Deformable HOG-based Shape Descriptor

Jon Almazán, Alicia Fornés, Ernest Valveny
Computer Vision Center – Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
Bellaterra, Barcelona, Spain
{almazan,afornes,ernest}@cvc.uab.es

Abstract—**In this paper we deal with the problem of recognizing handwritten shapes. We present a new deformable feature extraction method that adapts to the shape to be described, dealing in this way with the variability introduced in the handwritting domain. It consist in a selection of the regions that best define the shape to be described, followed by the computation of histograms of oriented gradients-based features over these points. Our results significantly outperform other descriptors in the literature for the task of hand-drawn shape recognition and handwritten word retrieval.**

I. Introduction

A lot of effort has been dedicated in Computer Vision and Pattern Recognitoon to the problem of shape recognition. It is at the core of many different applications, such as object retrieval or sketch recognition. In the Document Analysis domain, the case of hand-drawn is a specially challenging problem since we have to deal with large variability coming from noise, distortions, inaccuracy in strokes and changes caused by different writting styles. The extraction of robust features is a critical point in this case. Thus, descriptors able to adapt to all this variability are necessary.

In the literature many different feature extraction methods have been proposed for shape description. We are interested in descriptors that have been applied to the recognition of shapes written or drawn by hand. Several generic shapes descriptors have been applied to this kind of shapes, and a general overview can be found in (1). Among them, we can highlight the curvature scale space (CSS) descriptor (2) which successively blurs the shape contour by convolving it with a Gaussian kernel, and the Shape Context (3), which selects n points from the contour of the shape and computes the distribution of the distance and angle between them. Both descriptors are robust to deformations, however, wether they can only deal with some specific shapes or they are computationally expensive, they can not be applied to all the tasks in the "handwritten" domain. Specially conceived for the specific case of hand-drawn symbol recognition, the Blurred Shape Model (BSM) (4) has shown to obtain good results in hand-drawing applications. It is based on computing the spatial distribution of shape pixels in a set of pre-defined image subregions and is able to handle a certain degree of deformation. However, due to the rigidity of the model, large deformations cause large differences in the spatial information encoded by the BSM.

In order to overcome the rigidity of the BSM, the cmiBSM feature extraction method (5) was presented as an extension of the BSM improving its robustness against large deformations. It consists in substituting the fixed regular grid of the BSM by a more flexible grid. A region partition algorithm adapts a given number of points to the shape to be described, and then the "pixels density" is computed in each one of them by the accumulation of *shape pixels*, just as the BSM does. This approach showed a good performance for recognizing shapes in difficult problems such as writer independent symbol recognition. However, when dealing with fine details, *e.g.* recognizing skilled forgery signatures, it presented some difficulties. This is mainly due to the simplicity of the features extracted. We argue that the intensity of foreground pixels is unsufficient to capture all the fine-grained details.

Another descriptor that has been recently applied to handwritten shapes (6) with excellent results is the well-known Histogram of Oriented Gradients (HOG) (7). HOG takes the pixel gradient information as the basis to extract features, which has been shown to be able to deal with fine-grained details and to capture more information than other kind of features, such as the "pixels density". It consists in dividing the image in a rigid grid of cells and computing a histogram of gradients in each one of them. Therefore, apart from the basis features, HOG is similar to the BSM in the sense of using a grid and computing a histogram in each cell. Thus, we argue that the main issues of this descriptor with hand-drawn shapes, as it happens to the BSM, come from his rigidity: allowing some deformation will let us focus on the most discriminative areas, *i.e.*, those that best define the shape. Commonly, handwritten shapes are composed of regions without meaningful information, and on the other hand, regions where all the information is concentrated. Thus, descriptors should focus mainly on these meaningful regions. Therefore, in this paper we propose to combine the deformable grid scheme of the cmiBSM approach with HOG-based features. In this way we plan to improve the HOG descriptor in order to focus the description on the most discriminative regions of the shape.

