AFFINE INVARIANT SALIENT PATCH DESCRIPTORS FOR IMAGE RETRIEVAL

Furkan Işıkdoğan and Albert Ali Salah

Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

ABSTRACT

Image description constitutes a major part of matching-based tasks in computer vision. The size of descriptors becomes more important for retrieval tasks in large datasets. In this paper, we propose a compact and robust image description algorithm for image retrieval, which consists of three main stages: salient patch extraction, affine invariant feature computation over concentric elliptical tracks on the patch, and global feature incorporation. We evaluate the performance of our algorithm for region-based image retrieval and image reuse detection, a special case of image retrieval. We present a novel synthetic image reuse dataset, which is generated by superimposing objects on different background images with systematic transformations. Our results show that the proposed descriptor is effective for this problem.

1. INTRODUCTION

The development of multimedia technologies has enabled us to easily create and store digital images, and consequently led to an emergence of large scale image datasets. Efficient detection and description of the important parts of images becomes crucial in order to provide functionality for content-based access to image datasets. Early content-based image retrieval (CBIR) methods used global features [1], which were computed over whole images and lacked object-level information. Local features, on the other hand, provide a more detailed description of images by extracting features around local keypoints. Many local feature based approaches have been proposed [2, 3, 4] and compared [5, 6, 7] in the literature. Even though local descriptors are quite successful in detecting correspondences between images, they typically produce large and computationally expensive image descriptors, which are not efficient to use for image retrieval in large scale datasets. Bag of visual words (BoW) approach is commonly adopted to represent images with smaller feature vectors by generating visual codebooks using keypoint based local features. However, as noted in [8], the vector quantization step in codebook generation constitutes a bottleneck and reduces the scalability of the system.

In this paper, we present a new method, called Affine Invariant Salient Patch (AISP) descriptors, for describing images with low dimensional feature vectors. The images are represented mainly by foreground regions in our descriptors. We make use of the global contrast based salient region detection method [9] as a good estimator of foreground regions in images. Then, we extract affine invariant features from the estimated foreground region, i.e. the salient patch, by dividing it into concentric elliptical tracks. We also add global features to our final descriptor in order to achieve an admissible level of context independence for foreground objects.

The rest of the paper is organized as follows: Section 2 gives an overview of the AISP algorithm; Section 3 describes the related work in the literature; Section 4 introduces the synthetic image reuse dataset and presents experimental results; finally, Section 5 gives conclusions and future work.

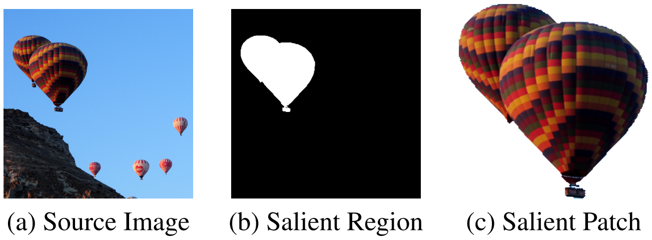
2. AFFINE INVARIANT SALIENT PATCH DESCRIPTORS

Our proposed approach consists of three main steps: salient patch detection (Section 2.1), affine invariant feature extraction (Section 2.2), and addition of global features (Section 2.3). The result is a compact descriptor obtained from a given image. Matching of the descriptor is achieved with standardized Euclidean distance, which balances out the contributions of different features.

2.1. Salient Patch Detection

We estimate the location of a foreground object by estimating image saliency, which refers to the prominence and uniqueness of a region relative to its neighbors. For salient patch extraction, we use the saliency map approach of Cheng et al. [9], which can be summarized as follows. For each pixel, histogram-based saliency values are computed, where saliency values are defined proportional to the total color distance to every other pixel in the image. In order to obtain region-level saliency information, the input image is segmented into regions. For each region, a weighted average of pixel saliency values are computed by incorporating distances to increase the importance of contrast to closer regions. The final saliency map is thresholded iteratively as described in [9], and the resultant binary image is used as a segmentation mask used for extraction of foreground region and computation of affine in-

This research is supported by Bogazici University project BAP 6531.



