IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON AI-DRIVEN BIG DATA PROCESSING: THEORY, METHODOLOGY, AND APPLICATIONS

Received March 30, 2019, accepted April 26, 2019, date of publication May 2, 2019, date of current version May 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914455

Shoe-Print Image Retrieval With Multi-Part Weighted CNN

ZHANYU MA[1], (Senior Member, IEEE), YIFENG DING[1], SHAOGUO WEN[2], JIYANG XIE[1], (Student Member, IEEE), YIFENG JIN[3], ZHONGWEI SI[2], AND HAINING WANG[4]

[1]Pattern Recognition and Intelligent Systems Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3]Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China
[4]School of Police Administration, People's Public Security University of China, Beijing 100038, China

Corresponding author: Yifeng Ding (dingyf@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61773071, in part by the Beijing Nova Program under Grant Z171100001117049, in part by the Beijing Nova Program Interdisciplinary Cooperation Project under Grant Z181100006218137, in part by the Open Projects of National Engineering Laboratory for Forensic Science under Grant 2018NELKFKT04, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2019109 and Grant XTCX201804.

ABSTRACT Identifying shoe-print impressions in the scene of crime (SoC) from database images is a challenging problem in forensic science due to the complicated impressing surface, the partial absence of on-site impressions, and the huge domain gap between the query and the gallery images. The existing approaches pay much attention to feature extraction while ignoring its distinctive characteristics. In this paper, we propose a novel multi-part weighted convolutional neural network (MP-CNN) for shoe-print image retrieval. Specifically, the proposed CNN model processes images in three steps: 1) dividing the input images vertically into two parts and extracting sub-features by a parameter-shared network individually; 2) calculating the importance weight matrix of the sub-features based on the informative pixels they contained and concatenating them as the final feature, and; 3) using the triplet loss function to measure the similarity between the query and the gallery images. In addition to the proposed network, we adopt an effective strategy to enhance the quality of the images and to reduce the domain gap using the U-Net structure. The experimental evaluations demonstrate that our proposed method significantly outperforms other fine-grained cross-domain methods on SPID dataset and obtains comparative results with the state-of-the-art shoe-print retrieval methods on FID300 dataset.

INDEX TERMS Cross-domain, image retrieval, shoe-print, scene of crime.

I. INTRODUCTION

Shoe-print identification is an important issue in forensics science, as the shoe marks are the most frequently left clues in a crime scene. This recognition problem is challenging due to the diversity in various types of crime scene (ranging from variety of impressing surface to partial absence) and the huge domain margin between the scene of crime (SoC) impressions and the shoe-print databases. Examples in Fig. 1 illustrates these challenges.

Given a SoC impression, in order to find the correct matches among a large set of candidates in a shoe-print database, three major tasks need to be addressed: 1) removing the distortions and enhancing the quality of images by pre-processing, 2) generating the discriminative features of both the query and the gallery images, and 3) matching the query samples with the whole database using suitable similarity metrics [1]. In addition to the aforementioned image pre-processing, a large group of works [2]–[6] have paid much attention to discriminative features generation using feature extraction techniques. Some other works [7]–[9] proposed new similarity metrics to measure the distance between the query samples and the gallery images. Recently with the progress in machine learning techniques, several learning-based techniques have been proposed. References [7], [10], [11] extract deep features using base models (*e.g.*, VGG19 [12], Resnet50 [13]) and train the networks for matching. However, the main problem of these methods is that few special structures have been designed for identifying shoe-print characteristics.

The associate editor coordinating the review of this manuscript and approving it for publication was Yi-Zhe Song.

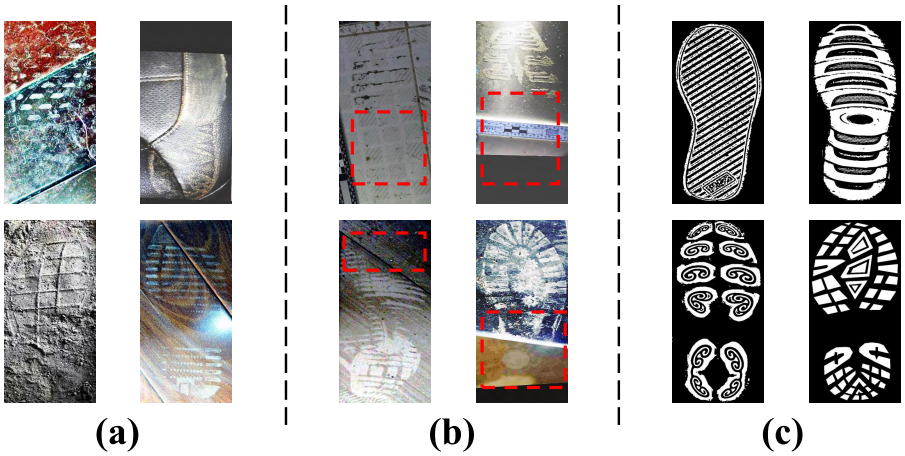


FIGURE 1. Shoe-print identification is challenging due to (a) variety of impressing surface (ceramic, leather, soil, and wood), (b) partial absence (the missing part is marked out in red dashed boxes), and (c) large domain gap when retrieving the database images.

