IMPORTANCE OF FEATURE LOCATIONS IN BAG-OF-WORDS IMAGE CLASSIFICATION

Nevena Lazic, Parham Aarabi

University of Toronto
The Edward S. Rogers Sr. Department of Electrical and Computer Engineering

ABSTRACT

The impact of image feature locations in the bag-of-words model for object classification is examined. It is demonstrated that a simple variance-based method works well and offers advantages over several other methods. In essence, the feature locations are selected intelligently, decreasing the redundancy and cost sometimes associated with feature extraction on dense grids. Classification results on two databases are presented, using a support vector machine classifier.

Index Terms— interest points, image classification

1. INTRODUCTION AND PREVIOUS WORK

Much of the previous work on image classification is based on a model analogous to the *bag-of-words* model for text document retrieval. This method consists of four basic stages: (i) a set of keypoints/regions is selected, (ii) keypoints are represented using local descriptors, (iii) descriptors are vector-quantized into a fixed-size codebook, and (iv) the image is represented as a histogram of the codewords it contains. The histogram representation is the input to the image classifier.

Numerous approaches to the outlined steps exist. Keypoint selection techniques include using every pixel [17], a dense regular grid [5, 3], randomly sampled points, segmentation based patches [1], and sparse sets of interest points or regions, including Lowe's difference-of-Gaussians (DoG) peaks [11], Harris affine-covariant regions [12] used in [4, 5, 14] and the Kadir & Brady saliency detector [8].

Once a set of locations is obtained, local descriptors are extracted. One of the most widely used local descriptors is SIFT [11], which is essentially a histogram of intensity gradient orientations, weighted by their magnitude and a Gaussian window. It is computed at different image scales, and the predominant gradient orientation is subtracted, to make it scale and rotation invariant. Several variations on the SIFT descriptors exist, including PCA-SIFT [10] and color SIFT [3]. Descriptors are also often computed by passing an image through filter banks, typically comprising of Gaussians, Gaussian derivatives, Laplacians, and wavelets [17, 16]. In [13], it is demonstrated that SIFT descriptors seem to be more robust than other descriptors, and dense sampling grids outperform other point detectors.

The collection of descriptors is vector-quantized into a dictionary of codewords, typically using K-means with Euclidean or Mahalanobis distance, and the size of the dictionary varies in the 200-4000 range. In [17], the optimal dictionary size and codewords are learned by pairwise merging from an initially large dictionary. An image is represented using a histogram of the codewords it contains.

Classification methods commonly used include naive Bayes [4], Gaussian mixture models (GMMs) [17], hierarchical Bayesian text models (probabilistic Latent Semantic Analysis and Latent Dirichlet Analysis)[3, 14, 2, 5], and support vector machines (SVMs) [4].

The focus of this paper is to evaluate the effect of keypoint selection methods on classifier performance, when the number of available keypoints per image is held constant. The paper is organized as follows. The next section describes a variance-based point selection the developed method. Section 3 provides classification details, and experimental results are shown in section 4. Discussion and conclusions are given in section 5.

2. FEATURE LOCATION SELECTION

What criteria can be used to select regions of an image that are useful in the classification context? In this paper, the relationship between patch content and the variance and entropy of its intensity histogram was explored. The approach was partially motivated by the information-theoretic salient region detector of Kadir & Brady [8], which measures saliency using the Shannon entropy of the intensity histogram of an image, over a range of scales (circles of radius s). The entropy around a point x at scale s is calculated as:

$$H_{s,x} = -\sum_i p_{s,x}(i) \log(p_{s,x}(i)) \qquad (1)$$

where $p_{s,x}(i)$ denotes the value of component i of the histogram of image intensities, in a circle of radius s around x. This method prefers regions where the image intensity varies to more uniform regions. However, since histogramming discards spatial information, any noise-like random permutation of the pixels results in the same entropy. To account for this, the entropy peaks are detected over different scales.

Here, a similar approach to detecting informative regions is explored. To decrease the effects of high-entropy noise, the image is first smoothed by a Gaussian filter with standard deviation σ. The effects of smoothing on the entropy and variance of an image patch are examined over a range of values of σ. Example images are given in Fig. 1. The first two rows show image patches and random permutations of their pixels, respectively. Initially, both the originals and the permutations have the same entropy and variance. As the images are smoothed, the variance of the originals becomes greater than that of the permutation, and this difference is much larger in informative image patches. Entropy does not always seem to provide enough clues.

