Person Re-Identification Using CNN Method With Combination of SVM and Semantic Segmentation

Kristian Adhi Kurniawan*[1], **Moch. Arief Soeleman**[2]
Department of Informatics Engineering, University of Dian Nuswantoro, Indonesia,
(024) 3517261
E-mail : p31201902267@mhs.dinus.ac.id[1], arief22208@gmail.com[2]*

**Corresponding author*

Abstract – *Person re-identification is a mechanized procedure of video investigation which has been widely studied in contemporary years. Research problems that are often raised in the field of a person's re-identification research are characteristic representations that are easily affected by closure (abhorrent to other objects). Furthermore, after extracting local features by means of a boundary box, the background image still contains and does not focus on the human body parts. This study comes up with a method combination of CNN, SVM classification, and semantic segmentation. CMC (Cumulative Matching Characteristics) and mAP (mean Average Precision) are measurements of assessment that will be utilized to measure the operation of re-identification. The ResNet + SVM + SSP-ReID technique performed best in the Market dataset, with a CMC increase of 3-10% (rank-1 through rank-20). The Market and CUHK03 (D) datasets both showed improvements of 1-4.1% in mAP.*

Keywords Person re-identification; Feature extraction; CNN; SVM; Semantic segmentation;

1. INTRODUCTION

Person re-identification is a fundamental technology of intelligent video surveillance, and has been deeply explored in the last several years [1]. Finding the same individual within a set of images from many cameras is the aim of person re-identification [2]. Human video monitoring demands a significant amount of resources and time, which negatively impacts the effectiveness of the surveillance [3]. Consequently, by using the re-identification process, inspection quality can be improved naturally and significantly [4].

The earliest phase of person re-identification was reported in 1997, and it expanded quickly around 2008 [5]. Both academic and business environments utilize person re-identification for purposes including safeguards, location-based tracking, and gesture investigation. Person re-identification has additional applications in several areas of automation, criminology, and digital media [3]. Due to the expanding demand for open security and camera frameworks in open places, person re-identification is becoming more prevalent [6].

When different methods are used for person re-identification, the overall procedure is: 1) the input of images, 2) the segmentation of images, 3) the extraction of local features, 4) the representation of features, 5) the storage of features, 6) Identification of individuals and things in the picture [7]. Normally derived from public data, input data is in the form of an image of a person. Then, image segments separate objects from the background and divide the human body area [8]. When specific body positions are selected, local feature extraction is comparatively superior to global feature extraction [9]. The entire image is utilized in order to get the feature representation [10], and computations are performed using many color schemes [7]. Files can serve as storage functions, including research experiments [11]. In order to identify individuals and objects in an image, it is necessary to locate people and determine the object types [12].

CNN is the most often suggested technique for creating feature representations [13], as demonstrated by its efficiency in visual categorization with massive-scale [14]. A unique kind of neural network called a CNN is used to handle image input in the form of a matrix [15]. The advancement of person re-identification research has been affected by CNN-based learning techniques [11]. The CNN model is utilized as a result of its exceptional performance [10]. Meanwhile, because deep learning methods require a lot of data access, there is generally a difficulty with significant annotation (data labeling) [11].

The problems with existing methods are spatial changes in the CNN model [16] and feature extraction for each image, which does not obtain knowledge from image pairs [17]. The CNN method has the advantage of parameter sharing, which can reduce the number of unique parameters and increase the network size without the need for an increase in training data [15]. However, CNN has weaknesses in models that experience spatial changes [16]. Spatial transformers networks can solve the problem of spatial changes in CNNs by explicitly spatially manipulating the data in the network [16]. Siamese Convolutional Neural Network (S-CNN) is superior in performance to most deep learning architectures, because S-CNN can progressively reduce the width of the feature map in the image, without reducing the height of layers 4 to 6 [17]. However, S-CNN has a weakness in feature extraction for each image which does not gain knowledge from image pairs [17].

Several related studies use CNN and SVM methods for feature extraction and classification. Varior et al. [17] proposed the Matching Gate method to overcome the weaknesses of S-CNN which were discussed previously. Zhong et al. [18] proposed a Feature Aggregation Network (FAN) method that can perform CNN feature extraction at many levels on a pair of images. Zhang et al. [19] proposed a Sample-Specific SVM method that can learn a suitable classifier for each person during the training stage. Kalayeh et al. [20] proposed the Human Semantic Parsing for Person Re-identification (SPReID) method, which uses semantic segmentation as an alternative to bounding boxes, for local feature extraction. Zheng et al. [21] which can reduce the effects of pose estimation errors and detail loss during PoseBox creation.