Nowadays, fine-grained cross-domain image retrieval methods have been introduced in related fields like matching aerial photos with geographic information system (GIS) map data [14]–[16], hand drawn sketches to real world images [17], [18], and historical architectural paintings to 3D models [19]. Generally speaking, these methods can also be applied in shoe-print identification. References [20], [21] improved the Siamese network [22] to achieve fine-grained retrieval across the sketch/image gap. Although promising results have been reported, the sketch based image retrieval (SBIR) task differs from the shoe-print identification in two aspects: 1) the SoC impressions often contain more noises in background than the hand draw sketches; 2) all sketches in the SBIR task are complete in shape while in SoC impressions, partial absence is a notable character that needs to be taken into consideration.

To deal with these challenges, we propose a novel multi-part weighted convolutional neural network (MP-CNN) for shoe-print image retrieval. The MP-CNN uses the Siamese network [22] as the base structure and adopts the triplet loss function as the similarity metrics. Given the partial absence on SoC impressions, we divide the input query images into top and bottom slices and individually send them into the VGG19 [12] models which share the same parameter set to extract their features. Next, we calculate the importance weight matrix of these two parts depending on the informative pixels they contained. By multiplying the weights on two sub-features, we concatenate them together as the final features which are then applied for similarity calculation. Through this operation, the triplet loss function can focus on matching the informative part of the query impressions with samples from the database. As a result, the MP-CNN can generate more discriminative features distinct from each other and improve the retrieval accuracy. Furthermore, in order to decrease the domain difference between the SoC impressions and the database, we propose a new strategy which utilizes U-Net [23] to enhance the quality of images. The SoC impressions are then converted to binary images (Bi-SoC) with environmental noises removed. Though some valuable information is lost through this procedure inevitably, the profit is markedly enormous as the aforementioned cross-domain problem is greatly solved. Note that this strategy only works in certain conditions since the U-Net structure requires precise pixel-level masks of shoe-print images in the training steps.

The main contributions of this paper are as follows: 1) we propose a novel MP-CNN architecture that focuses on the informative parts of the input images and 2) we introduce a new strategy to decrease the domain gap between the SoC and the shoe-print database. Experimental results on two datasets (our own shoe-print identification database (SPID) and footwear impression database (FID300)) show that the proposed method can achieve significant improvement on shoe-print identification tasks.

The rest of the paper is organized as follows. Section II describes the related work. Section III introduces the proposed method. Section IV provides the evaluation and analysis, followed by the conclusions in Section V.

II. RELATED WORK

In this Section, we briefly review the previous works regarding both shoe-print identification and some related fine-grained cross-domain image retrieval methods.

A. SHOE-PRINT IDENTIFICATION

The widespread success of automatic fingerprint identifications systems [24] has inspired many attempts to similarly automate shoe-print identification. Existing state-of-the-art techniques mainly differ in the way to extracted features.

Some methods seek to process shoe-print image in a holistic way. De Chazal *et al.* [25] calculated the 2D discrete Fourier transform (DFT) features to yield a periodogram estimate of the power spectral density (PSD). Gueham *et al.* [26] exploited Fourier-Mellin transformed features obtained by a log-polar mapping followed by a DFT. In addition, Zhang and Allinson [2] introduced multiresolution features

