COMPARING SIFT DESCRIPTORS AND GABOR TEXTURE FEATURES FOR CLASSIFICATION OF REMOTE SENSED IMAGERY

Yi Yang and Shawn Newsam

Electrical Engineering and Computer Science
University of California
Merced, CA 95344
yyang6, snewsam@ucmerced.edu

ABSTRACT

A richer set of land-cover classes are observable in satellite imagery than ever before due to the increased sub-meter resolution. Individual objects, such as cars and houses, are now recognizable. This work considers a new category of image descriptors based on local measures of saliency for labelling land-cover classes characterized by identifiable objects. These descriptors have been successfully applied to object recognition in standard (non-remote sensed) imagery. We show they perform comparably to state-of-the-art texture descriptors for classifying complex land-cover classes in high-resolution satellite imagery while being approximately an order of magnitude faster to compute. This speedup makes them attractive for realtime applications. To the best of our knowledge, this is the first time this new category of descriptors has been applied to the classification of remote sensed imagery.

Index Terms— Image classification, interest points, texture features, remote sensed imagery

1. INTRODUCTION

Novel geographic information platforms such as Google Earth and Microsoft Virtual Earth have enabled increased access to remote sensed imagery. These systems, however, only support visualization of the raw image data. Techniques for automatically annotating the image content would allow for richer user interaction and support exciting new applications.

Remote sensed image classification remains a challenging problem. While there have been noted successes over the last several decades in using descriptors such as spectral, shape, and texture features, significant opportunities remain.

In this paper, we compare established approaches to image classification based on Gabor texture features with classification based on a new category of low-level image analysis termed interest point descriptors. These descriptors have enjoyed surprising success recently when applied to a range of challenging computer vision problems. There has been little research, however, on applying them to remote sensed imagery. In previous work [1, 2], we explored content-based image retrieval of remote sensed imagery using interest point descriptors. In this paper, we turn to image classification.

We consider a diverse set of land-cover classes in which individual objects, such as cars and houses, are recognizable. We compare classification rates for Gabor texture features and Scale-Invariant Feature Transform (SIFT) descriptors [3], which were shown to outperform other interest point descriptors in an image matching task [4]. We present results using both support vector machines and maximum a posteriori classification using a large manually constructed ground truth dataset. We show this new category of image descriptors performs comparably to the state-of-the-art texture features while being approximately an order of magnitude faster to compute.

2. IMAGE DESCRIPTORS

2.1. SIFT Descriptors

SIFT based analysis involves detecting salient locations in an image and extracting descriptors that are distinctive yet invariant to changes in viewpoint, illumination, etc. We use the standard SIFT interest point detector and the standard SIFT histogram-of-gradients descriptor. These 128 dimension descriptors can be thought of roughly as summarizing the edge information in an image patch centered at an interest point.

We term the 128 dimension descriptors the *local SIFT descriptors* for an image. We also compute a single *global SIFT descriptor*. This global descriptor is a frequency count of the quantized local descriptors. We use the k-means algorithm to cluster a large collection of SIFT descriptors and label each local descriptor with the id of the closest cluster center. The global SIFT descriptor is then computed as

$$SIFT_{GLOBAL} = [t_0, t_1, \ldots, t_{k-1}] \ , \qquad (1)$$

where t_i is number of occurrences of the quantized SIFT features with label i. $SIFT_{GLOBAL}$ is similar to a term vector in document retrieval. The global SIFT descriptors are normalized to have unit length to account for the varying number of local SIFT descriptors per image.

2.2. Gabor Texture Features

Gabor texture features have proven to be effective for analyzing remote sensed imagery [5]. They were standardized in 2002 by the MPEG-7 Multimedia Content Description Interface [6] after they were shown to outperform other texture features in which one of the evaluation datasets consisted of remote sensed imagery.