The contribution in this research is applying two additional methods (SVM and semantic segmentation) to the CNN method. The SVM method, especially the linear SVM model, was proposed because it can express the maximum difference between two classes [19]. Meanwhile, the semantic segmentation method was proposed because it is accurate down to the image pixel level, and is more tolerant of changes in human body movement position compared to bounding boxes [36]. The aim of the SVM method combined with CNN is to complement CNN in terms of solving binary classification problems. So, in this research, the CNN method will be applied, which uses SVM classification and uses semantic segmentation for local feature extraction. This is done to solve the problem of feature extraction for each image without knowledge of image pairs, thereby producing more accurate predictions.

2. RESEARCH METHOD

2.1. Proposed Method

This research was carried out using a quantitative approach and experimental methods. As shown in Figure 1, the proposed additional methods have been marked with colors and dotted arrows. In the initial stage of data processing (Pre-processing), the semantic segmentation method is added. Meanwhile, at the classification stage, the SVM method was added.

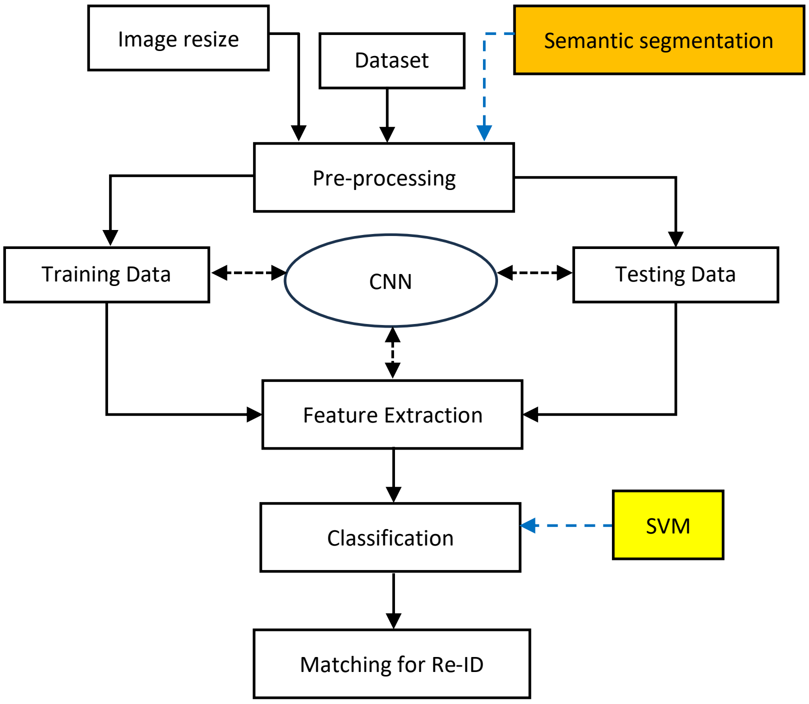


Figure 1 Research Flow of the Proposed Method

In the diagram shown in Figure 1, there are six main stages of the method proposed in this research. In the first stage, image datasets (CUHK03, Market, and Duke) are prepared first. The second stage, pre-processing is carried out on the image using image resizing and semantic segmentation techniques. The third stage, training and testing data from pre-processing results using the CNN method. The fourth stage, feature extraction by applying the CNN method from the training and testing results. The fifth stage, classification uses the SVM method from the feature extraction results. In the sixth stage, the identity of the same person is matched to be re-identified. CNN and SVM are combined by adding SVM as an output receiver from the CNN model.

2.2. Datasets

The datasets used in this research come from public datasets: CUHK03, Market1501 (Market), and DukeMTMC-reID (Duke). These three image datasets are often used for Person Re-identification research. Figure 2 shows several images that were selected as a dataset sample. The CUHK03 dataset has 14,096 images consisting of 767 identities of training data, and 700 identities of testing data. The Market dataset has 32,668 images consisting of 751 identities of training data, and 750 identities of testing data. The Duke dataset has 36,411 images consisting of 702 identities of training data, and 702 identities of testing data.