The main contribution of this work is the extension of the HOG descriptor for the specific case of handwrtting, combining gradient features and a flexible and adaptable grid. We use the *region partitioning* algorithm for the detection of shape regions where information is concentrated in combination with HOG, a feature extraction method able to capture fine and discriminative details. In this sense, we will show that gradient-based features performs better with hand-drawn symbols than density-based features encoded by the BSM, and that the flexibility of the deformable grid improves the results of the rigid grid that the HOG uses.

Finally, we will show that the new descriptor can solve one of the common problems (also related to the rigid grid) encountered when applying the HOG descriptor to images that

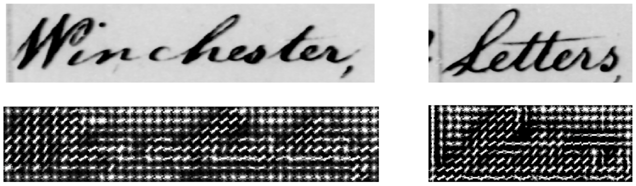


Fig. 1: HOG features.

have different aspect ratios. In order to compare two images using HOG, both should have the same size, otherwise, the dimension of the feature vector may result different. This makes a warping to a fixed image size necessary, which even deforms the shape contained or adds background space without meaningful information. This also provokes that corresponding HOG cells may not contain the same regions of the shape, so it will negatively affect the matching process. However, the approach that we propose, as a side effect of combining the HOG descriptor with the deformable grid, is able to deal with changes in the aspect ratio of the images. That is, as a result of the region partition algorithm, focuses will be located in similar regions of the shape independently of the aspect ratio of the image.

The ability of our method for the description of handwritten shapes has been evaluated for two different, but related, tasks: hand-drawn symbol recognition and handwritten word retrieval. In the latter we consider the handwritten word as a shape to be described and retrieved from the dataset, so it is not related with the typical word spotting approach. Both tasks will test the feature extraction method against writer independent configurations and also against images with different scales and aspect ratios.

The rest of the paper is organized as follows: Section II describes the method proposed. The explanation of the experiments, including the datasets used and the experimental protocols is conducted in Section III. Then, Section IV is devoted to show performance results, as well as the comparison with other approaches. Finally, Section V concludes the paper and porposes a future work line.

II. Deformable HOG

The deformable HOG-based feature extraction approach is based on the computation of HOG features in a given set of $k \times k$ points, denoted as *focuses*, over the shape to be described. These focuses, which can also be seen as an adaptable mesh, are automatically positioned with the objective of being distributed along the shape pixels. Therefore, this approach can be divided in two sequential steps: a first step devoted to compute the location of the focuses following an *iterative region partitioning* algorithm (8) and a second step where regions centered over the focuses are *extracted and described* using HOG features.

A. HOG Features

HOG descriptor was first introduced by Dalal and Triggs (7), but we use Felzenszwalb *et al.* implementation (9), which includes some improvements over the original approach. It consists in first computing for every pixel in the image the orientation and the magnitude of the intensity gradient. Then, the image is divided in an uniform grid of cells and for each one of them a histogram of gradients is computed using “soft binning”. Finally, a dimensionality reduction is performed, resulting in a 31-dimensional vector for each cell: 27 dimensions corresponding to different orientation channels (9 contrast insensitive and 18 contrast sensitive), and 4 dimensions capturing the overall gradient energy in square blocks of four cells around. An example of the HOG features extracted from two different words can be seen in Figure 1. As we can see, these gradient-based features contains enough discriminative and fine-grained information to be able to differentiate between both words.

