BAG OF FISHER VECTORS REPRESENTATION OF IMAGES BY SALIENCY-BASED SPATIAL PARTITIONING

Abin Jose and Iris Heisterklaus

Institut für Nachrichtentechnik, RWTH Aachen University, Aachen, Germany
jose@ient.rwth-aachen.de, heisterklaus@ient.rwth-aachen.de

ABSTRACT

In content-based image retrieval systems, visual content of the image is the criterion for measuring image similarity. We propose a method to solve the problem of loss of spatial information of objects when local descriptors from an image with multiple objects are aggregated to form a global representation. In our approach, after saliency-based spatial partitioning, local feature descriptors from distinct sub-regions are aggregated to form a bag of Fisher Vectors representation. This helps in suppressing the information from background clutter in scenes while forming the global descriptor. The retrieval performance was evaluated in synthetic and real datasets. The evaluation results show that the bag of Fisher Vectors representation gives better performance compared to baseline approach using Fisher Vectors.

Index Terms— Fisher Vectors, Content-based image retrieval, Saliency detection, Spatial partitioning.

1. INTRODUCTION

In content-based image retrieval systems (CBIR) [1], images are retrieved from the database based on their visual content. The visual content is uniquely stored in the corresponding feature descriptors for each image. The global descriptor is formed from these feature descriptors and using a similarity measure, the images in the database are ranked and retrieved. The main problem associated with local feature descriptors is that they are large in number and the descriptor processing may become memory intensive and computationally complex. There are various methods to overcome this problem and compress descriptors to form a global image representation.

Some of the prominent global representation methods are Bag of Visual Words (BoVW) [2], Vector of Locally Aggregated Descriptors (VLAD) [3], and Fisher Vectors (FV) [4]. In BoVW model, the image is represented as a histogram of visual words. The main shortcoming of BoVW representation is that the local statistics of feature vectors are never taken under consideration. In addition there is a loss of information due to lossy descriptor quantization. The VLAD model tries to surmount these shortcomings to a large extent. It is almost similar to the BoVW representation including the generation of visual words. After the generation of visual words, instead of just embedding the information about the particular visual words to which feature points are closer to, the distance vector to the nearest visual word is computed and stored. Thus, this model tries to encode the first order statistics. Nevertheless, the VLAD model fails in incorporating a probabilistic visual vocabulary as it uses K-Means clustering to form the visual words. In the FV model, a Gaussian Mixture Model (GMM) is used to model the feature space. It incorporates first-order and second-order statistics which helps in storing more information about the distribution of feature vectors in the feature space by measuring the deviation from the GMM model w.r.t the mean and the covariance of the distribution.

Even though global image representation summarizes the visual content, it suffers from some limitations especially when the image contains an object in the foreground with heavy background clutter or multiple objects. Some of the limitations are: (1) Spatial information is completely lost, (2) global descriptors ignore the geometry of objects present in the image, (3) global descriptors mix the foreground and background information, and (4) object features are not always captured in presence of background clutter.

The above mentioned limitations led to exciting research in improving global descriptors with spatial information. Lazebnik et al. [5] proposed a method which is popularly known as the Spatial Pyramid Matching (SPM) which tries to encode spatial relationship of features at different pyramid levels. Grzeszick et al. [6] imposed spatial information by extending feature descriptor by including spatial coordinates. Zhang et al. [7] exploited the similarity in geometry of objects belonging to same category. A joint distribution between the low-level descriptors and patch locations are considered in [8].

In a retrieval system, which focuses on objects present in a visual scene, it is preferable to extract more local features from the objects than the background. Saliency detection models provide excellent cues about the possible object locations in an image with less computational overhead. It closely models the selective processing of human visual system and thus helps in identifying regions in an image which tend to be object locations.

For natural images, the saliency detection accuracy is affected if the foreground or background contains lot of high-contrast patterns. When the image contains small scale and large scale patterns, the saliency map will generate lots of possible object locations which will make less contribution for an object based image retrieval system. In addition, for natural images the presence of high textured background is always a hindrance in detecting the most salient region in the image. These problems of saliency extraction in a visual scene are tackled to a great extent by an efficient and robust multilayer approach proposed by Yan et al. [9]. We extract the saliency cues to identify possible spatial locations of objects using this hierarchical model.

