https://doi.org/10.2352/ISSN.2470-1173.2018.15.COIMG-201
© 2018, Society for Imaging Science and Technology

Deep Gang Graffiti Component Analysis

He Li, Joonsoo Kim, Edward J. Delp
Video and Image Processing Laboratory (VIPER), School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA

Abstract

Gangs are a serious threat to the public safety in the United States. We have developed a system known as Gang Graffiti Automatic Recognition and Interpretation (GARI) to help law enforcement identify, track, and analyze gang activities. Gang graffiti components are the segmented graffiti content including symbols, digits, and characters. In this paper, we propose a deep convolutional neural network to classify the graffiti components. We make a comparison between our proposed deep learning method and our previous traditional method. Experimental results show the proposed method reaches 89.3% accuracy with dropout regularization.

Introduction

Gangs pose a serious issue to public safety in the United States. Gangs migrate from city to suburb constantly. Gang members usually paint graffiti onto different types of surfaces to claim territory, exchange information, or battle rival gangs. Law enforcement is interested in automatic ways to help investigate gang activities. We developed a system known as Gang Graffiti Automatic Recognition and Interpretation (GARI) [1] based on mobile device and back-end server that helps law enforcement to identify, track, and analyze gang activities. The image analysis of GARI system includes segmentation, matching, retrieval, and classification of gang graffiti images and gang tattoo images [1, 2, 3, 4]. In this paper, we propose a new method for gang graffiti component classification for the GARI system. A gang graffiti component is defined as a segmented individual "graffiti content" that can include digits, characters, and symbols. Examples of gang graffiti components can be found in Figure 1. Note that the gang graffiti components are hand drawn (or hand sprayed). The graffiti components can be further interpreted as semantic resources to form a language among gangs. Analyzing the gang graffiti components can help law enforcement have a better understanding of gang activities and regional situations so they can make a response. Our previous work [5] takes a query image as input and segments out each individual component into black and white image. This paper emphasizes on classifying the segmented graffiti components into different classes.

Background and Related Work

We take the assumption that the individual gang graffiti components have been segmented. After the individual graffiti components are segmented, we want to classify the components into different classes. Many methods have been proposed for general image classifications. Traditional methods employ some forms of feature extraction, typically using hand-crafted image feature descriptors, such as Scale Invariant Feature Transform (SIFT) [6], Histogram of Oriented Gradients (HOG) [7], and Speeded up Robst Features (SURF) [8]. The codebook based bag-of-feature model along with a spatial pyramid matching strategy [9] has been a key success in image classification for many years. In bag-of-feature methods, a codebook is formed by extracting local features from a set of database images and clustering through K-means [10]. Even with the spatial pyramid matching, a large amount of spatial information is still lost due to the quantization effect of the bag-of- features. Recent research in deep convolutional neural network (CNN) has achieved impressive results in segmentation, object detection, and image classification [11, 12, 13, 14, 15]. A CNN exploits spatial correlation and homogeneity of low/mid/high level features of natural images [12]. These features can be enriched by stacking convolution layers. Many deep neural network architectures have been proposed to improve the classification accuracy, including Alexnet [13] and Resnet [14].

In this paper, we propose a CNN architecture for gang graffiti component classification. The difficulty in our application scenario is all of our images are the real gang graffiti images and they are collected by police officers. There is a lack of large data sets and ground truth data for the CNN. We argue in the paper even with small amounts of data, using data augmentation and regularization to address overfitting can still obtain good classification accuracy. Section describes our previous method and our proposed new method for gang graffiti component classification. Section discusses the results between the two methods. We additionally compare our proposed method with some benchmark tests on the CIFAR10 [16] dataset for generality purpose. We draw a conclusion in section .

Gang Graffiti Component Classification

Overview

The goal of our method is to classify query images into different classes based on a set of gang graffiti components. This process is known as "Gang Graffiti Component Classification". For our study here, 14 classes are created from real graffiti images including digit *0*, *1*, *2*, *3*, *4*, *8*, character *E*, *G*, *s*, *x*, symbol *5-point star*, *6-point star*, *arrow*, *pitchfork*. Figure 1 shows examples of the query images. We describe our previous method and our proposed deep learning method for classifying gang graffiti component and make a comparison.

