Sketchable Histograms of Oriented Gradients for Object Detection

Ekaterina Zaytseva[1], Santi Seguí[1,2], and Jordi Vitrià[1,2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona
{ezaytseva,ssegui}@cvc.uab.es, jordi.vitria@ub.edu
[2] Dept. de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona

Abstract. In this paper we investigate a new representation approach for visual object recognition. The new representation, called sketchable-HoG, extends the classical histogram of oriented gradients (HoG) feature by adding two different aspects: the *stability* of the majority orientation and the *continuity* of gradient orientations. In this way, the sketchable-HoG locally characterizes the complexity of an object model and introduces global structure information while still keeping simplicity, compactness and robustness. We evaluated the proposed image descriptor on publicly Catltech 101 dataset. The obtained results outperforms classical HoG descriptor as well as other reported descriptors in the literature.

Keywords: Object Recognition, Feature stability, Centrality measures, Histogram of Oriented Gradients.

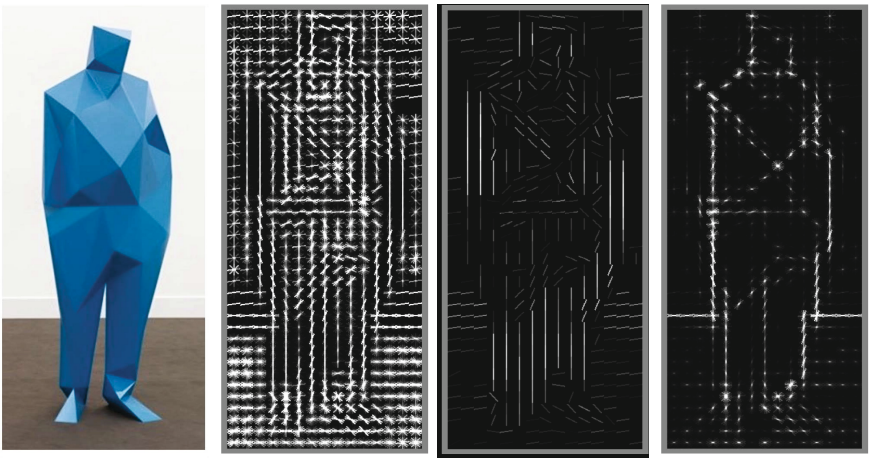
1 Introduction

In his seminal book, David Marr [8] conjectured that the path to object recognition could benefit from the use of a first level representation of images in terms of simple features. This representation, called primal sketch, was supposed to be a parsimonious but sufficient to reconstruct the original image without much perceivable distortions.

State of the art object recognition systems are not making use of this concept and, instead, are based on representing visual information with local edge statistics or patch based features. This paper focuses on applying the concept of sketchable representation to one of the most used representations: the histogram of oriented gradients (HoG) [4]. This image descriptor has been widely used and it is represented in several forms in the most of state of the art methods for object recognition. In this context, our goal is to go one step beyond the use of local edge statistics by considering additional information to improve upon this cutting-edge descriptor. In spite of this addition, the resulting descriptor keeps the simplicity and robustness of the original one.

The original structure of a HoG is implemented by dividing the image into a set of small connected regions and for each region, or cell, compiling a histogram of gradient directions for the pixels within the cell. This structure is shown in figure 1(b). HoG descriptor can be then built by concatenating the values of the

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 374–381, 2012.



(a) (b) (c) (d)

Fig. 1. (a) Original image; (b) HoG features of (a); (c) Stable orientations of (b); (d) Cells of (b) with high continuity values

bins of all histograms, getting a high-dimensional vector $I = (x_1, \ldots, x_n)$ that represents the image. The final step when using this representation for object recognition is to feed the descriptors into a discriminative learning method such as Support Vector Machine.

In this paper we propose the addition of two characteristics for each histogram cell that represent a more abstract feature of the image: (i) a measure of the *stability* of the most probable orientation in a cell and (b) a measure of the *continuity* of the cell orientation with respect to the whole model. The first characteristic identifies for each object model those cells of the HoG representation that clearly represent an oriented edge of the object. The second one reinforces those cells that represent a continuous oriented gradient field with respect to the neighboring cells.

In figure 1 we show (a) an image of an object and its corresponding (b) HoG model as defined in [4]. In 1(c) we show a new version of (b) where each orientation has been weighted by a value proportional to its stability. It can be easily seen that the stability feature reduces noise but highlights the most well defined edges of the model. Finally, in 1(d) we show a version of (b) where each orientation has been weighted according to its continuity value. As can be easily seen, the resulting models are simpler but sufficient representations of the object and for this reason we call it *sketchable* histograms of oriented gradients.

In summary, this paper shows a way for deriving sketchable histograms from images and that this new feature constitutes a more abstract layer of visual information that allows to build better models even from a few images.