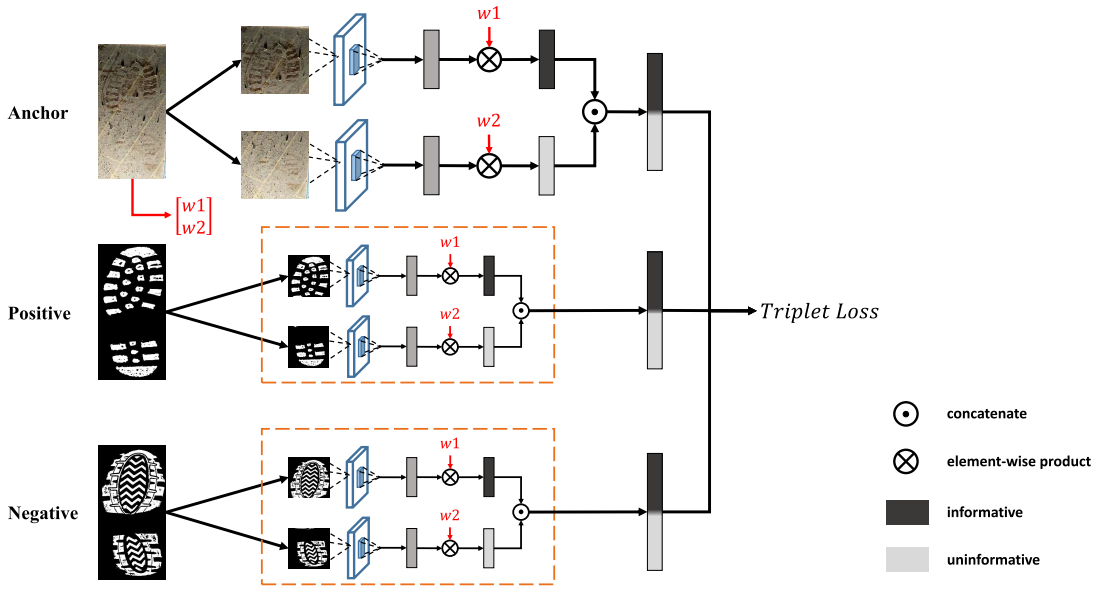


FIGURE 2. Illustration of the proposed MP-CNN framework. The proposed network is a Siamese network of three CNN branches with different input, corresponding to the anchor SoC impression, the positive database image and the negative database image, respectively. For each branch, the input image is divided into two slices to extract features individually. We concatenate these two features together with weights and send three new features to triplet loss layer to measure the similarity between them. Please refer to Section III for details.

using Gabor transform. Li *et al.* [5] combined the integral histogram of the Gabor features with the Euclidean distance and histogram intersection. Kong *et al.* [27] extracted Gabor and Zernike features combined with normalized correlation for matching. With the progress in machine learning techniques, several learning-based techniques have been proposed. Kortylewski and Vetter [7] suggested a probabilistic compositional active basis model for shoe-print identification. The recent study of Kong *et al.* [10], [11] introduced a multi-channel normalized cross-correlation to match multi-channel deep features extracted by pre-trained convolutional neural networks.

Another group of works tried to extract some discriminative features from local shoe-print regions. Pavlou and Allinson [28] exploited the maximally stable extremal region (MSER) to detect the points of interest from the gradient location, the orientation histogram (GLOH) and the scale invariant feature transform (SIFT) features. In the same context, Rathinavel and Arumugam [29] extracted the discrete cosine transform (DCT) coefficients for overlapped blocks, further combined with the principal component analysis (PCA) and the Fisher linear discriminant (FLD). Wei *et al.* [30] combined SIFT features with cross-correlation matching. Wang *et al.* [31] exploited the Wavelet-Fourier transform features. Almaadeed *et al.* [8] combined the Harris and the Hessian point of interest detectors with SIFT descriptors.

B. SIMILARITY METRIC

The similarity metrics in shoe-print identification task including the Euclidean distance [2], [5], [6], [29], [32], the 2D correlation [25], [26], [31], the mean square noise error method (MSNE) [33], and the normalized cross-correlation (NCC) [10], [11]. Other methods in person re-identification and face recognition usually use the cosine distance or the vector angle to measure feature similarity.

C. FINE-GRAINED CROSS-DOMAIN IMAGE RETRIEVAL

Differentiating shoe-prints from each other is much harder when compared with the conventional category-level classifications, since the visual differences between the shoe-print images are often subtle. Meanwhile, huge domain gap occurs from SoC impressions to database samples. So this problem can be considered as an implementation in the area of fine-grained cross-domain image retrieval tasks. Li *et al.* [34] solved the fine-grained sketch-based image retrieval (FG-SBIR) task by employing deformable part-based model (DPM) which learns mid-level representation for both sketches and images. Yu *et al.* [20] and Sangkloy *et al.* [35] evaluated the two-branch CNNs with pairwise verification loss and three-branch CNNs with triplet ranking loss aims to learn both the feature representation and the cross-domain matching function jointly. Song *et al.* [21] proposed a deep spatial semantic attention model for FG-SBIR by introducing an attention modeling and a shortcut connections on it. While all these methods suggest good ways to deal with the fine-grained cross-domain tasks, the SoC images are more complex that requires more complicated networks.

Our proposed MP-CNN model differs from the aforementioned methods by introducing a novel way to deal with

the absence in the SoC impressions. Through the proposed model, more discriminative features can be learned and higher accuracy can be reached.

