DOI 10.1590/s1982-21702021000300017

ORIGINAL ARTICLE

PEDESTRIAN SEGMENTATION FROM COMPLEX BACKGROUND BASED ON PREDEFINED POSE FIELDS AND PROBABILISTIC RELAXATION

Segmentação de pedestres em imagens com fundo variável baseada em análise de pose e relaxamento probabilístico

Caisse Amisse[1,2] - ORCID: 0000-0001-9458-5510

Mario Ernesto Jijón-Palma[1] - ORCID: 0000-0003-4890-2997

Jorge António Silva Centeno[1] - ORCID: 0000-0002-2669-7147

[1] Universidade Federal do Paraná, Programa de Pós-graduação em Ciências Geodésicas, Curitiba - Paraná, Brasil.
E-mail: caamisse@gmail.com; majijpa@hotmail.com; centeno@ufpr.br

[2] Universidade Rovuma, Departamento de Ciências Naturais, Nampula, Moçambique.

Received in 2nd July 2020
Accepted in 15th June 2021

Abstract:

The wide use of cameras enables the availability of a large amount of image frames that can be used for people counting or to monitor crowds or single individuals for security purposes. These applications require both, object detection and tracking. This task has shown to be challenging due to problems such as occlusion, deformation, motion blur, and scale variation. One alternative to perform tracking is based on the comparison of features extracted for the individual objects from the image. For this purpose, it is necessary to identify the object of interest, a human image, from the rest of the scene. This paper introduces a method to perform the separation of human bodies from images with changing backgrounds. The method is based on image segmentation, the analysis of the possible pose, and a final refinement step based on probabilistic relaxation. It is the first work we are aware that probabilistic fields computed from human pose figures are combined with an improvement step of relaxation for pedestrian segmentation. The proposed method is evaluated using different image series and the results show that it can work efficiently, but it is dependent on some parameters to be set according to the image contrast and scale. Tests show accuracies above 71%. The method performs well in other datasets, where it achieves results comparable to state-of-the-art approaches.

Keywords: image processing; background suppression; pedestrian segmentation; probabilistic relaxation.

How to cite this article: AMISSE, C.; JUJÓN-PALMA, M.E.; CENTENO, J.A.S. Pedestrian segmentation from complex background based on predefined pose fields and probabilistic relaxation. *Bulletin of Geodetic Sciences*. 27(3): e2021017, 2021.

This content is licensed under a Creative Commons Attribution 4.0 International License.

1. Introduction

The current capacity to capture images and videos is huge because of the advances in terms of low-cost digital cameras (Portmann et al., 2014; Ma et al., 2016). It is common to monitor people using cameras within a known environment, such as a supermarket, or it is also possible to record a video sequence from a moving platform (Davis et al., 2000), such as an unmanned aerial vehicle (drone). In the first example, the background may be "a priori" known and the detection of a person can be performed comparing the current image to a reference image of the empty scenario (Cai et al.,1995). In the second case, the platform is moving and the background changes along time (Davis et al., 2000). According to Chen et al. (2016), the second case introduces challenges to an autonomous monitoring system, as different problems arise, such as occlusion (caused by another object as lamps or trees), deformation of the object (the person changes his pose as he moves and his appearance also depends on the position of the system camera/drone, motion blur, and scale variation (when the person moves away from the camera, for example).

Although image segmentation has been a topic of research for several years, there is no general solution to the problem, as the methods are developed to satisfy specific needs and under different conditions (Zouagui et al., 2004). Once the remote sensing community started using image segmentation, the methods were based on the analysis of color differences or uniformity, applying region growing or edge detection algorithms (Pavlidis and Liow, 1990). As the spatial resolution of satellite images increased and the availability of other image sources rose up, new methods, such as the fractal net evolution approach (FNEA) were proposed (Baatz and Schäpe, 2000). Since then, more improvements have been introduced, as the use of convolutional neural networks (Kemker et al., 2018; Minaee et al., 2020).

In the context of video surveillance, the goal of segmentation is to segment an object instance, a person, in the video frame. This can be performed in automatic or interactive manner. Automatic methods segment the person in the image without human interaction (Friedman and Russell, 2013; Papazoglou and Ferrari, 2013; Guo et al., 2017). The main advantage is that the task can be performed without any prior knowledge of the target, e.g., initial object masks. Automatic Video Object Segmentation (VOS) can be performed by trajectory clustering (Brox and Malik, 2010; Wang et al., 2018), motion analysis (Jain et al., 2017; Zhuo et al., 2019) object proposal ranking (Vijayanarasimhan and Grauman, 2012; Koh and Kim, 2017), saliency (Wang et al., 2015; Hu et al., 2018) or optical flow (Papazoglou and Ferrari, 2013). Recent trends to address automatic video object segmentation use deep learning methods based on Deep Neural Networks (Tokmakov et al., 2017; Goel et al., 2018). For example, Wang et al. (2019) propose the use of attentive graph neural network for zero-shot video object segmentation, while Zhou et al. (2020) employ two-stream networks to extract appearance and motion features of objects. There are many papers that deal with image segmentation.

The survey of Minaee et al. (2020) enables a good and updated view of the most recent literature in image segmentation and discusses more than a hundred deep learning-based segmentation methods. While fully automated solutions are desired, there is still not a general solution, but a wide range of approaches, as proved by Minaee et al. (2020). Current automatic methods cannot handle multiple object segmentation because they are affected by motion (object and camera) and dynamic background. When the methods do not have a prior knowledge of the target object, they fail identifying various moving objects and have difficulties dealing with cluttered backgrounds. Therefore, there is the trend to use semi-automated methods that require a lower human interaction to separate the human figure from the background in at least one image (Ren and Malik, 2007; Caelles et al., 2017).

In the interactive methods, the user marks the object on the first image, and using features derived from this image, the system tries to find the object in the following frames (Bibby and Reid, 2008; Caelles et al., 2017). The user input mark is performed either by providing a pixel-accurate mask (Caelles et al., 2017), clicks (Maninis et al., 2018), or scribbles (Lin et al., 2016). Numerous interactive algorithms have been developed, many of which yield

very promising results. Background subtraction (Elgammal et al., 2000; Lim, 2017) is the most used interactive VOS, which extracts the foreground silhouettes of objects by computing the difference between a video image frame and the background model. It renders satisfying performance (Bouwmans, 2012; Sobral and Vacavant, 2014), however, it is incapable of handling a dynamic background. Statistical-based methods such us, single Gaussian model (Wren et al., 1997), Gaussian Mixture Models (GMM) (Stauffer and Grimson 1999), and kernel density estimation-based modeling (Elgammal et al., 2000) have achieved promising results on foreground object segmentation for many outdoor environments. Some approaches (Heikkilä et al., 2004) perform VOS by using texture-based features, such as a local binary pattern, which adapts to illumination changes in the background. Others (Yoshinaga et al., 2011; St-Charles et al., 2014) extract features and combine them with background models to enhance the robustness of background changes. Statistical-based methods have been proved robust for VOS if the data meets the distribution assumption (e.g., a Gaussian distribution). Even though an appropriate annotation and proper data sets are employed to generate foreground segmentation. Interactive approaches usually require a large amount of high-quality reference data for the labeling process. More recent VOS approaches that use deep models overcome the need of labeling large amount of reference data either by propagating labels from the first frame (Voigtlaender et al., 2019) or using reinforcement learning (Varga and Lőrincz, 2020) to reduce human effort in interactive video segmentation annotation.

