USING PYRAMID OF HISTOGRAM OF ORIENTED GRADIENTS ON NATURAL SCENE TEXT RECOGNITION

Zhi Rong Tan, Shangxuan Tian, and Chew Lim Tan

Department of Computer Science, School of Computing, National University of Singapore
Computing 1, 13 Computing Drive, Singapore 117417
Email: {tzr1791, tians, tancl}@comp.nus.edu.sg

ABSTRACT

Because of the unconstrained environment of scene text, traditional Optical Character Recognition (OCR) engines fail to achieve satisfactory results. In this paper, we propose a new technique which employs first order Histogram of Oriented Gradient (HOG) through a spatial pyramid. The spatial pyramid can encode the relative spatial layout of the character parts while HOG can only include the local image shape without spatial relation. A feature descriptor combining these two can extracts more useful information from the image for text recognition. Chi-square kernel based Support Vector Machine is employed for classification based on the proposed feature descriptors. The method is tested on three public datasets, namely ICDAR2003 robust reading dataset, Street View Text (SVT) dataset and IIIT 5K-word dataset. The results on these dataset are comparable with the state-of-the-art methods.

Index Terms— Text recognition, Support vector machines, Shape, Feature extraction

1. INTRODUCTION

The proliferation of social media and technology has empowered the mass to communicate and exchange information via digital format. Aside from text, images are also taken regularly by users to share with others. This leads to large amount of images that are hard to organize or search for, hence a need for systems like automatic annotation and image retrieval to manage the data. Since many of these images include text, it's important to let the computer understand the text so as to easily annotate and organize such images. For example, in order to know the specific model of an airplane, we need to extract features and match the features with existing airplane models to decide. However, an easier way is to read the text on the airplane to get the model name. There are many works that adopt text recognition to help improve object recognition tasks [1] [2]. Unfortunately, those text images often comprise of natural scenes, which come in many colors, background noise, fonts, illuminations etc. as shown in Figure 1. Such unconstrained environment renders the traditional Optical Character Recognition (OCR) methods unable to work well since OCR requires texts appearing on a clean background and that the texts itself do not vary much in font, color, size and so on.

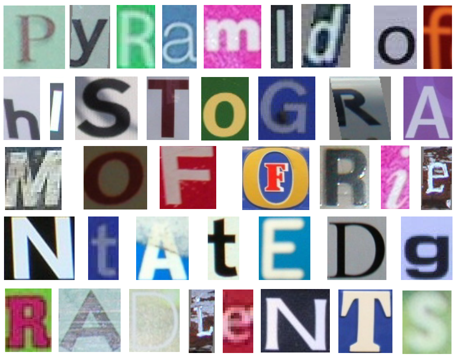

Figure 1: Images from the ICDAR 2003 character dataset

Considering these, it is necessary to solve the problem using a feature extraction technique that is robust to these variations in natural scenes. This paper outlines a feature extraction method that aims to take into account the spatial structure of the text images and find the similarities between the different shapes to distinguish the characters represented.

2. RELATED WORKS

Presently, the types of natural scene text recognition methods can be broadly grouped into those that require pre-processing like segmentation and those that do not. Amongst those that do not have pre-processing, an example is the cooperative multiple-hypothesis framework [3] which leverages on the current OCR engine and prunes unwanted detections and fills in the missing parts. An automatic recognition method based on convolutional neural network [4] is independent of pre or post processing or even tuning parameters. In a similar work, the pairing of multi-scale character recognition with linguistic knowledge [5] applied on a convolutional neural network is also independent of segmentation. For curved text recognition using Hidden Markov Models (HMM) [6], segmentation is also avoided. The Hough Forests [7] uses 'cross-scale binary features' for mapping of characters to omit text segmentation process.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

Still, there are a lot more text recognition that relies on the pre-processing of segmentation. Some examples are the Bayesian classifier with boundary growing method [8], hypotheses verification framework with the Maximally Stable Extremal Regions (MSERs) features [9], gradient direction features [10], region based and connected component based [11] methods. Another method in [12] proposes multiple segmentations using the Markov random field (MRF) to create multiple hypotheses, and background removal by using connected component analysis before applying grayscale consistency constraint on the text characters. The detection based on saliency cues and context fusion [13] is independent of tuning. More common features include Scale Invariant Feature Transform (SIFT), K-means clustering, k-nearest neighbor etc. A popular feature extraction method is the Histogram of Oriented Gradients (HOG) [14] feature that was adapted from human detection, but it lacks representation of the spatial information of an image. As such, variations to HOG, like integrating HOG with Boosted cascade and Waldboost classifier to generate a text confidence map [15], co-occurrence of the HOG [16], tries to overcome the limitation of the HOG features.