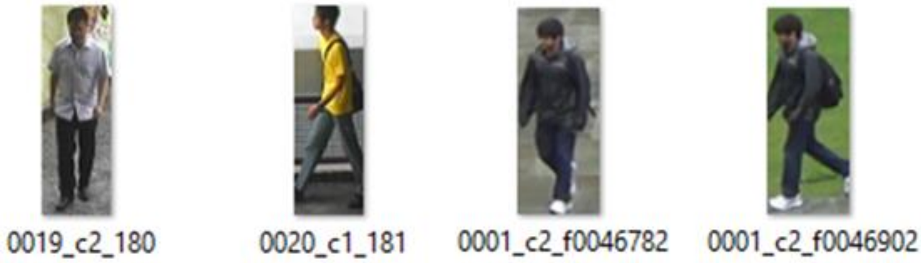


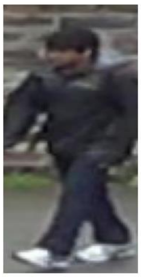
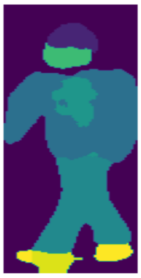
Figure 2 Dataset Sample

2.2. Pre-processing

The image will be processed first before entering the next stage. The pre-processing techniques used are resizing and semantic segmentation. The resizing technique involves changing the size of the image to be smaller than its original size, so that it can be used in the CNN architecture. Meanwhile, the semantic segmentation technique is carried out after the image resizing stage to divide areas of the human body.

An outline illustration of image processing is shown in Table 1. The original sample image, measuring 61 x 204 pixels, was resized to 64 x 128 pixels. After that, a semantic segmentation technique was carried out on the image to give different colors to each member of the human body (head, upper body, lower body, and shoes). So, in the result of semantic segmentation, the body parts of a person can be distinguished from each other. The environmental background around that person can also be properly marked by the darkest color.

Table 1 Pre-processing stage

Original image (61 x 204)	Result of image resize (64 x 128)	Result of semantic segmentation
		

2.3. CNN (Convolutional Neural Network)

Feature extraction in this research uses the CNN method. Image features are extracted at the testing stage to calculate the similarity of image pairs [22]. The pre-trained CNN-based model applied for feature extraction is ResNet50 which has been trained on the CUHK03, Market, and Duke datasets. Figure 3 shows the feature extraction process from image samples.

Semantic & Saliency detection is carried out on each image. Then feature extraction is carried out based on the type of detection that has been carried out. Semantic global feature extraction produces features in the form of human body parts separately. Meanwhile, Saliency global feature extraction produces features in the form of objects that can be used as clues to the image.

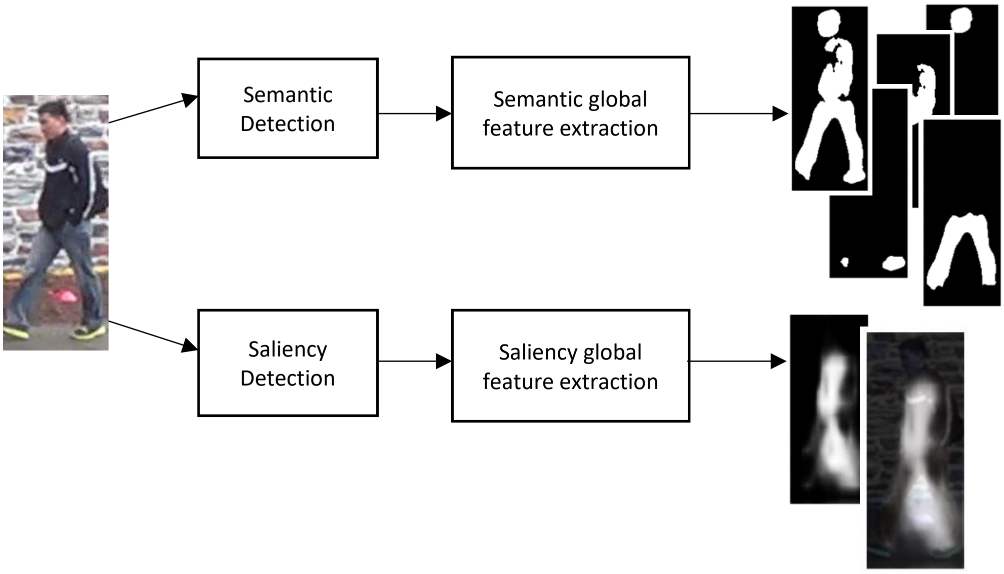

Semantic Detection
Semantic global feature extraction
Saliency Detection
Saliency global feature extraction

Figure 3 Feature extraction stage

The framework of the Resnet50 model is shown in Figure 4. The ResNet50 process flow generally consists of five stages. In the first stage, convolution is equipped with batch normalization, Relu layers, and max pool. Meanwhile, the second to fifth stages are structured by convolution blocks and identity blocks. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. After the fifth stage before output, average pool is used to calculate the average value of each filter shift in max pool. In the final step, flattening (feature map reshaping) in the form of a multidimensional array needs to be transformed into a vector, so that it can be used as input for the Fully-Connected (FC) layer.

