EFM-HOG: Improving Image Retrieval in the Wild

Sugata Banerji

Lake Forest College
555 North Sheridan Road
Lake Forest, IL 60045, USA
Email: banerji@lakeforest.edu

Atreyee Sinha

Edgewood College
1000 Edgewood College Drive
Madison, WI 53711, USA
Email: asinha@edgewood.edu

***Abstract*—The problem of retrieving images from a dataset, which are similar to a query image is an important high-level vision problem. Different tasks define similarity based on different low-level features like shape, color or texture. In the presented work, we focus on the problem of retrieval of images of similarly shaped objects, with the query being an object selected from a query image at runtime. Towards this end, we propose a novel shape representation and associated similarity measure, which exploits the dimensionality reduction and feature extraction methods of Principal Component Analysis (PCA) and Enhanced Fisher Model (EFM). The effectiveness of this representation is demonstrated on large-scale image datasets for the task of object retrieval and the performance is compared to Histograms of Oriented Gradients (HOG).**

Keywords–Computer Vision; Principal Component Analysis; Fisher Linear Discriminant; Enhanced Fisher Model; Histogram of Oriented Gradients; Image Search.

I. Introduction and Background

With the enormous popularity of digital devices equipped with cameras, along with the wide access to high speed Internet and cloud storage, several applications based on image search and retrieval have emerged. Such applications include augmented reality, geo-localization, security and defense, educational uses, to name a few. Billions of images are uploaded and shared over social media and web sharing platforms everyday, giving rise to a greater need for systems that can retrieve images similar to a query image from a dataset. Traditional approaches of content-based image retrieval are based upon low level cues such as shape, color and texture features. In this paper, we are trying to address the problem of retrieving images that have similarity in the shapes. Specifically, we select a window from a query image surrounding an object of interest and want to be able to retrieve similarly shaped objects from other images in the dataset, which are taken "in the wild", i.e., user generated content without any control. Towards that end, we investigate and propose a novel representation and retrieval technique that is based on shape features, dimensionality reduction and discriminant analysis and is robust to the slight changes in the window object selection.

The Histograms of Oriented Gradients (HOG) feature vector [1], originally proposed for pedestrian detection, is very popular among researchers for shape matching. It has successfully been combined with other techniques [2] and fused with other descriptors [3] for scene image classification. HOG has also given rise to other extremely successful object detection techniques, such as Deformable Part Models (DPM) [4]. More complicated descriptors [5] have been used for image retrieval with reasonable success. However, such methods are time consuming and more processor-intensive as compared to simple HOG matching. In recent years, handcrafted features have declined in popularity due to the success of deep neural networks in object recognition [6]–[8], but such methods are not without their drawbacks. Deep neural networks require a lot of processor time and run better on specialized hardware. They also require far greater number of training images than are available in a small or medium-sized dataset. For these reasons, enhancing simple handcrafted features like HOG can be effective for solving small-scale retrieval problems more effectively than more complex methods.