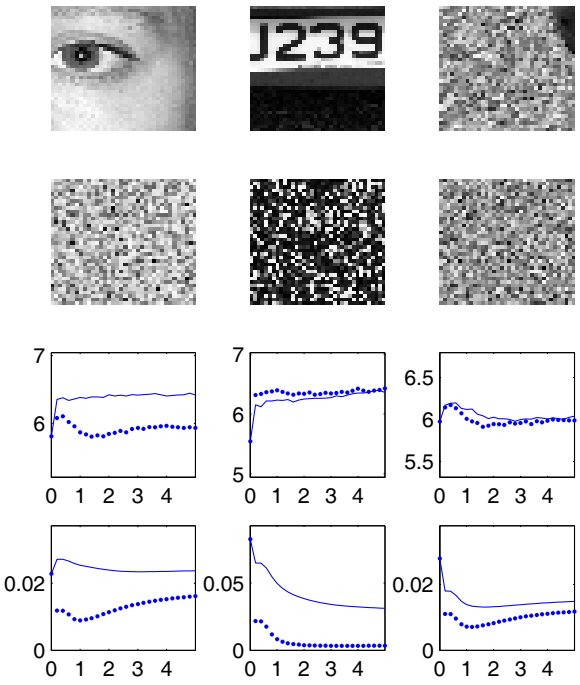

Fig. 1. The first row shows three image patches, and the second row shows random permutations of their pixels. The third and fourth row show entropy and variance respectively, as a function of smoothing. Solid lines are used for original images and dotted lines for pixel permutations

These observations led to the following approach to selecting descriptor locations in images. The image is first smoothed by a Gaussian filter with standard deviation σ, and the variance of half-overlapping square smoothed image patches p, $V_{p,\sigma}$ is calculated. A total of N points are available, and the number of points N_p allocated to each patch is proportional to its variance, i.e. $N_p \propto V_{p,\sigma}$. σ was set to 5 pixels, and patch size to 36x36 pixels, as these parameters yielded the most intuitive results among several parameter settings.

The described method was compared to a number of other point selection techniques: regular grid, random sampling, Lowe's DoG [11], and assigning points in proportion to patch entropy. Fig. 2 shows examples of selected points with the high-variance and DoG methods. The total number of points in the corresponding images is kept the same.

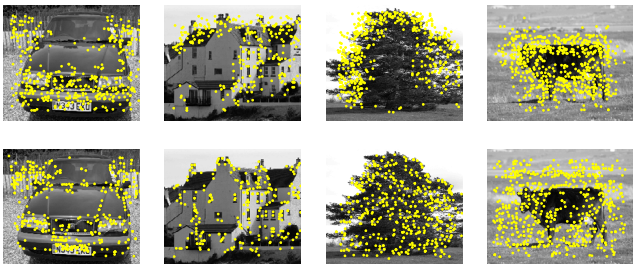

Fig. 2. Examples of points detected with high variance (top) and DoG (bottom) methods

3. CLASSIFICATION

3.1. Data representation

The selected locations were described using the PCA-SIFT descriptor [11, 10] of the grayscale image components. The descriptors were vector-quantized using K-means with K=500 and Euclidean distance. Each image was represented by a histogram of the quantized descriptors, normalized to the [0,1] range. Additionally, one variation on the histogram representation was explored - each histogram entry was raised to the power of 0.5. This was done to reduce the 'peakiness' of the histograms, and give more emphasis to less frequent codewords.

3.2. Support Vector Machines

Classification was performed using Support Vector Machines (SVMs). The SVM classifier separates two-class data by finding a maximum *margin* hyperplane between the two classes [15].

Let $\{\mathbf{x}_i, y_i\}, i = 1, \ldots, l, \mathbf{x}_i \in \mathbf{R}^d, y_i \in \{-1, 1\}$ denote the training data points and the corresponding labels. In the simple case of a linear machine and linearly separable data, there exists a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ separating the two classes. The maximum margin hyperplane is found by minimizing $||\mathbf{w}||^2$ such that $y_i(\mathbf{x} \cdot \mathbf{w} + b) - 1 \geq 0 \forall i$, a convex optimization problem. In the test phase, the data is classified based on the sign of $f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w} + b$. The data points lying closest to the separating hyperplane are called *support vectors*.