(a) Source Image (b) Salient Region (c) Salient Patch

Fig. 1. Example of salient patch extraction.

variant elliptical tracks over the region (Section 2.2). An example of salient patch extraction is shown in Fig. 1.

2.2. Affine Invariant Feature Extraction

In order to achieve affine invariance, we define elliptical concentric tracks over the salient patch on the image. We use moments of the binary shape generated by the salient region detector to fit an ellipse to the salient patch. Given a binary shape S the centroid (m_x, m_y) of the salient patch is obtained. Then, the central moments are computed as:

$$\mu_{ij} = \frac{\sum_x \sum_y (x - m_x)^i (y - m_y)^j S(x,y)}{\sum_x \sum_y S(x,y)} \tag{1}$$

The principal orientation of the shape is calculated as follows:

$$\theta_m = \frac{1}{2} \tan^{-1} \left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right) \tag{2}$$

Then, θ_m is assigned to the descriptor as the main orientation. The semimajor and semiminor axis lengths of the best fitting ellipse are calculated as follows [10]:

$$a = \sqrt{\mu_{20} + \mu_{02} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}} \tag{3}$$

$$b = \sqrt{\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}} \tag{4}$$

After defining the fitting ellipse of the salient patch, we divide the ellipse into four equal breadth concentric tracks as illustrated in Fig. 2. We assume in this work that the images are in the HSV color space, and for each track, we compute 8-bin color histograms for each of the color channels in addition to 8-bin edge-orientation histograms. Hue, saturation and brightness components of the image are first used for color histogram computation, then the brightness component is kept for edge orientation calculation.

In the edge orientation histogram computation step, similar to the Histogram of Oriented Gradients [11] approach, each pixel in the salient patch votes in its corresponding track for its orientation, weighted with the edge magnitude. In other terms, let $\mathbf{t_i}$ be a vector that contains the x and y indices of the pixels in the i^{th} elliptical track. A histogram, $H(i, \theta)$, can be defined as:

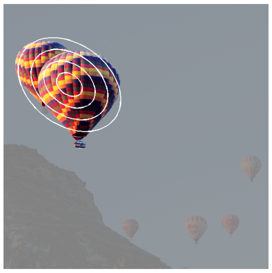


Fig. 2. Illustration of AISP descriptors on an image.

$$\mathbf{t_i} = [\mathbf{x_i y_i}] \tag{5}$$

$$H(i, \theta_k) = \sum_j M(\theta_{k-1} < \Theta(\mathbf{x_{i,j}}, \mathbf{y_{i,j}}) \leq \theta_k) \tag{6}$$

where θ_k denotes the upper boundary angle for the k^{th} bin ($\theta_0 = 0$). $M(x, y)$ and $\Theta(x, y)$ refer to the edge magnitude and orientation of the corresponding pixel, respectively, calculated as

$$M(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{7}$$

$$\Theta(x,y) = \tan^{-1}(G_y(x,y)/G_x(x,y)) - \theta_m \tag{8}$$

$$G_x(x,y) = B(x+1,y) - B(x-1,y) \tag{9}$$

$$G_y(x,y) = B(x,y+1) - B(x,y-1) \tag{10}$$

where G_x and G_y are the horizontal and vertical brightness image (B) gradients, which are computed by using a discrete derivative filter. We achieve rotation invariance by subtracting the main orientation (Eqn. 2) from orientation values for each pixel (Eqn. 8).

Our descriptors are mostly rotation invariant. However, the edge-based features that we use are not robust to 180 degree rotations and flips, since we define the orientation assignment variable over the binary shape. We integrate both color and edge based features into our descriptor so that one type of feature can compensate for the weaknesses of the other.

