A New Color SIFT Descriptor and Methods for Image Category Classification

Abhishek Verma, Sugata Banerji, and Chengjun Liu
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA
{av56, sb256, chengjun.liu}@njit.edu

Abstract — **We first propose in this paper a new oRGB-SIFT descriptor, and then integrate it with other color SIFT features to produce the Color SIFT Fusion (CSF) and the Color Grayscale SIFT Fusion (CGSF) methods for image category classification. The effectiveness of our proposed representation and methods are evaluated on three representative, large scale, and grand challenging datasets. The experimental results show that (i) our oRGB-SIFT descriptor improves recognition performance over other color SIFT descriptors; (ii) both the CSF method and the CGSF method perform better than the other color SIFT descriptors or the methods combining color features and SIFT.**

Keywords- oRGB-SIFT descriptor, Color SIFT Fusion (CSF), Color Grayscale SIFT Fusion (CGSF), image category classification

I. INTRODUCTION

Color features provide powerful information for object and scene classification, indexing and retrieval. For certain types of applications like identifying flowers, animals, certain natural scene categories and geographical features from satellite images, color can be a highly discriminative feature for retrieval.

Two important criteria for color feature detectors are that they should be stable under varying viewing conditions, such as changes in illumination, shading, highlights, and they should have high discriminative power. In this paper, we first propose a new oRGB-SIFT feature representation, and then integrate it with other color SIFT features to produce the Color SIFT Fusion (CSF) and the Color Grayscale SIFT Fusion (CGSF) methods for image category classification. The effectiveness of our proposed representation and methods are evaluated on three representative, large scale, and grand challenging datasets. The experimental results show that (i) our oRGB-SIFT descriptor improves recognition performance over other color SIFT descriptors; (ii) both the CSF method and the CGSF method perform better than the other color SIFT descriptors or the methods combining color features and SIFT.

Recently, there has been much emphasis on the detection and recognition of locally affine invariant regions for image category classification [1]-[5]. Affine region detectors when combined with intensity Scale-Invariant Feature Transform (SIFT) descriptor [2] has been shown to outperform many alternatives [4]. We extend this descriptor to different color spaces, including the recently proposed oRGB color space [6]. In order for us to be able to accurately measure the discriminative power of these descriptors, we run our experiments on three different widely varying datasets and perform multiple cross validations. Our results show that the color descriptors almost always outperform the grayscale descriptor on all datasets and the fusion of color descriptors improves over the grayscale descriptor by a good margin. We setup the performance baseline using the dense color histogram and make a comparison with the sparse color SIFT descriptor. Our results show that the local color descriptors significantly improve the recognition performance.

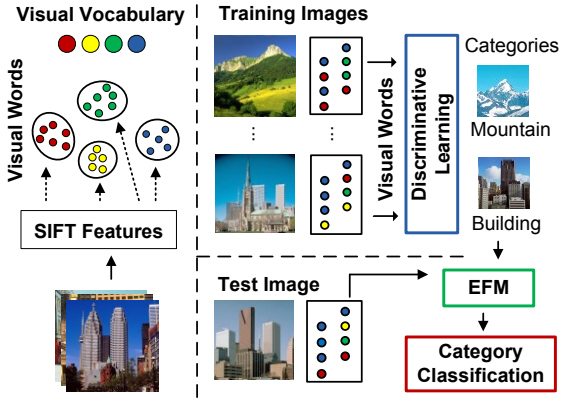

Figure 1. An overview of feature extraction after Harris-affine region detection and visual vocabulary formation on SIFT features, visual words, learning, and classification stages.

II. COLOR DESCRIPTORS, FEATURE EXTRACTION, AND CLASSIFICATION

The choice of an appropriate color space is of prime importance for color image recognition. We measure the performance of descriptors in the RGB, HSV, rgb, oRGB, and YCbCr color spaces.

A. Dense Color Histogram, Color SIFT Descriptor and Feature Extraction

We perform feature extraction based on two different methods. The first method constructs a dense color histogram. The system starts with a color image as an input and first splits it into three separate color component images. Next step is to compute histograms from each of the color channels. After normalization the individual histograms are concatenated to form a compact fixed length feature vector.