In the case of non-separable data, it is necessary to relax the constraints, and introduce a cost C into the objective

function, penalizing misclassified data in proportion to their distance from the hyperplane. Furthermore, the data can be mapped into another space in which it may make more sense to use a linear classifier. If there exists a mapping $\Phi(\cdot)$ of the data into another Euclidean space, and a *kernel function* K, such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x_j})$, the training algorithm only needs K, without even knowing what $\Phi(\cdot)$ is. In the test phase, a point can be classified by computing the sign of:

$$f(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b \qquad (2)$$

where $\mathbf{s}_i$ are the support vectors. Such a kernel needs to satisfy Mercer's condition [15]. Some of the kernels commonly used are:

- Polynomial $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$, resulting in a classifier that is a polynomial of degree p
- Gaussian, $K(\mathbf{x}, \mathbf{y}) = e^{\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$, resulting in a radial basis function (RBF) classifier

In the case of M-way classification, M detectors $f_m(\mathbf{x})$ are trained on positive and negative examples of each class. The class label is assigned to be $m^* = \arg\max_m f_m(\mathbf{x})$.

4. EXPERIMENTAL RESULTS

Classification experiments were conducted on two image databases. The first database is comprised of images from the MSRC-B1 database [9], containing images of 7 classes of objects: building, tree, cow, aeroplane, face, car, and bicycle. The object segmentation information was not used; instead, each image was labeled as containing one of the 7 objects. There were 30 images of each class; half of the images were used as a training set, and the other half as a testing set. The performance metric used was the mean of the normalized diagonal of the confusion matrix. Since there were an equal number of data for each object class, this is equivalent to the fraction of correctly classified images.

Classification was performed using linear, polynomial, and Gaussian SVM kernels, and the SVM_{light} [7] software package. The MSRC-B1 results are displayed in Figs. 3 and 4. In all cases, polynomial kernels worked the best. Raising histogram entries to the power of 0.5 improved the accuracy rate slightly. In most cases, over all tested classifiers and a range of the number of keypoints, choosing the high-variance descriptor locations offers an improvement over other methods. There does not seem to be a significant difference between the regular grid, high entropy, and random locations. Note that the DoG result only has a single point in the figures, as it cannot detect an arbitrary number of keypoints; the result is plotted over the average number of keypoints points per image.

Experiments were also performed on 4 classes of images from the Caltech101 [6] database, containing airplanes, cars, motorbikes, and faces. Again, no segmentation information was used; 60 images of each class were used to train the classifier, and 60 to test. The performance of high variance point selection was compared to regular grid and DoG point selection, raising the histogram entries to the power of 0.5 and using a cubic SVM kernel. However, on this data set, the choice of points did not make much difference in the overall classification results, as shown in Fig. 5. On a per-class basis, the high variance points performed better on all classes except airplanes. We suspect that this is due to the fact that the background points captured by the regular grid aid classification performance. The background in the airplane images usually consists of grass and sky, and backgrounds tend to be different in other image classes. When performing classification on three classes only (cars, motorbikes and faces), for lower total number of points, the high variance points give better results, as shown in Fig. 6.

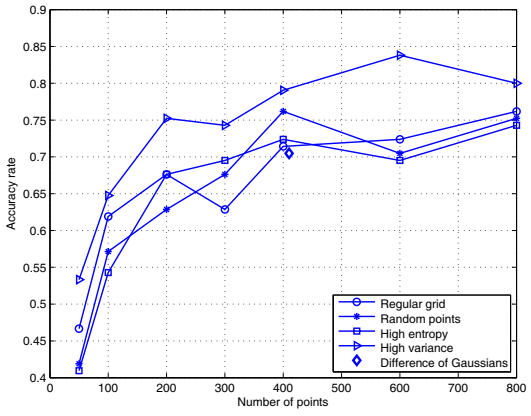

Fig. 3. MSRC-B1 results: cubic kernel SVM, plain histograms

5. CONCLUSIONS AND FUTURE WORK

An intuitive method of selecting descriptor locations in an image was presented, and compared to other methods in an image classification scenario. It was shown that a simple variance-based point selection method can be more effective than using a regular grid, random points, or difference-of-Gaussians. Possible extensions to variance-based point selection include: application to color spaces (as opposed to just image intensity), scale invariance (e.g. by using patches of varying size), using allocation functions that are different functions of variance, and automatically selecting an optimal number of points. Also, the classification accuracy can be improved by using more sophisticated models.