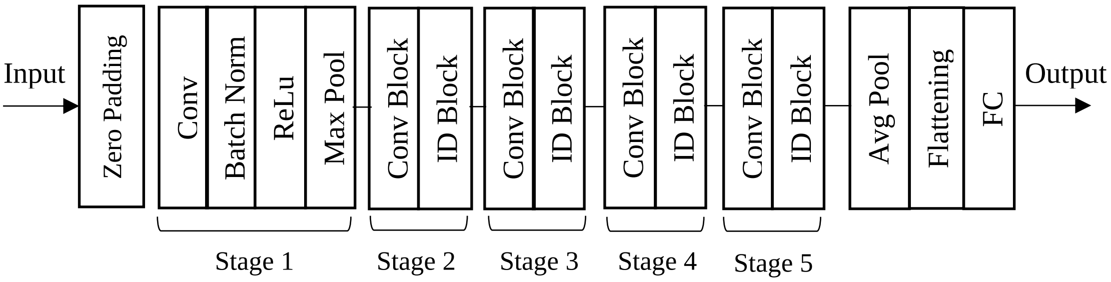


Figure 4 Resnet50 Framework Model

2.3. Framework of Proposed Method

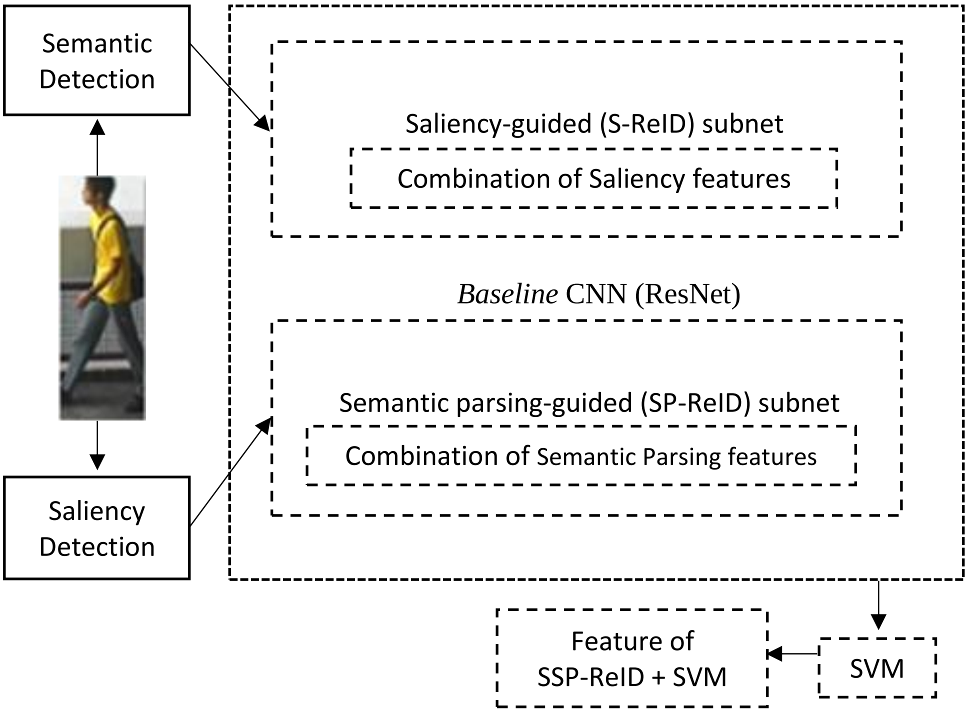

Semantic Detection
Saliency-guided (S-ReID) subnet
Combination of Saliency features
Baseline CNN (ResNet)
Semantic parsing-guided (SP-ReID) subnet
Combination of Semantic Parsing features
Saliency Detection
Feature of SSP-ReID + SVM
SVM