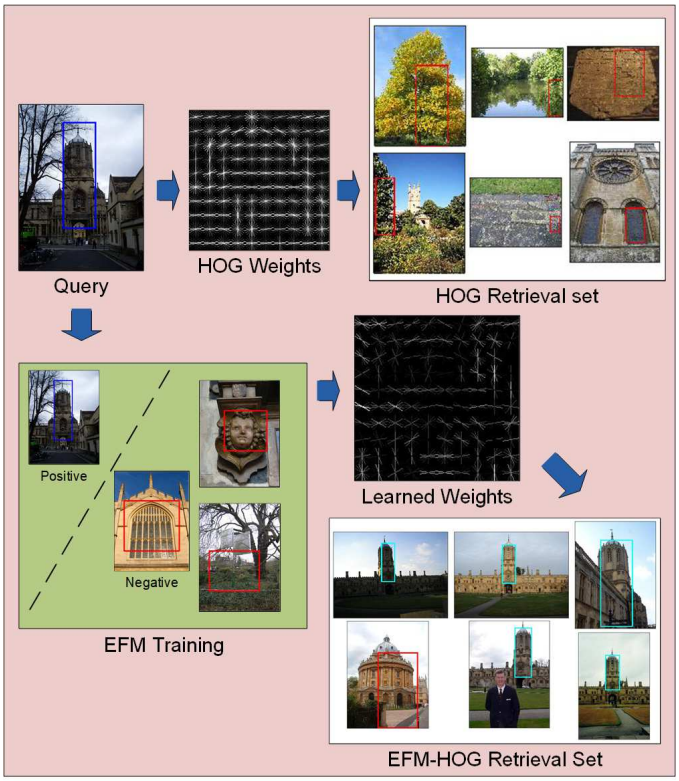

Figure 1. The proposed image representation aims at enhancing the HOG-based retrieval set by training an EFM-based classifier. The method is described in more detail in Section II

Simple HOG matching, however, poses significant challenges in effective image retrieval due to the fact that the apparent shape of the query object may change considerably

 ISBN: 978-1-61208-716-0

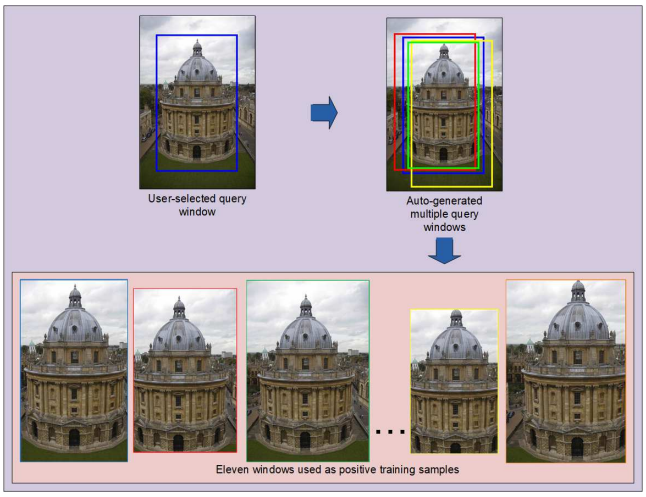

Figure 2. Auto-generation of offset windows to be used as positive training samples during querying. The window dimensions and offsets shown are only representative.

between images due to differences in lighting, viewing angle, scale and occlusion. This is particularly true for content generated by users in the wild. In effect, every query image is an exemplar of its own class and a retrieval system must be trained to treat it that way. In [9], this idea is handled using a Support Vector Machine (SVM) [10]. Instead of an SVM, here we introduce the novel idea of enhancing the HOG features by the EFM process [11] because it produces a low-dimensional representation, which is important from the computational aspect. Principal Component Analysis (PCA) has been widely used to perform dimensionality reduction for image indexing and retrieval [11]. The Enhanced Fisher Model (EFM) feature extraction method has achieved good success rates for the task of image classification and retrieval [3]. In the proposed method, which is represented schematically in Figure 1, we show this method to be effective in isolating the query object from the background.

The rest of this paper is organized as follows. Section II outlines in detail the method proposed in this paper. The datasets used and the experiments performed are detailed in Section III. Finally, we list our conclusions and directions for future research in Section IV.

II. Proposed Method

A. Window Generation

We start with generating objectness windows from each image. We use the method used by [12], which designs an objectness measure and explicitly trains it to distinguish windows containing an object from background windows. This method uses five objectness cues - namely, multi-scale saliency, color contrast, edge density, superpixels straddling, and location and size - and combines them in a Bayesian framework. We select the 25 highest-scoring windows from each image in our dataset and extract HOG features from these windows.

While testing our system, the user generates a window on the query image manually roughly enclosing the object of interest. Then, we automatically select 10 slightly offset versions of this window. Eight of these are generated by moving the user-selected window to the right, left, up, down, up-right, up-left, down-right and down-left by 5%, respectively. Two windows are generated by expanding and contracting the user's selection by 5%, respectively. Features are now extracted from these 10 as well as the original window for further processing. This process is represented in Figure 2.

