Saliency Based Spatial Partitioning for Global Image Representations

Abin JOSE[1], Iris HEISTERKLAUS[1]

[1]Institute of Communications Engineering, RWTH Aachen University, 52056 Aachen, Germany

abin.jose@rwth-aachen.de, heisterklaus@ient.rwth-aachen.de

Abstract. *Visual content of the image is the main criterion for measuring image similarity in a content based image retrieval system. A major problem with the existing global image representation models such as Fisher Vectors is the loss of spatial information about the objects present in the image. In order to solve this problem, possible object locations are identified using a saliency detector and the saliency maps are thresholded followed by morphological operations. Two types of spatial partitioning such as contour and rectangular partitioning are employed in our experiments. After the spatial partitioning, local feature descriptors from each sub-regions are aggregated to form a global representation. This suppresses the local features from background and the global representation will be a unique representation of the object features. Fisher Vector model is used in our experiments and an object based image retrieval system is evaluated in synthetic dataset and real dataset. The experimental results gave superior performance compared to existing Fisher Vector based approaches without spatial partitioning.*

Keywords

Saliency detection, Spatial partitioning, Fisher Vectors, Global representation, Local features

1. Introduction

In content-based image retrieval, the visual content of the image is the criterion for retrieval rather than metadata such as textual descriptions associated with the image. The better the description of visual content, the better the retrieval accuracy. So it is crucial to have strong descriptors such as SIFT (Scale-Invariant Feature Transform) [1] to represent the visual content in an image. The drawback of the local feature descriptors is that they are large in number and high dimensional. In order to find the most similar image to the query image, it requires lot of comparisons which is memory intensive and computationally complex. To overcome this problem, the feature descriptors are aggregated to form global image representations such as Fisher Vectors [2]. The main problem involved in the global image representations is the loss of spatial information about the objects present in the image. Predominantly in the case of object based retrieval, we are mostly interested in the object location and orientation in the image. In a multi-object image, if the query is for a single object present in the image, the image will not be in the top retrieved list since the global representation contains information from all the objects when the global descriptor is formed.

Different approaches have been proposed in the literature to address this issue. A popular method known as Spatial Pyramid Matching (SPM) [3], encodes the spatial relationship of features in different pyramid levels. Spatial coordinates were added to the feature descriptor by Grzeszick et al. [4]. The similarity in geometry of objects belonging to same category was modeled by Zhang et al. [5]. The joint distribution between the low-level descriptors and location of a patch was considered to include the spatial information in [6].

In this paper, we have addressed the problem of loss of spatial information of objects when global descriptor is formed. Image sub-regions were formed by saliency based object recognition followed by local thresholding, morphological operations and contour detection. The local features from each sub-region is aggregated to form a set of Fisher Vectors for each image. Two types of spatial partitioning such as rectangular and contour partitioning were also explored in this work.

The paper is organized as follows. In section 2, the Fisher Vector model is explained. In Section 3, we explain how saliency cue is incorporated in identifying the object locations, followed by a discussion on the post processing operations for spatial partitioning. The retrieval results for single-object images and multi-object images using Fisher Vectors with and without spatial partitioning is explained in Section 4. The concluding remarks are drawn in Section 5.

2. Global representation of images

A global representation model which generates a probabilistic visual vocabulary is the Fisher Vector (FV) model. In FV model, a Gaussian Mixture Model (GMM) is used to

model the feature space. It also incorporates second-order statistics which helps in storing more information about the distribution of feature vectors in the feature space.

Consider a set of J descriptors in A, where $A = \{a_j, j = 1....J\}$. The descriptors are of D-dimensions which will be 128 in the case of SIFT descriptors. Let p_λ represents the probability density function which models the way in which the descriptors are generated, where λ denotes the parameters of the generative model. The contribution of these parameters to the generative model can be measured by computing the gradient of the log-likelihood of the data on the model which is given by:

$$G_\lambda^A = \bigtriangledown_\lambda \log p_\lambda (A). \tag{1}$$

This describes the way in which the parameters of the model are to be modified to better fit the data under consideration.