Therefore, this paper will explore another variation to the HOG method, by taking a pyramid of the HOG features at different levels of cell dimensions to represent the spatial-shape of the image.

3. PROPOSED METHOD

The Pyramid of Histogram of Oriented Gradients (PHOG) is a variation to the original HOG, so in the first part, we will briefly explain how HOG works and in the second part, we will show in detail how to extend PHOG and use it for scene text recognition.

3.1. Histogram of Oriented Gradients

The original HOG features were applied to human detection. Due to its robustness to illuminations and local geometric and photometric changes [16], it became widely used in object detection. HOG features consist of dividing the image into small cells (usually 8x8 pixels) and computing the magnitude of the orientations of pixels and interpolating them into a histogram of orientation bins of 20 degrees each [14]. After which, the cells are grouped into overlapping blocks and normalized, and finally the normalized histograms are joined together to form the features for an image as summarized in Figure 2. However, HOG is limited as it only accounts for the orientation of individual pixels, without regards to the spatial distribution of the image.

3.2. Pyramid of Histogram of Oriented Gradients

To better present the spatial relationship of the oriented gradients, the Pyramid of Histogram of Oriented Gradients (PHOG) [17] was proposed for object categorization.

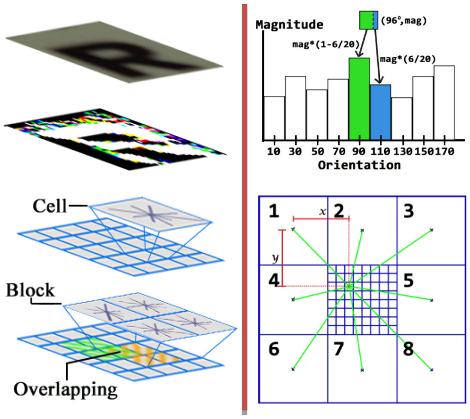

Figure 2: Brief overview of the HOG method on the left showing the process of getting the gradient of the image, dividing the image into cells and normalizing the overlapping blocks; right side illustrates bilinear followed by tri-linear interpolation performed between the orientation bins and across the cells respectively.

Inspired by this work, we propose to adopt PHOG to extract more information by encoding HOG feature in a spatial pyramid. The main idea for PHOG is to represent the image shape and its spatial layout so that the correspondence between two shapes can be calculated by chi-square kernel.

For a given image, instead of having fixed cell size as in HOG, PHOG divides the image into different cells at different pyramid levels, with 2^l x 2^l cells at the l^{th} level. The final feature of an image, as summarized in Table 1, will thereby be the HOG features for each of these levels combined together, having a total dimension of $K \cdot \sum_{l \in L} 4^l$, where $K = 20$ orientation bins, L is the total number of levels which is limited to no more than 3 to prevent over-fitting. From Table 2, it can be observed that PHOG features at each level for different characters are represented differently. The final feature is then normalized like in HOG to prevent unequal weighting for images with various illuminations and contrast.

Since PHOG divides the cell into different resolution that gets increasingly smaller to form the pyramid structure, the pyramid structure at the lower resolutions allows for focus onto the region of interest. Specifically, on top of the orientation information captured by HOG, the spatial matching [18] of the pyramid structures allows for the geometric matching of the orientations at finer resolution.

The variation made to the original PHOG method is to follow the original HOG method with the overlapping block normalization so that each cell can contribute to more than one component of the final feature where each cell is normalized to their respective different blocks [14]. Bi-linear interpolation of the pixels is also implemented for each level, whereas tri-linear interpolation kicks in when $l \geq 2$ and $cell - size \leq 8$, as interpolation helps to decrease the artifacts and distortions. The cell-size restriction is to limit the noise from pixels far from the neighboring cells.