Figure 5 Proposed method framework developed from R. Quispe *et al.* [25]

As shown in Figure 5, there is a general flow of the proposed research method framework (ResNet + SVM + SSP-ReID). First, objects in the image are detected using two methods separately: saliency and semantic parsing. Saliency detection can identify objects that are unique or attract human attention for the first time visually [23]. Semantic parsing detection can divide human images into semantic parts that contain important information [24]. Second, the detection results are processed by the main CNN network (ResNet), which has two sub-networks (subnets): S-ReID (Saliency) and SP-ReID (Semantic Parsing). The two subnets have the same architecture as CNN's main network, namely ResNet. Third, there is a combination of features on each subnet. The S-ReID subnet functions to obtain global features from unique objects in the image. The SP-ReID subnet functions to obtain global features in the semantic area [25]. Fourth, classification is carried out using the SVM method. Learning strategy from SVM by maximizing the classification interval. The result of the training process is to make the correct category score higher by at least one interval, compared to the incorrect category score. Then, at the final stage of this framework flow, combined features from the SSP-ReID and SVM methods are produced.

2.4. Evaluation Metric

Re-identification performance is measured using two evaluation metrics: Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP). CMC evaluation aims to calculate the accuracy and probability of the correct identity appearing at the top 1, 5, ..., k ranks (rank-k) in a collection of images [26]. Meanwhile, mAP evaluation aims to calculate the average precision value of all images that are matched with training data. In Table 2, there is a formula for mAP and CMC calculation:

Table 2 Formula of evaluation metric

Parameter	Formula
Precision	$\frac{TP}{TP+FP}$ (1)
Recall	$\frac{TP}{TP+FN}$ (2)
Average Precision	$AP = \frac{1}{N}\sum_{Recall_{(i)}} Precision(Recall_{(i)})$ (3)
Mean Average Precision	$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$ (4)
Cumulative Matching Characteristics	$CMC(R = n) = \sum_{i=1}^{N}\sum_{R=1}^{n}\sum_{j=1}^{M} B\left(R_i, y_j^i, k^i\right)\frac{1}{M}\frac{1}{N}$ (5)

Based on Table 2, three formulas of Precision, Recall, Average Precision are used as the basis formulas of mAP and CMC. Both Precision and Recall formulas are focused on the calculation of true / false in prediction results. True Positive (TP) is the number of positive samples that the model can successfully detect. While False Positive (FP) refers to the number of negative samples that incorrectly classified as positive. False Negative (FN) refers to positive samples that were incorrectly categorized as negative. While True Negative (TN) shows the model's correct identification of negative samples. The formula of Average Precision is used to calculate the sum of Recall and Precision, which is averaged by the class number. While mAP and CMC formulas were already explained in the previous paragraph.

3. RESULTS AND DISCUSSION

For the CUHK03 dataset, CUHK03-NP (New Protocol) dataset was used, which consists of two sub-datasets: CUHK03 (D) - Detected and CUHK03 (L) - Labeled. After testing, accuracy results were obtained, as shown in Table 3.

Table 3 Test results from previous research and current research

Method	Dataset							
	Market		CUHK03 (D)		CUHK03 (L)		Duke	
	mAP (%)	rank-1 (%)	mAP (%)	rank-1 (%)	mAP (%)	rank-1 (%)	mAP (%)	rank-1 (%)
ResNet [27]	72.9	88.1	52.9	55.6	56.7	58.8	62.1	77.7
ResNet + SP-ReID [25]	72.4	87.8	53.5	56.2	55.5	57.3	62.7	78.0

ResNet + S-ReID [25]	73.0	87.6	53.4	56.0	54.4	55.9	63.1	78.9
ResNet + SSP-ReID [25]	75.9	89.3	57.1	59.4	58.9	60.6	66.1	80.1
ResNet + SVM	52.4	70.9	29.9	30.6	42.7	43.0	22.6	38.2
ResNet + SVM + SP-ReID	56.1	74.1	26.7	26.3	35.4	35.6	46.3	64.9
ResNet + SVM + S-ReID	55.7	73.5	39.7	39.3	43.5	45.2	46.3	66.2
ResNet + SVM + SSP-ReID	73.9	88.4	57.0	59.3	60.5	61.9	63.4	78.5

Based on Table 3, the test results obtained from research [25] show an improvement compared to previous research [27]. The combination of S-ReID and SP-ReID (SSP-ReID) methods proposed by research [25] shows better results, when compared with research [27]. Results on the Market dataset increase up to 3% for mAP values, and 1.2% for rank-1 values. Results on the CUHK03 (D) dataset increased up to 4.2% for mAP values, and 3.8% for rank-1 values. Results on the CUHK03 (L) dataset increased up to 2.2% for mAP values, and 1.8% for rank-1 values. Results on the Duke dataset improve by 4% for mAP values, and 2.4% for rank-1 values.