There is also another approach, the so-called weakly-supervised VOS (Yao et al., 2020). In this method the ground truth masks for the first frame are given, so there is one reference to identify which objects must be segmented and how they look. For instance, Liu et al. (2014) introduce a neighbor-based label transfer scheme for weakly supervised video segmentation. Araslanov and Roth (2020) propose weakly supervised semantic segmentation approach, which comprises a single segmentation network trained in one round.

Superpixels have also been successfully applied in image segmentation (Achanta, 2012). In Milan et al. (2015) it is suggested a conditional random field (CRF) model that exploits high-level detector responses and low-level superpixel information to jointly track and segment multiple objects. Classical probabilistic neural network architecture has been used to improve the object segmentation (Doulamis et al., 2003) and perform the classification of segmented objects (Tian et al., 2000). Recently, Chaibou et al. (2020) have suggested to learn superpixels from CNN for foreground segmentation.

In the present paper it is introduced a method to segment a pedestrian from the background in previously selected image regions. It is assumed that the background may vary and is not uniform. The method is based on an initial color segmentation, followed by pose estimation using a pose library and posterior refinement by probabilistic relaxation. Probabilistic relaxation has a long history in remote sensing and computer vision as it has demonstrated to be very useful for numerous applications including scene labelling (Rosenfeld et al.,1976) and label relaxation (Zhu et al., 2019), threshold selection (Rosenfeld and Smith, 1981), image segmentation (Parvin and Bhanu, 1983), template matching (Kittler and Illingworth, 1985), and pattern recognition (Christmas et al., 1995). To the best of our knowledge, no previous studies have been found in the literature, proposing a method based on probabilistic fields computed from possible pose figures and including an improvement step based on a relaxation process for the purpose of pedestrian segmentation.

After this introduction, Section 2 presents the proposed approach, while Sections 3 and 4 presents respectively the experiments and results. Finally, Section 5 summarizes the main conclusions.

2. Method

The input of the workflow is a set of selected Regions of Interest (ROI) where persons (pedestrians in an urban scene) are present, as displayed in Figure 1. It is not the intention to discuss the detection methods because

it was a matter of a previous paper of the same authors. Therefore, in this paper it is only discussed the problem of the segmentation of the pedestrian, given a small image region that contains him but with varying background. The problem in hand is to separate the pedestrian figure from the background, considering that the background is variable. The whole framework is shown in Figure 1. The workflow is divided into the following steps: initial image segmentation; segmentation improvement by clustering; foreground estimation; and refinement by relaxation. These steps are described below.

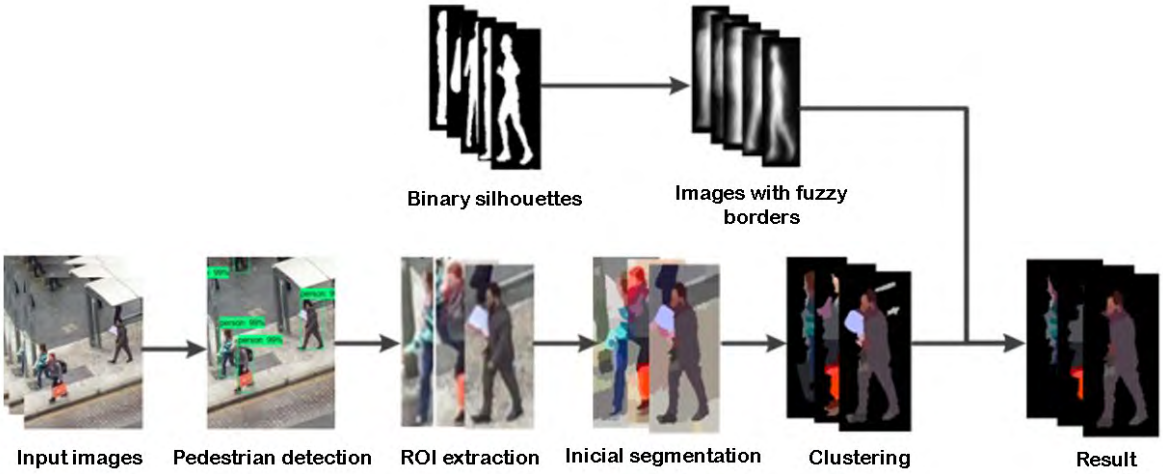

Figure 1: Framework of the proposed method. Given the input images. Areas with pedestrians are detected and extracted. Then, initial segmentation followed by the clustering, improvement based on predefined pose fields and relaxation are performed to get the final segmentation result.

2.1 Initial segmentation

In the segmentation step, the image is divided in small uniform regions (superpixels) that will be used as image element instead of the pixel. The superpixel is defined as a small, but uniform in terms of color, group of pixels (Neubert and Protzel, 2012). The concept has been extensively used in various fields of image analysis because, compared to pixel-based approaches, it reduces the number of image elements and speeds up the succedent steps, like object recognition.

The Simple Linear Iterative Clustering (SLIC) is the alternative proposed by Achanta et al. (2012) to compute superpixels from an image, without requiring much computational power. According to Stutz et al. (2018), SLIC has also advantages as compactness, consistency, robustness, and depth information, and gives superior performance in terms of boundary recall. SLIC begins by choosing cluster seeds regularly spaced along the image, which is equivalent to resampling the image to a lower resolution, taking, initially, the central pixel value. The user needs to choose the size of the regular grid, setting the grid interval (S). The initial number of clusters (K) is provided by Equation 1.

$$\mathrm{S} = \sqrt{\mathrm{N}/\mathrm{K}} \tag{1}$$

here, N is the number of all pixels in the image, and S the regular grid interval between the superpixels.

The center of each region is then used to cluster the neighboring pixels. The clustering process is based on a similarity criterion, given in Equation (2), that measures color similarity in L*a*b space (d_c) and pixel proximity in x, y space (d_s). The latter is normalized by the grid interval (G). The compactness and regularity of the superpixels is controlled with the constant m.

$$D = \sqrt{\left(\frac{d_c}{m}\right)^2 + \left(\frac{d_s}{G}\right)^2} \quad (2)$$

where d_c and d_s are respectively computed according to Equations 3 and 4:

$$d_c = \sqrt{\left(\bar{L}_\iota - \bar{L}_J\right)^2 + \left(\bar{a}_\iota - \bar{a}_j\right)^2 + \left(\bar{b}_\iota - \bar{b}_j\right)^2} \quad (3)$$

$$d_s = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \quad (4)$$

Then, the seed is moved to the next position with the lowest gradient. This decision is taken considering five features that include color and spatial position. Once the seeds are defined, each pixel in the image is assigned to the nearest center.

2.2 Segmentation improvement by clustering

The division of the image in superpixels produces small compact regions that are regularly distributed in the image. Nevertheless, the method produces adjacent regions that may belong to the same image object and therefore have a similar color. To reduce such situations and achieve a better estimate of the image objects, adjacent, similar superpixels were combined to build up a new larger uniform region.

Considering that Ns superpixels were obtained, the neighboring relation between regions was stored in a matrix, the adjacency matrix (A, NsxNs). This is a square, binary matrix whose dimension is equal to the number of segments (Ns x Ns). The neighborhood of two regions is represented by each cell, (1) if two regions are neighbors or (0) if not represents the neighborhood.

A second matrix is also computed to describe the color affinity of two neighboring regions (a and b). First, the mean color values (ma) of each segment are computed. Then the Euclidean Distance (ED) between the mean values of the two neighboring regions (a and b) is computed as:

$$ED(a,b) = \sum_{i=1}^{3} (m_{ai} - m_{bi})^{1/2} \quad (5)$$

where m_{ai} and m_{bi} stands respectively for the mean value of region “a” and “b” in the i-th band.