III. APPROACH

The main characteristics that make the shoe-print identification distinct from other fine-grained cross-domain image retrieval task are two-folds: 1) SoC impressions often contains more complex noise in the background and 2) scattered SoC impression is common in real world cases. In this section, we adopt a new strategy which uses the U-Net [23] to enhance the quality of the images and to reduce the domain gap between the SoC impressions and the database images. Furthermore, we propose a novel multi-part weighted CNN named as MP-CNN to cut input images into pieces and extract features individually. By providing the sub-features with different weights, we enforce the similarity metric to pay more attention on the most informative parts of the query.

A. TWO STRATEGIES

For shoe-print image retrieval task, one natural thought is following the fine-grained sketch-based image retrieval (FG-SBIR) problems to treat it as a cross-domain task. A Siamese network with base networks like VGG [12], ResNet [13] is proven effective in extracting images from different domains [22]. Therefore, the first strategy directly uses the SoC impressions to match the images in the database through a siamese liked framework. We choose the VGG19 as the base network to extract the features of images from two domains and use the triplet loss function to measure their similarities. Moreover, we propose another strategy using U-Net to enhance the quality of images and then conduct image retrieval. The procedure of these two strategies is shown in Fig. 3. The U-Net architecture contains a convolutional auto-encoder with lateral connections between corresponding layers from the encoder to the decoder. It can be trained end-to-end with very few images and performs significant result in segmentation task. In this paper, we use U-Net to extract the meaningful information from the SoC impressions, which is then converted to binary images (Bi-SoC) with the environmental noises removed. Implementation details are shown in Section IV-B. Though some valuable information are lost through this procedure inevitably, the profit is markedly enormous as the cross-domain problems are solved. We compared this strategy with the origin one to demonstrate its significance. Note that this strategy only works in certain conditions since the U-Net network requires precise pixel-level masks of shoe-print images in the training steps.

B. MULTI-PART WEIGHTED CNN

The architecture of the proposed multi-part weighted convolutional neural network (MP-CNN) is illustrated in Fig. 2. It is a triplet training network with three branches which share the same parameter set. Each branch contains a VGG19

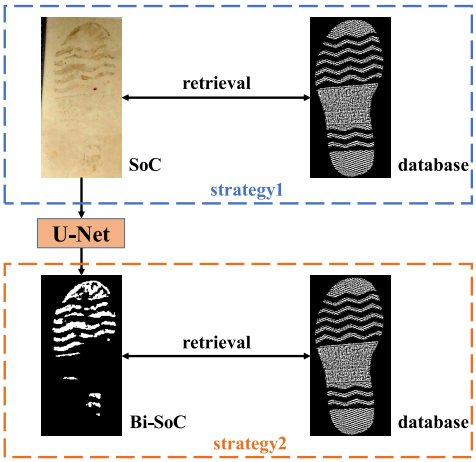


FIGURE 3. The procedure of two aforementioned strategies. Strategy 1 (in blue dashed box) directly uses the SoC impressions to match the images in the database. Strategy 2 (in red dashed box) uses U-Net to convert the SoC impressions to Bi-SoC images and then conduct image retrieval.

network with a single fully-connected (FC) layer which extracts features and converts them into a 128-dimensional vectors. Three input images correspond to an anchor SoC impression, a positive database image and a negative database image, respectively. The anchor-positive-negative triplet which is defined as $\{\boldsymbol{X}^o, \boldsymbol{X}^+, \boldsymbol{X}^-\}$ is selected according to the matching relationship, *i.e.*, the true match image is the positive and any false match can be used as the negative. Take the anchor branch for example, given an input image $\boldsymbol{X} \in \mathbb{R}^{W\times H}$ (W and H are the width and height of $\boldsymbol{X}$ respectively), we first divide it vertically into two sub-parts $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{W\times(H/2)}$ and extract the deep features by feeding them into convolutional layers and the FC layer individually. The extracted sub-features are denoted as $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. Meanwhile, we set an extra stream which takes the anchor impression $\boldsymbol{X}^o$ as input and calculate a weight matrix $\boldsymbol{W}^o = \{W_1^o, W_2^o\}$ depending on the amount of information these two sub-parts contains. The weight matrix $\boldsymbol{W}^o$ can be defined as:

$$W_i^o = \frac{g(\boldsymbol{x}_i^o)}{\sum_{j=1}^{2} g(\boldsymbol{x}_j^o)}, \quad i = 1, 2, \tag{1}$$

where $g(\cdot)$ computes the percentage of informative pixels which is defined as

$$g(\boldsymbol{x}_i^o) = \frac{\sum_{j=1}^{W}\sum_{k=1}^{H/2} \boldsymbol{x}_{i,j,k}^o}{W \times H/2}. \tag{2}$$