Gabor texture features are extracted by applying a bank of scale and orientation selective Gabor filters to an image [7]. A filterbank with R orientations and S scales results in a total of RS filtered images

$$f'_{11}(x,y), \ldots, f'_{RS}(x,y) \ . \tag{2}$$

Considered differently, this data cube represents an RS dimension feature vector at each pixel location. We term these the set of *local Gabor texture features* for an image. We form a single *global Gabor texture feature* by computing the mean and standard deviation of the filtered images. A $2RS$ dimension feature vector, $Gabor_{GLOBAL}$, is formed as

$$Gabor_{GLOBAL} = [\mu_{11}, \sigma_{11}, \mu_{12}, \sigma_{12}, \ldots, \mu_{RS}, \sigma_{RS}] \ , \tag{3}$$

where μ_{rs} and σ_{rs} are the mean and standard deviation of $f'_{rs}(x,y)$. Finally, to normalize for differences in range, each of the $2RS$ components is scaled to have a mean of zero and a standard deviation of one across a dataset.

3. CLASSIFICATION METHODS

3.1. Maximum A Posteriori

Image classification based on local features is performed using maximum a posteriori (MAP) classifiers. An image with the set of local features, $\mathbf{x}$, is assigned to class c^* where

$$c^* = \arg\max_{1 \le c \le C} P(c|\mathbf{x}) \ . \tag{4}$$

The feature distributions of the classes are modelled by Gaussian mixtures so that the posterior probabilities, $p(c|\mathbf{x})$, are computed using Bayes' rule where the class-conditioned probabilities, $p(\mathbf{x}|c)$, are

$$p(\mathbf{x}|c) = \sum_{j=1}^{J} P(j|c)\, p(\mathbf{x}|j,c) \ . \tag{5}$$

The class- and mixture-conditioned probabilities for a single feature vector are

$$p(x|j,c) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \, \mathrm{e}^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j (x-\mu_j)} \tag{6}$$

where μ_j is the mean vector and Σ_j is the covariance matrix of the j^{th} mixture for class c. The local features are considered to be independent so the joint probability of a set of features is computed as the product of the individual probabilities. The Gaussian mixture model (GMM) parameters, μ_j and Σ_j, are learned from a training set using the expectation-maximization (EM) algorithm [8].

Design decisions for the MAP classifiers include the number of mixtures in the GMMs and the form of the covariance matrices. We investigate these in the experiments below.

3.2. Support Vector Machines

The global features are classified using support vector machines (SVMs). When applied to classification, SVMs seek the optimal separating hyperplane between two classes, typically in a higher dimensional space than the original features. In our multi-class problem, we use a "one-against-one" strategy wherein a binary classifier is trained for each pair of classes. Unknown samples are classified using a majority voting strategy among the binary classifiers. We use the LIBSVM package [9] in the experiments below.

4. DATASET

A collection of 1-m panchromatic IKONOS satellite images is used for evaluation. A ground truth dataset consisting of ten sets of 100 64-*by*-64 pixel images was manually extracted from the IKONOS images for the following land cover classes: aqueduct, commercial, dense residential, desert chaparral, forest, freeway, intersection, parking lot, road, and rural residential. Examples are shown in Figure 1.

128 dimension local SIFT descriptors were extracted from the images. Figure 2 shows the locations of these features as determined by the SIFT detector. A large set of SIFT descriptors randomly sampled from the full IKONOS image set was clustered using the k-means algorithm. The local SIFT descriptors were quantized by assigning the label of the closest cluster center. The frequency counts of these labels form the global SIFT descriptors (see Equation 1). Previous work [1] showed that 50 clusters was optimal for content-based image retrieval using quantized SIFT descriptors. Our global SIFT features thus have dimension 50.

Gabor texture features were extracted from the images using a filterbank tuned to five scales and six orientations [7]. The 30 dimension feature vectors, one at each pixel location, form the local Gabor texture features. The mean and standard deviation of each filter output form the 60 dimension global Gabor texture features (see Equation 3).