The Euclidean distance is then truncated to the range between 0 and 255 and a color similarity measure (S) derived, according to Equation (6), which is normalized to the range 0-1.

$$S(a,b) = 1 - \frac{ED(a,b)}{255} \quad (6)$$

Clustering is performed within an iterative process. At each iteration, the best fusion is evaluated, according to the color compatibility of two neighboring segments. The color compatibility is estimated as the Euclidian color distance, weighted by the adjacency between the regions, as displayed in Equation 7. If two regions don’t share a common boundary, they are not considered similar.

$$A(a,b) * S(a,b) > t_1 \quad (7)$$

here, t_1 is the minimum similarity that is chosen by the user. Normally, values between 0.85-9.95 are recommended. The process stops when there are no more possible fusions that satisfy Equation (4).

2.3 Foreground estimation

The result of the segmentation is a set of regions of different colors, but it is still unknown which regions belong to the background and the object (Foreground). The aim now is to classify the segments in these two classes but no previous information about the color of the classes is available. It must also be considered that the background is not uniform and may be composed of more than a surface.

To estimate the probability that a segment belongs to the pedestrian (foreground), the probability of a pixel belonging to the foreground was estimated from a series of available binary images. Several binary human silhouettes in different poses (Figures 2a) were downloaded from https://publicdomainvectors.org/ for this purpose. Given the articulated nature of the human body: people can adopt a wide range of poses (e.g.: walking, standing, sitting, lying, jumping, dancing, running, kneeling, squatting, or crouching) and can be captured in different viewpoints from single or multiple cameras. For the experiments, and to keep a reduced number of poses to speed up the process, five different poses: “walking1”; “stand side”; “front”; “walking2” and “sitting” were considered (Figure 2c). From the available set, 70 silhouettes for each class were selected. Figure 2b displays a small set of silhouettes used as samples of the class “walking”. These selected binary images were resized to 140x70 pixels and added to obtain a new image (as displayed in Figure 2c) that stores the frequency of the foreground in the image.

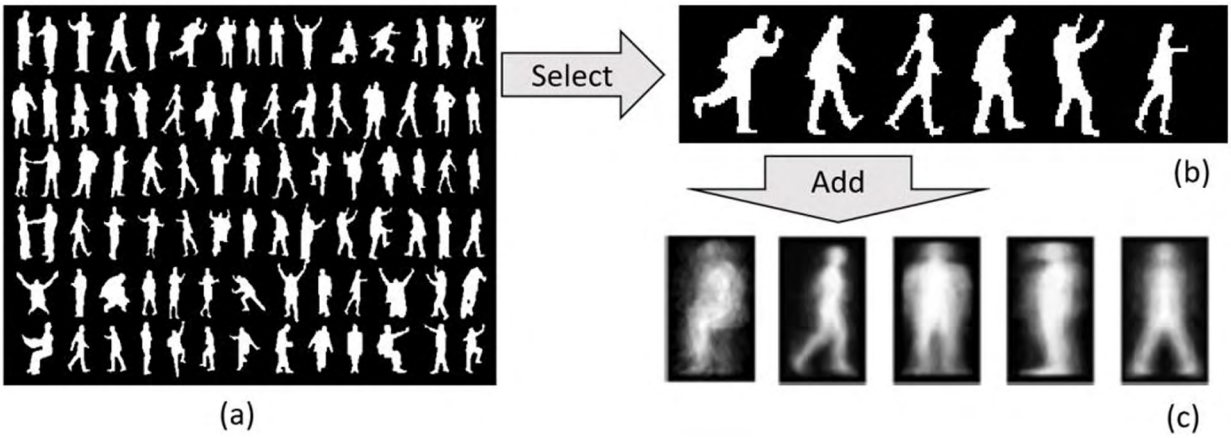

Figure 2: Composition of pose images with fuzzy borders. (a) original image downloaded from publicdomainvectors.org; (b) binary image of selected pose; (c) Five examples of the combination of multiple binary images with different poses: “sitting”; “walking1”; “front”; “stand side”; and “walking2”.

The addition of the binary images (Equation 8) enables composing a fuzzy image to map the density of the active pixels. The resulting image is later normalized to the range 0-1. This procedure was repeated for five different possible poses, as displayed in Figure 2c. This field is the description of the probability of a pixel belonging to the foreground. In this paper it will be referred as density field.

$$M(i,j) = \frac{1}{NB}\sum_{p=1}^{NB} BI_p(i,j) \tag{8}$$

In the following step, it is estimated the probability that a segment belongs to the foreground, based on the density fields. The hypothesis is that if a segment belongs to the foreground, it will cover a region associated to high density values. For this purpose, the segmented image is overlaid to the probability fields of each pose (Figure 2c) and a mean value (probability) is computed for each segment. The mean hangs on the quality of the segmentation and the size of the segments. A fine over-segmentation would produce small segments that would be better defined

in terms of probability but, on the other hand, wrong classified segments would be more difficult to correct.

To access which pose model better fits the image, the mean probability of the segments is computed for each pose and the pose with higher mean is chosen. The selected pose is then used to obtain the initial classification of the image, separating background from foreground regions. The segments with probability above 0.5 are considered part of the foreground. This result is not optimal but is an initial situation for the next step.

2.4 Refinement by relaxation

In the previous step, the mean probability of each segment was computed but, as the probability fields do not include all possible poses, errors are expected, especially when the mean probability lies close to 50%. So, in the next step, the probability values are updated using the information contained in the neighboring segments.

The hypothesis here is that, if a segment that should belong to the foreground has relative low mean value, below 50%, and it is not classified as foreground, its mean can be corrected considering the mean values of the surrounding segments if there is high color compatibility between the regions. For example, if a segment has low value but it is surrounded by high-value segments, and the segment has a similar color as the neighbors, then its neighborhood would contribute to increase its probability. On the other hand, if it is surrounded by lower values, it is logical to decrease the mean. This is performed within an iterative process where the probabilities are updated according to the rule displayed in Equation 9:

$$P_i{}^{t+1}(cl) = \frac{1}{L}\left(P_i{}^{t} * \left[1 + \frac{1}{nv}\sum_{j=1}^{nv}\left\{ R(cl,i,j) * P^{t}{}_{j}(cl) * K(cl,j) \right\} \right] \right) \tag{9}$$

here:

$P_i{}^{t}$: is the mean probability of the ith-segment at iteration t;

R (cl, I, j): is the compatibility in terms of color between segments i and j;

cl: denotes the class: foreground or background;

K: is a binary value that is one when region j belongs to class cl;

L: is a normalizing factor to keep the probabilities in the range between 0 and 1.

In the iterative process, the probability of each segment is updated considering the information provided by the neighbors and the process stops when the classification reaches a steady situation when no segment is reclassified in a different class.

The compatibility function should range between -1 (fully not compatible) and 1 (fully compatible) and it is derived from the Euclidean distance between the color of the segments. The color similarity function (S) proposed in Equation (6), is used to describe the color compatibility (CC) between two segments. First, the similarity is scaled to the range -1 and 1(Equation 10) and then the sigmoid function is applied using Equation (11). The use of the constant c = 5 is necessary to adapt the sigmoid function to the range (-1,1). The use of the sigmoid function is recommended because it has higher slope for values close to zero and lower variation in the extremes, which allows changing the classification of dubious situations, as shown in Figure 3.

$$x(a,b) = 2c * (S(a,b) - 0.5) \tag{10}$$

$$CC(a,b) = \frac{c * 1}{1 + e^{-x(a,b)}} \tag{11}$$

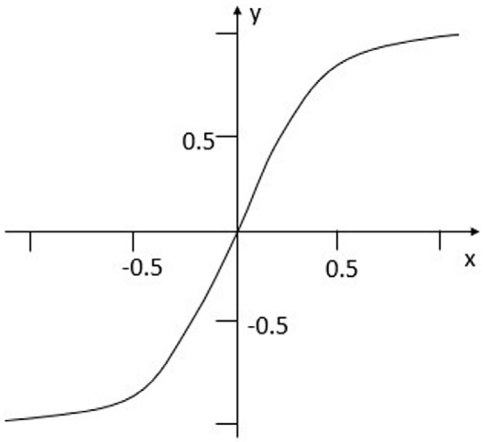

Figure 3: sigmoid function adapted to the range between -1 and 1.