In addition, no weights specific to the levels were given, because although lower levels has less dissimilar features, but they already weigh less in terms of feature dimension, hence additional weights to penalize the lower levels will make these features meaningless; but the lower levels are useful as they serve to capture the more general outline of the image and is more robust to noises in the image compared to the higher levels at lower resolution.

4. EXPERIMENTS

The support vector machine, Libsvm [19] is used for the training and testing.

4.1. Chi-Square Kernel

Instead of using SVM with the normal linear or radical-based function (RBF) kernel, the chi-square kernel is used. The chi-square kernel calculates the similarity between two image features, by taking two times the square difference divided by the sum of the two features,

$$k(x,y)=1-\sum_{i=1}^{n}\frac{2\cdot(x_i-y_i)^2}{(x_i+y_i)}$$

where x and y are the features. As PHOG features capture the spatial layout of an image, modeling the overlaps between features using chi-square kernel to encompass the relation between images will perform better than using linear and RBF kernel. Linear and RBF calculates Euclidean distance, thereby requiring the individual features to capture more correlational data of the image itself in order to map similar features together. The results of testing with other kernel are recorded in Table 3.

4.2. Dataset

The training data consist of ICDAR2003 [20], chars74k [21], IIIT5k [22] and synthetic data from computer fonts, amounting to a total of 24k characters. The testing datasets are from 3 sources: ICDAR2003, SVT and IIIT5k dataset, which contains around 5k, 3k and 15k characters respectively (ICDAR2003 and IIIT5k provides separate training and testing dataset). The labels consist of the 62 classes including digits, upper and lower case characters. Each of the images is then resized into 32x32 pixels.

4.3 Result

The testing result in Table 3 shows that PHOG performed comparatively close to the result obtained from Co-HOG. One main advantage compared with Co-HOG is that PHOG has a much smaller feature dimension (4K vs 100K) but it extracts an almost equal amount of feature information. Additionally, computation complexity is lower for PHOG over Co-HOG (O^2 vs O^3). This is crucial for real-time mobile applications which require less memory and processing time. Besides, the accuracy of all 3 datasets is

Table 1: PHOG combines the HOG features of the image at each level divided into different cells to represent the shape of the image

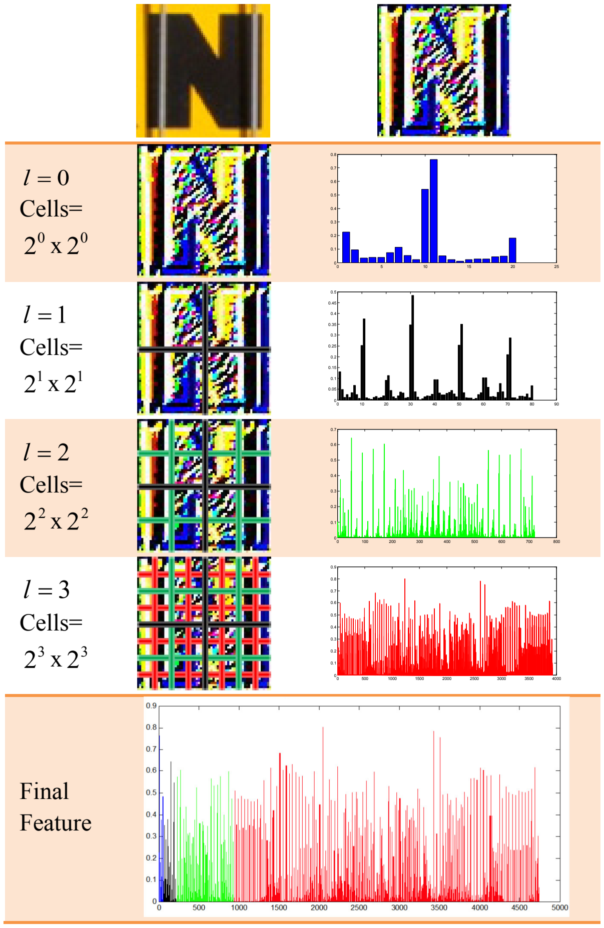

Table 2: PHOG features at different levels for different characters: different characters on similar background has vastly different PHOG shape features, whereas same character on different background have similar PHOG shape features