Each ground truth image is thus represented by:

- A set of 128 dimension SIFT descriptors. Each image has 59.1 descriptors on average.
- A 50 dimension global SIFT descriptor.
- A set of 30 dimension local Gabor texture features, one at each pixel location.
- A 60 dimension global Gabor texture feature.

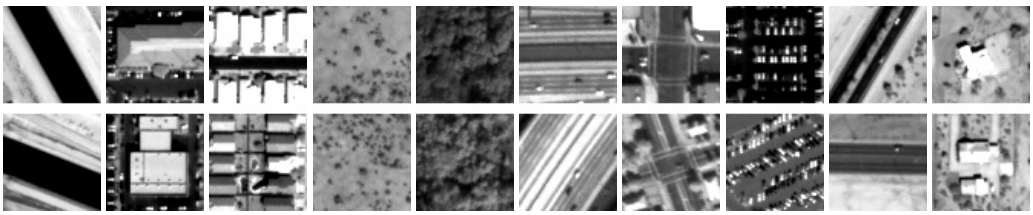

Fig. 1. Two examples from each of the ground truth classes. (a) Aqueduct. (b) Commercial. (c) Dense residential. (d) Desert chaparral. (e) Forest. (f) Freeway. (g) Intersection. (h) Parking lot. (i) Road. (j) Rural residential.

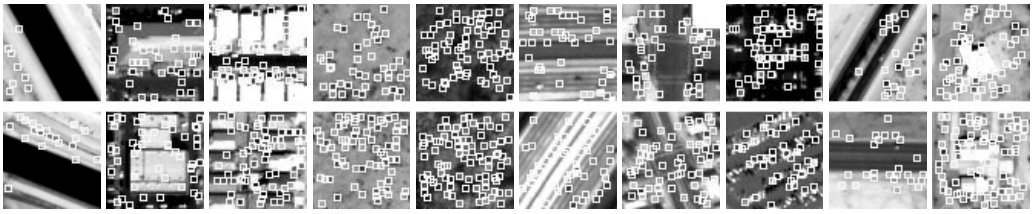

Fig. 2. The interest point locations for the ground-truth images in Figure 1.

The difference in feature extraction times is significant. It took approximately 353 seconds to extract the Gabor texture features from the 1,000 images in the ground truth dataset (using a typical desktop workstation). By comparison, it took only approximately 51 seconds to extract the SIFT descriptors. While the extraction software was not optimized and the timing measurements were not controlled, we believe this order-of-magnitude difference between the two features is to be expected. Efficient extraction is a noted strength of the SIFT descriptors.

5. EXPERIMENTS AND RESULTS

The feature and classifier combinations are evaluated by ten-fold cross validation. The ground truth dataset is split into ten partitions each containing ten images from each of the ten classes. Ten rounds of training and testing are performed in which nine partitions are used for training and the remaining partition is used for testing. Each round uses a different partition for testing. A single classification rate is computed indicating the percent of the 1,000 images that are assigned to the correct class.

The MAP classifiers use the local features. A separate classifier is trained for each class. All of the local SIFT descriptors for an image are used in training and testing. Due to the large number of local Gabor texture features–4,096 for a 64-*by*-64 pixel image–only a random sampling of 100 features per image is used. Using a larger number of samples did not have a significant effect on the classification rates.

We first investigated the number and shapes of the Gaussians in the mixture models. We found that the classification rates did not vary significantly between spherical Gaussians (diagonal covariance matrix with the same value at each entry), elliptical Gaussians with axes aligned with the dimensions of the feature space (diagonal covariance matrix with possibly different values), and elliptical Gaussians with axes at any orientation (covariance matrix with possibly non-zero off diagonal entries). Since the minimum description length principle favors fewer parameters, spherical Gaussians are used in the remainder of the results. Training via the EM algorithm is also significantly faster in the spherical case.

Figure 3 plots the MAP classification rate versus the number of mixtures in the GMMs. The rate peaks at around five mixtures for both features. We therefore use five mixtures in the remainder of the results.