As shown in Table 3, for the current research on the ResNet + SVM + SSP-ReID method, better results were obtained compared to the ResNet + SSP-ReID method in research [28]. The improvement is especially visible when tested with the CUHK03 (L) dataset. Results on the CUHK03 (L) dataset for the current study show an increase in mAP values of up to 1.6%, and an increase in rank-1 values of up to 1.3%. The ResNet + SVM + SSP-ReID method also obtained improved results, when compared with the ResNet method in research [27]. The results on the Market dataset have increased by up to 1% (mAP), and 0.3% (rank-1). The results on the CUHK03 (D) dataset increased by up to 4.1% (mAP), and 3.7% (rank-1). The results on the CUHK03 (L) dataset increased by up to 3.8% (mAP), and 3.1% (rank-1). Results on the Duke dataset have increased by up to 1.3% (mAP), and 0.8% (rank-1).

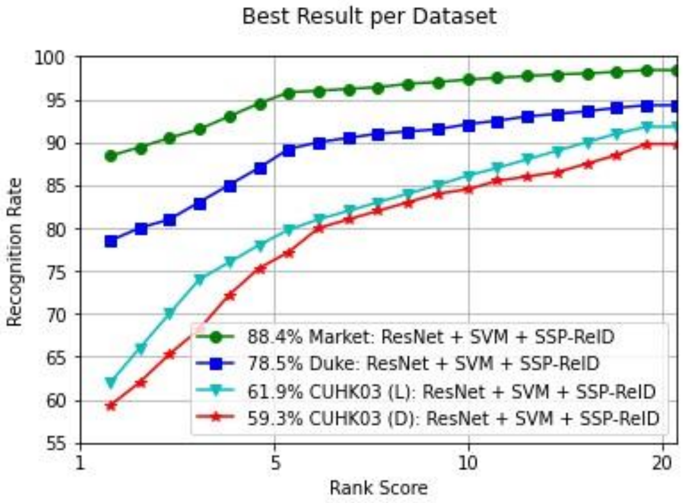


Figure 6 The CMC curve results from the best performance of each dataset

All ranking results from the Market, Duke, CUHK03 (D), CUHK (L) datasets are taken by one method per dataset with the best performance, as shown in Figure 6. The best performance of the four datasets was obtained with values ranging from 59.3% to 88.4%. All of these performance values can be achieved with a combination of the ResNet, SVM, and SSP-ReID methods. In the CUHK03 (D) and CUHK03 (L) sub-datasets, from the main CUHK03 dataset source and the same method (SSP-ReID), it produces almost the same recognition rate in the value range of ± 80-85% from rank-5 to rank-10 . Apart from that, the two sub datasets from rank-1 to rank-5 experienced an increase in the rate range of ± 59.3% to 80%. The two sub datasets CUHK03 (D) and CUHK03 (L) at rank-10 to rank-20 experienced an increase in the rate range of ± 85% to 92%. Since the initial performance of rank-1 to rank-20, the trend on the curve has experienced a relatively parallel increase in the two methods: SSP-ReID Market (green) and SSP-ReID Duke (dark blue). The lowest value range of the two methods is rank-1, namely between ± 78% to 88%. This indicates that at rank-1 of the lowest value range for the two methods there is a rate increase of around 10%. Meanwhile, the highest value range for the two methods is rank-20, namely between ± 95% and 98%. This indicates that at rank-20 of the highest value range for the two methods, the rate increased by around 3%.