3. Experiments

For evaluation of the proposed method, different images from public Penn-Fudan dataset (Wang et al., 2007) and the Cityscapes dataset (Cordts et al., 2015), were used. An example of the images is displayed in Figure 4a. The background where the pedestrians move is relatively complex as it is not uniform and varies from dark to light and from unsaturated to saturated colors.

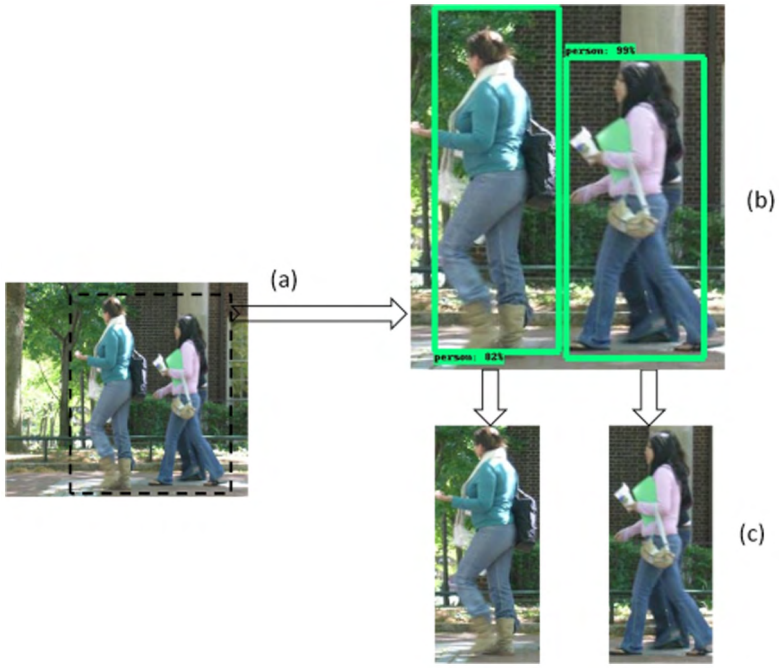

Figure 4: Example of an image from Penn-Fudan dataset used in the experiment.

In a previous work, Amisse et al. (2021) described a method to detect pedestrians in the scene, being, the location of the persons in the image available as a bounding box, as displayed in Figure 4b. The size of the initial image regions varies around 160 rows x 90 columns, depending on the pose of the pedestrian and the distance to the camera. For the initial segmentation, the grid size was chosen considering the width of the legs of a pedestrian in such images. As it can be seen on Figure 4b, a leg comprises about 1/6 of the image width, so, a grid size around 10

pixels is recommended. According to Equation 1, and considering a mean size of 160x90 pixels, this leads to divide the image into 144 regions. We rounded the value to 150. As there are large uniform regions when the pedestrian wears uniform clothes, the number of resulting regions is lower after the segmentation, in mean around 100.

The SLIC segmentation divides the image into uniform regions, but also produces adjacent regions with similar color. Therefore, adjacent regions with similar color need to be grouped to better estimate the human figure in the image. This was done applying the region growing approach, described in Equation 7, with the similarity factor set to 90%.

To evaluate how the proposed method deals with multiple instances in images datasets, experiments was conducted on the Penn-Fudan and Cityscapes datasets which consists of challenging images with appearance and pose variations, occlusion, and background clutters. Therefore, it can be considered that different background configurations were studies.

The quality of the result was measured by comparing the result to a reference image previously obtained by visual analysis. Two binary images were compared: Reference (R) and obtained result (O). The total number correct classified pixels (True Positive TP) were counted, as well as the total of false positives (FP) and false negatives (FN). The total number of pixels labeled as pedestrians (TT) include the correct classified (CC) pixels as well as the false positive TT=CC+FP. These quantities were then used to compute some quality metrics as suggested by Lee et al. (2003): the branching factor (Bf) that reveals the rate of incorrectly labeled pedestrian pixels and the Miss factor (Mf) the rate of missed pedestrian pixels.

$$Bf = \frac{FP}{TP} \tag{12}$$

$$Mf = \frac{FN}{TP} \tag{13}$$

These two quality parameters describe two types of error in the results, omission, and commission errors, but are relatively simple. Therefore, the detection success was also measured using the recall, precision, and the Intersection over Union (IoU) (Minaee et al., 2020).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{17}$$

Recall is the user's accuracy when dealing with the confusion matrix of a classical classification evaluation. It describes the probability that a pixel labeled as pedestrian in the resulting image is correct. Precision, on the other hand, describes the rate of correct pixels within the set of pixels labeled as pedestrian. F1-score describes the segmentation quality. Finally, the precision and recall are balanced by IoU.

4. Results

In this section, we qualitatively and quantitatively evaluate our proposed method on the Cityscapes and Penn-Fudan datasets. Meanwhile, in order to demonstrate the effectiveness of the proposed algorithm, we also compare it with the state-of-the-art methods.

4.1 Experimental results of the proposed method

4.1.1. Qualitative Evaluation

To illustrate the results, some examples using images downloaded from the Cityscapes and Penn-Fudan datasets are displayed in Figure 5. In Figure 5a it is displayed a set of the different steps (rows) of the procedure for six examples (columns). For example, the different steps of one isolated pedestrian from Figure 4c are shown in the first column. The first image is the original RGB input image. Note that the person seems to hold a light green object, but it can be also understood as part of the background. In the second row, the result of the initial segmentation and clustering is displayed. The image has a small number of segments and one can note that the borders of the person are well delineated in the segmentation. The third row is the result of the initial classification. It is noticeable that the classification based on the density fields is good, but some mistakes are visible, like the inclusion of two segments that are part of the background (like street and pole). The result also includes the object that the woman holds in her left hand. This can be or not considered an error, depending on the manual reference segmentation. The fourth row displays the result after the relaxation step. Comparing the third and fourth rows (For example, the fourth and third rows of the first column of Figure 5a) it is visible that errors like the light segments of the ground (lower right corner) were removed.