2 Image Description

In order to represent visual object models we propose the addition to the classical HoG representation of two new image features that can be readily derived from it. In the first case, the derived feature is called *stability* and it assigns a value that represents the homogeneity of the gradient vector field that corresponds to a HoG cell. In the second case, the feature is called *continuity* and it represents the level of continuity of the object edges that are represented by that cell.

Our main hypothesis is that these features constitute a parsimonious and sufficient representation of the Histogram of Oriented Gradients. Image representation, obtained by concatenation of stability, continuity and original HOG values, is called sketchable-HoG descriptor. The use of this representation should increase the performance of classifiers because more abstract information is readily accessible to it. In the following sections we defined both features and check the expected increment of performance in a standard dataset.

2.1 Stability

In a HoG representation, each pixel contributes a weighted vote for an edge orientation based on the orientation of the gradient element centered on it. Votes, which are a function of the gradient magnitude at the pixel, are accumulated into orientation bins h_i over local spatial regions that we call cells.

Let $H = (h_1, \ldots, h_9)$ be the vector representing the values of the bins of a standard HoG cell. To define its *stability* we can use the Hoeffding inequality [9], which relates the empirical mean of a bounded random variable to its true mean.

Let $x_1, x_2, \ldots, x_k$ be iid random variables with values in $[0, 1]$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$,

$$E[x] \leq \frac{1}{k} \sum_{i=1}^{k} x_i + \sqrt{\frac{\ln(1/\delta)}{2k}} \tag{1}$$

In our case, and considering that each pixel contributes with its gradient magnitude to its corresponding orientation bin, we can state, with probability at least $1 - \delta$, that we can associate to h_i a confidence interval $[h_i^{LB}, h_i^{UB}]$, derived from (1), within we expect the true value of h_i to be.

We will say that a cell H is *stable* if its most voted orientation i is *admissible*, in the sense that its estimated value incurs at most β times as much error as any other orientation:

$$H \text{ is stable } \iff \exists i, (1 - h_i^{LB}) < \beta \cdot min_{i' \neq i}(1 - h_{i'}^{UB}) \tag{2}$$

In our experiments the value of β parameter was fixed at 1. It means that we consider orientation i admissible if the lower bound of the most voted orientation is greater than the upper bound of the second most voted orientation (see figures 2 and 3).

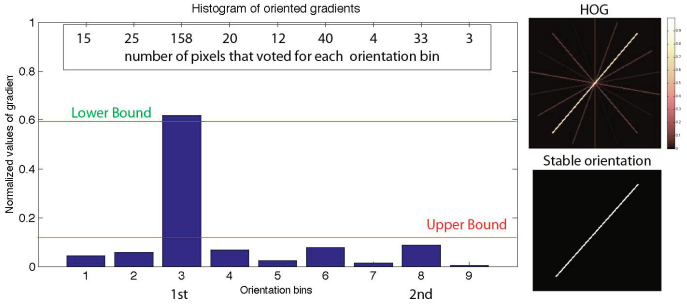


Fig. 2. In the case of a well defined distribution of gradient orientations, the lower bound of the most voted orientation is higher than the upper bound of the second most-voted orientation

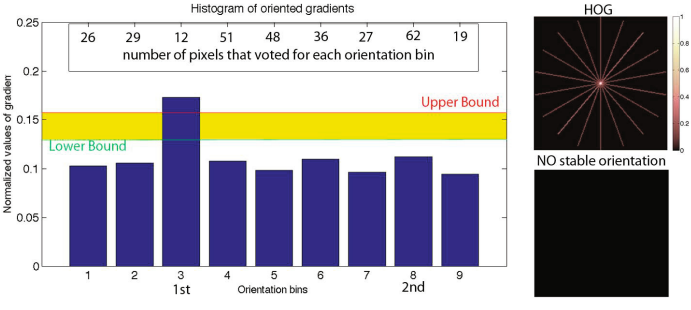


Fig. 3. In the case of an ill defined distribution of gradient orientations, the lower bound of the most voted orientation is lower than the upper bound of the second most-voted orientation. Please note that for the estimation of the lower bound of the fist most probable orientation bin and the upper bound of the second most probable orientation bin both the gradient histograms and the number of voted pixels for each histogram bins are important.

The concept behind the stability indicates that easier visual classes are characterized by higher stability values whilst harder classes have fewer admissible orientations. To check this hypothesis, we have considered the twenty object classes of the PASCAL VOC Challenge 2009. For each class we have computed its HoG model and its mean stability. The complexity of the class has been represented by the average precision (AP) value obtained by state of the art detection methods in the PASCAL VOC Challenge 2009. Finally, we have computed the Spearman's rank correlation coefficient ρ between mean stability of the model and the average precision (AP) value of the class, resulting in a $\rho = 0.81$. As it is shown in figure 4, this value corresponds to a significant correlation between the mean stability and the complexity of the class, corroborating the idea that, when considering state of the art methods [5], easier classes to detect are those presenting higher mean stability values.