B. Region Partitioning Procedure

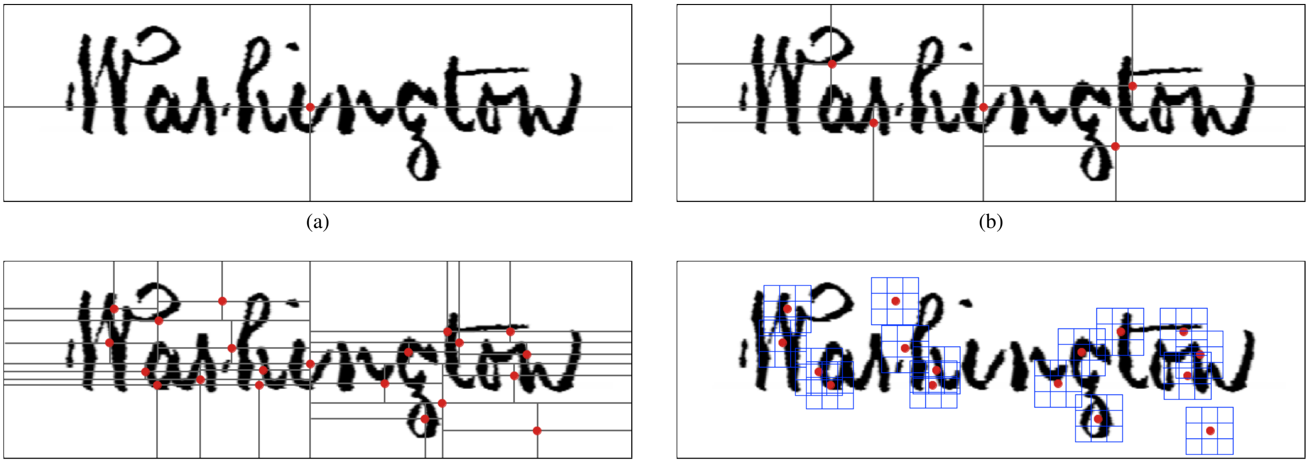
The **region partitioning** procedure consists in subdiving the image into regions centered on the geometrical centroid of the corresponding region of the previous level. The location coordinates of the resulting geometrical centroids will be the points, denoted as *focuses*, where features will be following extracted. Next, we give a brief description of this procedure in order to introduce some notation. For further details, we refer the reader to (8), where this procedure was originally proposed, and (5), where was first used as an adaptable mesh for the extraction of features.

We denote the set of *shape pixels* of the binary image as S and their number as N. The region partitioning procedure will work by obtaining a series of subregions of the image at successive levels. Furthermore, we define as $R_i^l,\ i = \{1, 2, \ldots, 4^l\}$ the i-th rectangular region obtained in the iteration (or 'level') l of the partitioning algorithm, and as $F^l \in \mathbb{R}^2$ the set of geometrical centroids of the regions in R^l. For each level l, the *region partitioning* procedure estimates the geometric centroid of all regions R_i^l and then splits each region into four sub-regions using the geometric centroid. The new sub-regions generated will form the new set of regions R^{l+1}. We consider R^0 as the whole image, and F^0 to contain the geometrical centroid of this region (Figure 2a).

Considering a separate cartesian coordinates system for each region R_i^l, the geometrical centroid F_i^l is computed using equations

$$\mathbf{x_c} = \frac{\sum_{(x,y)\in S_i^l} \mathbf{x}}{\mathbf{N_i^l}},\ \mathbf{y_c} = \frac{\sum_{(x,y)\in S_i^l} \mathbf{y}}{\mathbf{N_i^l}}, \tag{1}$$

where N_i^l denotes the number of shape pixels set S_i^l in the processed region R_i^l, and $\mathbf{x}$, $\mathbf{y}$ are the pixel coordinates. This iterative procedure finishes when a termination level L is reached. Then, the final coordinates of the *focuses* will be only the geometrical centroids of the level L, *i.e.*, F^L. Thus, the number of focuses to represent the shape 4^L can be determined using this termination level L. These focuses can be seen as the representation of a deformable grid adapted to the shape to be described. Examples of the distribution of focuses for levels L equal to 0, 1 and 2 are shown in Figures 2a, 2b and 2c respectively.


(a)

(b)

(c)

(d)

Fig. 2: (a-c) Regions and focuses resulted in the *region partition* procedure for different levels with L equal to 0, 1 and 2. (d) HOG features extracted from level L equal to 2 using a 3×3 cells grid.