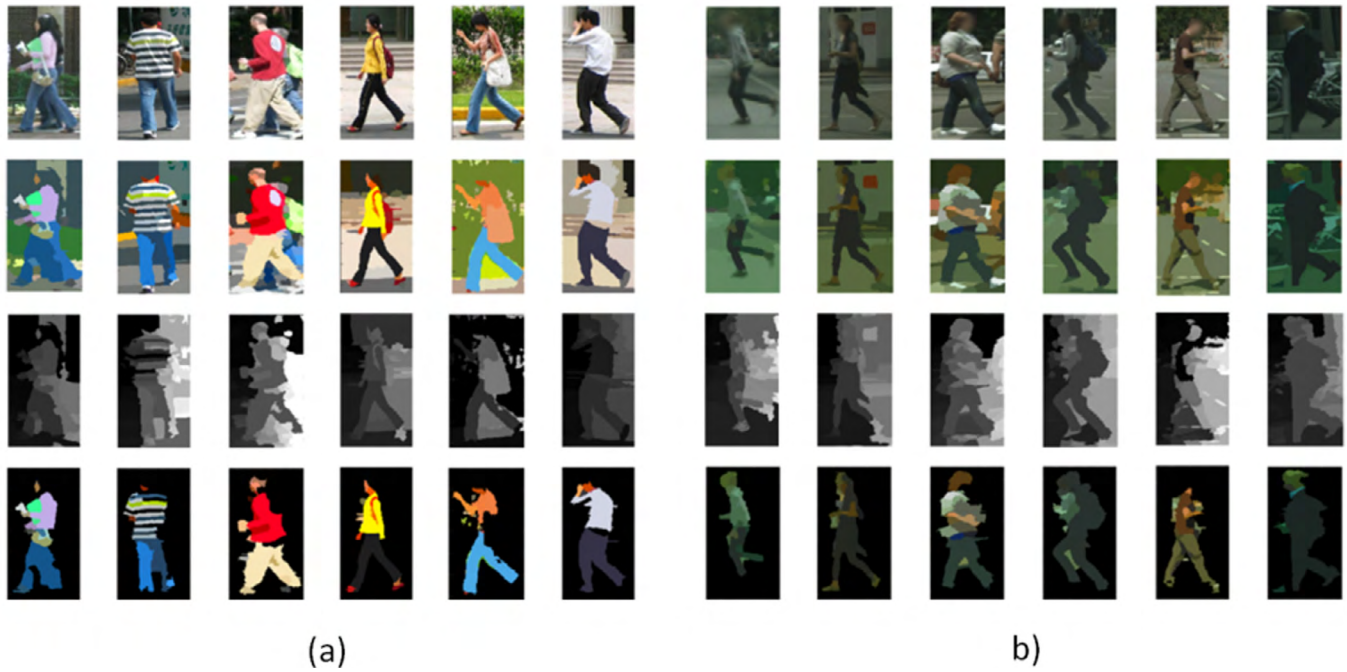

Figure 5: Results of segmentation method on the Penn-Fudan (left six) and Cityscapes (right six) dataset. From top down: input images, the initial segmentation, the classification by clustering and the probabilistic relaxation.

The second example in the second column of Figure 5a is more challenging, as the person wears a striped shirt that is not uniform. The segmentation performed well, even considering the color variety. After classification and relaxation, the person is partially identified, but his arm is missing. Two facts contribute to the error: First, the arm is more similar, in terms of color, to the background than to the clothes and, second, the probability values are lower at the borders of the probability field.

The third, fourth and sixth examples display a side view of walking persons. Here it is visible that the segmentation performed good because the persons wear uniform color clothes and there is a good contrast

between the person and the background. Nevertheless, some parts are lost, like the backpack of the fourth example or part of the back of the third person. In the last case, the white spot on the clothes reduced the color compatibility between the segments.

The possible errors are more visible in the fifth column. Here, the initial segmentation is good but the bag of the person is lost during the relaxation process because the bag occupies a larger region that includes pixels with low probability in the probability field. As the bag is not compatible with other neighboring regions, its classification cannot be improved.

The second set of images (Figure 5b) were obtained from Cityscapes and show pedestrians walking on the street but the images have lower contrast. Despite the low contrast, the detection procedure performed well.

A common problem is the removal of the extremes of arms and feet because such regions lie close to the borders, where the probability fields have low values. The probability fields (Figure 2) do not include a good description of the arms due to their variable position. In the examples, it is noticed a tendency to confuse skin with the background. This happens also due to the lack of saturation of the skin color. As the saturation is low, it is easily confused with the background.

4.1.2. Quantitative Evaluation

The results can be compared to a manually obtained reference, as illustrated in Figure 6, where the reference image is displayed as a binary image. The commission errors are displayed in yellow and the omission errors in red. Figure 6b displays the comparison of the result of the first segmentation and the clustering. For the computation of the quality parameters, the total pixels omission (red) and commission (yellow) were counted. The Black area displays the pixels that were correctly classified (agreement).

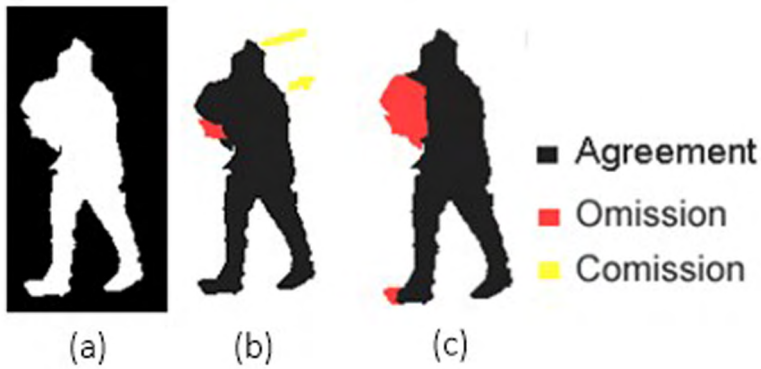

Figure 6: Example of comparison of the results from (a) manually obtained reference, (b) first segmentation and the (c) clustering.

To evaluate the result with a reasonable set of images, five series of images obtained at the city of Curitiba were used. A reduced set of samples of two series is displayed in Figure 7. The faces of the pedestrians are partially covered by masks, but they were also blurred to avoid the face recognition. The images were obtained at different days using cell phone cameras that were carried by a person or installed in a moving car. Figure 7 displays some images of two series. The first one was obtained from a window, oblique. The second set was obtained from a moving car. The pedestrians in the images were also manually delineated, as displayed in Figure 7. These binary images were not used in the previous steps but only to measure the quality of the results.

Figure 7: Example of the samples used to evaluate the method. a) original RGB images b) manually obtained ground truth.

From each series, ten samples were collected so that fifty images were analyzed. For each image, the Branch Factor, Miss Factor, Precision, Recall, F1-score, and IoU were computed. First, the results obtained without the relaxation step are presented to allow evaluating the contribution of this step. They are summarized in Table 1. The results can be considered good considering the quality parameters of Branch factor, Miss factor, recall, precision, F1-score, and IoU.

Table 1: Performance of the classification measured by six quality parameters (%): Branch factor (Bf), Miss factor (Mf), precision, recall, F1-score, and IoU.

	Bf	Mf	Recall	Precision	F1-score	IoU
Mean	28.91	36.80	82.04	79.90	80.96	67.48
Minimum	2.33	1.31	19.14	47.23	24.24	17.87
Maximum	111.73	>100	98.71	97.73	98.22	93.88

Table 2 displays the same statistics computed for the final image, after the probabilistic relaxation. A comparison of these tables reveals that the results are improved by the relaxation step. The mean values of the Branch and Miss factors are reduced by the relaxation, proving that some errors are corrected. The Bf without relaxation ranges from 29 to 111%, and between 15 and 59% after the relaxation, for example.

The mean recall decreased slightly, which can be a bad signal. Nevertheless, the minimum recall increased significantly, which indicates that some errors were removed. The values of the precision and F1-score increased and the minimum values went respectively, from 47 to 62, and 24 to 34. The mean and minimum IoU values increased too, but the maximum was reduced. The statistics after relaxation have smaller ranges. It is interesting to analyze the range of these parameters because it indicates that extreme low values are corrected. Based on the comparison, it can be stated that the relaxation introduces improvements in the results.

Table 2: Performance after the improvement by relaxation, measured by six quality parameters (%): Branch factor (Bf), Miss factor (Mf), precision, recall, F1-score, and IoU.