Jaakkola and Haussler [7] have proposed a measure to find the similarity between 2 samples A and B using the Fisher Kernel $K_{FK}(A, B)$ which is defined as:

$$K_{FK}(A, B) = G_\lambda^{A^T} F_\lambda^{-1} G_\lambda^B \tag{2}$$

where F_λ is the Fisher Information Matrix (FIM). Since FIM is positive semi-definite, it can be decomposed using Cholesky decomposition, $F_\lambda^{-1} = C_\lambda^T C_\lambda$. This helps in modifying 2 as a dot-product,

$$K_{FK}(A, B) = g_\lambda^{A^T} g_\lambda^B \tag{3}$$

where g_λ^A is given by:

$$g_\lambda^A = C_\lambda G_\lambda^A. \tag{4}$$

The normalized gradient vector g_λ^A is called the FV of A. The main advantage of representing the kernel function as a dot-product is that it is similar to defining a Euclidean space where the distances between feature vectors can be calculated by using the kernel function.

A GMM is used to model the generative model. The GMM parameters are estimated using a large pool of local descriptors obtained from a training dataset using the Expectation Maximization (EM) algorithm [8]. The GMM model can be represented as the weighted sum of K Gaussians as:

$$p_\lambda(a) = \sum_{k=1}^{K} w_k p_k(a) \tag{5}$$

where p_k represents Gaussian k and is represented as:

$$p_k(a) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} e \left\{ -\frac{1}{2} (a - m_k)^T \Sigma_k^{-1} (a - m_k) \right\}. \tag{6}$$

The main parameters λ of this model are given by $\lambda = \{w_k, m_k, \Sigma_k, k = 1....K\}$ where w_k, m_k, Σ_k denotes the weight, mean and covariance matrix of Gaussian k respectively.

The GMM assigns each descriptor a_j to the k-th mode of the GMM with the posterior probability:

$$\phi_j(k) = \frac{w_k p_k(a_j)}{\sum_{l=1}^{K} w_l p_l(a_j)} \tag{7}$$

The deviation measure of a descriptor a_j w.r.t the mean and covariance are:

$$g_{m_k}^A = \frac{1}{J\sqrt{w_k}} \sum_{j=1}^{J} \phi_j(k) \frac{a_j - m_k}{\sigma_k} \tag{8}$$

$$g_{\sigma_k}^A = \frac{1}{J\sqrt{2w_k}} \sum_{j=1}^{J} \phi_j(k) \left[\frac{(a_j - m_k)^2}{\sigma_k^2} - 1 \right] \tag{9}$$

in which $g_{m_k}^A$ and $g_{\sigma_k}^A$ are vectors of size D. The final FV is a concatenation of deviations $g_{m_k}^A$ and $g_{\sigma_k}^A$ for K modes of the Gaussian and is therefore of dimension 2×D×K.

The FV representation has many advantages over other global representation models such as BoVW [9] and VLAD [10] representation. FV gives us the flexibility to define a kernel from a generative process which basically considers how the data is generated. This is better than just modeling using a fixed kernel. BoVW can be considered as a particular case of FV in which the deviation is measured only w.r.t the GMM weight parameter. Another major advantage of FV comes when the computational cost is taken into consideration. Since FV can be computed from smaller vocabularies, the computational cost is lower.

3. Saliency detection and spatial partitioning

When we are considering a retrieval system in which the image to be retrieved contains an object of interest, it is very important that the global representation gives more importance to the objects present in the foreground image than the background information. An effective way of doing this is by partitioning the image space into object and non-object regions and pooling the local features from the object region to form the global descriptor. This would automatically separate the foreground and background information. In addition, in the case of images containing multiple objects, this would partition the image space into multiple sub-regions. Partitioning of image space into sub-regions helps in keeping the local features associated within each sub-region together. This would in turn, retain the spatial information of the object independent of the object location in the image.

3.1. Saliency detection

The perceptual quality that makes an object or region in an image stand out with respect to its neighbors and which grabs the attention of the observer is called Visual Saliency.

Saliency detection closely models the selective processing of human visual system and thus helps in identifying regions in an image which tend to be object locations. A model which closely models the human visual system and tackles the problem of salient small scale structures was proposed by Yan et al. [11] in an efficient and robust multi-layer approach. We have used this saliency maps to identify possible object locations and thus partitioning the image space.

3.2. Thresholding and morphological operations

In order to generate the object masks, from the saliency maps, thresholding is applied. In the case of images with single objects, a unique global threshold T^* was determined by using the Otsu thresholding [12] algorithm. For images with multiple objects, the use of a global threshold value will not serve the purpose. It might generate inaccurate object masks since the saliency maps generated depends on the local object contrast with the background. To avoid this problem, we have computed local threshold value in a sub-window based on the size of the image. The sub-window was chosen as $w/5 \times h/5$ where w and h corresponds to the width and height of the image with multiple objects. The window chosen in our method is a circular region of interest (ROI). The window is moved across the image and Otsu thresholding is computed in each sub-window and thus it takes into account the local pixel intensities instead of just generating a single global threshold.