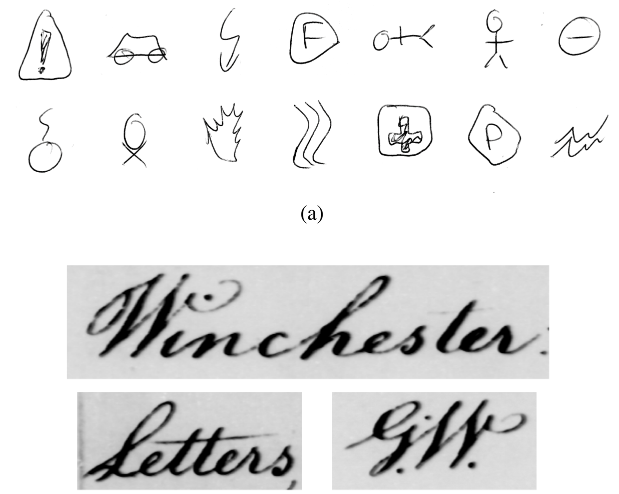
C. Feature Extraction

Once focuses locations have been calculated, the **feature extraction** is computed according to the coordinates of focuses in set F^L. For every focus f_i, $i = \{1, 2, \dots, 4^L\}$ we extract a sub-image I_i centered on their (x, y) coordinates. The size of the sub-images depends on the number $\mathbf{c}$ of HOG cells and the size in pixels of every cell. For our experiments we fix the cell size to 16 pixels and analyze the performance with different number of $\mathbf{c} \times \mathbf{c}$ cells grid. As we said, we follow (9) and use HOG histograms of 31 dimensions to represent each cell. The final vector descriptor v of a given image results from the concatenation of all the histograms from all the focuses. Thus, its dimension depends on the number of cells $\mathbf{c}$ and the termination level L, *i.e.*, the number of focuses 4^L: $v \in \mathbb{R}^d$, $d = 4^L \cdot \mathbf{c}^2 \cdot 31$.

III. Experiments

The new descriptor proposed has been experimentally evaluated for two different purposes in two different datasets: NicIcon dataset (10) for hand-drawn symbol recognition and the George Washington dataset (11; 12) for handwritten word retrieval. We compare its performance with the cmiBSM and the HOG descriptor (fixing the cell size to 16 pixels) for both tasks.

The NicIcon dataset (Figure 3a) is composed of 26,163 handwritten symbols of 14 classes from 34 different writers with on-line and off-line data available. The dataset is divided in three subsets (training, validation and test) for two different settings: writer dependent and writer independent. Every symbol has been cropped and size-normalized in an image of 256×256 pixels. We have selected the off-line data with the writer independent configuration for the symbol recognition task and we have used two different classifiers: Nearest Neighbor-based and Suppor Vector Machine with an exponential χ^2 kernel, whose cost and gamma parameters have been experimentally validated. For comparison we report the classification accuracy.


(a)

(b)

Fig. 3: (a) NicIcon dataset for shape recognition. (b) George Washington dataset for word retrieval.

The George Washington dataset (Figure 3b) is comprised of 20 pages and 4,846 words. We apply a pre-processing for noise removal and slant correction and use the groundtruth information to segment the words. For word retrieval purposes we use the following protocol: each word is considered once as a query and used to rank the rest of the words using cosine as a similarity distance between words. We report the mean Average Precision of all the queries, which is a standard measure in retrieval systems and can be understood as the area below the precision-recall curve. For the original HOG descriptor we resize all the images to a fixed size of 180×270, which is the mean value of height and width respectively.

Concerning the selection of parameters, we have fixed them with the aim of reaching a trade off between performance and dimensionality. However, we will further explore their influence in an extensive analysis, showing that performance of the nrHOG can be considerably increased at a cost of increasing the dimension of the feature vector. As it was shown in (5), cmiBSM performance reaches a plateu at the termination level L equal to 5, so we set it to this value for our experiments. So, in order to have a comparable dimensionality, we set L in the nrHOG equal to 2, and we use a grid of 3×3 to compute HOG features in the focuses. Finally, both HOG and nrHOG use a size bin equal to 16 pixels.

