2011 International Conference on Document Analysis and Recognition

Multiscale Histogram of Oriented Gradient Descriptors for Robust Character Recognition

Andrew J. Newell
Department of Computer Science
University College London
a.newell@ucl.ac.uk

Lewis D. Griffin
Department of Computer Science
University College London
l.griffin@cs.ucl.ac.uk

***Abstract*—Characters extracted from images or graphics pose a challenge for traditional character recognition techniques. The high degree of intraclass variation along with the presence of clutter makes accurate recognition difficult, yet the semantic information conveyed by sections of text within images or graphics makes their recognition an important problem.**
Previous work has shown that, on the two most commonly used datasets of such characters, Histogram of Oriented Gradient (HOG) descriptors have outperformed other methods. In this work we consider two extensions of the HOG descriptor to include features at multiple scales, and evaluate their performance using characters taken from images and graphics. We demonstrate that, by combining pairs of oriented gradients at different scales, it's possible to achieve an increase in performance of 12.4% and 5.6% on the two datasets.

***Index Terms*—Histograms, Oriented Gradients, HOG, Character Recognition, Oriented Gradient Columns**

I. Introduction

When documents contain graphics or images, the text within these may contain a significant amount of semantic information. The extraction and subsequent recognition of this text is therefore of use to many modern applications. This work concentrates on the second of these aspects, namely the recognition of characters taken from images.

It is not straightforward to quantify the exact differences between characters taken from printed text and those taken from images in terms of inter and intraclass variation. However, we might imagine certain differences to be apparent. Whereas in a printed document the text has generally been placed with the intention to convey information, and therefore has some consideration for clarity, text found in natural images may have arisen through different motivations. For example, text may be used for aesthetic appeal or to capture attention. There may be greater variation in terms of presentation angle, lighting conditions and clutter, as well a far greater range of writing styles.

We formulate the problem as one of object categorisation with a substantial degree of intraclass variation, and thus it can be used to test methods of character recognition that are robust. In order to develop methods that are genuinely able to deal with this variation we concentrate on testing with a low number of training images per class.

The structure of this paper is as follows. First, we summarise the previous work that has been done on the recognition of characters from images and related work on Histogram of Oriented Gradient (HOG) descriptors. Next, we present our multsicale extensions to the HOG descriptor and describe how they can be applied to the problem of character recognition. These are then tested on two datasets and compared to the performance of other methods. We also provide a brief investigation into the sensitivity of the key parameters in our method. Finally, we offer a brief discussion and conclusions.

II. Related Work

A. Recognising Characters from Images

Many methods exist for extracting text from images, however individual character recognition has generally been approached by using methods that have previously shown success in recognising shape. In a comparison of several methods deCampos et al [6] showed that two such methods, Geometric Blur [2] and Shape Context [1], used in conjunction Nearest Neighbour (NN) classifier, performed better than other object recogntion methods such as SIFT [9] and leading optical character recognition software. In the same work the authors showed that a further improvement in performance could be gained when using a more advance machine learning apparatus, namely Multiple Kernel Learning (MKL)[12].

Wang et al subsequently showed[13], using the same evaluation framework as de Campos, that performance could be further improved by using Histograms of Oriented Gradients [4], in conjunction with an NN classifier.

B. Histograms of Oriented Gradients

Histograms of Oriented Gradients were initially described by Dalal et al. in the context of person detection in images [4] and in video [5]. Multiscale extensions to the HOG descriptor have been previously considered. For example, He et al. [7] demonstrated that, by combining HOG descriptors at multiple scales into a single encoding, that performance could be improved in the context of person detection. Felzenszwalb et al. used HOG pyramids, where HOG descriptors are combined at two scales, also showing improved performance in pedestrian detection. Similarly Bileschi [3] considered extensions that demonstrated improved performance over single scale HOG.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.219

III. METHODS

We consider two schemes. The first scheme simply extends the histograms across scale space, in a way related to previous multiscale methods. The second scheme incorporates incorporates a novel step, where pairs of oriented gradients across scale are combined to form features we refer to as oriented gradient columns. These are then used to produce histograms of oriented gradient columns.

For the first scheme, oriented gradients are calculated using Derivative-of-Gaussian filters. For each location in the image, at a given scale, a single orientation is assigned along with a weight, which is calculated from the response of the DoG filters. Next, for a given block size, we calculate the total strength for each orientation across the block and across all scales. This is repeated across multiple overlapping blocks within the image. Each histogram is then normalised so that the total weight across all orientations sums to one. All histograms are then concatenated to make a single descriptor for the image. The algorithm is given in Algorithm 1.