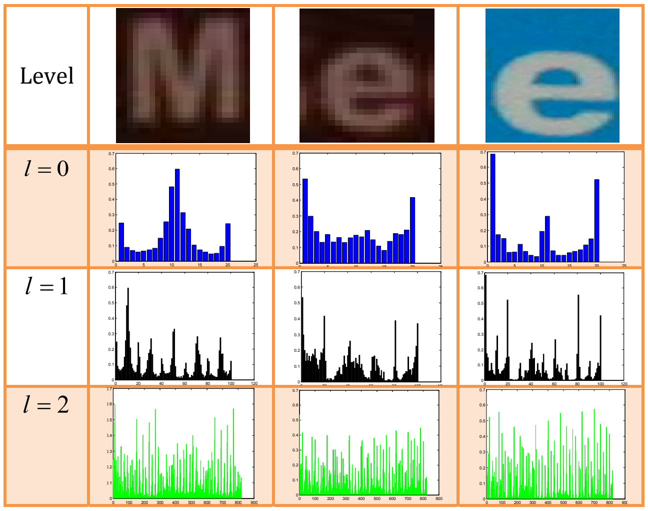

Table 3: Results on ICDAR2003, SVT and IIIT5k datasets

Testing Method	Testing Dataset Accuracy (%)		
	ICDAR	SVT	IIIT
ABBYY FineReader 10 [23]	26.6	15.4	33.9
GB+NN [24]	41.0	-	-
HOG+NN [25]	51.5	-	-
NATIVE+FERNS [26]	64.0	-	-
MSER [9]	67.0	-	-
HOG+SVM [27]	74.5	61.9	-
Co-HOG [16]	79.4	75.4	-
Co-HOG (case insensitive)[16]	83.6	80.6	-
PHOG (Linear Kernel)	76.5	76.4	72.6
PHOG (Chi-Square Kernel)	**79.0**	**74.7**	**75.8**
PHOG (Chi-Square Kernel) (case insensitive)	**82.7**	**80.1**	**81.7**

also comparatively close to each other, especially for case insensitive testing, showing that the method is stable across different natural scene text images.

One of the features that are hard to distinguish in natural scene text will be the upper and lower cases, especially for characters like ‘c’, ‘p’, ‘z’ etc. This can also be observed from the confusion matrix in Figure 3, where two distinctive light blue lines running parallel to the main line corresponds to the upper and lower casing classified wrongly into its counterpart. Hence, when tested without the case-sensitivity, the accuracy result shows marked improvement.

Another feature that is hard to distinguish is for ambiguous characters like ‘l’ (love), ‘I’ (Indigo) and ‘1’ (one) or ‘O’ (orange) and ‘0’ (zero) etc. It is made worst in natural scene text where characters come in various fonts which express these ambiguous characters differently. Such ambiguities can only be resolved if the context of the character is known, i.e. the whole word is needed.

The pre-processing segmentation is also very crucial to the recognition result. For the same image, if the segmentation is not appropriate, it can be labeled wrongly, as observed between Figure 4 and Figure 5, especially if the orientation is very skewed when the original character already has a noisy background, or the original character has abnormal shape (e.g. extra-long tail of ‘R’) and the segmentation did not bound the character properly.

Still, the recognition for some disjoint characters or characters in noisy background performed well. Some of these characters in Figure 4 are hard to even visually recognize due to poor lighting or low contrast etc. Whereas in Figure 5, some characters are labeled incorrectly due to the segmentation or ambiguity problem, while others are too distorted for visual recognition. Sometimes, the distortion could be due to the size of the original image, which might be very small. Hence, the resizing is constrained to 32x32 pixels, and by enlarging the images, it will inevitably lead to lose of image quality and hence affect the resultant feature.