978-981-08-6846-8/$26.00 ©2010 IRAST

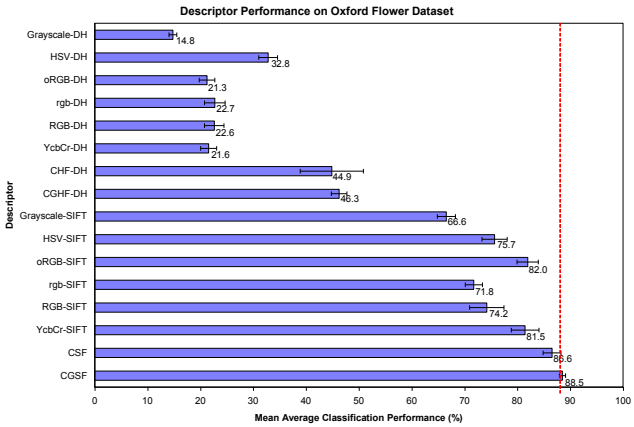

Figure 2. Descriptor performance on Oxford Flower dataset averaged over 17 categories. Error bars indicate the standard deviation in mean average rate. Dashed line indicates the lower bound of the CGSF confidence interval.

The second method extracts the SIFT descriptor from scale invariant key points [1]. We compute the key points on the intensity channel by the Harris-affine point detector as it has previously shown good performance [3]. The color SIFT descriptor is formed by computing SIFT descriptors on each of the three color channels independently. This results in three 128 dimensional vectors from each color channel, after concatenation we get a 384 dimensional descriptor for each key point. The SIFT descriptor is able to describe the local spatial information and it is more robust to transformations.

B. *Clustering, Visual Vocabulary Tree and Visual Words*

The SIFT descriptors are quantized with the vocabulary tree using the bag-of-words model [5]. Thus, each image is represented as a fixed length feature vector of visual words, provided from a visual vocabulary. The visual vocabulary tree defines a hierarchical quantization, which is constructed with hierarchical *k*-means clustering on a large set of randomly chosen features from the training images. We build a vocabulary tree of 6561 leaf nodes and $k = 9$. See fig. 1 for an overview of the processing pipeline.

C. *Learning and Classification*

We perform learning and classification using Enhanced Fisher Linear Discriminant Model (EFM) [7]. The EFM method first applies Principal Component Analysis (PCA) to reduce the dimensionality of the input pattern vector. A popular classification method that achieves high separability among the different pattern classes is the Fisher Linear Discriminant (FLD) method. The FLD method, if implemented in an inappropriate PCA space, may lead to over fitting. The EFM method, which applies an eigenvalue spectrum analysis criterion to choose the number of principal components to avoid over fitting, thus improves the generalization performance of the FLD. The EFM method thus derives an appropriate low dimensional representation from the color histogram or color SIFT descriptor and further extracts the EFM features for pattern classification. We compute similarity score between a training feature vector and a test feature vector using the cosine similarity measure.

TABLE I
Comparison with Other Methods on Oxford Flower Dataset

Our method		**[8]**		**[9]**	
RGB-SIFT	74.22	Color	73.7	Shape	68.88
HSV-SIFT	75.69	Shape	71.8	Color	59.71
YCbCr-SIFT	81.47	Texture	56.0*	Texture	59.00
oRGB-SIFT	81.96				
CSF	**86.57**				
CGSF	**88.53**	Fusion	81.3	Fusion	82.55

All values are in %. * Approximate value inferred from fig. 12 in [8]

III. EXPERIMENTAL RESULTS

A. *Assessment of Color Descriptors, the CSF and CGSF Methods on the Oxford Flower Dataset*

The Oxford Flower dataset [8] consists of 17 species of flowers with 80 images per category. All the images are in color, JPEG format and the mean image size is 560x560 pixels. There are species that have a very unique visual appearance, e.g. Fritillaries and Tigerlilies, as well as species with very similar appearance, for example Dandelions and Coltsfoot. The large intra-class variability and the small inter-class variability make this dataset very challenging.

Experimental setup consists of three sets of 40 training images and 20 test images per class (same data splits as in [8]). See fig. 2 for classification performance (Dense Histogram (DH), Color Histogram Fusion (CHF), Color Gray Histogram Fusion (CGHF), Color SIFT Fusion (CSF), and Color Gray SIFT Fusion (CGSF)).