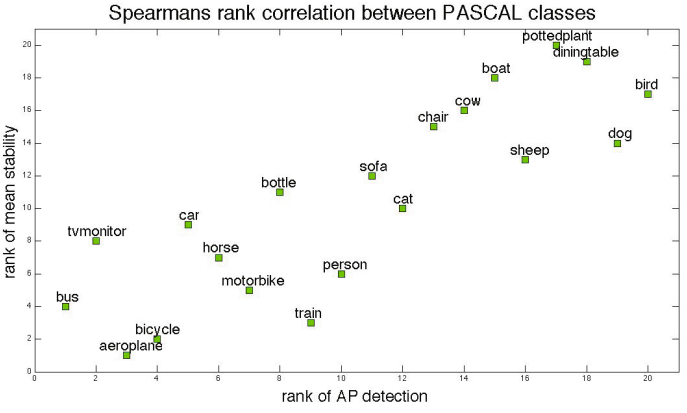


Fig. 4. The Spearman's rank correlation coefficient between the mean stability of a HoG model and the average precision (AP) value of its class in the PASCAL VOC Challenge 2009 is 0.81, which clearly shows a relationship between visual class complexity and stability of its HoG features

2.2 Continuity

In order to represent continuity, we will consider the centrality measure of the nodes of a graph G defined from a HoG descriptor $\{H^j\}_{j=1,\ldots,m}$, where m is the total number of cells of the descriptor. Each node v_j of the graph G corresponds to a cell H^j and the edges $e_{i,k}$ connect neighboring HoG cells H^i and H^k (we have considered a 8-connectivity structure).

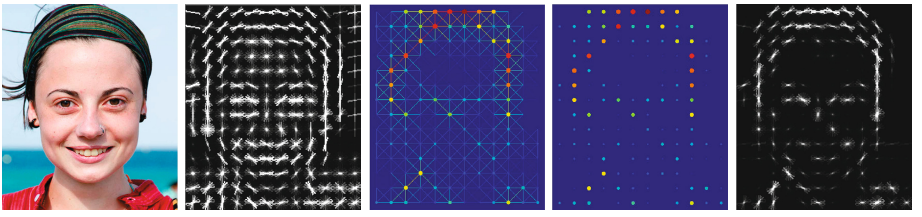
The cost of an edge $e_{i,k}$ is assigned by taking into account the values of the histogram bins that correspond to angles which are similar to the angle between both HoG cells. That is, if the node v_i corresponds to an image location that is oriented by angle α with respect to the node v_k, the cost of $e_{i,k}$ is computed by adding the value of the bins from H^i and H^k that correspond to the following angles: $\alpha - 22.5°$ to $\alpha + 22.5°$. In this way, edges which correspond to neighboring cells that represent aligned gradient fields will get higher values (see figure 5).

We have selected the *betweenness* measure proposed in [7] to represent the centrality of each node $v_j \in V$ of the graph . The betweeness measure of a node v_j is equal to the number of shortest paths from all vertices to all others that pass through v_j. This value is then normalized by dividing through the number of pairs of vertices not including v_j. Formally, this continuity measure can be defined as:

$$C(v_j) = \sum_{s \neq v_j \neq t \in G} \frac{\sigma_{st}(v_j)}{\sigma_{st}} \tag{3}$$

where σ_{st} is the total number of paths from node s to node t and $\sigma_{st}(v_j)$ is the total number of shortest paths from node s to node t that pass through v_j. The resulting descriptor is a vector C of size m.

In order to estimate the betweeness we use the Brandes' algorithm [3]. This algorithm is, up to date, the fastest exact algorithm with a complexity $\mathcal{O}(VE + V^2 log(V))$. The method solves n SSSP (Single Source Shortest Path) problems by using Dijkstra's algorithm, where n is the number of vertices, and then adds counter values from the leaves to the root.



(a) (b) (c) (d) (e)

Fig. 5. (a) Original image; (b) HoG features of (a); (c) Graph constructed using (b) indicating with jet colormap (blue/lower value to red/higher value) the obtained continuity values; (d) Continuity value of cells from (b); (e) Cells of (b) with continuity values

3 Results

In order to evaluate the discriminative power of the proposed feature descriptor, the performance of this descriptor was evaluated by training a One-against-All Support Vector Machine (SVM) [10] classifier for the Caltech-101 dataset [6]. This dataset consists of images from 101 different classes, and contains 30 to 800 images per class. The performance was obtained by averaging results from 10 different trials. In each trial of the evaluation, 15 random training images and 50 random testing images were selected per class. In those classes with less samples than needed fewer images were used for test set. We use this protocol to be able to compare with results reported in [1] and [2].