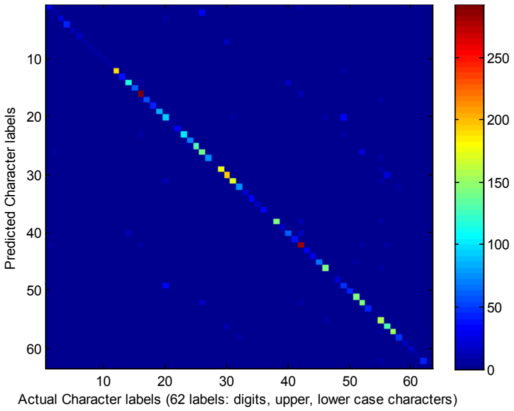

Figure 3: Confusion Matrix of the PHOG method on ICDAR2003 dataset (1 to 62 on the axes represent’0-9A-Za-z’ respectively) .

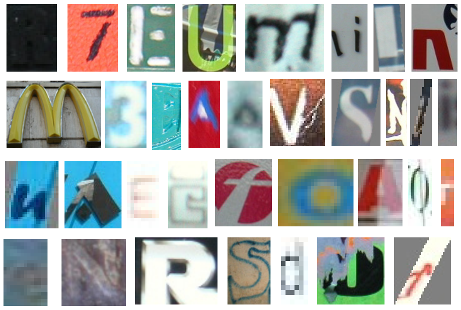

Figure 4: Examples of correctly labeled images.

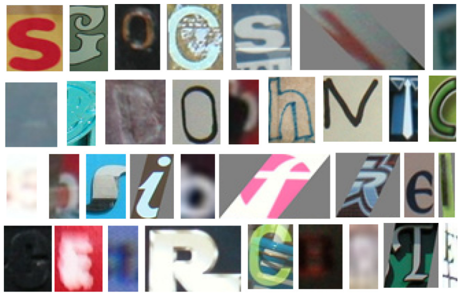

Figure 5: Examples of incorrectly labeled images.

5. CONCLUSION AND FUTURE WORK

The Pyramid of Histogram of Oriented Gradients (PHOG) breaks an image into levels of different cell size to encode the spatial structure and find the relevance between images using the chi-square kernel. The result is comparable to the current state-of-the-art feature extraction techniques. Although chi-square kernel can calculate similarity between images, the comparison could be improved if the spatial co-relation of each image is incorporated to further enhance the current performance. These possible future works to explore comprises of methods like encoding co-occurrence between the pixels [16], co-occurrence of adjacent local binary patterns [28] or using second order HOG [29].

6. REFERENCES

[1] Q. Zhu, M.C. Yeh and K.T. Cheng, “Multimodal fusion using learned text concepts for image categorization,” *ACM International Conference on Multimedia*, Santa Barbara, pp. 211–220, 2006.

[2] H.C. Wang and M. Pomplun, “The attraction of visual attention to texts in real-world scenes,” *Journal of Vision 12.6*, pp. 1-17, 2012.

[3] R. Huang, P. Shivakumara, Y. Feng, and S. Uchida, “Scene Character Detection and Recognition with Cooperative Multiple-Hypothesis Framework,” *IEICE TRANSACTIONS on Information and Systems Vol.E96-D No.10*, IEICE, Fukuoka, pp. 2235-2244, Oct. 1, 2013.

[4] Z. Saidane, and C. Garcia, “Automatic Scene Text Recognition using a Convolutional Neural Network,” *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR 2007)*, pp. 100-106, 2007.

[5] K. Elagouni, C. Garcia, F. Mamalet, and P. Sébillot, “Combining multi-scale character recognition and linguistic knowledge for natural scene text OCR,” *Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 120–124, Mar. 27-29, 2012.

[6] S. Roy, P.P. Roy, P. Shivakumara, G. Louloudis, C.L. Tan, and U. Pal, “HMM-based Multi Oriented Text Recognition in Natural Scene Image,” *ACPR2013*, 2013.

[7] G. Yildirim, R. Achanta, and S. Süsstrunk, “Text Recognition in Natural Images Using Multiclass Hough Forests,” *Proceedings of the 8th International Conference on Computer Vision Theory and Applications*, vol. 1, pp. 737-741, 2013.