On dense histogram, HSV features give us the best success rate of 32.8%. Combined color histograms reach the rate of 44.9% and fusing color and gray histogram reaches 46.3%. On sparse SIFT descriptor, oRGB-SIFT gives us the best performance of 82%, followed by YCbCr-SIFT at 81.5%. Fusion of five color SIFT descriptors reaches 86.6%. Fusion of color and gray SIFT yields a recognition rate as high as 88.5%.

Table I shows a comparison of our results with those obtained by Nilsback [8] and Varma [9]. Our technique

TABLE II
Descriptor Performance Split-out for Oxford Flower Categories

Category*	**CGSF**	**CSF**	**oRGB SIFT**	**YCbCr SIFT**	**HSV SIFT**	**Gray SIFT**
sunflower	100.0	100.0	100.0	100.0	100.0	95.0
daisy	98.3	98.3	100.0	98.3	96.7	93.3
tigerlily	98.3	96.7	98.3	95.0	71.7	76.7
windflower	98.3	91.7	91.7	91.7	93.3	90.0
bluebell	93.3	90.0	83.3	76.7	78.3	48.3
coltsfoot	93.3	95.0	90.0	93.3	81.7	83.3
dandelion	93.3	91.7	91.7	91.7	88.3	81.7
pansy	93.3	86.7	75.0	78.3	83.3	75.0
cowslip	90.0	86.7	81.7	88.3	70.0	45.0
lilyvalley	90.0	90.0	81.7	80.0	78.3	78.3
buttercup	85.0	81.7	83.3	81.7	71.7	48.3
fritillary	85.0	81.7	80.0	83.3	76.7	75.0
iris	85.0	80.0	73.3	70.0	73.3	78.3
daffodil	80.0	83.3	76.7	71.7	60.0	43.3
snowdrop	80.0	76.7	60.0	56.7	55.0	63.3
tulip	73.3	73.3	63.3	70.0	56.7	33.3
crocus	68.3	68.3	63.3	58.3	51.7	23.3
Mean	**88.5**	**86.6**	**82.0**	**81.5**	**75.7**	**66.6**

All values are in %. *Sorted on CGSF results.

Figure 3. Examples of correctly classified images of the Bluebell (top) and Lily Valley (bottom) categories from the Oxford Flower dataset.

outperforms the state of the art on this dataset even without combining color descriptors or considering texture and shape features independently. Each of the four color SIFT descriptors outperform descriptors in [8]-[9]. Combining SIFT descriptors (CSF & CGSF) improves over the fusion result in [8] and SVM 1-vs-All fusion result in [9], previously the best result on this data set.

Table II shows the success rate of various descriptors on the different flower categories. It can be seen that the CGSF recognition rate for the top ten categories lies between 90% and 100%. Fig. 3 shows some test images from this dataset that were classified correctly by our technique. Note the large intra-class variation.

B. Assessment of Color Descriptors, the CSF and CGSF Methods on the Caltech 256 Dataset

The Caltech 256 dataset [10] holds 30,607 images divided into 256 object categories and a clutter class. The images have high intra-class variability and high object location variability. Each category contains at least 80 images, a maximum of 827 images and 119 mean number of images per category. The images represent a diverse set of lighting conditions, poses, backgrounds, and sizes. Images are in color, in JPEG format with only a few in grayscale. The average size of each image is 351x351 pixels.

On this dataset, we perform two sets of experiments; one

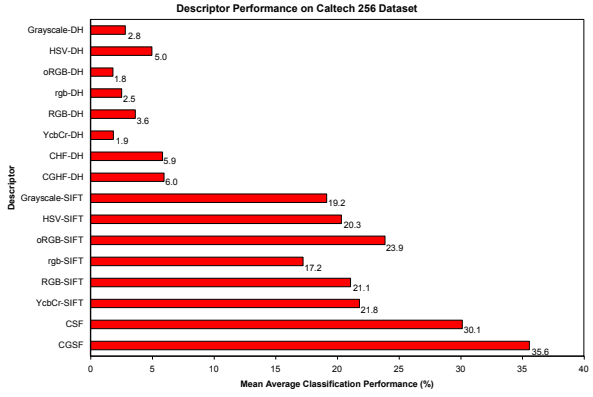

Figure 4. Descriptor performance on Caltech 256 dataset averaged over 256 object categories.