Intuitively, $\mathbf{W}$ reflects the importance weight of two sub-parts. We multiply the sub-features with the weight matrix and concatenate them to generate the final representation feature of $\mathbf{X}$ as

$$\boldsymbol{f}^\xi = [\boldsymbol{f}_1^\xi \cdot W_1^o, \boldsymbol{f}_2^\xi \cdot W_2^o]^{\mathrm{T}}, \quad \xi = o, +, -. \tag{3}$$

With the same process on other two branches we obtain the triplet feature $\{\boldsymbol{f}^o, \boldsymbol{f}^+, \boldsymbol{f}^-\}$ and send them into the triplet loss metric. The similarities between the triplet images are measured by the Euclidean distances between $\boldsymbol{f}^o, \boldsymbol{f}^+, \boldsymbol{f}^-$. The triplet loss function requires that distance of pair $(\boldsymbol{f}^o, \boldsymbol{f}^-)$ be

TABLE 1. Statistics of datasets.

Datasets	#query	#gallery
SPID	3959	3138
FID300	300	1175

larger than that of the pair $(\boldsymbol{f}^o, \boldsymbol{f}^+)$ by a predefined margin. In this work, the triplet loss is calculated as

$$L = d(\boldsymbol{f}^o, \boldsymbol{f}^+) - d(\boldsymbol{f}^o, \boldsymbol{f}^-) + \alpha, \tag{4}$$

where α is the predefined margin and the distance function $d(\cdot, \cdot)$ is the Euclidean distance defined as

$$d(\boldsymbol{f}^o, \boldsymbol{f}^+) = \|\boldsymbol{f}^o - \boldsymbol{f}^+\|^2. \tag{5}$$

The MP-CNN structure provides a large weight on the informative part while ignoring the scattered part in a SoC impression, through which the triplet loss function can focus on comparing the subtle difference of informative part on SoC impressions with database images.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. DATASETS AND SETTINGS

1) DATASETS

We evaluate the proposed MP-CNN network on two datasets. One is the shoe-print identification database (SPID). The other is the footwear impression database (FID300). Detailed information of the datasets is provided in Table 1. In shoe-print retrieval task, there exists many-to-many relationship between the SoC impressions and the database images. In SPID we group the SoC impressions with same matches into one class and get 1843 classes in total. Then we randomly divide them into training group with 1614 classes and test group with 229 classes. As for FID300, the probe images have been digitized with a scanner after being lifted with a gel foil from the ground. Therefore, the relationship turns to be many-to-one, *i.e.*, every query only has one image matched in the gallery. So we simply using an 8/2 splitting ratio for training and test.

2) IMPLEMENTATION DETAILS

To make fair comparison, we conducted experiments with the same settings. Especially, in order to maintain the rough ratio of the Shoe-print images, we resized every input to a size of 128×256 and conducted flipping with 0.5 probability. Next, we extracted features with VGG19 pretrained on the ImageNet classification dataset. We added a fully-connected (FC) layer and converted the features into a 128-dimensional vectors. We chose the batch-all strategy for triplet loss function and set the margin α in Equation (4) as 0.3. The learning rates of all layers are initially set as 0.001, and it is multiplied by 0.9 every 20 epochs. We trained our model for 500 epochs and the weight decay value is kept as 1×10^{-4}.

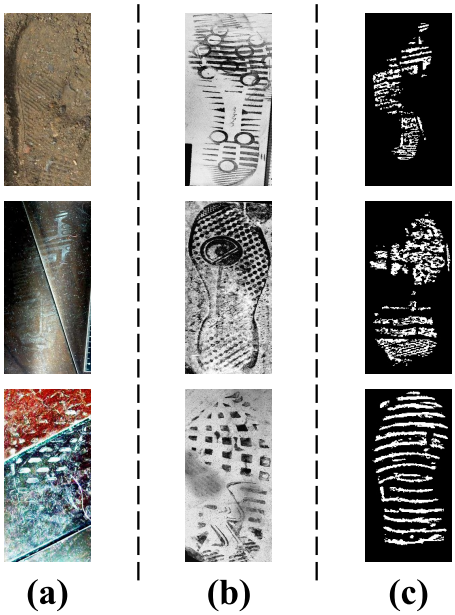


FIGURE 4. Example of (a) the SoC impressions in SPID, (b) the probe shoe-print images in FID300, and (c) the Bi-SoC images obtained by strategy 2.