In this paper, we address the problem of loss of spatial information of objects when a global descriptor is formed by forming a Bag of Fisher Vectors (BoFV) representation of images. Spatial regions are identified by using saliecy maps which are further processed by local thresholding, morphological closing operation and contour detection. Two types of spatial partitioning such as rectangular and contour partitioning are proposed. The local features from each sub-region are aggregated to form BoFV. Images with single objects are used as queries. The FV corresponding to the most similar sub-region is retrieved which in turn maps to the corresponding

parent image with multiple objects. We have conducted our experiments on both synthetic and real datasets and have compared our result with Fisher Vector based representation.

2. FISHER VECTOR MODEL

Fisher Vector model generates a probabilistic visual vocabulary. The main advantage of this model over BoVW and VLAD is that FV offers the flexibility to define a kernel from a generative process which in turn considers how the data is generated. BoVW model is a particular case of FV in which the deviation is measured only w.r.t weights of the generative process which is a Gaussian Mixture Model (GMM). The parameters of this model are estimated using a large pool of local descriptors obtained from a training dataset using the Expectation Maximization (EM) algorithm [11]. The GMM model which is basically the weighted sum of N Gaussians is represented as $p_\lambda(a) = \sum_{n=1}^{N} w_n p_n(a)$, where p_n represents Gaussian n. The parameters λ of this model are given by $\lambda = \{w_n, m_n, \Sigma_n; n = 1...N\}$ which denotes the weight, mean and covariance matrix of Gaussian n respectively. The FV model embeds the deviation of local descriptors of an image from this generative model. Consider a set of K descriptors in A, where $A = \{a_k, k = 1...K\}$. The descriptors are of D-dimensions. The gradient of the log-likelihood of the descriptors on the model 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 local descriptors under consideration. Jaakkola and Haussler [10] 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 F_λ is positive semi-definite, it can be decomposed using Cholesky decomposition, as $F_\lambda^{-1} = C_\lambda^T C_\lambda$. This helps in modifying equation 2 as:

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

where g_λ^A is expressed as $C_\lambda G_\lambda^A$. 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 $g_\lambda^{A^T} g_\lambda^B$ is that it is similar to defining a Euclidean space where the distances between feature vectors can be calculated by using the kernel function.

The GMM assigns each descriptor a_k to the n-th mode of the GMM with the posterior probability:

$$\phi_k(n) = \frac{w_n p_n(a_k)}{\sum_{l=1}^{N} w_l p_l(a_k)}. \tag{4}$$

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

$$g_{m_n}^A = \frac{1}{K\sqrt{w_n}} \sum_{k=1}^{K} \phi_k(n) \frac{a_k - m_n}{\sigma_n} \tag{5}$$

$$g_{\sigma_n}^A = \frac{1}{K\sqrt{2w_n}} \sum_{k=1}^{K} \phi_k(n) \left[\frac{(a_k - m_n)^2}{\sigma_n^2} - 1 \right] \tag{6}$$

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

3. SALIENCY DETECTION

For spatial partitioning of images, saliency cues are extracted to identify possible object locations and the different stages involved in the hierarchical model is explained below.

3.1. Extraction of image layers at different scales

Image layers are representation of images with different degrees of details. In the lowest layer, fine details are retained while in higher layers, larger structures are preserved. In the preprocessing stage, an initial over-segmentation using watershed segmentation [12] method is employed. A region can be defined to be homogeneous, if it can be encompassed by a $s \times s$ bounding box such that the ratio of number of similar pixels to number of non-similar pixels within the bounding box is above a threshold.

Once the scale s of a region is determined, neighboring segments within each layer are merged in an iterative manner. In order to accomplish this, each region is sorted initially according to the scale in an ascending order. For each layer, scale thresholds are chosen such that finer details are retained in first layer. The higher layers are generated from the lower layers using similar procedure.

