# Recognition of human body parts in real-time camera images*

Kamila Raś
*Poznan University of Technology*
Poznań, POLAND
Email: kamila.ras@student.put.poznan.pl

Aleksandra Świetlicka
*Poznan University of Technology*
*Institute of Automatic Control and Robotics*
Poznań, POLAND
Email: aleksandra.swietlicka@put.poznan.pl

***Abstract* - This article presents a system for recognizing human body parts in real-time camera images. The solution uses an edge-processing camera, on which image acquisition and preliminary person detection are performed, whereas detailed body analysis is carried out on the host computer. The system is based on a two-stage approach: first, a person is detected and tracked in the frame, and then human parsing segmentation is performed for a selected region of interest. Person detection was implemented using the MobileNet-SSD model run with the DepthAI platform, whereas body-part segmentation uses the Self-Correction Human Parsing method in the Pascal Person-Part variant. The result is presented as a mask of anatomical classes overlaid on the image and as rectangles corresponding to selected body fragments. The verification covers operation of the online application, tracking stability of the selected object, readability of the visualization, and host resource load. The presented results show that a hybrid architecture combining on-device detection with ROI-based segmentation makes it possible to obtain an interpretable real-time result, while the limitations mainly concern multi-person scenes, poor lighting, dynamic motion, and division of limbs into left and right sides.**

***Keywords—human parsing, body-part recognition, semantic segmentation, MobileNet-SSD, SCHP, DepthAI, computer vision***

## I. INTRODUCTION

Human body analysis is an important problem in computer vision because it allows a camera image to be transformed into information about a person's position, orientation, motion, and selected body parts. Such solutions can be used in human-computer interaction systems, surveillance, robotics, rehabilitation, activity analysis, and applications operating on edge devices. In real-time applications, it is particularly important that the system not only interpret the image correctly but also operate with the lowest possible latency and limited computational cost. Person detection alone in an image is not sufficient when the goal is body-part recognition. A detector usually returns a rectangular bounding box around the silhouette, but it does not indicate where exactly the head, torso, arms, or legs are located. A more detailed representation is provided by body segmentation, in which image pixels are assigned to specific anatomical classes. In this view, the system output is not only information about the presence of a person, but a map of regions corresponding to selected body parts.

This study uses a two-stage approach. In the first stage, a person is detected in the frame, and their position is stabilized across consecutive frames using tracking based on the IoU metric. In the second stage, body segmentation using a human parsing method is performed for the selected region of interest (ROI). Restricting segmentation to the ROI reduces the number of analyzed pixels and helps maintain application responsiveness because the most computationally expensive stage is not performed on the entire frame. The article is organized as follows. Section II describes the methodology, Section III presents the verification procedure, Section IV discusses the results, and Section V summarizes the conclusions.

## II. METHODOLOGY

Recognition of human body parts in a video image can be treated as a sequence of related computer vision tasks: person detection, object tracking over time, region-of-interest determination, and semantic segmentation of the silhouette. In the literature, one-stage detectors such as Single Shot MultiBox Detector (SSD) are often used in real-time systems because object localization and classification are performed in a single network pass [9]. SSD operates by analyzing image feature maps and simultaneously predicting bounding-box positions and class probabilities for multiple object candidates. As a result, it does not require a separate stage for generating region proposals, which reduces latency and makes it well suited for video streams.

Combining SSD with the lightweight MobileNet architecture reduces the cost of feature extraction, which is particularly important in edge systems [7], [9]. MobileNet is a convolutional neural network designed for devices with limited computational resources. It uses lightweight convolutional operations, which allows it to generate the feature maps required by the detector at a lower cost than classic, large convolutional networks. Human parsing methods, in turn, extend classical semantic segmentation because they are not limited to separating the person from the background; instead, they assign pixels to specific anatomical regions [3], [5], [8]. This article uses the Self-Correction Human Parsing (SCHP) method [8], implemented with the publicly available SCHP extractor [4], which enables a multi-class body-part mask to be obtained for a selected image fragment. SCHP is a human silhouette segmentation method in which each pixel of the analyzed area receives a label of one body part, such as the head, torso, or limbs The self-correction mechanism iteratively improves the quality of the labels used during training, which increases the model's robustness to ambiguous or noisy pixel-level annotations [8].

In related work, human body analysis is commonly addressed using three main groups of methods: object detection, human parsing, and pose estimation. Detection-based approaches, such as SSD, focus on locating the person in the image by predicting a bounding box, which is computationally efficient but does not provide information about individual body parts [9]. Human parsing methods extend semantic segmentation by assigning pixels to anatomical regions and are therefore more suitable when a dense body-part representation is required [3], [5], [8]. Another popular approach is pose estimation, in which the human body is represented as a set of keypoints connected into a skeleton rather than as a pixel-level mask. For example, Cao proposed a real-time multi-person pose estimation method based on Part Affinity Fields, which associates detected body parts with individual people in the scene [2]. In the proposed system, human parsing was selected because the goal is not only to estimate the position of joints, but also to obtain interpretable pixel regions corresponding to selected body parts. The adopted architecture separates processing stages according to their computational cost and timing requirements. Person detection is performed on the edge device, whereas detailed body-part segmentation is performed on the host side only for the selected region of interest. DepthAI simplifies integration by combining image acquisition, model input resizing, inference, and result transfer into one pipeline [10]. The division between edge-device and host-side processing is shown in Fig. 1.

In the first stage, the image frame is passed in parallel to two paths: preview and person detection. The detection path resizes the preview image to the input size required by MobileNet-SSD and passes it to the detection network. A configurable confidence-threshold parameter is used, and only person-class detections are returned to the host as normalized bounding boxes. Confidence values are used at the network filtering stage, whereas the tracker receives only bounding-box coordinates. As shown in Fig. 1, this stage does not yet recognize body parts, but provides information on