upper limbs, and from below by a line passing through the protruding points of the shoulder blades and chest), waist clothing (clothes resting on the lower supporting surface of the body, bounded at the top by the waist line, and at the bottom by the hip line), hats and shoes. The average computation time for one frame is 1.5 seconds.

Our main contributions are

- developing the new system of image cutting in ecommerce sphere based on the deep neural network modified by the techniques that help a “model-intraining” notice important things more effectively;
- extend the existing OpenPose system using Attention model that helped VGG architecture used in OpenPose to learn and detect more image features.

We achieve state-of-the-art results on standard benchmarks including the COCO dataset.

II. RELATED WORK

The classical approach to the problem of a human body parts extraction and pose estimation, presented in [3], includes representation of the object as a set of "parts" located in a deformable configuration (not rigid). Most of the newer pose estimation systems use

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

convolutional neural networks as the main building block, largely replacing hand-crafted functions and graphical models; this strategy has significantly improved standard approaches.

DeepPose [4] is the first deep convolutional neural network architecture applied to the problem of human pose estimation. It achieved the performance of advanced algorithms and outperformed existing models. In this approach, pose estimation is formulated as a convolutional network regression problem to determine the joints of the human body. The work also uses a cascade of such regressors to refine and obtain more accurate estimates of the pose. However, the disadvantage of the model is the complexity of training due to the specifics of regression, which weakens generalization and, therefore, does not work well in certain regions.

Newer techniques transform the pose estimation problem into a heatmap estimation problem, where each heatmap indicates the reliability of the location of the n-th key point of the human body. The work [5] is based on this approach.

The work [5] is based on the architecture that uses a convolutional neural network ConvNet [6] and a refinement model. In the method, heatmaps are created by parallel running an image rendered at different resolutions to capture objects at different scales at the same time. The disadvantage of this approach is the lack of structural modeling.

In this article, the OpenPose architecture [7] that is divided into feature selection block modified by the attention model [8] and maps generation block is used to identify parts of the human body.

OpenPose [7] is a deep feed forward neural network. This method can effectively detect 2D positions of human body parts in real-time RGB images. It does this by detecting and associating body parts using part similarity fields (PAFs) and confidence maps. A confidence map is a probability density function for a new image that assigns a probability to each pixel of a new image, which is the probability of a pixel belonging to a body part in an object in the previous image. The detection of body parts occurs in a sequential manner, performing bottom-up prediction using spatial context.

Attention Model (AM), first introduced in 2015 for Machine Translation [8] has now become a predominant concept in neural network literature. Attention has become enormously popular within the Artificial Intelligence (AI) community as an essential component of neural architectures for a remarkably large number of applications in Computer Vision [9].

The main purpose of the attention model is to use attention maps. An attention map is a scalar matrix representing the relative importance of activation layers at different 2D spatial positions with respect to the target. Attention model uses maps to define and use effective spatial support of visual information used by the convolutional network in decision making.

The constructed in this work model makes it possible not only to structure parts of the human body due to the fields of similarity of parts, but also to highlight parts of the human body in more detail due to stimuli that enhance significant and suppress insignificant objects in the image. This result is achieved due to the construction of a two-dimensional matrix of estimates for each heat map.

III. METHOD

We developed the method allowing to crop the image automatically without usage of the human resources. It calls Smart Cropping and includesthe following components:

- Feature selection module;
- Vector and heatmap building module;

Position relations calculation module; Cutting according to the specified rules module.

The architecture of the Smart Cropping method is shown in Fig. 1.