B. HOG

The idea of HOG rests on the observation that local features such as object appearance and shape can often be characterized well by the distribution of local intensity gradients in the image [1]. HOG features are derived from an image based on a series of normalized local histograms of image gradient orientations in a dense grid [1]. The final HOG descriptors are formed by concatenating the normalized histograms from all the blocks into a single vector.

Figure 3 demonstrates the formation of the HOG vector for a window selected from an image. We use the HOG implementation in [13] for both generating the descriptors and rendering the visualizations used in this paper.

C. Dimensionality Reduction

PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation. Specifically, let $\mathcal{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is defined as follows [14]:

$$S = \mathcal{E}\{[\mathcal{X} - \mathcal{E}(\mathcal{X})][\mathcal{X} - \mathcal{E}(\mathcal{X})]^t\} \quad (1)$$

where $\mathcal{E}(\cdot)$ represents expectation and t the transpose operation. The covariance matrix S is factorized as follows [14]:

$$S = \Phi \Lambda \Phi^t \quad (2)$$

where $\Phi = [\phi_1 \phi_2 \cdots \phi_N]$ is an orthogonal eigenvector matrix and

$$\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$$

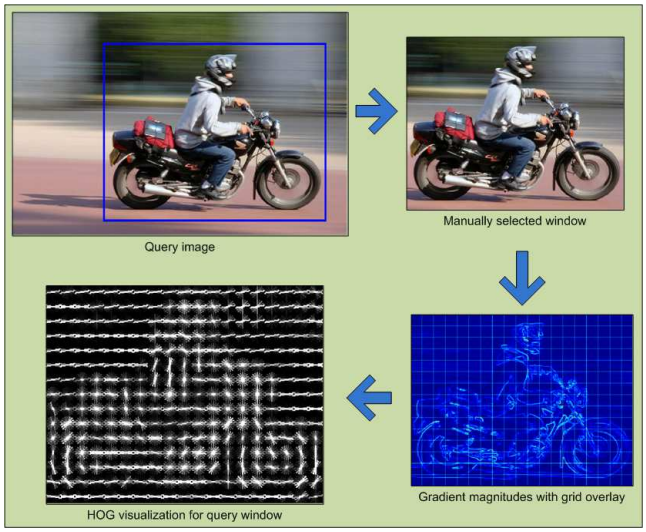

Figure 3. Formation of the HOG descriptor from a query image window.

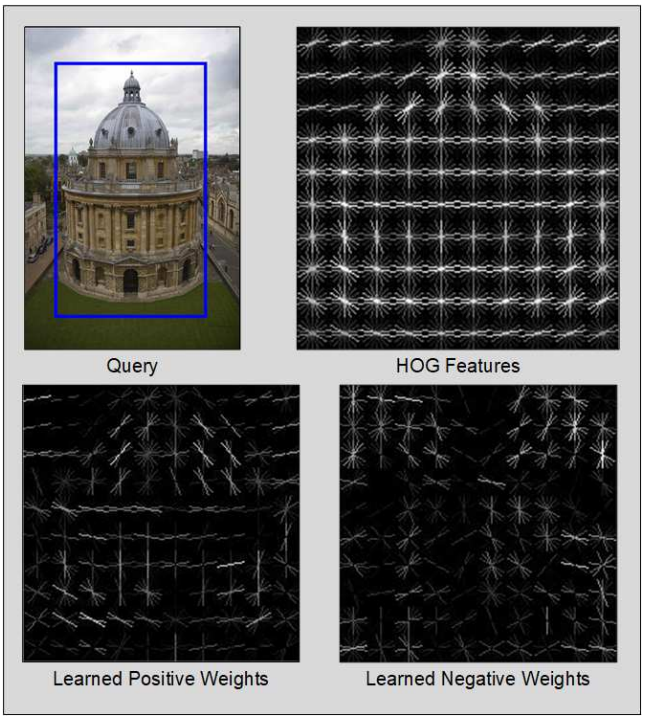

Figure 4. The positive and negative weights learned from the HOG features through the EFM discriminative feature extraction process.

a diagonal eigenvalue matrix with diagonal elements in decreasing order. An important application of PCA is the extraction of the most expressive features of $\mathcal{X}$. Towards that end, we define a new vector $\mathcal{Y}$: $\mathcal{Y} = P^t\mathcal{X}$, where $P = [\phi_1\phi_2 \dots \phi_K]$, and $K < N$. The most expressive features of $\mathcal{X}$ thus define the new vector $\mathcal{Y} \in \mathbb{R}^K$, which consists of the most significant principal components.