B. PERFORMANCE COMPARISON WITH TWO STRATEGIES

The results on SPID are displayed in Table 2. Two typical baseline models in FG-SBIR task were chosen for comparison, namely the Sketch-a-Net (SAN) and the deep spatial-semantic attention (DSSA) network. We used the ratio of correctly predicting the true match at top1%, top5% and top10% as the evaluation metrics. Test time argumentation (TTA) performs random modifications to the test images, and takes the average of the predictions of each corresponding image as the final prediction. We both test our model with/without TTA operation to judge its performance in this task. The results suggest that 1) strategy 2 using U-Net to segment shoe-print from background which convert SoC to Bi-SoC can largely improve the performance on all the algorithms (SAN, DSSA, and MP-CNN) and 2) the proposed model significantly outperforms all the baseline models on all the evaluation metrics. The MP-CNN is trained with strategy 1 achieves the best accuracy of 62.69% at top5% while MP-CNN trained with strategy 2 outperforms corresponding baseline by a margin of 10%.

The results obtained by two proposed strategies demonstrate that strategy 2 is not absolutely better than strategy 1. Though the U-Net converts SoC to Bi-SoC which helps in reducing the domain gap between the query images and the gallery images, some valuable information are lost during this procedure. Note that we conduct image annotation on SoC impressions to get the dataset for the U-Net training. Hence the performance largely depends on the amount of training data and the annotation quality. In Fig. 5 (a), we visualize the top-15 retrieved test impressions on our MP-CNN model from SPID.

TABLE 2. Experimental results on the SPID dataset, the best results are marked in bold fonts.

SoC	top1%	top5%	top10%	top1%_TTA	top5%_TTA	top10%_TTA
SAN(strategy1)	18.41	42.79	58.21	23.88	45.77	56.22
DSSA(strategy1)	22.89	52.74	67.16	24.88	57.71	70.65
MP-CNN(strategy1)	26.87	**58.71**	69.15	32.84	**62.69**	73.63
SAN(strategy2)	19.21	43.67	60.70	24.02	49.78	64.19
DSSA(strategy2)	24.45	46.72	65.50	27.07	52.40	64.19
MP-CNN(strategy2)	**28.38**	55.46	**70.74**	**37.12**	59.83	**75.11**

TABLE 3. Experimental results on FID300 dataset, the best results are marked in bold fonts.

FID	top1%	top5%	top10%
ACCV14	27.10	59.40	74.40
BMVC2016	21.50	47.00	57.80
LoG2016	58.37	71.44	79.28
MCNCC(BMVC17)	72.55	82.30	87.45
MCNCC(IJCV19)	**79.67**	**86.33**	89.30
Ours(strategy1)	51.00	72.00	79.00
Ours(strategy2)	47.33	66.67	77.67
Ours(w retrain)	61.02	81.36	**89.83**

C. EXPERIMENTS ON FID300

In addition to the internal dataset SPID described in Section IV-B, we also evaluated our approach on a publicly available dataset (FID300) [4]. This database contains 1175 gallery and 300 probe shoe-print images. The results on FID300 are displayed in Table 3. Note that the probe shoe-print images in FID300 have been digitized with a scanner after being lifted with a gel foil from the ground, which is different from SPID. In this case, we conducted two experiments for comparison. Firstly we directly tested the pre-trained models on the FID300 benchmark without retraining (refer to "MP-CNN w/o retrain" in Table 3). In addition, we divided FID300 using an 8/2 splitting ratio and trained it by ourselves (refer to "MP-CNN w retrain" in Table 3). The MP-CNN achieves 61.02%, 81.36%, 89.83% accuracies on top1%, top5%, top10% evaluations, respectively. The results significantly outperform existing approaches (ACCV [4], BMVC [7], LoG [36]) with a considerable margin. At the meantime, it also achieves an competitive accuracy compared with MCNCC [10], [11]. One reason that the MP-CNN does not perform the best on the FID300 dataset can be explained by the retraining operation. If we test the model without retraining, the results largely depend on the similarity between the training dataset and the test dataset. The query images in FID300 can be considered as the middle state from SoC to Bi-SoC (shown in Fig. 4) which has effect on the performance. In Fig. 5 (b), we visualize the top-15 retrieved test impressions for a subset of crime scene query prints from FID300.

TABLE 4. Effect of two strategies, the results presented in top1% accuracy.

	SAN	DSSA	MP-CNN
strategy1	18.41	22.89	26.87
strategy2	**19.21**	**24.45**	**28.38**

D. ABLATION STUDY

We conduct ablation studies to show the effect of our proposed multi-part weighted structure and U-Net based strategy. The experiments are based on the SPID dataset.

1) EFFECT OF STRATEGY 2

We investigate the effect of Section III-A by adopting two strategies on different networks (SAN, DSSA, MP-CNN). The results are displayed in Table 4. Strategy 2 uses U-Net to convert the SoC images to Bi-SoC images. It largely reduces the difficulty on comparing the similarity with images from two domains which are extremely different. The top1% accuracy has been increased by 0.80%, 1.56% and 1.51% on three test structures, respectively.