[8] R. Sreedhar, T.Q. Phan, S. Lu and C. Tan, “Multi-Oriented Video Scene Text Detection through Bayesian Classification and Boundary Growing,” *Circuits and Systems for Video Technology, IEEE Transactions on (Volume:22 , Issue: 8)*, pp. 1227 – 1235, 2012.

[9] L. Neumann, and J. Matas, "A method for text localization and recognition in real-world images," *Proc. of the 10th Asian Conf. on Computer Vision*, Queenstown, New Zealand, pp. 770-783, Nov, 2010.

[10] A. Gonzalez, L.M. Bergasa, J.J. Yebes, and S. Bronte, “A Character Recognition Method in Natural Scene Images,” *21st International Conference on Pattern Recognition (ICPR 2012)*, Tsukuba, pp. 621-624, Nov. 11-15, 2012.

[11] R. Mohanabharathi, K. Surender, and C. Selvi, “Detecting and Localizing Color Text in Natural Scene Images Using Region Based & Connected Component Method,” *International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1*, pp. 331-335, Jan-Feb, 2013.

[12] J.M. Odobez, and D. Chen, “Robust Video Text Segmentation and Recognition with multiple Hypotheses,” *Proc. of IEEE International Conference on Image Processing 2002 vol. II.*, Rochester, NewYork, pp. 433-436, 2002.

[13] S. Karaoglu, J.C. Gemert, and T. Gevers, “Object reading: text recognition for object recognition,” *ECCV'12 Proceedings of the 12th international conference on Computer Vision - Volume Part III*, pp. 456-465, 2012

[14] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 886-893, June, 2005.

[15] Y.F. Pan, X. Hou, and C.L. Liu, “Text localization in natural scene images based on conditional random field,” *ICDAR 2009: Proc. of the 2009 10th International Conference on Document Analysis and Recognition*, pp. 6–10, 2009.

[16] S. Tian, S. Lu, B. Su, and C.L. Tan, “Scene Text Recognition using Co-occurrence of Histogram of Oriented Gradients,” *Document Analysis and Recognition (ICDAR), 2013 12th International Conference*, Washington, DC, pp. 912 – 916, Aug. 25-28, 2013.

[17] A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel,” *CIVR*, 2007.

[18] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” *CVPR*, 2006.

[19] C.C Chang and C.J Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology, http://www.csie.ntu.edu.tw/~cjlin/libsvm,* pp. 2:27:1--27:27, 2011.

[20] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “Icdar 2003 robust reading competitions,” *Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. 2*, pp. 682–687, 2003.

[21] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” *Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE*, pp. 1457–1464, 2011.

[22] A. Mishra, K. Alahari and C.~V. Jawahar, "Scene Text Recognition using Higher Order Language Priors," *MishraBMVC12,* 2012

[23] ABBYY FineReader 10, http://www.abbyy.com/.

[24] T. E. de Campos, B. R. Babu, and M. Varma, “Character recognition in natural images,” *VISAPP (2)*, pp. 273–280, 2009.

[25] K. Wang and S. Belongie, “Word spotting in the wild,” *Computer Vision–ECCV 2010*, pp. 591–604, 2010.

[26] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” *Computer Vision (ICCV), 2011 IEEE International Conference on IEEE*, pp. 1457–1464, 2011.

[27] A. Mishra, K. Alahari, and C. Jawahar, “Top-down and bottom-up cues for scene text recognition,” *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on IEEE*, pp. 2687–2694, 2012.

[28] R. Nosaka, Y. Ohkawa and K. Fukui, “Feature Extraction Based on Co-occurrence of Adjacent Local Binary Patterns,” Ho, Y.-S. (ed.) PSIVT 2011, Part II. LNCS, vol. 7088, pp. 82–91, 2011.

[29] H. Cao, K. Yamaguchi, T. Naito and Y. Ninomiya, "Pedestrian Recognition Using Second-Order HOG Feature," *ACCV'09 Proceedings of the 9th Asian conference on Computer Vision - Volume Part II*, pp. 628-634, 2009.