D. EFM

The features obtained after dimensionality reduction by PCA as discussed in Section II-C are the most expressive features for representation. However, they are not the optimum features for classification. Fisher's Linear Discriminant (FLD), a popular method in pattern recognition, first applies PCA for dimensionality reduction and then discriminant analysis for feature extraction. Discriminant analysis often optimizes a criterion based on the within-class and between-class scatter matrices S_w and S_b, which are defined as follows [14]:

$$S_w = \sum_{i=1}^{L} P(\omega_i)\mathcal{E}\{(\mathcal{Y} - M_i)(\mathcal{Y} - M_i)^t|\omega_i\} \quad (3)$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M)(M_i - M)^t \quad (4)$$

where $P(\omega_i)$ is *a priori* probability, ω_i represent the classes, and M_i and M are the means of the classes and the grand mean, respectively. One discriminant analysis criterion is J_1: $J_1 = tr(S_w^{-1}S_b)$, and J_1 is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1}S_b$ [14]:

$$S_w^{-1}S_b\Psi = \Psi\Delta \quad (5)$$

where Ψ, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1}S_b$, respectively. The discriminating features are defined by projecting the pattern vector $\mathcal{Y}$ onto the eigenvectors of Ψ:

$$\mathcal{Z} = \Psi^t\mathcal{Y} \quad (6)$$

$\mathcal{Z}$ thus contains the discriminating features for image classification.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix. The EFM improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices [11]. The simultaneous diagonalization demonstrates that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. As shown by [11], the small eigenvalues tend to encode noise, and they cause the whitening step to fit for misleading variations, leading to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance). For this work the number of eigenvalues was empirically chosen.

E. Training

The EFM feature extraction method uses positive and negative training samples to find the most discriminative features.

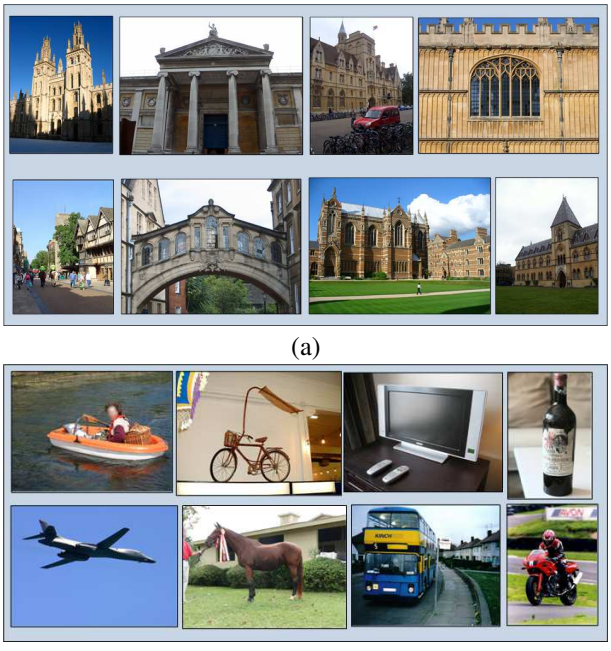

Figure 5. Some sample query images from (a) the Oxford Buildings dataset, and (b) the PASCAL VOC 2012 dataset.

In our setting, there is only one query image to be used as a positive sample. This is similar to the Exemplar-SVM training scenario used by [9], but to make the training more robust to selection error by the user and to prevent overfitting, we use 11 windows instead of just the one selected by the user as described in Section II-A.