2.3. Addition of Global Features and Normalization

In many cases, complete context independence of the foreground object might not be desired. Thus, in addition to the features computed over the salient patch, we also incorporate global features into our final descriptor. Addition of global features also decreases the dependence on the salient object detection if the salient object is large enough (i.e. used as a background). We compute the color and edge orientation histograms over the whole image, treating it like a single track. These features are added to the final descriptor having the same weight as a single track. As a final step all histograms are normalized to be a unit vector. Although the number of tracks and bins are tunable parameters, in our experiments we use 4-track 8-bin descriptors. Including the global features, we obtain a 160-dimensional $((\#tracks + 1) \times \#histograms \times \#bins)$ feature vector for each image.

3. RELATED WORK

The closest work to ours is probably the rotation invariant feature transform (RIFT) by Lazebnik et al. [4], which defines circular rings on normalized affine regions. However, there are significant differences between [4] and the presented method including the definition of salient regions and extraction of features. Lazebnik et al. extract a set of elliptical regions from an image, whereas we extract one relatively bigger salient region in which the foreground object is likely to appear. In order to accomplish affine invariance, the RIFT approach transforms the ellipses into circles and extracts features from the circles. The authors define the descriptors rotationally invariant, instead of performing an orientation assignment, since the affine regions they compute are not suitable for dominant orientation estimation. However, a salient patch enables us to compute the main orientation using image moments. Our method does not require image remapping, since the orientation assignment is achieved by subtracting the main orientation from the angles of the corresponding pixels. Another major difference of our method from RIFT features and other local descriptors is that our method produces significantly smaller descriptors. Since object detection is not our concern here, we focus on improving accuracy versus descriptor size.

A similar problem, partial duplicate image search, is addressed by Zhou et al. [12], where the authors use BoW with SIFT descriptors. In our experimental results, we compare the BoW - SIFT approach with our method in general, instead of addressing a single previous work, since it is a common approach that is adopted by many CBIR systems.

4. DATASET AND EXPERIMENTAL RESULTS

In this section we first introduce an artificial image dataset, and then we evaluate our algorithm for region-based image retrieval by comparing it to the BoW approach. As a benchmark, we also use the COREL database [13], which is a 80-category and 10,800-instance image database.

Our synthetic dataset consists of 1530 modified images, which are produced by overlaying a set of foreground objects onto background images with random positions. The background images are crawled from the web, and the foreground objects are extracted from the images in the 2012 PASCAL Visual Object Classes [14] dataset using the segmentation information of the dataset.

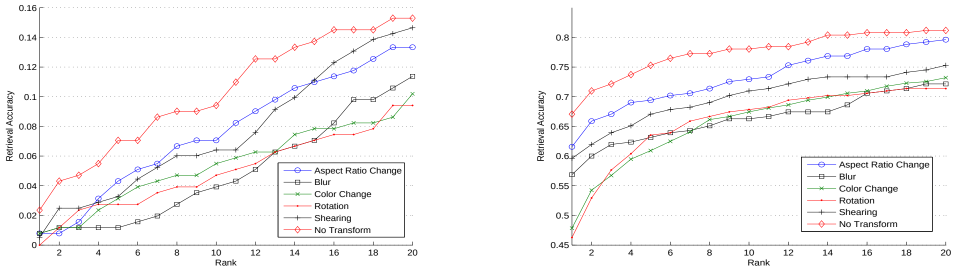
In order to measure the robustness of our descriptors to various modifications, we define six types of changes: rotation, aspect ratio change, shearing, blur, color change, and translation with no transformation. For each type, we apply corresponding transformations with randomized parameters to a random set of foreground objects, and we generate an image set by superimposing foreground objects on randomly selected background images. To prevent any possible memorization of a specific linear transform due to overtraining, transformation parameters are selected randomly for each instance within the following ranges: [0 90] degrees for rotation, [0.5 2] for aspect ratio, [0 0.5] for shear factor, [0 10] sigma and 5-pixel radius for Gaussian blur, and for color change [0 50] percent circular value shift for each of the color channels. Each category in the dataset has 255 images, and each image has only one corresponding foreground image in the folder of original foreground images. To test the retrieval performance, AISP features are computed for 1530 modified images and their corresponding original foreground images. For each modified image, original image descriptors are ranked according to their similarity, which is defined inversely proportional to the standardized Euclidean distance, where the feature dimensions are scaled by their standard deviations.