Algorithm 1 The simple multiscale HOG encoding

1) For a given scale, σ, measure filter responses $c_{1,0}$ and $c_{0,1}$ of 1st order derivative-of-Gaussian filters, and from these calculate the scale normalised filter responses $s_{i,j} = \sigma c_{ij}$
2) Assign orientation by quantising $atan2(s_{0,1}, s_{1,0})$
3) Calculate weight according to $\sqrt{s_{10}^2 + s_{01}^2}$
4) Repeat for range of σ
5) For each block in the image sum weights across all positions and all σ for each orientation and normalise
6) Concatenate all blocks in the image to make overall encoding

In the second scheme, oriented gradients are calculated at two scales, the base scale σ_{BASE} and a coarser scale, $r\sigma_{BASE}$, where r is referred to as the scale ratio. Then, for each location in the image, an orientation vector is assigned comprising the orientations at each scale, and a weight equal to the product of the weight at each scale. These features are referred to as oriented gradient columns. Histograms of these oriented gradient columns are then calculated across multiple blocks and base scales, and then normalised as before. The algorithm is described in Algorithm 2

For both schemes classification is performed with a Nearest Neighbour classifier using the Bhattacharyya distance [8].

Algorithm 2 The HOG Column encoding

1) For a given σ_{BASE} and scale ratio, r, measure filter responses $c_{1,0}$ and $c_{0,1}$ of 1st order derivative-of-Gaussian filters at scales σ_{BASE} and $r\sigma_{BASE}$, and from these calculate the scale normalised filter responses $s_{i,j} = \sigma c_{ij}$
2) Calculate quantised orientations, θ_1 and θ_2, according to $atan2(s_{0,1}, s_{1,0})$ at both scales
3) Compute weight, w_1 and w_2, according to $2\sqrt{s_{10}^2 + s_{01}^2}$ at both scales
4) Combine across scales to form a feature with orientation (θ_1, θ_2) and weight $w_1 w_2$
5) Repeat for range of σ_{BASE}
6) For each block sum weights across all positions and σ_{BASE} for each orientation vector and normalise
7) Concatenate all blocks in the image to make overall encoding

IV. RESULTS

A. Datasets

We tested the two schemes on the two most commonly used datasets. First, the chars74k dataset [6], which contains 62 classes consisting of digits and upper and lower case letters. Each image contains a single main character, although a visual inspection indicated that there was a substantial amount of clutter in the images and some had significant sections of other characters present. Example images from the chars74k dataset are shown in Figure 1.

The second dataset, referred to as ICDAR03-CH[10], is the character recognition dataset from the robust reading challenge from ICDAR 2003. The dataset is similar to the chars74k set, except for the inclusion of punctuation symbols.

B. Preprocessing

All images were to converted to grayscale and scaled so that they were all the same size. In order to be invariant to whether a character was dark on light or light on dark we performed a simple test using the Laplacian to see whether, at a coarse scale, the image tended to a dark patch on light or a light patch on dark, in which case images were inverted.

C. Dataset Splits

In order to be able to compare our results to previously published results we concentrated on testing the chars74k dataset using 5 and 15 training images per class, referred to as chars74k-5 and chars74k-15, and the ICDAR03-CH dataset with 5 training images per class.

For the chars74k the dataset was split by first selecting, at random, 30 images per class from which to draw training and test tests. The remaining images were used to tune the parameters. This tuning set contained a variable number of images per class, with some classes only containing a single image.

The ICDAR03-CH dataset comes split into training and test sets. For each run we selected 5 training images, at random, per class and used the whole test set. The same parameter values were used as in the chars74k testing.

D. Performance

The performance of the two schemes for each testing regime is given in Table I. This is given alongside previously published results, including SIFT and the OCR software ABBYY. Each score is the mean performance over 50 runs for the

Fig. 1. Example images from the chars74k dataset (top) and the ICDAR03-CH dataset (bottom).

chars74k dataset and 10 runs for the ICDAR03-CH-5. For both schemes, 16 orientations were used and the block size was set to 20 pixels, which was approximately half the object size, with an overlap of 15 pixels between neighbouring blocks. For the second scheme, the tuning process gave an optimal scale ratio of 3. For both schemes we used arithmetically spaced scales between 1 and 7.

From the table it can be seen that the first scheme offers a small improvement over the single scale HOG on the chars74k dataset, but a decrease in performance on the ICDAR03-CH dataset. The second scheme, using oriented gradient columns, shows an improvement in performance on both datasets.

The full performance graph for the chars74k dataset, with training sets of between 1 and 29 images per class, is shown in Figure 2 (a).

E. Confusion Matrix

The confusion matrix for the chars74k-15 problem is shown in Figure 3. It is interesting to note the lines running parallel to the main diagonal which show, for certain letters, a high level of confusion between upper and lower case examples. With the testing regime we've used, each image is tested in isolation. Therefore it is possible that with some letters, for example the letter 'c', that upper and lower case examples may be identical. In order to try gauge performance without this effect we also looked at how the two schemes compared when tested using only digits, or only upper or lower case letters. In each of these comparisons we would not expect examples from any two classes to be identical. These graphs are shown in Figure 2 (b) to (d).