We rank all objectness windows from all images in the dataset in terms of Euclidean distance in the HOG space from the original query window. For the negative training samples, we use 110 windows that are ranked low, i.e., are very distant in the HOG space. Experimentally, we found that the last ranked windows are not very good candidates for negative training samples, since they are often outlier windows that contain large blank areas like the sky. Instead, windows that have a rank 1000 to 5000 perform well. We also tried training the system with different numbers of negative samples and found a number close to 100 performs the best. These windows are mostly background regions like ground and vegetation. The positive and negative weights for the HOG features learned by this method can be seen in Figure 4.

For an n-class problem, the EFM process for discriminatory feature extraction reduces the dimensionality of any vector to $n-1$. Since our problem is a two-class problem, EFM produces one feature per window. We compute the score of each window by finding the absolute value of the difference between the window EFM feature and the average positive training set EFM feature. Ranking the images by their best-scoring windows gives us the retrieval set.

III. Experiments

A. Dataset

We have used the two datasets shown in Figure 5 for this work. First, we evaluate the retrieval performance of the proposed method on images gathered in the wild. For this, we use the Oxford Buildings dataset [15], which consists of 5062 images of 11 different Oxford landmarks and distractors collected from Flickr [16]. 55 images from this dataset were used as queries for testing our retrieval system. Flickr images are completely user-generated, which means there is a great variation in camera type, camera angle, scale and lighting conditions. This makes this dataset very difficult for image retrieval in general and landmark-identification in particular (the results of which are shown in Figure 6). Figure 5(a) shows some of our query images from this dataset. For each query, the images that contain the query landmark are further classified into *good*, *OK* and *junk* categories, with progressively poorer views of the query landmark. Table I shows the landmark-wise distribution of *good*, *OK* and *junk* images in this dataset.

TABLE I. The number of images containing each landmark in the Oxford Buildings dataset

Landmark	Good	OK	Junk
All Souls Oxford	24	54	33
Ashmolean Oxford	12	13	6
Balliol Oxford	5	7	6
Bodleian Oxford	13	11	6
Christ Church Oxford	51	27	55
Cornmarket Oxford	5	4	4
Hertford Oxford	35	19	7
Keble Oxford	6	1	4
Magdalen Oxford	13	41	49
Pitt Rivers Oxford	3	3	2
Radcliffe Camera Oxford	105	116	127

We also test retrieval performance on the PASCAL VOC 2012 dataset [17]. We only use the training/validation data from this dataset to test our retrieval algorithm. This data consists of 17,125 images from 20 classes. We create five random test sets of size 100 each from the original image set and perform a five-fold cross-validation on all our experiments. Figure 5(b) shows some images from this dataset.

B. The Retrieval Task

The proposed image representation is tested on two different tasks the first of which is retrieval. Here, an image is used as a query to retrieve similar scenes from the dataset. For this, the user selects a rectangular region of interest from the query image, and HOG features from this rectangular window is matched with the 25 highest scoring objectness windows from each image in the database, both in the raw HOG space and after the proposed training and feature extraction procedure. The closest matches based on Euclidean distance are retrieved

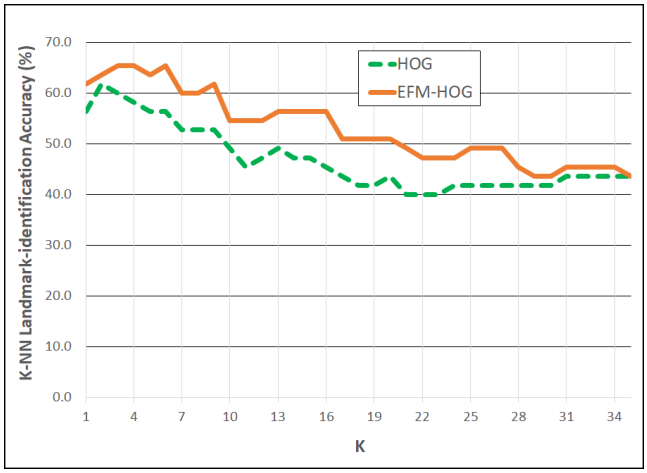

Figure 6. The mean landmark-identification performance by using the K-nearest neighbors method with varying K.