The SVM multi-class classifiers use the global features. One classifier is trained and tested during each round of the ten-fold cross validation. We use a linear kernel. Initial investigation into using other kernels, such as polynomial and radial bases function produced similar classification rates.

Table 1 shows the classification rates for four different combinations of features and classifiers: 1) local SIFT descriptors classified using MAP classifiers; 2) local Gabor texture features classified using MAP classifiers; 3) global SIFT descriptors classified using SVMs; and 4) global Gabor texture features classified using SVMs.

6. DISCUSSION

The ground truth dataset used in the experiments contains substantial within class variability, as illustrated in Figure 1, so classification rates nearing 90% are significant. A re-

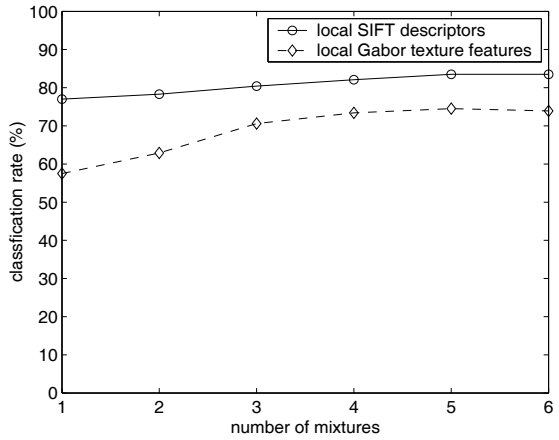

Fig. 3. MAP classification rate versus number of mixtures.

Table 1. Classification rates for different feature and classifier combinations.

	SIFT	**Gabor**
MAP	84.5%	73.9%
SVM	76.2%	89.8%

cent retrospective on satellite image classification reports that the average classification rate over the last 15 years is only 76.19% with a standard deviation of 15.59% [10]. Of course, the classification rate depends on the difficulty of the problem so talking about an average rate across problems is not that meaningful. Nonetheless, some of the feature and classifier combinations presented above have rates at the top of this distribution, a result underscored by the fact that only spatial information is used. Incorporating spectral information would certainly improve the results.

The best results are achieved by Gabor texture features and SVM classification. However, the next best combination, SIFT descriptors and MAP classification, perform comparably so that the order of magnitude difference in feature extraction speed could be a deciding factor especially for realtime application. These two best results are also achieved with features that are fundamentally different representations. A global Gabor texture feature is associated with a well-defined region, usually rectangular in shape. The local SIFT descriptors can represent more general regions. While both are applied here to classifying entire images, they could enable different techniques when applied to classifying regions within larger images. We are exploring the application of SIFT descriptors to labelling arbitrary shaped regions in remote sensed imagery. The controlled experiments using the ground truth datasets presented in this paper represent an important foundation for this future work.

7. ACKNOWLEDGEMENTS

This work used Robert Hess' OpenCV implementation of the SIFT detector and descriptor. The IKONOS images were made available by a grant from Lockheed Martin Corporation and are copyright Lockheed Martin Corporation, all rights reserved. This work was funded in part by an Early Career Scientist and Engineer Award from the Department of Energy.

8. REFERENCES

[1] S. Newsam and Y. Yang, "Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery," in *ACM International Symposium on Advances in Geographic Information Systems (ACM GIS)*, 2007.

[2] S. Newsam and Y. Yang, "Geographic image retrieval using interest point descriptors," in *International Symposium on Visual Computing (ISVC)*, 2007.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[5] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Journal of Applied Optics: Information Processing*, vol. 43, no. 2, pp. 210–217, 2004.

[6] B. S. Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG7: Multimedia Content Description Interface*, John Wiley & Sons, 2002.

[7] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[9] "Libsvm–a library for support vector machines," http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[10] G. G. Wilkinson, "Results and implications of a study of fifteen years of satellite image classification experiments," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 433–440, 2005.