Once we have the thresholded image, the main objective is to generate object masks such that we can form distinct sub-regions. After thresholding, there will be some unwanted structures such as holes inside the possible object locations and bridges which connect two different sub-regions which affects the spatial partitioning of images. Such holes and bridges are eliminated by proper morphological operations. The main morphological operation used here is the closing operation which is basically the combination of two basic operations, dilation [14] followed by erosion [14]. There are two main advantages of using the closing operation on the binary images obtained after the local thresholding. Closing helps in removing the holes inside the object masks, thus keeping more intact object shapes and information about the objects. Another advantage is that closing helps in keeping spatially closer regions together instead of creating unwanted partitioning.

After morphological operation, we have the binary mask of the image. This binary mask is used to find the contours of the objects present in the image from the edges of the binary image. Once the contour points are identified, the number of points in the contour is reduced for faster processing by approximating the contour points to the closest polygonal approximation using the Douglas-Peucker [13] algorithm. This algorithm finds a similar curve with fewer points and the similarity between the original curve and the approximated curve is measured using the Hausdorff distance. After finding the contours around the possible object locations, a bounding box is formed. This bounding rectangle creates sub-images which can be used for partitioning the image space.

3.3. Rectangular and contour partitioning

In rectangular partitioning of the image space, each rectangular region around the contour forms a sub-region. Thus, there will be as many sub-regions as the number of bounding boxes formed. These sub-regions are processed individually to obtain separate Fisher Vector representations. We will have as many Fisher Vectors as there are sub-regions in the image. The original image is represented as a collection of Fisher Vectors corresponding to each spatially partitioned region in the database. Even though spatial partitioning using rectangular partitioning eliminates the information from the other objects present in the image while forming the global representation, the background information is still present in the sub-region. Thus the Fisher Vector formed will contain this background information. It is important to remove this information and concentrate more on the information present in the object identified in the sub-region. This is achieved by contour partitioning. In contour partitioning, the contour boundary is used to sub-divide the image. Thus each sub-image obtained will consist of only the information which is within the contour which corresponds to the object present in that sub-region. This in turn adds the advantage that the background information is completely suppressed while forming the Fisher Vector for a particular sub-region.

4. Results

4.1. Spatial partitioning of images with single object - Evaluation

In this section, the retrieval results obtained for spatial partitioning of images containing single object are evaluated. The two test databases considered were Caltech-256 [15] objects and 17 category flower dataset from Oxford [16]. 5 categories were considered from each database with 50 images in each category. The total number of images considered were 250 images. Each image was queried against the 250 images and the Mean Average Precision (MAP) was computed at different retrieval numbers.

Fig. 1 shows the evaluation results for the two datasets. We can observe that the spatial partitioning improves the retrieval results for both the datasets. There is an improvement of 10% for the Oxford flowers dataset and around 5% for the Caltech-256 dataset when the top 10 closest matches to the query image were retrieved. From the graphs, it is evident that the performance is better for the Oxford flowers

dataset. The main reason for this is that the saliency maps which were used for spatially partitioning the images were better for this dataset.

Fig. 2 shows the precision recall curve for the Caltech-dataset. It was computed by averaging the precision and recall values for each query. The results were evaluated for image retrieval with and without spatial partitioning. We can clearly observe that without spatial partitioning precision drops to 0.34 when we have a recall of 0.1. For the same recall, when we employ spatial partitioning, we have higher precision of around 0.42. Thus, by spatial partitioning, for each query image, we have more correct retrievals in each category.

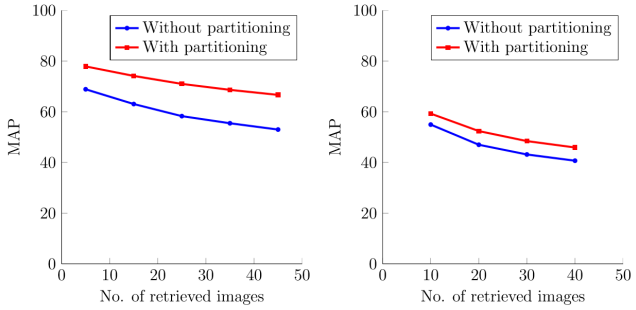

Fig. 1. (a) MAP vs Number of retrieved images for Oxford owers dataset (left), Caltech256 objects dataset (right)