The achieved results are presented in Table 1. In this table we compare the results obtained using the proposed sketchable HoG descriptor with results of classical HoG descriptor and Berg in [1] and Bileschi in [2]. In order to do a comparison with reported results we used a Linear SVM. As it can be seen, the proposed sketchable HoG descriptor obtains the best performance ($52.71\% \pm 0.80\%$) outperforming all other image descriptors. Table 1 also shows results, obtained by sketchable and classical HoG descriptors using a SVM with rbf-kernel. Sigma parameter of rbf kernel was fixed as the $1/d$ where d is the dimensionality of the image descriptor. In this case, the sketchable HoG descriptor gets an accuracy of $58.80\% \pm 0.57\%$ outperforming classical HoG by more than 4%.

Figure 3 shows the confusion matrix for sketchable HoG and classical HoG descriptors using SVM with rbf-kernel. The confusion matrix denotes the absolute difference in the accuracy between classes. For simplicity, the figure only presents the 5 classes for which the performance increases the most, and the 5 classes for which the performance decreases the most. Moreover, for each of these classes

Table 1. Caltech 101 classification results

Method	% Accuracy
Lineal SVM	
Berg et.al.[1]	45
Bileschi et.al.[2]	$48.26\% \pm 0.91\%$
HOG	$47.71\% \pm 0.80\%$
sketchable-HoG	$52.71\% \pm 0.61\%$
RBF SVM	
HOG	$54.35\% \pm 0.80\%$
sketchable-HoG	$58.80\% \pm 0.57\%$

we show its most confusing class. As it can be seen in this confusion matrix, for some classes the performance increases on 20%, when we use a sketchable HoG descriptor instead of classical HOG, while, in those classes where the accuracy decreases the most, the decrease is only 2%.

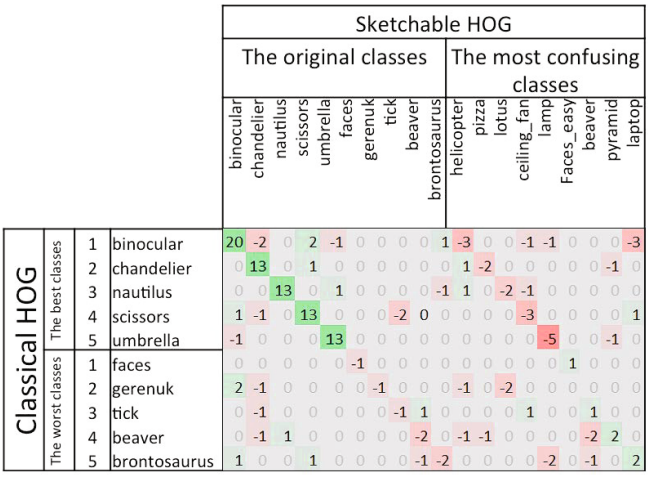


Fig. 6. Confusion matrix for Caltech-101 dataset using sketchable-HoG and classical HoG with RBF-SVM. We show the 5 classes for which the performance increases the most, and the 5 classes for which the performance decreases the most. Each of these classes are paired with its most confusing class.

4 Discussion and Future Work

The most similar approach to our method was proposed by Bileschi and Wolf in [2]. This method defines 4 different image features based on continuity, symmetry, closure and repetition. The main difference between this method and the proposed sketchable-HoG is that this image representation, extends the classical (HoG) descriptor by adding two different aspects: the stability of the majority

orientation and the continuity of gradient orientations. As can be seen in Table 1, the proposed sketchable-HoG outperforms in a significant way the results obtained by classical HoG descriptor and the method proposed by Bileschi and Wolf.

These results point towards a clear objective for object recognition systems: to go one step beyond classical local gradient descriptors and develop mid level image descriptors that represent more abstract object characteristics such as stability and continuity. Our option has been not to define a brand new descriptor, but to leverage a powerful one, the HoG. As it is observed in [2], the idea that more meaningful image representations can produce significantly better recognition results is attractive, but it is not trivial to demonstrate. We think that in this work contributes to this research objective in a clear way.

Acknowledgements. This work was partially supported by MEC grants TIN2009- 14404-C02-01 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

1. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondences. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 26–33 (2005)
2. Bileschi, S.M., Wolf, L.: Image representations beyond histograms of gradients: The role of gestalt descriptors. In: CVPR. IEEE Computer Society (2007)
3. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25, 163–177 (2001)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge (VOC2009) Results (2009), `http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html`
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision (2004)
7. Freeman, L.C.: A Set of Measures of Centrality Based on Betweenness. Sociometry 40(1), 35–41 (1977)
8. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc., New York (1982)
9. Serfling, R.J.: Probability Inequalities for the Sum in Sampling without Replacement. The Annals of Statistics 2(1), 39–48 (1974)
10. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York (1995)