Table III
Descriptor Performance Split-out for Top 15 Caltech 256 Categories

Category*	CGSF	CSF	oRGB SIFT	YCbCr SIFT	RGB SIFT	Gray SIFT
leopards	100	100	96	88	100	100
car-side	100	100	96	76	100	100
faces-easy	100	100	88	84	100	88
sunflower	96	96	80	60	92	80
hibiscus	92	88	56	76	72	56
h-simpson	92	84	48	48	52	44
tower-pisa	92	84	80	84	76	60
brain	88	76	56	76	48	48
chessboard	88	88	84	80	80	80
frenchhorn	88	80	60	48	64	72
s-a-knife	64	60	32	48	44	28
fire-truck	64	52	48	32	48	20
school-bus	64	64	44	48	72	64
l-mower	64	64	28	40	32	40
zebra	64	64	48	56	60	32

All values are in %. *Sorted on top 15 CGSF results.

set for the dense histogram and another set for sparse SIFT descriptors from five different color spaces, grayscale and their fusion. For each class, we make use of 50 images for training and 25 images for test. The data splits are the ones that are provided on the Caltech website [10]. Fig. 4 shows the detailed performance of our EFM based classification technique on this dataset. The best recognition rate that we obtain is 35.6%, which is a very respectable value for a dataset of this size and complexity. Note that dense histograms perform poorly on this dataset as the intra-class variability is very high. Also, in several cases the object occupies a small portion of the full image. oRGB-SIFT achieves the classification rate of 23.9% and once again outperforms other color descriptors. YCbCr-SIFT comes close second with 21.8% recognition rate. Fusion of color SIFT descriptors improves over grayscale SIFT by 11%. Grayscale SIFT shows more distinctiveness than rgb-SIFT, interestingly it also improves the fusion (CGSF) result by a good 5% over CSF.

Table III shows the descriptor performance for the top 15 categories from this dataset. Fig. 5 shows some sample images from the Bat and Swiss Army Knife categories that were classified successfully. Note the variations in viewpoint, illumination and scale across different images of the same category. The CGSF recognition rate for the top 10 categories lies between 88% and 100% with three categories having full success rate.

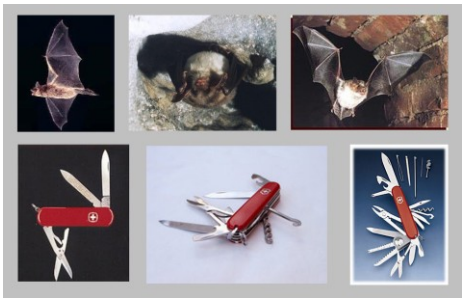

Figure 5. Examples of correctly classified images of the Bat (top) and Swiss Army Knife (bottom) categories from the Caltech 256 dataset.

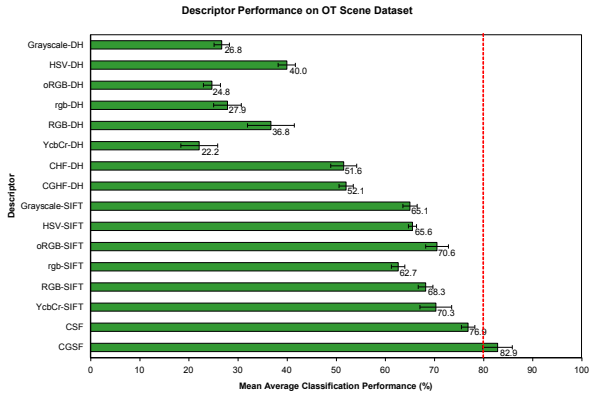

Figure 6. Descriptor performance on OT Scene dataset averaged over 8 categories. Error bars indicate the standard deviation in mean average rate. Dashed line indicates the lower bound of the CGSF confidence interval.

C. *Assessment of Color Descriptors, the CSF and CGSF Methods on the OT Scene Dataset*

The Olivia and Torralba (OT) Scene dataset [11] has 2,688 images classified as eight categories: 360 coast, 328 forest, 374 mountain, 410 open country, 260 highway, 308 inside of cities, 356 tall buildings, and 292 streets. All of the images are in color, in JPEG format, and the average size of each image is 256x256 pixels. There is a large variation in light, pose and angles, along with a high intra-class variation. The sources of the images vary (from commercial databases, websites, and digital cameras) [11].