Table 4 Saliency detection sample

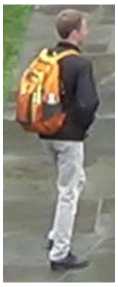
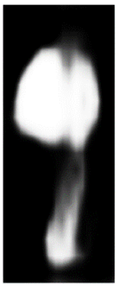
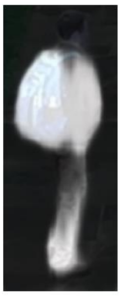
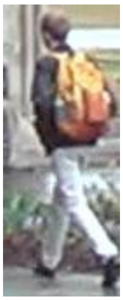
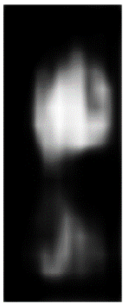
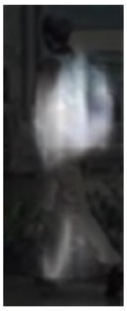
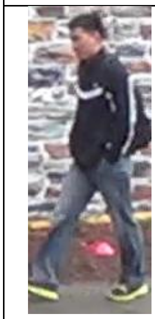
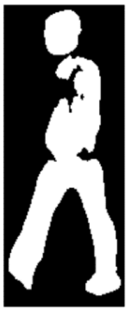
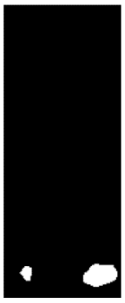
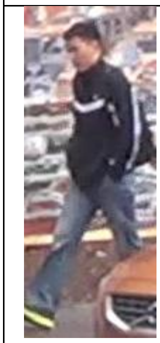
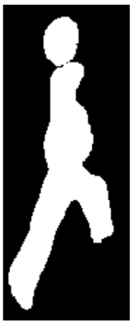
Original image	**Saliency area from original image**	**The result of saliency area which is placed on top of the original image**
		
		

Table 4 shows the saliency detection images obtained from the current research sample. The sample consists of two images that have different perspectives, but still refer to the same person. The clues in the image are an orange backpack and a white trouser, which are marked in the white area using saliency detection. Then, when the detection results are combined with the original image, distinctive characteristics can be seen on the two objects (bag and trousers). Even though these samples show different bag positions and were recorded from two different cameras, saliency detection in the final stage can still identify the same person from two different images.

Table 5 Semantic parsing detection sample

Original image	Semantic area on the human body				
	Complete body	Head	Upper body	Lower body	Shoes
					
					

There is a sample of semantic parsed images in Table 5 obtained from current research. This sample shows two images of the same person from different perspectives. All semantic regions are marked in white, and the background is black. Semantic parsing can recognize all human body objects and divide them into five body parts (semantic regions), consisting of the head, upper body, lower body, and shoes. In the second original image example, from the shank to the right shoe is blocked by the front of a car, but the semantic parsing method is able to handle this occlusion (obstacle). As a result, the car is perceived as a separate object with a color that matches the black background.

4. CONCLUSION

The addition of the SVM method to CNN, semantic segmentation, and saliency produces a combination of proposed methods for this research. Experiments have been performed on four methods on each tested dataset (Market, CUHK03, and Duke). The general research results from this experiment show that the ResNet + SVM + SSP-ReID method has the best performance compared to the other three proposed methods. Specifically in the Market dataset, the ResNet + SVM + SSP-ReID method can outperform the other datasets, increasing the CMC by 3-10% (ranks 1–20). While the mAP of Market and CUHK03 (D) datasets, both improved by 1-4.1%. Based on those performance data summaries from the experimental results, it can be concluded that the combination of these three methods has an impact in terms of increasing the performance of recognition rate, and mAP of all datasets that have been tested. The problem of feature extraction for each image that does not obtain knowledge from image pairs in the S-CNN algorithm, can be solved when the CNN, SVM, semantic segmentation and saliency methods are applied in this research. The SVM method added to this research can separate matching image pairs from all images that do not match. The saliency method can extract global features from unique objects in the image. The semantic segmentation method can perform global feature extraction in semantic areas.

REFERENCES

[1] X. Y. Jing *et al.*, “Super-Resolution Person Re-Identification with Semi-Coupled Low-Rank Discriminant Dictionary Learning,” *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, 2017, doi: 10.1109/TIP.2017.2651364.

[2] L. Wu, Y. Wang, Z. Ge, Q. Hu, and X. Li, “Structured deep hashing with convolutional neural networks for fast person re-identification,” *Comput. Vis. Image Underst.*, vol. 167, no. December 2016, pp. 63–73, 2018, doi: 10.1016/j.cviu.2017.11.009.

[3] A. Bedagkar-Gala and S. K. Shah, “A survey of approaches and trends in person re-identification,” *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, 2014, doi: 10.1016/j.imavis.2014.02.001.