2) EFFECT OF MULTI-WEIGHTS

In order to judge the influence of multi-weights on two sub-features, we set 1) MP-CNN with equal weight on two sub-features which are set as 1. 2) MP-CNN with a weight matrix $\mathbf{W}^o$ mentioned in Section III-B. The results are displayed in Table 5. The MP-CNN with weighted features achieves a overall improvement upon the one with equal

(a)

(b)

FIGURE 5. Retrieval results on (a) SPID and (b) FID dataset. Red boxes indicate the corresponding ground truth test impression.

TABLE 5. Effect of weight on sub-features.

Method	top1%	top5%	top10%
MP-CNN	28.38	55.46	70.74
MP-CNN(weighted)	**33.19**	**60.70**	**75.55**

weight by 4.81%, 5.24% and 4.81% on top1%, top5%, top10% accuracy, respectively. The results confirm that, with different weight on sub-features, the network can concentrate on the informative part of the input images, and therefore, learns more discriminative representation which achieves better performance.

V. CONCLUSIONS

In this paper, we propose a novel multi-part weighted convolutional network (MP-CNN) for shoe-print image retrieval. By extracting the sub-features individually using a two-stream structure, we multiply them with the importance weights which reflect the amount of meaningful information contained by their corresponding sub-images. The proposed MP-CNN pays more attention on the subtle differences from the informative part of images. In addition, we adopt a new strategy to deal with this task that can reduce domain gap by removing background noise from the crime scene images. Experimental results on two datasets demonstrated the good performance of our methods. In the future, we will further study the proposed MP-CNN on two directions: 1) adopting flexible division strategies upon different inputs instead of the fixed two-parts division, and 2) extending our framework to other tasks such as the sketch based image retrieval problem.

REFERENCES

[1] I. Rida, S. Bakshi, H. Proença, L. Fei, A. Nait-Ali, and A. Hadid. (2019). "Forensic shoe-print identification: A brief survey." [Online]. Available: https://arxiv.org/abs/1901.01431
[2] L. Zhang and N. Allinson, "Automatic shoeprint retrieval system for use in forensic investigations," in *Proc. UK Workshop Comput. Intell.*, vol. 99, 2005, pp. 137–142.
[3] P. M. Patil and J. V. Kulkarni, "Rotation and intensity invariant shoeprint matching using Gabor transform with application to forensic science," *Pattern Recognit.*, vol. 42, no. 7, pp. 1308–1317, 2009.
[4] A. Kortylewski, T. Albrecht, and T. Vetter, "Unsupervised footwear impression analysis and retrieval from crime scene data," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 644–658.
[5] X. Li, M. Wu, and Z. Shi, "The retrieval of shoeprint images based on the integral histogram of the gabor transform domain," in *Proc. Int. Conf. Intell. Inf. Process.*, 2014, pp. 249–258.
[6] C.-H. Wei and C.-Y. Gwo, "Alignment of core point for shoeprint analysis and retrieval," in *Proc. Int. Conf. Inf. Sci., Electron. Elect. Eng.*, vol. 2, Apr. 2014, pp. 1069–1072.
[7] A. Kortylewski and T. Vetter, "Probabilistic compositional active basis models for robust pattern recognition," in *Proc. BMVC*, 2016, pp. 1–12.
[8] S. Almaadeed, A. Bouridane, D. Crookes, and O. Nibouche, "Partial shoeprint retrieval using multiple point-of-interest detectors and SIFT descriptors," *Integr. Comput. Aided Eng.*, vol. 22, no. 1, pp. 41–58, 2015.

[9] S. Alizadeh and C. Kose, "Automatic retrieval of shoeprint images using blocked sparse representation," *Forensic Sci. Int.*, vol. 277, pp. 103–114, Aug. 2017.

[10] B. Kong, J. Supancic, III, D. Ramanan, and C. Fowlkes, "Cross-domain forensic shoeprint matching," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–17.

[11] B. Kong, J. Supančič, III, D. Ramanan, and C. C. Fowlkes, "Cross-domain image matching with deep feature maps," *Int. J. Comput. Vis.*, Jan. 2019.

[12] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[14] T. Senlet, T. El-Gaaly, and A. Elgammal, "Hierarchical semantic hashing: Visual localization from buildings on maps," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 2990–2995.

[15] D. Costea and M. Leordeanu. (2016). "Aerial image geolocalization from recognition and matching of roads and intersections." [Online]. Available: https://arxiv.org/abs/1605.08323

[16] M. Divecha and S. Newsam, "Large-scale geolocalization of overhead imagery," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2016, p. 32.

[17] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 124:1–124:10, Dec. 2009.

[18] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Trans. Graph.*, vol. 30, no. 6, p. 154, Dec. 2011.