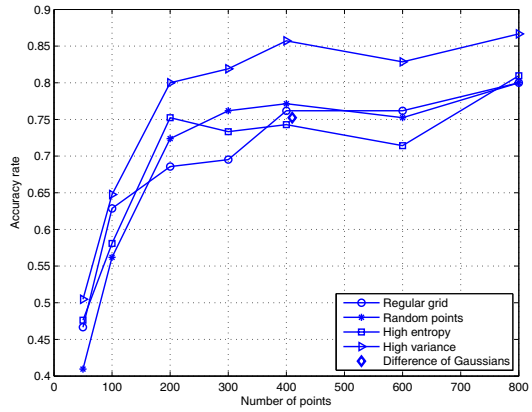

Fig. 4. MSRC-B1 results: cubic kernel SVM, histogram entries raised to power 0.5

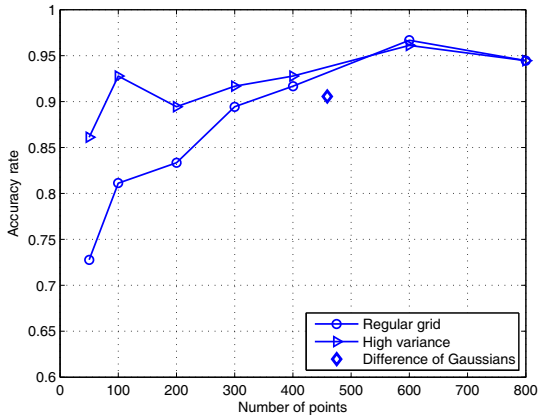

Fig. 6. Caltech 101 results: cubic kernel SVM, 3 classes, histogram entries raised to power 0.5

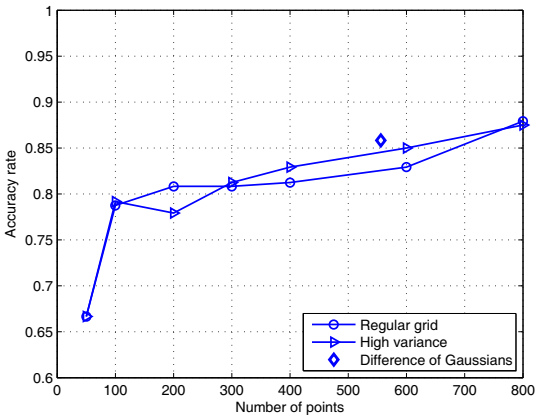

Fig. 5. Caltech 101 results: cubic kernel SVM, 4 classes, histogram entries raised to power 0.5

6. REFERENCES

[1] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, "The effects of segmentation and feature choice in a translation model of object recognition" , IEEE CVPR, pp II: 675-682, 2003.

[2] D. Blei. Probabilistic Models of Text and Images. PhD thesis, U.C. Berkeley, Division of Computer Science, 2004.

[3] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification via pLSA", Proceedings of the ECCV, 2006.

[4] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints", In Proc. of the 8th ECCV, Prague, May 2004.

[5] L. Fei-Fei and P. Perona, A Bayesian hieararchical model for learning natural scene categories". In Proc. of IEEE CVPR, 2005.

[6] L. Fei-Fei, R. Fergus and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories", IEEE CVPR, Workshop on Generative-Model Based Vision, 2004.

[7] T. Joachims, "Making large-scale SVM learning practical". Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[8] T. Kadir, M. Brady, "Saliency, Scale, and Image Description ", IJCV, vol. 45, no. 2, pp. 83-105, 2001.

[9] http://research.microsoft.com/vision/cambridge/recognition/default.htm

[10] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors", Proceedings of CVPR, Washington DC, pp. 66-75, 2004.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," IJCV, vol. 60, no. 2, pp. 91-110, 2004.

[12] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector", In Proc. ECCV. Springer-Verlag, 2002.

[13] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", IEEE Trans. on PAMI, vol. 27, pp. 1615-1630, 2005.

[14] J. Sivic and A. Zisserman, "Efficient Visual Content Retrieval and Mining in Videos", Pacific-Rim Conference on Multimedia, (PCM 2004), Tokyo, Japan, 2004.

[15] V. N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[16] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence", In Proc. ECCV, vol. 3, pp 255-271, Springer-Verlag, May 2002.

[17] J. Winn, A. Criminisi, and T. Minka, "Object Categorization by Learned Universal Visual Dictionary", In Proc. ICCV, Beijing, China, 2005.