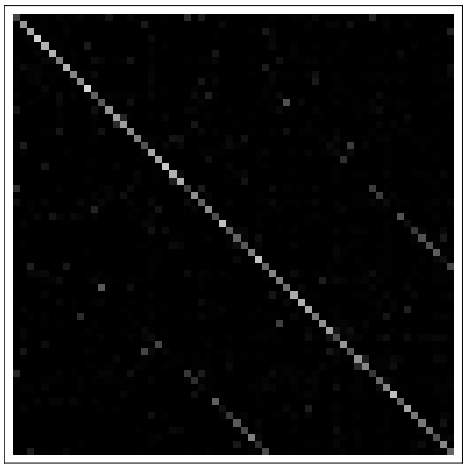


Fig. 3. The confusion matrix when using HOG Columns on the chars74k dataset with 15 training images per class.

F. Parameter Sensitivity

As our better performing scheme, using oriented gradient columns, used the scale ratio parameter that is not found in other implementations of HOG we were keen to see how it affected performance. To do this we used the chars74k-15 test and looked at how performance changed as we varied the scale ratio from 1, which is equivalent to the first of our schemes, up to 7. The results are shown in Figure 4. As the graph shows, there is a sharp increase in performance as the ratio increases above 1, with a peak at a scale ratio of 3, followed by a slow drop off.

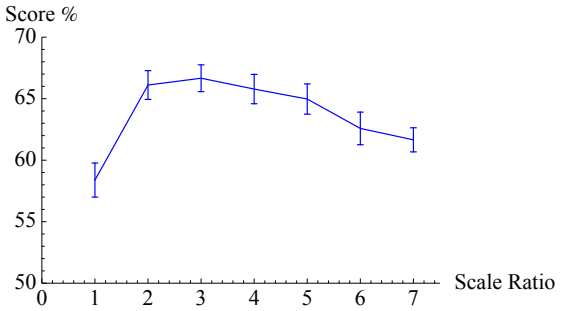
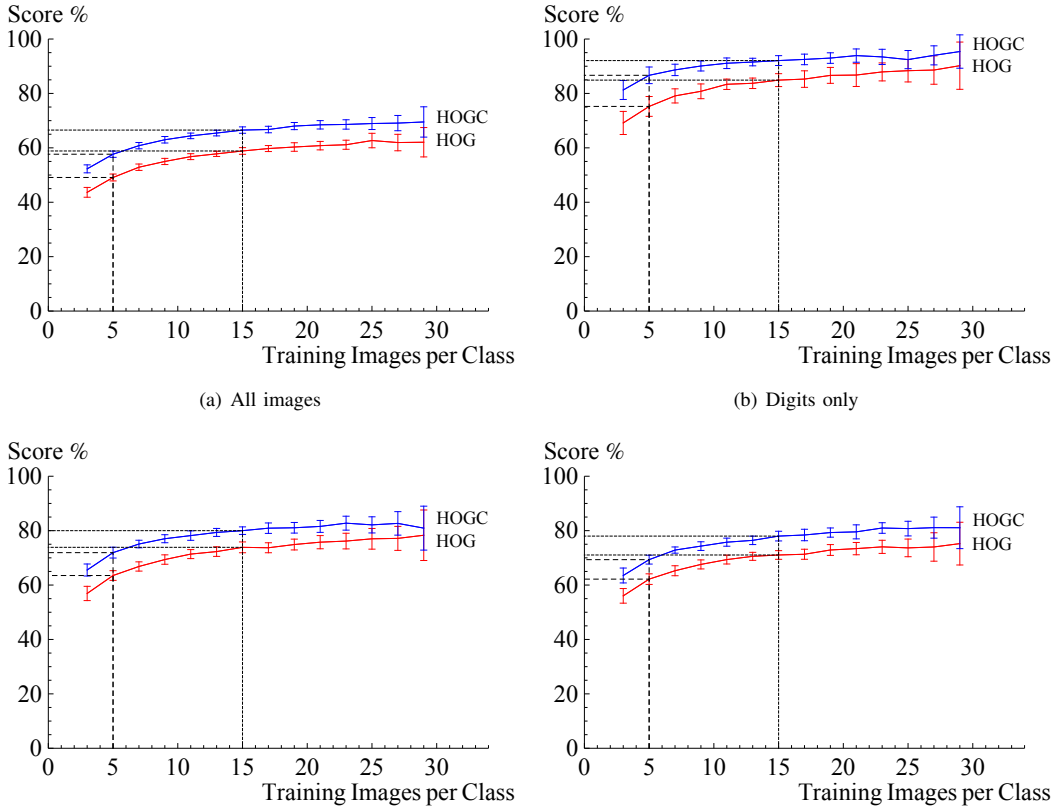