[19] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales, "Automatic alignment of paintings and photographs depicting a 3D scene," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Nov. 2011, pp. 545–552.

[20] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 799–807.

[21] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5551–5560.

[22] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[24] R. E. Gaensslen, R. Ramotowski, and H. C. Lee, *Advances in Fingerprint Technology*. Boca Raton, FL, USA: CRC Press, 2001.

[25] P. De Chazal, J. Flynn, and R. B. Reilly, "Automated processing of shoeprint images based on the Fourier transform for use in forensic science," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 341–350, Mar. 2005.

[26] M. Gueham, A. Bouridane, D. Crookes, and O. Nibouche, "Automatic recognition of shoeprints using Fourier-mellin transform," in *Proc. NASA/ESA Conf. Adapt. Hardw. Syst.*, Jun. 2008, pp. 487–491.

[27] X. Kong, C. Yang, and F. Zheng, "A novel method for shoeprint recognition in crime scenes," in *Proc. Chin. Conf. Biometric Recognit.*, 2014, pp. 498–505.

[28] M. Pavlou and N. M. Allinson, "Automatic extraction and classification of footwear patterns," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, 2006, pp. 721–728.

[29] S. Rathinavel and S. Arumugam, "Full shoe print recognition based on pass band DCT and partial shoe print identification using overlapped block method for degraded images," *Int. J. Comput. Appl.*, vol. 26, no. 8, pp. 16–21, 2011.

[30] C.-H. Wei, Y. Li, and C.-Y. Gwo, "The use of scale-invariance feature transform approach to recognize and retrieve incomplete shoeprints," *J. Forensic Sci.*, vol. 58, no. 3, pp. 625–630, 2013.

[31] X. Wang, H. Sun, Q. Yu, and C. Zhang, "Automatic shoeprint retrieval algorithm for real crime scenes," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 399–413.

[32] L. Ghouti, A. Bouridane, and D. Crookes, "Classification of shoeprint images using directional filter banks," in *Proc. IET Int. Conf. Vis. Inf. Eng.*, 2006, pp. 167–173.

[33] A. Bouridane, A. Alexander, M. Nibouche, and D. Crookes, "Application of fractals to the detection and classification of shoeprints," in *Proc. Int. Conf. Image Process.*, vol. 1, Sep. 2000, pp. 474–477.

[34] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," Queen Mary Univ. London, London, U.K., Tech. Rep., 2014.

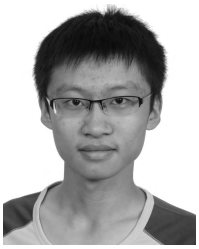
[35] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, p. 119, Jul. 2016.

[36] A. Kortylewski, "Model-based image analysis for forensic shoe print recognition," Ph.D. dissertation, Dept. Math. Comput. Sci., Univ. Basel, Basel, Switzerland, 2017.

ZHANYU MA received the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH Royal Institute of Technology. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He has also been an Adjunct Associate Professor with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics. He is a Senior Member of IEEE.

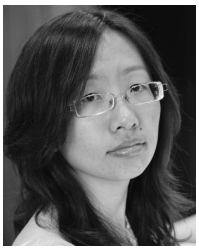
YIFENG DING received the B.E. degree in information engineering from the Beijing University of Posts and Telecommunications (BUPT), China, in 2018, where he is currently pursuing the degree. His research interests include pattern recognition and deep learning.

SHAOGUO WEN received the B.E. degree in information engineering from the Beijing University of Posts and Telecommunications (BUPT), China, in 2017, where he is currently pursuing the degree. His research interests include image processing, pattern recognition, machine learning, and deep learning.

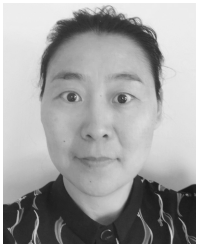


JIYANG XIE received the B.E. degree in information engineering from the Beijing University of Posts and Telecommunications (BUPT), China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in image processing, data mining, and deep learning.

YIFENG JIN received the master's degree in mechanical engineering from Hunan University, China, in 2012. In 2012, he joined the Institute of Forensic Science, Ministry of Public Security. His research interests include marks examination, marks informatization, crime scene investigation, and crime scene reconstruction.



ZHONGWEI SI received the Ph.D. degree from the KTH Royal Institute of Technology, Sweden, in 2013. In 2013, she joined the Beijing University of Posts and Telecommunications, where she is currently an Associate Professor. Her research interests include wireless communication, information theory, and data mining.



HAINING WANG received the MBA degree from the Xi'an University of Technology, China. In 2005, she joined the People's Public Security University of China. From 2010 to 2011, she was a Visiting Scholar with Ohio University. Her research interests include scientific decision making, public security informationization, and information theory.