Figure 1: Sample images for each class.

Previous Method

Our previous method [5] uses SIFT keypoints based spatial location to create Local Shape Context (LSC) descriptors described in [17] and [4]. LSC takes the spatial location of the SIFT points and puts them into different bins to create a histogram. The bins of LSC are large enough to compensate for the local shape distortions and orientation variations [5]. The SIFT-based LSC are used along with hierarchical k-means clustering to build the vocabulary tree [18, 19] classifier shown in Figure 2. This approach is similar to the bag-of-feature approach since the vocabulary tree represents the codebook.

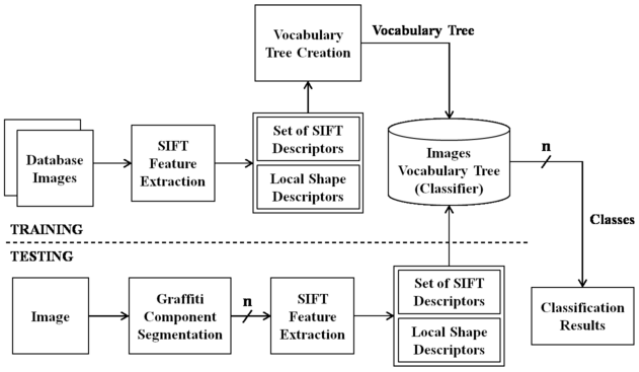

Figure 2: Our Previous Method for Gang Graffiti Component Classification

Data Augmentation

For our proposed method, since our dataset only contains 257 images for training, we need to generate more data in order for the CNN perform well. We use a series of operations in the data augmentation process, including Gaussian blur, image sharpening, add values, multiplications, contrast normalization, rotation. Detailed parameters to each operation can be found in Table 1. All operations are randomly selected and applied to the training images. For every training image, we create 500 augmented images. In total we obtain 128757 images for training.

Operation	Parameter
Gaussian Blur	$\sigma \in (1.0, 3.0)$
Image Sharpening	$\alpha \in (0, 1.0)$
Add Value	$\delta \in (-10, 10)$
Multiplication	$\beta \in (0.5, 1.5)$
Contrast Normalization	$\varepsilon \in (0.5, 1.0)$
Rotation	$\theta \in [90°, -90°, 180°, -180°]$

Table 1: Parameters used for Data Augmentation. σ is variance, α is sharpening factor, δ is added value, β is multiplication factor, ε is contrast normalization ratio, θ is rotation angle

Proposed Method

We propose a Deep Convolutional Neural Network architecture for classifying the gang graffiti components. Our network consists of 5 convolutional layers followed by 3 fully connected layers and an output softmax layer. Table 2 shows the detailed information about each layer. The first two convolutional layers (**conv1** and **conv2**) are followed by a max pooling layer (**pool1** and **pool2**) and normalization layer (**norm1** and **norm2**). The Rectified Linear Units (ReLU) [20] nonlinearity has been used to model the neuron's output. Let x be the input and f be the output, then the nonlinearity described by $f(x) = max(0,x)$ has advantages over $tanh(x)$ or *sigmoid* since it provides a constant learning rate over $x > 0$ while the others see significant drop of learning rate after some iterations of training. Many normalization techniques have been proposed to act like some sort of regularization. Our normalization use local response normalization in some layers after applying ReLU nonlinearity. According to [13], local response normalization implements a form of lateral inhibition similar to the real neurons, therefore creating better performed outputs with different kernel size.

Proposed Network Layers
conv1 (11,11,1,64)
pool1
norm1
conv2 (7,7,64,64)
norm2
pool2
conv3 (5,5,64,192)
conv4 (3,3,192,256)
conv5 (3,3,256,256)
pool3
fc1 (256,1152)
fc2 (1152,384)
fc3 (384,192)
softmax (14)

Table 2: For conv layers, the first two numbers are kernel size, the third number is the input channel and the last number is the output size. For fc layers, the first number is input and the second number is the output. For softmax we get the prediction for each class

During the training phase, 128757 images created from data augmentation are used. Every image is resized to 48x48 and changed to grayscale. These images are treated as database images. The images are then gone through the network. After 100000 iterations of training, we obtain the trained network. During the testing phase, each test image is also resized to 48x48 and then enters the trained network. The softmax layer will output 14 activation numbers and the class corresponding to the maximum activation will be selected as the predicted label. The process is illustrated in Figure 3.