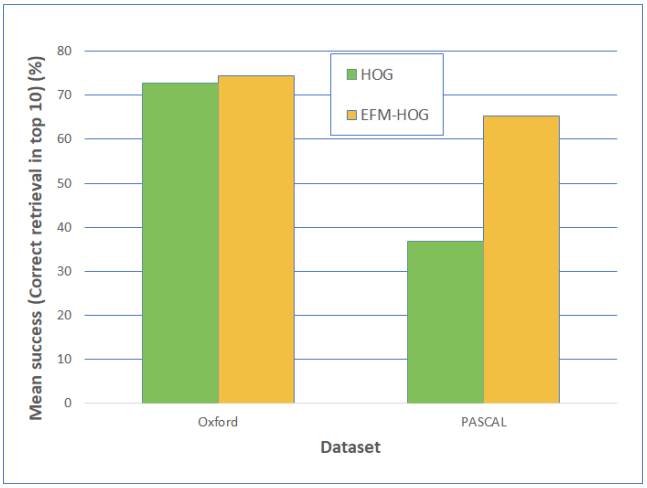

Figure 7. Mean retrieval accuracy (measured by the presence of a relevant image in the top 10 retrieved images).

 ISBN: 978-1-61208-716-0

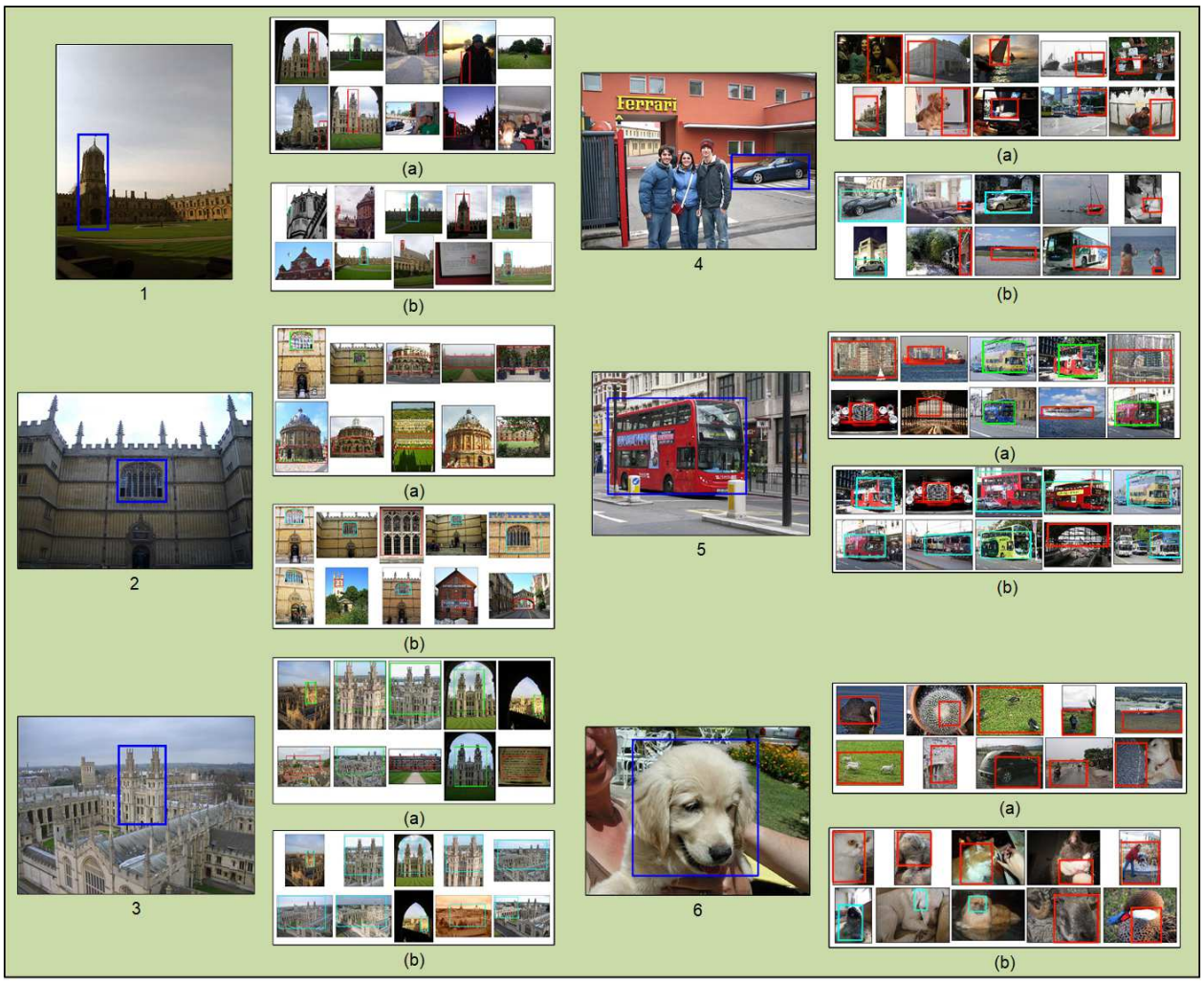

Figure 8. Comparison of image retrieval results for HOG and the proposed EFM-HOG. Images 1, 2 and 3 are from the Oxford Buildings dataset. Images 4, 5 and 6 are from the PASCAL VOC 2012 dataset. In each case, (a) shows top ten images retrieved by HOG, and (b) shows top ten images retrieved by EFM-HOG. Red rectangles indicate images that do not represent the same landmark or object class as the query.

in order of their distance from the query window. Finding an instance of the query in the top 10 retrieved images is considered a success. Figure 7 compares the retrieval success rates of the HOG descriptor and the proposed EFM-HOG representation. Specifically, in 41 cases out of 55 queries in the Oxford buildings dataset, the query landmark is retrieved within top 10 images by the proposed method, as opposed to 40 by HOG. This is actually a very small difference, but this can be explained by the nature of this dataset. For all landmark query images in this dataset, there are at least some images in the dataset that show clear views of the landmarks with no occlusions. HOG is actually pretty effective at retrieving these images. To actually understand the effectiveness of the proposed method, we repeat this experiment with just the *junk* files for each query. In this experiment, we find that the HOG method retrieves a relevant image in the top 10 only once out of all 55 queries, while the proposed EFM-HOG method achieves this 5 times out of the 55.

For PASCAL VOC, the experiment is performed on all five random splits and the average success rate is found to be 65.2% for EFM-HOG as compared to 36.8% for HOG. We also find that the conventional HOG performs quite well for clearly segmented objects, such as airplanes in the sky, but the EFM-HOG performs much better for images of objects with a cluttered background. Some HOG and EFM-HOG retrieval results are shown in Figure 8. Figure 9 shows another interesting aspect of our retrieval technique. Here, we show the image means of the first 100 windows retrieved by both HOG and EFM-HOG on PASCAL VOC. The figure shows that the EFM-HOG means contain clearer shapes, which indicates that the EFM-HOG retrieves more similar shapes than HOG, even when the results are irrelevant to the query.

C. The Landmark-identification Task

Some images in our Oxford Buildings dataset belong to one of the eleven landmarks listed in Table I, the others belong to none of the classes and are used as distractors. The second experiment that we performed with the new EFM-HOG descriptor was a landmark-identification task where the system tries to label each query image with its correct landmark label. This is done by retrieving relevant images in a manner similar to the retrieval task, and then performing the K-