From each class, we use 100 images for training, 50 images for testing the performance, and perform multiple cross validations. HSV histogram achieves 40% success rate. Combining histograms reaches a rate of 52%. oRGB-SIFT is the best color descriptor at 70.6%. The combined descriptors CSF and CGSF give a mean average performance of 76.9% and 82.9% respectively. See fig. 6 for details. Fig. 7 shows the success rate of different descriptors across all eight categories. Seven categories achieve a success rate of over 80% for CGSF. Fig. 8 shows some example images that were classified correctly by our system.

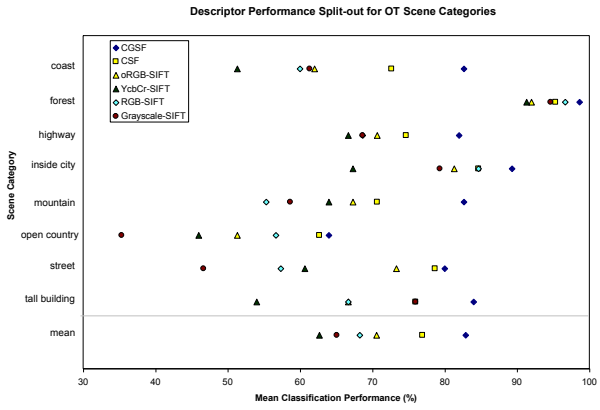

Figure 7. Descriptor performance split-out for OT Scene categories.

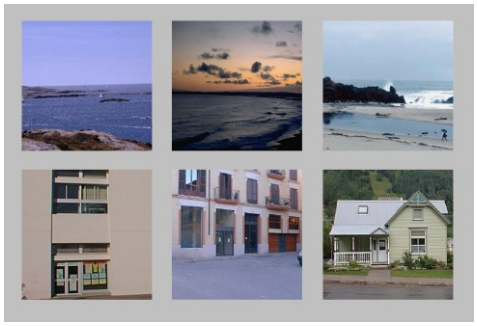

Figure 8. Examples of correctly classified images of the Coast (top) and Inside City (bottom) categories from the OT Scene dataset.

IV. CONCLUSION

We proposed a new oRGB-SIFT feature representation, and then integrated it with other color SIFT features to produce the Color SIFT Fusion (CSF) and the Color Grayscale SIFT Fusion (CGSF) methods for image category classification. Experimental results on three large representative datasets show the effectiveness of the proposed methods for image category classification.

REFERENCES

[1] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," Int. Journal of Computer Vision, vol. 73, Jun. 2007, pp. 213-238, doi: 10.1007/s11263-006-9794-4.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int. Journal of Computer Vision, vol. 60, Nov. 2004, pp. 91-110, doi:10.1023/B:VISI.0000029664.99615.94.

[3] K. Mikolajczyk, et al., "A Comparison of Affine Region Detectors," Int. Journal of Computer Vision, vol. 65, Nov. 2005, pp. 43-72, doi: 10.1007/s11263-005-3848-x.

[4] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, Oct. 2005, pp. 1615-1630, doi: 10.1109/TPAMI.2005.188.

[5] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," IEEE Int. Conf. on Computer Vision, vol. 2, Oct. 2003, pp. 1470-1477, doi: 10.1109/ICCV.2003.1238663.

[6] M. Bratkova, S. Boulos, and P. Shirley, "oRGB: A Practical Opponent Color Space for Computer Graphics," IEEE Computer Graphics and Applications, vol. 29, Jan. 2009, pp. 42-55, doi: 10.1109/MCG.2009.13.

[7] C. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases," IEEE Trans. on Image Processing, vol. 9, Jan. 2000, pp. 132-137, doi: 10.1109/83.817604.

[8] M. E. Nilsback and A. Zisserman, "A Visual Vocabulary for Flower Classification," Int. Conf. on Computer Vision and Pattern Recognition, vol. 2, Jun. 2006, pp. 1447-1454, doi: 10.1109/CVPR.2006.42.

[9] M. Varma and D. Ray, "Learning the Discriminative Power-Invariance Trade-Off," IEEE Int. Conf. on Computer Vision, Oct. 2007, pp. 1-8, doi:10.1234/12345678.

[10] http://www.vision.caltech.edu/Image_Datasets/Caltech256/

[11] A. Olivia and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," Int. Journal of Computer Vision, vol. 42, May 2001, pp. 145-175, doi: 10.1023/A:101113963172