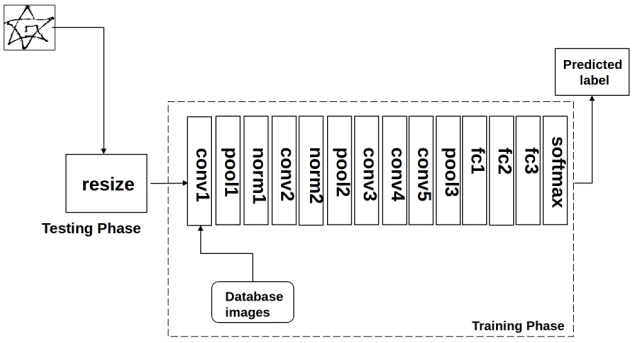

Figure 3: Proposed Convolutional Neural Network for Graffiti Component Classification

Experimental Results

For our previous method, using SIFT feature yields a 41.07% accuracy. By using LSC [5] we can achieve best accuracy 89.3%. For the proposed method, the precision and recall of the test images are listed in Table 3. The precision is computed by $\frac{\text{tp}}{\text{tp+fp}}$ and the recall is computed by $\frac{\text{tp}}{\text{tp+fn}}$, where tp is true positive, fp is false positive, fn is false negative [21]. The overall accuracy is 82.1%. Table 3 shows most of the classes can reach 100% precision. For the digit 3 and character *s*, the precision is lower.

label (class)	Size	Precision	Recall
0 (0)	4 (7.14%)	100%	100%
1 (8)	4 (7.14%)	100%	75%
2 (G)	4 (7.14%)	80%	100%
3 (3)	4 (7.14%)	66.7%	100%
4 (E)	4 (7.14%)	100%	50%
5 (s)	4 (7.14%)	66.7%	100%
6 (2)	4 (7.14%)	100%	75%
7 (1)	4 (7.14%)	80%	100%
8 (x)	4 (7.14%)	100%	100%
9 (5 point star)	4 (7.14%)	80%	100%
10 (6 point star)	4 (7.14%)	100%	75%
11 (4)	4 (7.14%)	100%	50%
12 (pitchfork)	4 (7.14%)	100%	100%
13 (arrow)	4 (7.14%)	100%	100%

Table 3: Precision and Recall of test data

We additionally use dropout to combat overfitting problem. Dropout is a simple technique to randomly disable (drop) some units during training time to prevent co-adapting [22]. We add dropout operations for layer **fc1** and **fc2**. Each of the dropout operation is controlled by a dropout rate that determines the probability of a unit being dropped. Table 4 shows the detailed experimental results of how different combinations of dropout rates affect the classification accuracy. It shows with dropout, the classification accuracy increases by about 7%.

d_2 \ d_1	0%	25%	50%	75%
0%	82.1%	82.1%	85.7%	87.5%
25%	**89.3%**	83.9%	83.9%	85.7%
50%	83.9%	85.7%	**89.3%**	85.7%
75%	83.9%	85.7%	87.5%	80.4%

Table 4: Classification accuracy with different combinations of dropout rates. d_1 denotes dropout rate for fc1, d_2 denotes dropout rate for fc2.

Since we do not have enough test images, to make our network fair, we also use it on another benchmark dataset. We choose CIFAR10 dataset [16] for our comparison. CIFAR10 dataset consists of 60000 32x32 RGB images in 10 classes. All the classes are uniformly distributed. There are 50000 training images and 10000 test images. We do not fine tune our network to adapt to CIFAR10 dataset because it defeats the purpose of generality. We only make the necessary modifications. We change the first layer to accept 3-channel inputs. We change the kernel size of the first 3 layers to adapt to the smaller image size. Specifically, we change the first layer to **conv1**$(5,5,3,64)$, the second layer to **conv1**$(5,5,64,64)$, and the third layer to **conv1**$(3,3,64,192)$. We train our network for 1 million epochs. We use the same dropout method to counter overfitting. We achieve 87% overall classification accuracy. Table 5 shows the comparison between our result and some benchmark results. It shows even without heavy modification, our network still produces competitive results.