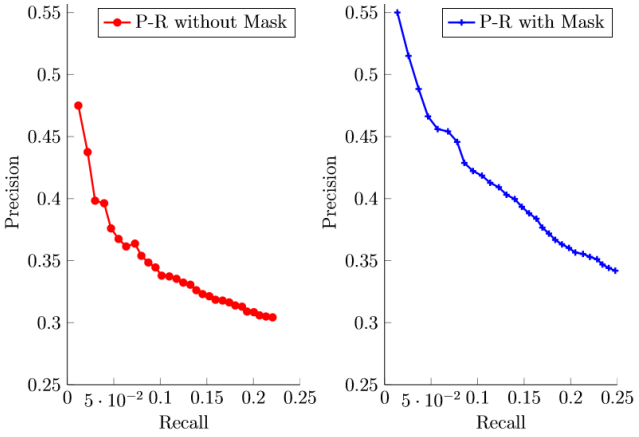

Fig. 2. Precision recall curve for Caltech-256 dataset without spatial partitioning (left), with spatial partitioning (right)

4.2. Spatial partitioning of images with multiple objects - Evaluation

When we query using image of any of the objects present in the multi-object image, the system should retrieve all the images containing that object. For evaluating this scenario, a synthetic dataset was created. The synthetic dataset is created from the Oxford 17 category flower dataset [16]. 1000 images are formed by combining 4 random images from 17 categories of the dataset.

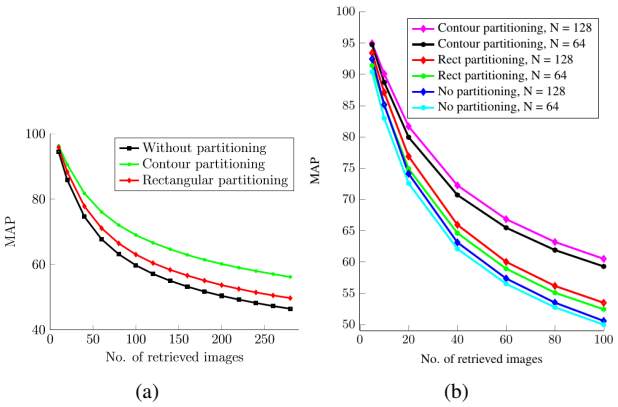

Fig. 3. (a) MAP vs No. of retrieved images for 3 different cases for the synthetic dataset created, (b) Comparison of retrieval results in the case of GMM model with 64 Gaussians and 128 Gaussians with contour partitioning, rectangular partitioning and without partitioning.

For testing this scenario, 1000 images from the synthetic dataset are considered. The number of Gaussians in the GMM is 128. 40 images from the 17 categories of the oxford flowers dataset are used as query images. Thus, we have in total 680 queries. For each query image, the retrieval results are evaluated at different values of N and the MAP is computed. Fig. 3 (a) shows the graphs where the variation of MAP with the number of retrieved images is shown.

The black curve shows the MAP when no spatial partitioning is applied. The other two curves shows the MAP when spatial partitioning is used. The main observation is that the retrieval results are better when spatial partitioning is applied to the images. The main reason for the improvement of results is due to the formation of separate global representations for the sub-regions. When we query for a single object present in the image, the most similar sub-image which contains that object is found and a mapping which says to which parent image that sub-region belongs helps in retrieving the parent image.

The two main types of spatial partitioning proposed are rectangular and contour partitioning. In the case of rectangular partitioning, the sub-regions are formed by splitting the image using the bounding box around the objects identified. This includes the background information around the object as well. Thus, the problem becomes similar to the case of retrieval of single object without spatial partitioning. For suppressing the background information, the method proposed is contour partitioning in which the contour around the objects are used for partitioning the images. In Fig. 3(a), the MAP is better when contour partitioning is used. For instance at $N = 50$, MAP is 75% without spatial partitioning and it improves to 78% when rectangular partitioning is used and even get better to 82% with contour partitioning.

4.3. Comparison of performance with different number of Gaussians in the GMM model

The performance of the retrieval system with different number of Gaussians to form the GMM model is discussed here. For testing this scenario, two GMMs are trained, one with 64 Gaussians and another model with 128 Gaussians. When we double the number of Gaussians, the dimension of the Fisher Vector also doubles. This would increase the computational costs involved in the generation of the Fisher Vectors and measuring the similarity between the Fisher Vectors. The memory requirements for storing the Fisher Vectors will also increase. However, the retrieval results become better when more number of Gaussians are used to model the GMM. Fig. 3(b) shows the graphs of MAP obtained at different retrieval number of images for three different cases with 64 Gaussians and 128 Gaussians in the GMM. The best results were obtained with 128 Gaussians in GMM and contour partitioning. However, GMM with 64 Gaussians and contour partitioning performed better than the GMM with 128 Gaussians and rectangular partitioning. Therefore, the effect of contour partitioning is higher than the effect of increasing the number of Gaussians in the GMM. The results are lower as expected without spatial partitioning.