IV. Results and Discussion

In Table I we show the classification accuracy in the NicIcon dataset for the three methods compared: *cmiBSM*, *HOG* and the proposed approach *non-rigid HOG*, denoted as *nrHOG*. We can see that for both classifiers used (Nearest Neighbor-based and SVM with exponential χ^2 kernel) HOG-based approaches outperform the *cmiBSM* descriptor. This confirms the need of capturing fine-grained details using more informative and discriminative features. Moreover, we also observe that the incorporation of an adaptative grid in the grid-based HOG improves the performance for the classification of shapes.

TABLE I: Results in the NicIcon dataset for the word symbol recognition task in the writer independent configuration

Method	NN accuracy (%)	SVM accuracy (%)
cmiBSM$_{\text{f+p}}$ (5)	89.42	90.62
HOG (9)	93.47	96.68
nrHOG	95.88	97.69

Then, we show in Table II the mean Average Precision for the word retrieval task over the George Washington dataset, where we extract a similar conclusion: HOG-based features are able to deal with fine details to discriminate between handwritten shapes, and its combination with a deformable mesh substituting the rigid grid leads to a significant performance improvement. In this task, where shapes are more complex and we have to deal with a larger number of classes, differences between descriptors are considerably larger. The cmiBSM is clearly not able to deal with the fine details to correctly differentiate words, and it is surpassed by HOG-based descriptors. The proposed nrHOG approach reports the best performance.

TABLE II: Results in the George Washington dataset for the word retrieval task

Method	mAP (%)
cmiBSM$_{\text{f+p}}$ (5)	8.51
HOG (9)	37.21
nrHOG	44.59

As we said in the introduction, the integration of the deformable mesh of focuses provides to the original HOG descriptor some invariance to changes in the aspect ratio. This can be specially appreciated in the results of the GW dataset, where we have big changes in aspect ratio for images of the same class, so the difference in performance between nrHOG and HOG is larger than in the NicIcon, which only contains squared images.

Like in (5), we could use the focus coordinates as a feature vector and perform an in-kernel fusion with the HOG features when pre-computing the exponential χ^2 kernel to improve the performance. However, in this case the improvement is unsubtantial and we do not consider worthy the extra computational time that this fusion requieres.

Finally, we show in Figure 4 some qualitative results comparing the three approaches for the word retrieval task over the George Washington dataset. There we can see that, even that HOG improves the results of the cmiBSM by retrieving words whose shape is more similar to the query ("October", "November" and "December" share most of the characters), it is not enough to be able to differentiate between handwritten words. For that, we need to focus the description over the discriminative regions as the nrHOG does, resulting in in this way in a better performance.

A. Parameters Analysis

The *nrHOG* has two main parameters: the termination level L, which determines the number of *focuses* used to describe the shape, and the value $\mathbf{c}$ of the $\mathbf{c} \times \mathbf{c}$ grid used to extract HOG features around every focus. In Figure 5 we explore the effect of these parameters. There we can see that increasing their values leads performance to increase. However, for both L and c, higher values means a larger vector dimension, so their value adjustment will be a trade-off between performance and dimensionality. Considering that the size of the cell has been fixed to 16 pixels, the dimension of the resulting feature vector of the HOG descriptor is equal to 3,906. In the case of the nrHOG, the first configuration that outperforms the HOG with the minimum dimensionality has a feature vector with dimension equal to 1,116. The configuration of the nrHOG that has the best performance in Figure 5 results in a vector with dimension equal to 198,400.

V. Conclusion and Future Work

In this work we have shown how a combination of a deformable grid and a fine-grained feature extraction method based on histrograms of gradients can be used to describe handwritten shapes and can be applied to shape recognition and retrieval. We have also shown its robustness against variability for different writting styles and different aspect ratios. This has resulted in a succesful adaptation of the well-known HOG descriptor to the handwritting domain. We have obtained excellent results when comparing to other shape descriptors. We plan to publish the MATLAB code implementation and the prepocessed datasets upon publication in the hope that it would provide a comparison framework for new shape descriptors.