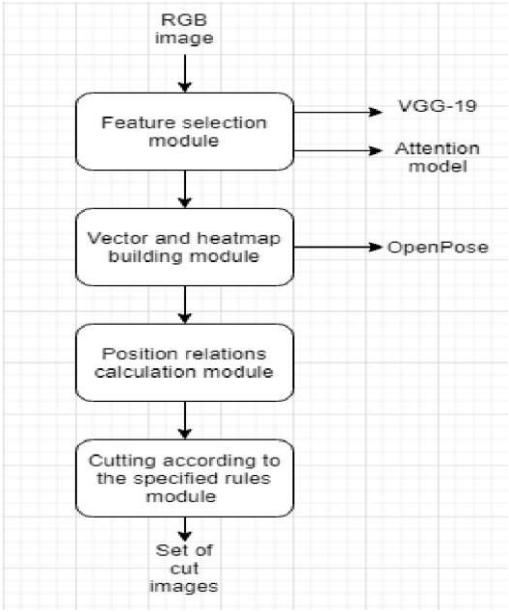

Fig. 1. Smart Cropping architecture

Due to theme of this article is extraction of human body parts in image using convolutional neural network and attention model, special attention will be for Feature selection module, other modules will be described briefly.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

A. *Feature Selection Module*

Feature selection module is the first step of the Smart Cropping method and the main focus of this article. In this module the image is processed with the help of the deep convolutional neural network VGG-19 that is a part of the OpenPose architecture. The novelty of the algorithm proposed in this article lies in the modification of the network VGG-19 by the attention model – the learned attention maps neatly highlight the regions of interest while suppressing background clutter.

VGG-19 [10] is a type of the VGG (Visual Geometry Group) model, which consists of 19 layers (16 convolutional layers, 3 fully connected layers, 5 MaxPool layers and 1 SoftMax layer). The architecture is shown in Fig. 2. The main idea behind VGG is to show that classification/localization can be improved by increasing the convolutional block and using a 3×3 convolution kernel to help highlight the features of the image.

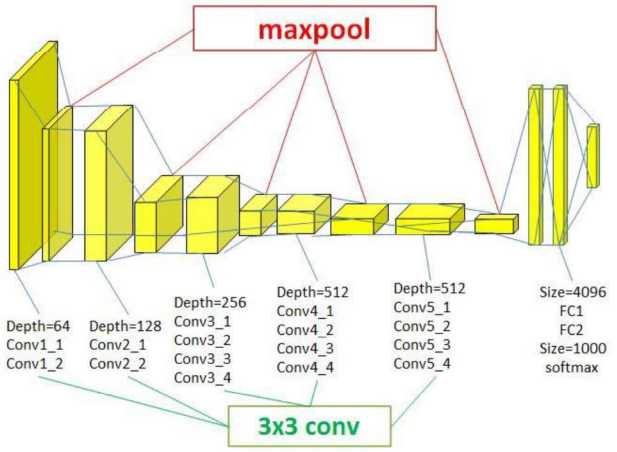

Fig. 2. VGG-19 architecture

One of the ideas of this work is to enforce the existing architecture by combining it with the attention model. The attention model engine learns during network training and should help the network focus on key image elements.

We developed the algorithm which is based on the hypothesis that there is an advantage in identifying significant areas of the image and enhancing their influence, while suppressing irrelevant and potentially misleading information in other areas. In particular, it is expected that providing more targeted and economical use of image information should help in generalizing changes in data distribution, as happens, for example, when training on one set and testing on another. In the standard convolutional network architecture, the global image descriptor g is obtained from the input image and traversed through the fully connected layer to obtain the prediction probabilities. The attention model expresses g through the mapping of input data into a multidimensional space in which observable visual concepts are presented in different dimensions to make classes linearly separable.

We included 2 key changes in VGG-19 architecture and the algorithm is the following (Fig. 3):

- after layers 7, 10, and 13 (highlighted in yellow in Fig. 3), attention estimators are inserted, on the basis of which a binary mask is calculated, where 0 is irrelevant information for the desired object, and 1 is important. The mask, represented by the matrix, is then multiplied by the original result of the layer for which it was calculated, for example, 7, thus overestimating attention;
- the last fully connected layer was replaced with a fully connected layer, the input of which is the results of 3 attention estimators.