3.2. Extraction and fusion of saliency cues

In the second stage of saliency computation, for each extracted layer, saliency cues are generated. The main 2 saliency cues used are contrast and location. The local contrast cue C_x for a region R_x with n neighboring regions is computed as the weighted sum of color differences from regions R_y where y varies from 1 to n and is given by:

$$C_x = \sum_{y=1}^{n} w(R_y) \varphi(x, y) \left\| c_x - c_y \right\|_2, \tag{7}$$

where c_x and c_y denote the color information of regions R_x and R_y, respectively. The number of pixels in the neighboring regions are accounted by the term $w(R_y)$. This term helps in weighing regions having more pixels with higher weight which in turn increases its contribution to the contrast cue. The term $\varphi(x, y) = \exp\{-d/\sigma^2\}$ controls the influence of spatial distance between the regions x and y. d is the Euclidean distance between region centers and σ controls the size of neighborhood. For instance in top layer, σ is large making a global comparison between all the regions.

The next cue information is location. Psychological studies [13] prove that humans tend to focus more on the central region of the image and the central pixels contribute more towards the saliency maps. The location cue H_x is modeled as:

$$H_x = \frac{1}{w(R_x)} \sum_{p_x \in R_x} \exp\left\{ -\lambda \left\| p_x - p_c \right\|^2 \right\}. \tag{8}$$

Here, p_c represents the coordinate of the image center and p_x models the coordinates of regions identified in that particular layer. The parameter λ in equation 8 is used when H_x is combined with the C_x to form the combined saliency cue as $\hat{s}_x = C_x H_x$. The single-layer saliency cue maps are then fused by a weighted average approach to form the final saliency map (Fig. 1 (a)).

3.3. Generation of object masks from saliency maps

Thresholding is applied to generate object masks from the saliency maps (Fig. 1 (b)). In the case of images with one object, a unique global threshold T^* is determined by Otsu thresholding [14]. For

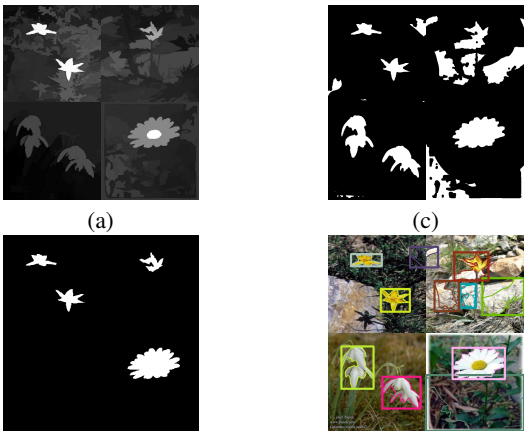

Fig. 1: (a) Saliency map, (b) Otsu thresholding, (c) Local thresholding and morphological closing, (d) Spatial sub-regions identified.

images with multiple objects, the use of a global threshold value might generate inaccurate object masks since the saliency maps generated depends on the local object contrast with the background. This problem is circumvented by computing local threshold value in a sub-window. A circular window is moved across the image and Otsu thresholding is computed in each sub-window (Fig. 1 (c)). After thresholding, unwanted structures such as holes and bridges are eliminated by closing operation. Closing helps in keeping spatially closer regions together instead of creating unwanted partitioning.

3.4. Detection of contours and elimination of false detections

After morphological operation, the edges of the binary mask are used to find the respective contours of the objects (Fig. 1 (d)). The bounding rectangle around the contour creates sub-images which can be used for partitioning the image space. Some false detections are identified at this stage. These false detections must be removed since it results in unwanted partitioning of the image space. For avoiding false detections of object locations, 2 constraints are imposed in the retrieval system during the formation of contours. They are: (1) Eliminate contours which are within a contour, (2) eliminate contours with radius of minimum enclosing circle below a threshold. The first criterion makes sure that only the most external contour is extracted when a set of contours are identified in a region to avoid unnecessary partitioning. The second criterion assures that small contours are eliminated. It is achieved by using a threshold for the radius of minimum enclosing circle around the contour.