5. Conclusions

In this work the problem of loss of spatial information while forming Fisher Vectors was addressed. Saliency detection was used to identify the possible object locations. Spatial partitioning helped in successful removal of feature points from the background during the formation of global image descriptor. 2 major types of spatial partitioning such as rectangular and contour partitioning were used in evaluation of the results. Contour partitioning gave superior retrieval results since the background information was completely suppressed. Furthermore, from experiments, it was clear that the effect of removing the feature points from the background is having a higher impact in increasing the retrieval accuracy than the increase in number of Gaussians in the GMM which models the feature space. Effectively combining or compressing the Fisher Vectors of each spatial region and incorporation of prior knowledge about the object size for effectively choosing the region of interest for local thresholding are future research directions.

References

[1] LOWE, D. G. Object recognition from local scale invariant features. *In Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, p. 1150 - 1157.

[2] SANCHEZ, J., PERRONNIN, F., MENSINK, T., and VERBEEK, J. Image classication with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013, vol. 105, no. 3, p. 222 - 245.

[3] LAZEBNIK, S., SCHMID, C., and PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, p. 2169 - 2178.

[4] GRZESZICK, R., ROTHACKER, L., FINK, G., et al. Bag-of-features representations using spatial visual vocabularies for object classication. *In Proceedings IEEE International Conference on Image Processing*, 2013, p. 2867 - 2871.

[5] ZHANG, E., and MAYO, M. Improving bag-of-words model with spatial information. *In Proceedings IEEE International Conference of Image and Vision Computing New Zealand*, 2010, p. 1 - 8.

[6] SANCHEZ, J., PERRONNIN, F., and CAMPOS, T. D. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*,2012, vol. 33, no. 16, p. 2216 - 2223.

[7] JAAKKOLA, T.S., and HAUSSLER, D. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*,1999, p. 487 - 493.

[8] MOON, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine*,1996, vol 13, no.6, p. 47 - 60.

[9] CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J., and BRAY, C. Visual categorization with bags of keypoints. *In: ECCV Workshop on Statistical Learning in Computer Vision.*, 2004, vol 1, no. 1-22, p. 1 - 2.

[10] JEQOU, H., PERRONNIN, F., DOUZE, M., SANCHEZ, J., PEREZ, P., and SCHMID, C. Aggregating local image descriptors into compact codes. *In: IEEE Transactions on Pattern Analysis and Machine Intelligence*,2012, vol 34, no. 9, p. 1704 - 1716.

[11] YAN, Q., XU, L., SHI, J., and JIA, J. Hierarchical saliency detection. *In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, p. 1155 - 1162.

[12] OTSU, N. A threshold selection method from gray-level histograms. *In: Automatica*, 1975, vol. 11, no. 285-296, p. 23 - 27.

[13] DOUGLAS, D., H., and PEUCKER, T. K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *In: Cartographica: The International Journal for Geographic Information and Geovisualization*, 1973, vol 10, no. 2, p. 112 - 122.

[14] GONZALEZ, R. C., WOODS, R. E., and EDDINS, S. L. Digital image processing using MATLAB. *Pearson Education India*, 2004.

[15] GRIFFIN, G., HOLUB, A., and PERONA, P. Caltech-256 object category dataset. *California Institute of Technology*, 2007.

[16] NILSBACK, M. E., and ZISSERMAN, A. A Visual Vocabulary for Flower Classication. *In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2006, vol 2., p. 1447 - 1454

About Authors...

Abin JOSE was born in Pulpalli, India in 1989. He received the Masters degree in Electrical Engineering, Information Technology and Computer Engineering from RWTH Aachen University, Aachen, Germany, in 2016 and is currently working as a Ph.D. student at the Institute of Communications Engineering, RWTH Aachen University focusing on image analysis and object detection and tracking.

Iris HEISTERKLAUS was born in Berlin, Germany in 1986. She received the Masters degree in Electrical Engineering, Information Technology and Computer Engineering from RWTH Aachen University, Aachen, Germany, in 2012 and is currently working as a Ph.D. student at the Institute of Communications Engineering, RWTH Aachen University. Her focus is on image and video analysis and content recognition.