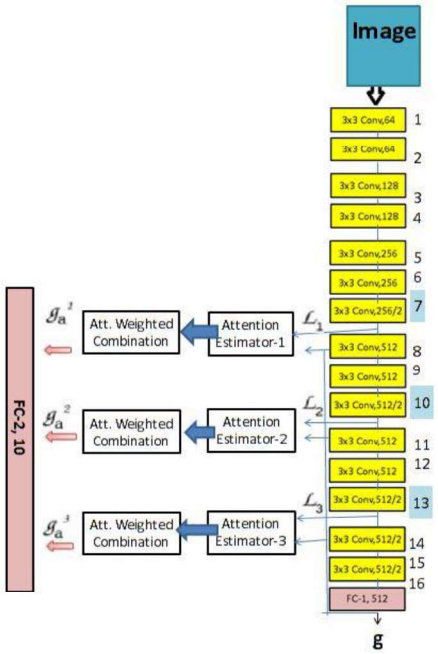

Fig. 3. Attention model architecture

B. *Vector and heatmap building module*

The second module in the Smart Cropping method is vector and heat maps building. Inspired by the work [7], we use an OpenPose system to produce the vector and heat maps to detect the position of human parts on the image. Since our work is an extension of their model, we will only present a very brief overview of the architecture.

The OpenPose input is an RGB image that forwards through the VGG-19. There are two branches at each stage: one for the heatmap detection and the other is for vector map detection. Obtaining heatmap and vector map, one could determine all the key points in the image. The OpenPose architecture is shown in Fig. 4.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

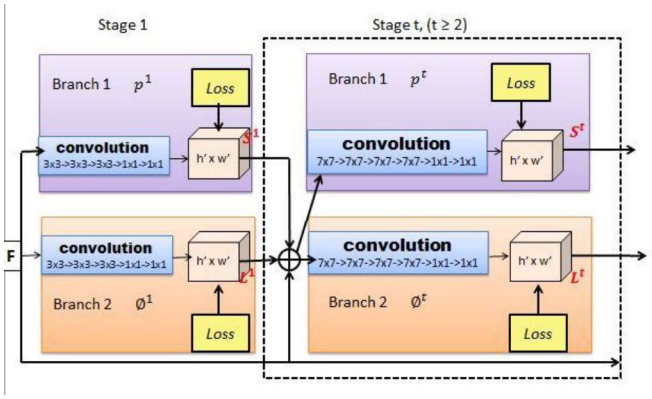

Fig. 4. OpenPose architecture

C. *Position relations calculating module*

To implement a cropping of the original image and create a set of images representing goods, it is needed to get the coordinates of the key points of the human body, and then calculate the positional relationships between them through the coordinates of each key point. The coordinates of three points should be known in order to calculate the angle formed by the three points. Then using the range of values of these angles the pose of the person is etimated and correct cropping is made. The formulas for calculating these angles are shown below.

Let 3 points $A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$ are known. The corresponding vectors are

$$\overrightarrow{AB}: (x_2 - x_1, y_2 - y_1),$$
$$\overrightarrow{AC}: (x_3 - x_1, y_3 - y_1),$$
$$\overrightarrow{BC}: (x_3 - x_2, y_3 - y_2).$$

Then

$$|AB| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$
$$|AC| = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2},$$
$$cos\angle A = \frac{(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1)}{|AB||AC|}.$$

System could detect 23 key points of the human body (e.g. right elbow, left hip, etc.).

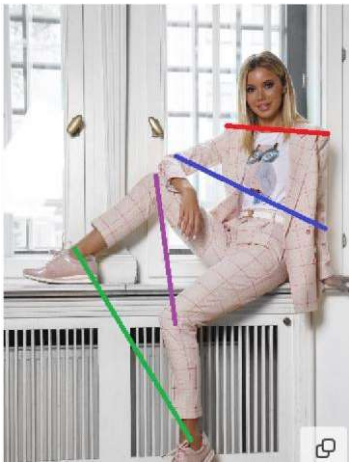

Fig. 5. Example of the staight lines construction

23 key points construct the straight lines that form the person's pose, for example, left and right shoulders, left and right hips, etc. An example of building straight lines is shown in Fig. 5.

IV. Experiments and Results

In e-commerce, a product catalog is an illustrated list of goods or services. The catalog is compiled for the needs of customers, buyers or other interested parties. The hierarchical structure of the catalog consists of categories and subcategories, which contain the actual information about the goods. An electronic catalog is a type of the product catalog where all the information is presented in electronic form.