	Bf	Mf	Recall	Precision	F1-score	IoU
Mean	15.12	31.97	79.72	87.86	83.59	71.64
Minimum	0.00	2.09	23.50	62.57	34.17	23.22
Maximum	59.81	>100	97.95	100.00	98.96	87.36

To understand why some quality parameters decreased, the histogram of Branch factor, Miss factor, precision, recall, and IoU parameters was analyzed. The histograms of these parameters displayed in Table 2 are presented in Figures 8-10. The histograms of the Bf and Mf show that the values concentrate around low values, although some high values are still present. The proposed method has limitations in solving all cases and, therefore, errors are reflected as high values. Analyzing the histograms displayed in Figure 8, it is noticeable that the Mf is higher than the Bf. This reveals a tendency to miss segments in the result, for example the arms of some pedestrians. In some cases, the Bf is almost zero, showing that the result is fully compatible with the reference image. Nevertheless, some bad results are also obtained, as it is indicated by the maximum Bf of 60%.

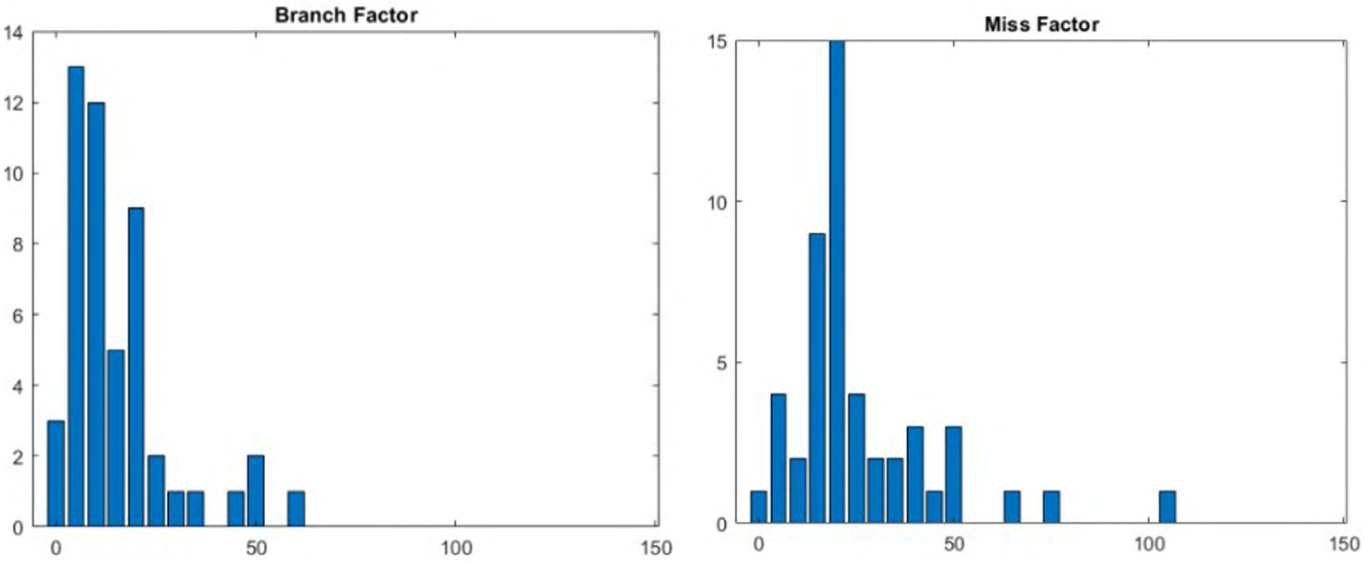

Figure 8: Histograms of the branch factor (left) and miss factors (right) of 50 samples.

The statistics of recall and precision can also be analyzed with the support of the Histograms displayed in Figure 9. The recall describes the rate of correct pixels within the total set of labeled pixels. The computed values concentrate around 80% and can be considered high. There is a pair of bad results that caused the mean to be lower, showing that the method is not able to solve all problems. It is worth to quote here that the density fields used are not able to model all possible poses and therefore not all the images can be correctly classified. The mean rate of correct pixels within the set of labeled pixels (precision), like the producer's accuracy concept, is 82% and, some results have achieved values closer to 100%.

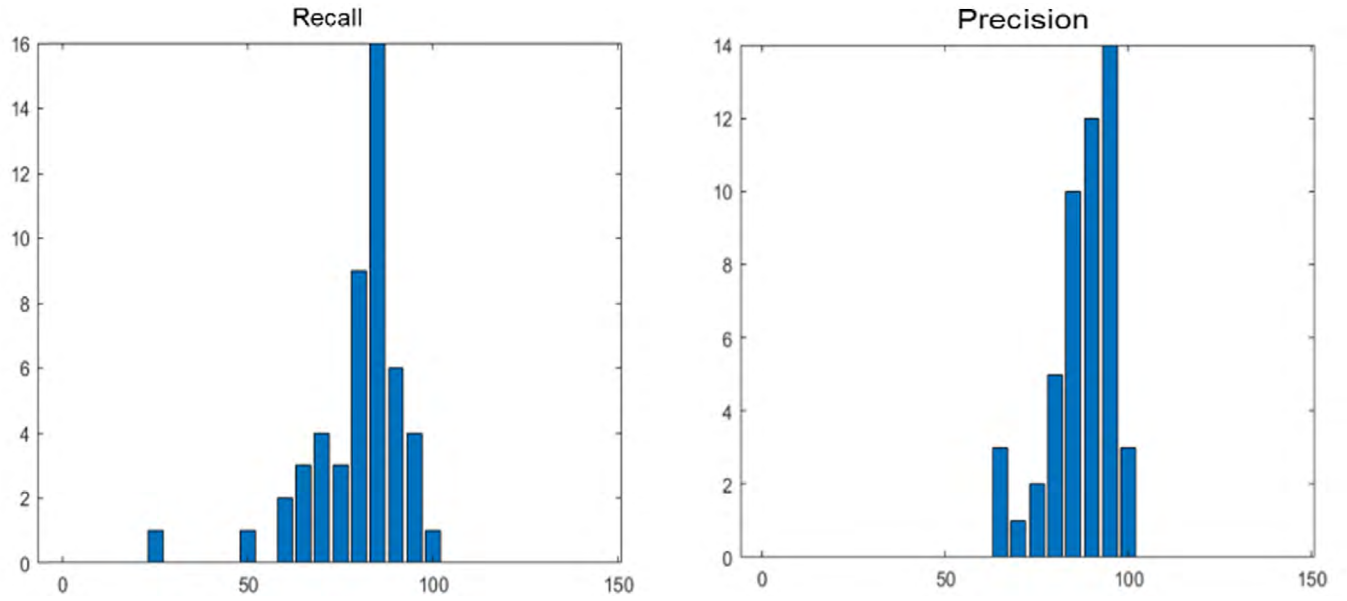

Figure 9: Histograms of the Recall (left) and Precision (right) indexes of the 50 samples.

Figure 10 displays the variation of the IoU. It is visible that the values are concentrated and close to 100, indicating a good IoU. The mean value is 71%, as some parts of the human figure are lost. Nevertheless, the IoU is good and proves that the method can be used to separate pedestrians from the background even when the background changes from image to image.