Method	Accuracy
Resnet [14]	93.57%
DSN [23]	91.78%
Maxout Network [24]	90.65%
Alexnet [13]	89%
Our Method	87%
Stochastic Pooling [25]	84.87%
Ex-CNN [26]	84.3%

Table 5: Benchmark results

Conclusion

We proposed a deep convolutional neural network to classify gang graffiti components. We compared our results with our previous method. It showed with regularization techniques like dropout, we can easily achieve competitive results. We also show our network is not just tuned for one dataset, it is suitable for other datasets such the CIFAR10 dataset with minor modification.

References

[1] A. Parra, B. Zhao, J. Kim, and E. Delp, "Recognition, segmentation and retrieval of gang graffiti images on a mobile device," *Proceedings of the IEEE International Conference on Technologies for Homeland Security*, pp. 178–183, November 2013, waltham, MA.

[2] J. Kim, H. Li, J. Yue, and E. Delp, "Tattoo image retrieval for region of interest," *Proceedings of the IEEE Symposium on Technologies for Homeland Security*, pp. 1–6, April 2016, waltham, MA.

[3] J. Kim, H. Li, J. Yue, J. Ribera, E. Delp, and L. Huffman, "Automatic and manual tattoo localization," *Proceedings of the IEEE Symposium on Technologies for Homeland Security*, pp. 1–6, April 2016, waltham, MA.

[4] J. Kim, A. Parra, J. Yue, H. Li, and E. Delp, "Robust local and global shape context for tattoo image matching," *Proceedings of the IEEE International Conference on Image Processing*, pp. 2194–2198, September 2015, quebec City, Canada.

[5] A. Parra, "Integrated mobile systems using image analysis with applications in public safety," Ph.D. dissertation, Purdue University, West Lafayette, IN, May 2014.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, June 2005, san Diego, CA.

[8] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Proceedings of the European Conference on Computer Vision*, pp. 404–417, May 2006, graz, Austria.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, June 2006, new York, NY.

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367, June 2010, san Fracisco, CA.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[12] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Proceedings of the European Conference on Computer Vision*, pp. 818–833, September 2014, zurich, Switzerland.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1097–1105, December 2012, Stateline, NV.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, las Vegas, NV.

[15] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," *Visual Communication and Image Representation*, vol. 41, pp. 406–413, November 2016.

[16] A. Krizhevsky, "Learning multiple layers of features from tiny images," April 2009, university of Toronto.

[17] E. Mortensen, H. Deng, and L. Shapiro, "A sift descriptor with global context," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 184–190, June 2005, san Diego, CA.

[18] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2161–2168, June 2006, washington, DC.

[19] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. Han, "Contextual weighting for vocabulary tree based image retrieval," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 209–216, 2011, barcelona, Spain.

[20] V. V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the International Conference on Machine Learning*, pp. 807–814, June 2010, haifa, Israel.

[21] D. M. W. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, December 2011.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[23] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 562–570, May 2015, san Diego, CA.

[24] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *Proceedings of the International Conference on Machine Learning*, pp. 1319–1327, June 2013, atalanta, GA.

[25] M. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," *Proceedings of the International Conference on Learning Representations*, vol. arXiv:1301.3557, May 2013, scottsdale, AZ.

[26] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, September 2016.

Author Biography

He Li received his BS in Computer Engineering from Purdue University (2014) and his Master in Electrical Engineering from Purdue University (2017). He has been working as an Algorithm Engineer on HDR algorithms at Omnivision Technology since then.

Joonsoo Kim received the B.S. in Electrical and Electronics Engineering (EE) and M.S. in Electrical and Electronics Engineering (EE) from Yonsei University, Seoul, Korea in 2003 and 2010 respectively. He also received the Ph.D. in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana in June 2017.

Edward J. Delp was born in Cincinnati, Ohio. He is currently The Charles William Harrison Distinguished Professor and Professor of Biomedical Engineering at Purdue University. His research interests include image and video compression, multimedia security, medical imaging, multimedia systems, communication and information theory. Dr. Delp is a Life Fellow of the IEEE, a Fellow of the SPIE, a Fellow of IS&T, and a Fellow of the American Institute of Medical and Biological Engineering.