Such catalogs are the most important, and often they are the only communication channel between the manufacturer or supplier of products or goods and the buyer. The main goal of the electronic catalog is to present information in such a way that the buyer has the ability to effectively search for the necessary information without any difficulties with its understanding and use.

During the creation of an electronic catalog for clothes, a person is usually photographed in full-length, then unnecessary clothes are cut off: for example, when generating a page for a jacket, tit should be cut from the head / neck of the model to the hips. This cropping is done manually and takes a long time.

The network for determining key points was trained on an NVIDIA T4 GPU using VGG-19 modified by the attention model for feature extraction, batch_size = 6, the number of iterations was 800,000.

The resulting accuracy is 86% for the dataset including images of e-commerce products.

Fig. 6. Example of the image cropping

Our algorithm is capable of cropping shoulder clothes (from eyes / nose / shoulders to wrists / hips), waist clothes (from wrists / elbows to toes / knees / ankles), hats (from top of image to shoulders), and shoes (from knees / ankles to toes / bottom of the image), as well as their combinations.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The time needed to prepare one catalog containing 10 products is 5 minutes.

Example of the image cropping is presented in the Fig. 6.

V. DISCUSSION AND CONCLUSION

In the course of the research, an algorithm based on the OpenPose architecture using VGG-19 modified by the attention model was developed. The algorithm showed improvement in the accuracy by 8%, and it is capable of recognizing 23 key points of the human body.

To identify 23 key points of the human body the COCO dataset [11] was used. However, for e-commerce tasks, it is also necessary to define points such as the chest, crown of the head, abdomen, which are not represented in COCO. Thus, the system can be improved by training on an extended dataset.

The resulting network became the foundation of the Smart Cropping system, which allows cropping images based on positional relationships between key points of the human body to create a catalog of e-commerce products. This allows preparing images that form the product catalog for the classes of shoulder, waist and outerwear clothes, as well as shoes and hats.

The accuracy of the trained model is 86% for the dataset represented by images of e-commerce products, due to the specifics of the area. Accuracy can be improved by introducing a block of generative adversarial networks that can predict the presence of a key point that is not explicitly present in the image (for example, a floor-length dress that covers the knees).

The developed algorithm can be improved by expanding the recognized key points, and also used to crop images of other product categories.

ACKNOWLEDGMENT

This research was partly supported under the project BRFFI F20KIGT-006.

REFERENCES

[1] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Computer Vision and Image Understanding, 2020, vol. 192.

[2] eCommerce Product Image Guide [Electronic resource]. Mode of access: https://www.threekit.com/blog/ecommerce-product-imageguide-2020. – Date of access: 25.03.2021.

[3] Y. Yang, D. Ramanan, Articulated Human Detection with Flexible Mixtures of Parts, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, vol. 35, no. 12, pp. 2878–2890.

[4] DeepPose: Human Pose Estimation via Deep Neural Networks / A. Toshev and C. Szegedy // IEEE Conference on Computer Vision and Pattern Recognition, June 24-27, 2014, Columbus, OH, USA, 2014, pp. 1653–1660.

[5] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Join training of a convolutional network and a graphical model for human pose estimation, Advances in Neural Information Processing Systems, 2014, pp. 1799–1807.

[6] J. Tompson [et al.] Efficient object localization using Convolutional Networks, 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015, pp. 648-656.

[7] Zhe Cao [et al.] Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 21-26 July 2017, Honolulu, 2017, pp. 1302–1310.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations.

[9] Feng Wang and David MJ Tax. 2016. Survey on the attention based RNN model and its applications in computer vision. arXiv preprint arXiv:1601.06823 (2016).

[10] Karen Simonyan and Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations, San Diego, May 7-9, 2015, San Diego, 2015, pp. 1137–1149.

[11] COCO dataset // COCO 2018 Keypoint Detection Task [Electronic resource]. – Mode of access: http://cocodataset. org/#overview – Date of access: 05.04.2019.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.