Fig. 4. The relationship between performance and scale ratio when using HOG Columns on the chars74k dataset

Scheme	Chars74k-5	Chars74k-15		ICDAR03-CH-5
Shape Context [6]	26.1±1.7	34.4		18.3
Geometric Blur [6]	36.9±1.0	47.1		27.8
Multiple Kernel Learning [6]	-	55.3		-
ABBYY [13]	18.7	18.7		21.2
SIFT [6]	-	20.8		-
HOG Features [13]	45.3±1.0	57.5		51.5
HOG multiscale	49.1±1.3	58.8±1.2		48.3±1.2
HOG Columns	**57.7±1.1**	**66.5±1.2**		**57.1±0.9**

TABLE I
THE PERFORMANCE OF THE TWO MULTISCALE HOG SCHEMES COMPARED TO PREVIOUSLY PUBLISHED RESULTS USING OTHER METHODS.



(a) All images

(b) Digits only

(c) Upper case letters

(d) Lower case letters

Fig. 2. Performance graphs for the chars74k dataset showing hsitograms of oriented gradients (HOG) and histograms of oriented gradient columns (HOGC).

V. DISCUSSION

The first scheme offers an improvement on one dataset but not on the other. As, in this scheme, histograms are effectively averaged over scale then this may offer a greater level of scale invariance than a single scale HOG system. As we have only resized the images, but made no effort to rescale the characters, then if there is a significant variation in the appropriate scale within a dataset then we might expect such a method to do better than a single scale one. However, if all characters were at the same scale then a single scale tuned system may offer better performance. Thus, the difference between the two schemes could be explained by a difference in the degree of scale variation between them.

The second scheme should also demonstrate this scale invariance, and thus we might expect a better relative performance on the same dataset. However, as the underlying features are diffrerent they are capable of a higher level of discrimination and therefore we see an increase in performance.

Whilst the better performing scheme shows an improvement when using the whole datasets, we believe that there may

be confusion between pairs of classes which can only be overcome using contextual clues. However, the performance when using only digits or upper or lower case letters, where we might expect an upper bound of 100%, shows performance of around 80%. This indicates that there is still considerable room for improvement.

A further point is that images in the chars74k dataset have not been aligned to a common orientation, as can be seen in Figure 1. In both the methods we have tested, there is no provision for recognising rotated versions of the same character and thus rotational variation within the dataset is likely to reduce the level of performance. However, visual inspection indicated that the number of images containing characters at a significantly unusual orientation was relatively small (approximately 5%) and we considered the cost of introducing rotational invariance, arising from the additional confusion of similar rotationally invariant characters such as 'p' s and 'd' s, would outweigh the performance gain.

If the methods were to be extended to datasets with a greater degree of rotational variation then the methods could be altered by using a polar form of spatial binning such as in other forms of rotationally invariant HOG like features (e.g. [11]). Individual oriented gradient column features could then be made rotationally invariant by aligning each to a common orientation at a particular scale, thus capturing the difference in orientation between scales at each location in the image.

VI. Conclusions

In this work we have explored two extensions of the HOG descriptor, both of which involve using oriented gradient features at multiple scales. When tested with datasets of characters taken from images we have shown that a significant improvement can be gained when using histograms of oriented gradient columns, which consist of pairs of orientations weighted by the product of their individual weights. This system has been shown to be stable with respect to the ratio of the two scales.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, volume 1, pages 26–33, 2005.

[3] S.M. Bileschi. A multi-scale generalization of the HoG and HMAX image descriptors for object detection. In *CSAIL Technical Report MIT-CSAIL-TR-2008-019*, 2008.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 428–441. Springer Berlin / Heidelberg, 2006.

[6] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In *Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal*, February 2009.

[7] Ning He, Jiaheng Cao, and Lin Song. Scale space histogram of oriented gradients for human detection. In *Proceedings of the 2008 International Symposium on Information Science and Engieering - Volume 02*, pages 167–170, Washington, DC, USA, 2008. IEEE Computer Society.

[8] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications Technology*, 15(1):52–60, 1967.

[9] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision (ICCV'99)*, pages 1150–1157, 1999.

[10] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. ICDAR 2003 robust reading competitions. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 682–687. IEEE Press, 2003.

[11] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod. Unified real-time tracking and recognition with rotation-invariant fast features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR10)*, 2010.

[12] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. *IEEE 11th Internationl Conference on Computer Vision*, 2007.

[13] Kai Wang and Serge Belongie. Word spotting in the wild. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV*, volume 6311, pages 591–604. Springer, 2010.