[4] P. H. Tu *et al.*, “An intelligent video framework for homeland protection,” *Unattended Ground, Sea, Air Sens. Technol. Appl. IX*, vol. 6562, p. 65620C, 2007, doi: 10.1117/12.729215.

[5] K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, “Survey on person re-identification based on deep learning,” *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 219–227, 2018, doi: 10.1049/trit.2018.1001.

[6] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person Re-identification: Past, Present and Future,” vol. 14, no. 8, pp. 1–20, 2016, [Online]. Available: http://arxiv.org/abs/1610.02984.

[7] E. Poongothai and A. Suruliandi, “Survey on colour, texture and shape features for person re-identification,” *Indian J. Sci. Technol.*, vol. 9, no. 29, 2016, doi: 10.17485/ijst/2016/v9i29/93823.

[8] L. Bazzani, M. Cristani, and V. Murino, “Symmetry-driven accumulation of local features for human characterization and re-identification,” *Comput. Vis. Image Underst.*, vol. 117, no. 2, pp. 130–144, 2013, doi: 10.1016/j.cviu.2012.10.008.

[9] J. Yin, A. Wu, and W. S. Zheng, “Fine-Grained Person Re-identification,” *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1654–1672, 2020, doi: 10.1007/s11263-019-01259-0.

[10] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, “Human Semantic Parsing for Person Re-identification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 227, pp. 1062–1071, May 2018, doi: 10.1016/S0140-6736(01)37157-X.

[11] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised Person Re-identification: Clustering and Fine-tuning,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 4, pp. 1–18, 2018, doi: 10.1145/3243316.

[12] X. Yang, Y. Tang, N. Wang, B. Song, and X. Gao, “An End-to-End Noise-Weakened Person Re-Identification and Tracking with Adaptive Partial Information,” *IEEE Access*, vol. 7, pp. 20984–20995, 2019, doi: 10.1109/ACCESS.2019.2899032.

[13] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, “Deep Representation Learning with Part Loss for Person Re-Identification,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, 2019, doi: 10.1109/TIP.2019.2891888.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[15] I. Goodfellow, Y. Bengio, and · Aaron Courville, *Deep Learning*, vol. 91, no. 5. 2012.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 2017–2025, 2015.

[17] R. R. Varior, M. Haloi, and G. Wang, “Gated siamese convolutional neural network architecture for human re-identification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9912 LNCS, pp. 791–808, 2016, doi: 10.1007/978-3-319-46484-8_48.

[18] W. Zhong, L. Jiang, T. Zhang, J. Ji, and H. Xiong, “Combining multilevel feature extraction and multi-loss learning for person re-identification,” *Neurocomputing*, vol. 334, pp. 68–78, 2019, doi: 10.1016/j.neucom.2019.01.005.

[19] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, “Sample-specific SVM learning for person re-identification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 1278–1287, 2016, doi: 10.1109/CVPR.2016.143.

[20] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, “Human Semantic Parsing for Person Re-identification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 227, pp. 1062–1071, 2018, doi: 10.1109/CVPR.2018.00117.

[21] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose-Invariant Embedding for Deep Person Re-Identification,” *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, 2019, doi: 10.1109/TIP.2019.2910414.

[22] Y. Zheng, H. Sheng, Y. Liu, K. Lv, W. Ke, and Z. Xiong, “Learning irregular space transformation for person re-identification,” *IEEE Access*, vol. 6, pp. 53214–53225, 2018, doi: 10.1109/ACCESS.2018.2871149.

[23] W. Wang, J. Shen, and L. Shao, “Deep Learning For Video Saliency Detection,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, 2018, doi: 10.1109/TIP.2017.2754941.

[24] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, “Look into Person: Self-supervised Structure-sensitive Learning and a new benchmark for human parsing,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6757–6765, 2017, doi: 10.1109/CVPR.2017.715.

[25] R. Quispe and H. Pedrini, “Improved person re-identification based on saliency and semantic parsing with deep neural network models,” *Image Vis. Comput.*, vol. 92, 2019, doi: 10.1016/j.imavis.2019.07.009.

[26] M. O. Almasawa, L. A. Elrefaei, and K. Moria, “A Survey on Deep Learning-Based Person Re-Identification Systems,” *IEEE Access*, vol. 7, pp. 175228–175247, 2019, doi: 10.1109/ACCESS.2019.2957336.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[28] C. Ding, T. Bao, and H. Huang, “Quantum-Inspired Support Vector Machine,” pp. 1–13, 2021.