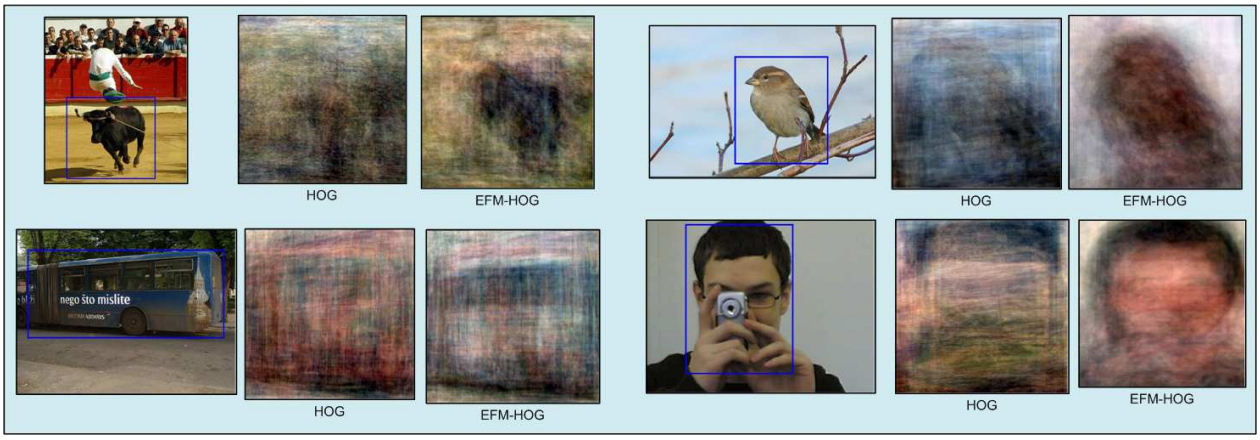

Figure 9. The means of the top 100 retrieved windows for HOG and EFM-HOG for 4 query images from the PASCAL VOC 2012 dataset.

nearest neighbors classification on the top K results. The same experiments are repeated for the conventional HOG descriptor as well. As can be seen from Figure 6, the proposed EFM-HOG outperforms HOG by a significant margin for nearly all values of K between 1 and 35. The highest EFM-HOG landmark-recognition performance of 65.5% is achieved at $K = 3$.

IV. Conclusion

We have presented in this paper a new image descriptor based on HOG and discriminant analysis that uses a novel approach to fetch scenes with similar shaped objects. We have conducted experiments using over $5,000$ images from the Oxford Buildings dataset and over $17,000$ images from the PASCAL VOC 2012 dataset and concluded the following: (i) HOG features are not always sufficiently discriminative to perform meaningful retrieval, (ii) the discriminative nature of HOG features can be improved with the EFM for feature extraction and dimensionality reduction, and (iii) HOG features perform well for clearly isolated objects with little background clutter, but the EFM-HOG performs better for real-world images with cluttered backgrounds.

We intend to use this method with more datasets in the future, so that a more thorough understanding of its strengths and weaknesses can be achieved.

Acknowledgment

The authors would like to thank Professor Jana Košecká at the Department of Computer Science, George Mason University, Fairfax, Virginia for some valuable input on the EFM-HOG method and the experiments conducted.

References

[1] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[2] S. Banerji, A. Sinha, and C. Liu, "Scene Image Classification: Some Novel Descriptors," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2012, pp. 2294–2299.

[3] A. Sinha, S. Banerji, and C. Liu, "Novel Color Gabor-LBP-PHOG (GLP) Descriptors for Object and Scene Image Classification," in Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, ser. ICVGIP '12. ACM, 2012, pp. 58:1–58:8.

[4] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, 2010, pp. 1627–1645.

[5] K. E. A. v. d. Sande, C. G. M. Snoek, and A. W. M. Smeulders, "Fisher and VLAD with FLAIR," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp. 2377–2384.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proceedings of the Twenty-sixth Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the 3rd International Conference on Learning Representations, 2015.

[8] C. Szegedy, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2015, pp. 1–9.

[9] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," in Proceedings of the International Conference on Computer Vision, 2011, pp. 89–96.

[10] V. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[11] C. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases," IEEE Transactions on Image Processing, vol. 9, no. 1, 2000, pp. 132–137.

[12] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the Objectness of Image Windows," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, Nov 2012, pp. 2189–2202.

[13] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008 [accessed 2019-04-18].

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. Academic Press, 1990.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[16] "Flickr," http://www.flickr.com, 2004, [accessed 2019-04-18].

[17] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, vol. 88, no. 2, 2010, pp. 303–338.

 ISBN: 978-1-61208-716-0