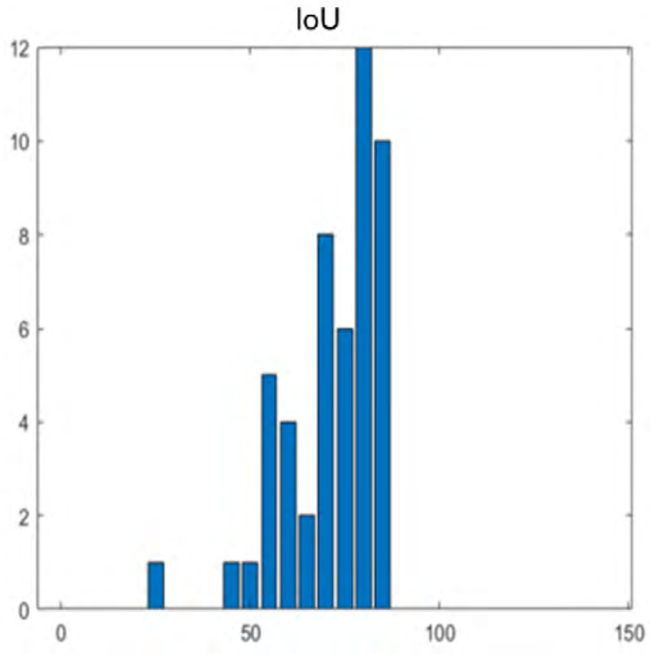

Figure 10: Histograms of the IoU computed from 50 samples.

The process is sensitive to occlusion, when the pedestrian is hidden by a pole or another pedestrian. But, if only a part of his body is hidden, it is still, possible to model the silhouette. Two examples are displayed in Figure 11. In the first example, a lamp hides part of the body. As the lamp lies in the central part of the image, it is included as part of the body. The ground between the legs of the person could also not be correctly classified

because the probability field in the center is high and there is another segment, the foot, that supports it because the color is similar. In the second situation (row 2) the lamp that hides the leg is close to the border and can be suppressed as background.

Figure 11: Examples of the results obtained with partial occlusion. Pedestrian hidden by a lamp (row 1) or pole (row2). From left to right: input images, the initial segmentation, the classification by clustering, the probabilistic relaxation, and the ground truth.

4.2 Comparison to the State-of-the-art Methods

4.2.1 Quantitative Evaluation

To assess the performance of the proposed method, the results obtained were further compared to those obtained using three state-of-the-art methods. Mask R-CNN (He et al., 2017), Yolact++ (Bolya et al., 2019), and DeepLabv3 (Chen et al., 2017) are machine learning algorithm trained on the COCO dataset and are reported to provide good segmentation performance across a wide-range of object classes, sizes, and colors. The three models are fully trainable end to end. For comparison purposes, the models were fine-tuned under the same experimental environments on Cityscapes and Penn-Fudan dataset using pretrained weights learnt on the COCO dataset. Mask R-CNN and DeepLav3 fine-turned models uses resnet 101 as backbone while Yolact++ uses resnet 50 as backbone. The performance of the fine-tuned models and the proposed model is summarized in Table 3.

Table3: Performance comparison of the proposed method and others on Cityscapes and Penn-Fudan datasets.

Dataset	Methods	Precision	Recall	F1-score	IoU
Cityscapes	Mask R-CNN	83.60	91.11	87.19	84.09
	Yolact++	94.03	93.54	93.78	96.13
	DeepLabv3	88.75	92.17	90.42	90.35
	Ours	94.27	93.95	94.11	96.06
Penn-Fudan	Mask R-CNN	79. 25	92.63	85.42	80.14
	Yolact++	92.20	94.02	93.10	95.01
	DeepLabv3	78.06	92.83	84.81	90.35
	Ours	94.31	93.91	94.11	96.75

According to Table 3, the accuracy indexes that can be achieved using the proposed method are compatible to the ones obtained with the other reference methods on both datasets Cityscapes and Penn-Fudan. Comparing the IoU index, which can be interpreted as an estimate of the global accuracy. The worst results were obtained using the Mask R-CNN approach, followed by DeepLabv3. The proposed method achieved a IoU index that is close to the one obtained with Yolact++, being superior in the Penn-Fudan experiment. The same difference and ranking are verified when comparing the recall and F1-score indexes. These two methods can be considered equivalent. The performance of Mask R-CNN, Yolact++, and DeepLabv3 are slightly high on Cityscapes than on Penn-Fudan dataset because the latter has only 170 images as these models depends on amount of training date. Contrarily, the performance of the proposed method does not change considerably as it does not need the training process. Considering the precision index, the difference between the methods is smaller. Nevertheless, the performance of the proposed method is the highest in both cases, being comparable to the accuracy obtained with Yolact++.

4.2.2 Qualitative Evaluation

Some segmentation results of the proposed method for the two datasets are shown in Figure 12. The proposed method yields encouraging results even in different situations such as columns one, three, and four (Figure 12, Penn-Fudan dataset), where the persons are occluded, and the background is very colorful, and on cityscapes dataset where the background color is similar to the foreground object. Overall, the proposed method performed well for the segmentation task as it improved the performance of human segmentation to a certain extent.

Figure 12: Comparison of our algorithm and other methods on the Penn-Fudan (left) and Cityscapes (right) dataset. For Mask R-CNN and Yolact++ the segmented regions are shown in red.

5. Conclusion

This paper introduced a method to separate human figures from the background based on predefined pose fields and probabilistic relaxation. The input is an image of a pedestrian with varying background. The method is composed by the following steps: segmentation, pose classification and refinement by probabilistic relaxation.

The method was tested with real images and the results prove that the method can be used to identify the segments related to the pedestrian with relatively good performance. The tests reveal a mean IoU around 71% with some cases reaching IoU values close to 100%. Nevertheless, the experiments also show that the method can fail. It occurs when the shape of the pedestrian cannot be modeled by the used density fields or, when there is low contrast between pedestrian and the background. The first situation can be improved including more templates with different poses, but it must also be considered difficult to model all possible poses and errors will persist.

The second problem, the contrast between clothes or skin of the pedestrian and the background, is more difficult to solve and causes difficulties even to a human interpreter. The low contrast that affects the segmentation and clustering steps cannot be improved in the following steps. The low contrast also can affect the refinement by probabilistic relaxation because it is based on the RGB color information.

The method performed well and it can be used to segment pedestrians from images obtained from moving cameras as it does not depend on a fixed known background and can be applied to images with different sizes, which makes it usable in the analysis of images with varying depth.

It is recommended to improve the first segmentation step, by including more information about the possible shape of the human figures or, look for other alternatives to solve this problem. Other alternative color spaces like CIELab or CIELuv not explored in this paper can be experimented.

ACKNOWLEDGEMENTS

The authors are grateful to CAPES/CNpq for financial support for this research.

AUTHOR´S CONTRIBUTION

Caisse Amisse: designed and implemented the experiments, and wrote the paper. All the authors contributed with preparation of draft manuscript, review of the proposed method, and formal analysis.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), 2274-2282.

Amisse, C., Jijón-Palma, M. E., Centeno, J. A. S. 2021. Fine-tuning deep learning models for pedestrian detection. Boletim de Ciências Geodésicas, 27.

Araslanov, N., Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4253-4262.

Baatz, M., Schäpe, A. 2000. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In: *XII Angewandte Geographische Informations-Verarbeitung*, pp. 12-23. Wichmann Verlag, Karlsruhe.

Bibby, C., Reid, I. 2008. Robust real-time visual tracking using pixel-wise posteriors. In *European Conference on Computer Vision*, pp. 831-844. Springer, Berlin, Heidelberg.

Bolya, D., Zhou, C., Xiao, F. Lee, Y.J., 2019. Yolact++: Better real-time instance segmentation. *arXiv preprint arXiv*:1912.06218.

Bouwmans, T., 2012. *Background subtraction for visual surveillance: A fuzzy approach*. Handbook on soft computing for video surveillance, 5, pp.103-138.

Brox, T., Malik, J. 2010. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision,* pp. 282-295. Springer, Berlin, Heidelberg.