We illustrate our experimental results with Cumulative Matching Characteristic (CMC) curves, which represent the cumulative retrieval accuracy against the rank of retrieved instances. For the AISP descriptors, Fig. 3 shows the contribution of the features that are extracted from the estimated salient object. For comparison, SIFT features are computed for the synthetic dataset, and a codebook is generated using the BoW approach. The number of clusters are selected as 160, 320, and 640 in order to make the dimensionality of codebook comparable to the default size of AISP descriptors. The results are shown in Fig. 4. Even though the SIFT descriptor do not employ color features, the results show that the retrieval performance is affected by color transformations, since an excessive change in a color channel may alter the edges. In our experiments on the COREL database, we compared each image with every other image in the database and considered the result positive if the retrieved image is in the same category. The first, tenth, and twentieth rank retrieval performances are observed as 25.3, 50.7, and 62.5 percent, respectively.

5. CONCLUSIONS AND FUTURE WORK

We proposed a low dimensional image description method for image retrieval. In particular, we focused on object-level retrieval, which puts more emphasis on the foreground object than the background scene. The major contribution of our method is that it provides a compact representation of images for object-level image access. Our algorithm performed significantly better than BoW approach with SIFT descriptors when the codebook size is set equal to the size of AISP descriptors.

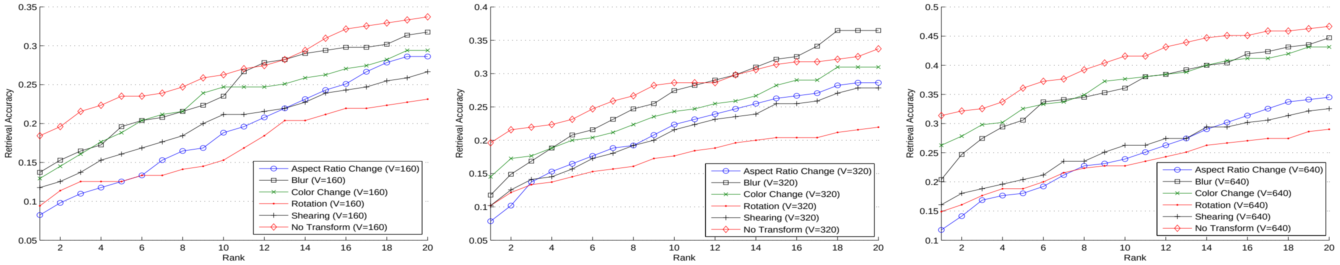
A potential future work can be developing a more compu-



(a) Global features only

(b) AISP descriptors

Fig. 3. Cumulative matching accuracies of AISP descriptors.



(a) 160 Clusters

(a) 320 Clusters

(c) 640 Clusters

Fig. 4. Cumulative matching accuracies of BoW with SIFT descriptors for 160, 320, and 640 clusters.

tationally efficient method for saliency-based object segmentation, since it forms the bottleneck in our algorithm. Another improvement can be parametrizing the importance of color and edge based features by using a weighting scheme to adjust the invariance to different types of transformations.

6. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349 –1380, 2000.

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

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1265–1278, 2005.

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

[6] L. Juan and O. Gwun, "A comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing*, vol. 3, no. 4, pp. 143–152, 2009.

[7] J. Li and N.M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10, pp. 1771–1787, 2008.

[8] K. E.A. van de Sande, T. Gevers, and C. G.M. Snoek, "Empowering visual categorization with the GPU," *Multimedia, IEEE Transactions on*, vol. 13, no. 1, pp. 60–70, 2011.

[9] M.-M. Cheng, G.-X. Zhang, N.J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 409 –416, 2011.

[10] M.R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, vol. 70, pp. 920–930, 1980.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Computer Vision and Pattern Recognition, IEEE Conference on*, vol. 1, pp. 886 –893 vol. 1, 2005.

[12] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," *Proceedings of the ACM Conference on Multimedia*, pp. 511–520, 2010.

[13] J.Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 9, pp. 947–963, 2001.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.