As future work we plan to integrate the new nrHOG descriptor in the word spotting framework proposed in (6). It will require a new sliding window procedure that takes into account the deformable grid characteristics and to adapt the matching process according to that. Moreover, we plan to perform an extensive analysis comparing different feature extraction techniques in combination with the deformable grid.

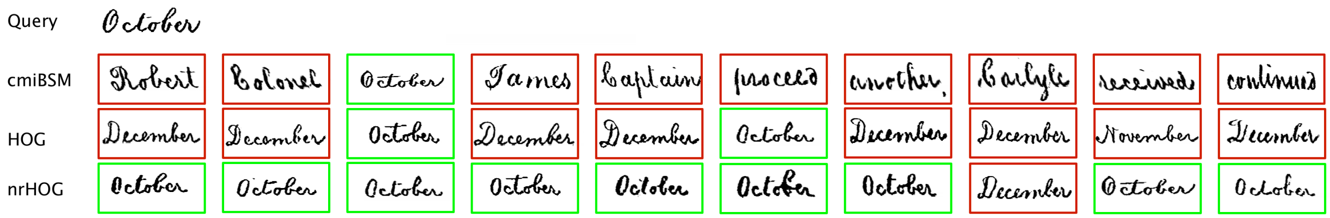


Fig. 4: Qualitateve results comparing cmiBSM, HOG and nrHOG for the word retrieval task over the George Washington dataset.

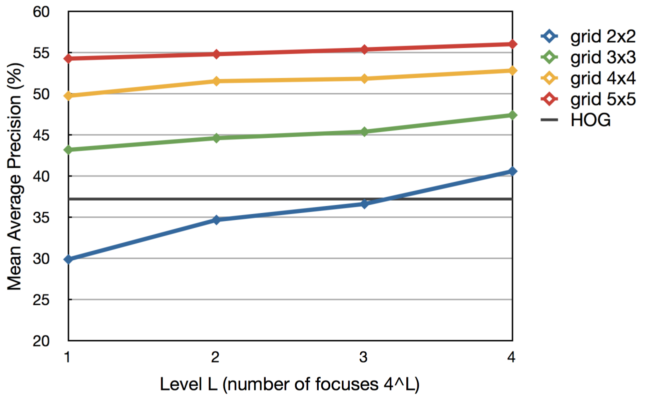


Fig. 5: Influence of level L in the George Washington dataset for different number of grid cells. The size of cell is fixed to 16 pixels.

References

[1] D. Zhang and G. Lu, "Review of shape representation and description techniques," *PR*, vol. 37, no. 1, pp. 1–19, 2004.

[2] F. Mokhtarian and A. Mackworth, "Scale-based description and recognition of planar curves and two-dimensional shapes," *TPAMI*, vol. 8, no. 1, pp. 34–43, 1986.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE TPAMI*, vol. 24, no. 4, pp. 509–522, 2002.

[4] S. Escalera, A. Fornés, O. Pujol, P. Radeva, and J. Lladós, "Blurred shape model for binary and grey-level symbol recognition," *PRL*, vol. 30, pp. 1424–1433, 2009.

[5] J. Almazán, A. Fornés, and E. Valveny, "A non-rigid feature extraction method for shape recognition," in *ICDAR*, 2011.

[6] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient Exemplar Word Spotting," in *BMVC*, 2012.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[8] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris, "Content-based binary image retrieval using the adaptative hierarchical density histogram," *PR*, vol. 44, no. 4, pp. 739–750, 2010.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramaman, "Object detection with discriminatively trained part based models," *IEEE TPAMI*, 2010.

[10] D. Willems, R. Niels, M. van Gerven, and L. Vuurpijl, "Iconic and multi-stroke gesture recognition." *PR*, vol. 42, no. 12, pp. 3303–3312, 2009.

[11] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *CVPR*, 2003.

[12] ——, "Word spotting for historical documents," *IJDAR*, 2007.