Caelles, S., Maninis, K. K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 221-230.

Cai, Q., Mitiche, A., Aggarwal, J. K. 1995. Tracking human motion in an indoor environment. In *Proceedings, International Conference on Image Processing,* vol. 1, pp. 215-218.

Chaibou, M. S., Conze, P. H., Kalti, K., Mahjoub, M. A., Solaiman, B. 2020. Learning contextual superpixel similarity for consistent image segmentation. *Multimedia Tools and Applications*, 79(3), 2601-2627.

Chen, L. C., Papandreou, G., Schroff, F., Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv*:1706.05587.

Chen, Y., Yang, X., Zhong, B., Pan, S., Chen, D., Zhang, H. 2016. CNNTracker: Online discriminative object tracking via deep convolutional neural network. *Applied Soft Computing*, 38, 1088-1098.

Christmas, W. J., Kittler, J., Petrou, M. 1995. Structural matching in computer vision using probabilistic relaxation. *IEEE Transactions on pattern analysis and machine intelligence*, 17(8), 749-764.

Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., ... Schiele, B. 2015. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, vol. 2.

Davis, L., Philomin, V., Duraiswami, R. 2000. Tracking humans from a moving platform. In *Proceedings 15th International Conference on Pattern Recognition*. *ICPR-2000*, vol. 4, pp. 171-178.

Doulamis, A., Doulamis, N., Ntalianis, K., Kollias, S. 2003. An efficient fully unsupervised video object segmentation scheme using an adaptive neural-network classifier architecture. *IEEE Transactions on Neural Networks*, 14(3), 616-630.

Elgammal, A., Harwood, D., Davis, L. 2000. Non-parametric model for background subtraction. In *European conference on computer vision*, pp. 751-767. Springer, Berlin, Heidelberg.

Friedman, N., Russell, S. 2013. Image segmentation in video sequences: A probabilistic approach. *arXiv preprint arXiv*:1302.1539.

Goel, V., Weng, J., Poupart, P. 2018. Unsupervised video object segmentation for deep reinforcement learning. *arXiv preprint arXiv*:1805.07780.

Guo, L., Cheng, T., Huang, Y., Zhao, J., Zhang, R. 2017. Unsupervised video object segmentation by spatiotemporal graphical model. *Multimedia Tools and Applications*, 76(1), 1037-1053.

He, K., Gkioxari, G., Dollár, P. Girshick, R., 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969.

Heikkilä, M., Pietikäinen, M., Heikkilä, J. 2004. A texture-based method for detecting moving objects. In *Bmvc*, vol. 401, pp. 1-10.

Hu, Y. T., Huang, J. B., Schwing, A. G. 2018. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 786-802.

Jain, S. D., Xiong, B., Grauman, K. 2017. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2117-2126.

Kemker, R., Salvaggio, C., Kanan, C. 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, 145, 60-77.

Kittler, J., Illingworth, J. 1985. Relaxation labelling algorithms—a review. *Image and vision computing*, 3(4), 206-216.

Koh, Y. J., Kim, C.S. 2017. Primary object segmentation in videos based on region augmentation and reduction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7417-7425.

Lee, D.S., Shan, J., Bethel, J.S. 2003. Class-guided building extraction from IKONOS imagery. *Photogrammetric Engineering and Remote Sensing*, 69 (2), 143–150.

Lim, K., Jang, W. D., Kim, C. S. 2017. Background subtraction using encoder-decoder structured convolutional neural network. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1-6.

Lin, D., Dai, J., Jia, J., He, K., Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3159-3167.

Liu, X., Tao, D., Song, M., Ruan, Y., Chen, C., Bu, J. 2014. Weakly supervised multiclass video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 57-64.

Ma, Y., Wu, X., Yu, G., Xu, Y., Wang, Y. 2016. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors*, 16(4), 446.

Maninis, K. K., Caelles, S., Pont-Tuset, J., Van Gool, L. 2018. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 616-625.

Milan, A., Leal-Taixé, L., Schindler, K. Reid, I., 2015. Joint tracking and segmentation of multiple targets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 5397-5406.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D. 2020. Image segmentation using deep learning: A survey. *arXiv preprint arXiv*:2001.05566.

Neubert, P., Protzel, P. 2012. Superpixel benchmark and comparison. In *Proc. Forum Bildverarbeitung*, vol. 6, pp. 1-12.

Papazoglou, A., Ferrari, V. 2013. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*, pp. 1777-1784.

Parvin, B. A., Bhanu, B. 1983. Segmentation of images using a relaxation technique. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 151-153.

Pavlidis, T., Liow, Y. T. 1990. Integrating region growing and edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3), 225-233.

Portmann, J., Lynen, S., Chli, M., Siegwart, R. 2014. People detection and tracking from aerial thermal views. In *2014 IEEE international conference on robotics and automation (ICRA)*, pp. 1794-1800.

Ren, X., Malik, J. 2007. Tracking as repeated figure/ground segmentation. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Rosenfeld, A., Hummel, R. A., Zucker, S. W. (1976). Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man, and Cybernetics*, (6), 420-433.

Rosenfeld, A., Smith, R. C. 1981. Thresholding using relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5), 598-606.

Sobral, A., Vacavant, A. 2014. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122, 4-21.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149),* vol. 2, pp. 246-252.

St-Charles, P. L., Bilodeau, G. A., Bergevin, R. 2014. Flexible background subtraction with self-balanced local sensitivity. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 408-413.

Stutz, D., Hermans, A., Leibe, B. 2018. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166, 1-27.

Tian, B., Azimi-Sadjadi,M. R., Haar, T. H. V., Reinke, D. 2000. Temporal updating scheme for probabilistic neural network with application to satellite cloud classification. In *IEEE Trans. Neural Networks*, vol.11, pp. 903-920.

Tokmakov, P., Alahari, K., Schmid, C. 2017. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4481-4490.

Varga, V., Lőrincz, A. 2020. Reducing human efforts in video segmentation annotation with reinforcement learning. *Neurocomputing*, 405, 247-258.

Vijayanarasimhan, S., Grauman, K. 2012. Active frame selection for label propagation in videos. In *European conference on computer vision*, pp. 496-509. Springer, Berlin, Heidelberg.

Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L. C. 2019. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 9481-9490.

Wang, L., Shi, J., Song, G., Shen, I. F. 2007. Object detection combining recognition and segmentation. In *Asian conference on computer vision*, pp. 189-199. Springer, Berlin, Heidelberg.

Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L., 2019. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9236-9245.

Wang, W., Shen, J., Porikli, F., Yang, R. 2018. Semi-supervised video object segmentation with super-trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 41(4), 985-998.

Wang, W., Shen, J., Shao, L. 2015. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11), 4185-4196.

Wren, C. R., Azarbayejani, A., Darrell, T., Pentland, A. P. 1997. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7), 780-785.

Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y. 2020. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4), 1-47.

Yoshinaga, S., Shimada, A., Nagahara, H., Taniguchi, R. I. 2011. Statistical local difference pattern for background modeling. *IPSJ Transactions on Computer Vision and Applications*, 3, 198-210.

Zhou, T., Wang, S., Zhou, Y., Yao, Y., Li, J., Shao, L. 2020. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence, v*ol. 34, No. 07, pp. 13066-13073.

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., Catanzaro, B. 2019. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8856-8865.

Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., Kankanhalli, M. 2019. Unsupervised online video object segmentation with motion property understanding. *IEEE Transactions on Image Processing*, 29, 237-249.

Zouagui, T., Benoit-Cattin, H., Odet, C. 2004. Image segmentation functional model. *Pattern Recognition*, 37(9), 1785-1795.