3.5. Spatial partitioning of images

For spatial partitioning of the images based on the contours identified, mainly 2 models are proposed. We have identified possible object locations and have formed bounding boxes around objects. These bounding boxes are used to partition the image space. In rectangular partitioning of the image space, each rectangular region around the contour forms a sub-region. These sub-regions are processed individually to obtain separate FV representations. Even though rectangular partitioning eliminates the information from the other objects present in the image, the background information is still present in the sub-region. For solving this problem, and to increase the retrieval accuracy, contour partitioning was introduced. In

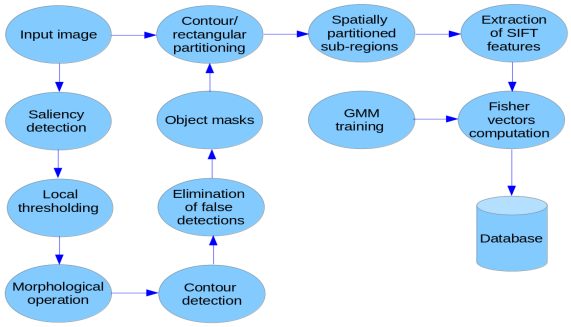

Fig. 2: Block diagram showing the generation of Bag of Fisher Vectors.

contour partitioning, the contour boundary is used to sub-divide the image. This assures that the background information is completely suppressed while forming the FV for a particular sub-region. Thus, each image in the database is represented as a BoFVs and if any of these FVs have a minimum Euclidean distance with the FV of query image, the image will be retrieved. The block diagram showing the formation of BoFV is shown in Fig. 2.

4. EXPERIMENTAL RESULTS

4.1. Evaluation criterion

Average precision (AP) helps in having a metric which can compare the performance of retrieval systems. The definition of AP is the average of precision when recall varies from 0 to 1. For each query image, the AP is computed and it is averaged over the total number of queries to obtain the Mean Average Precision (MAP). In order to compare two retrieval systems, MAP is measured at different values of retrieved images (N).

4.2. Evaluation of retrieval results - Single object

First we measured the retrieval performance on images with single objects to see how the removal of background clutter affects the retrieval performance. The retrieval results obtained for Caltech dataset [15] for 8 different categories is discussed in this section. Fig. 3 shows the bar graph showing the average precision (AP). For the categories Hibiscus, Ibis, Hot-air-ballon, and Bonsai, BoFV (foreground and background FV) representation improved the results as the saliency maps generate good spatial partitioning between the object and background. There is decrease in APs by 3.43, 4.98 and 4.75 for the categories Brain, Buddha and Eiffel-tower respectively due to spatial partitioning. This slight decrease in performance is due

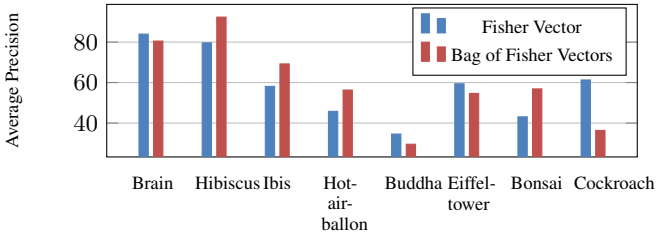

Fig. 3: AP in % for 8 categories from the Caltech dataset with contour partitioning and baseline approach.

to the loss of information from objects when we have poor saliency maps. For Eiffel-tower, which represents a static scene, background might add information about the scene as background remains constant in almost all images. Thus, removal of background in this case decreases the performance. For the category which contains cockroach images, the saliency maps are good but still there is a huge drop in AP by a factor of 24.92 with spatial partitioning. The reason for this might be the lack of texture in the objects and loss in information from the minute structures such as legs and antennas of the object which are lost when the saliency maps are formed.

4.3. Evaluation of retrieval results - Multiple objects

We have conducted our experiments on both synthetic and real datasets. 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 to form a multi-object dataset. Query images consist of 40 single object images from 17 categories making 680 query images. The performance of the retrieval system on synthetic dataset with different number of Gaussians to form the GMM model is evaluated. The retrieval results become better with higher number of Gaussians at the expense of increased FV size. Table 1 shows the MAP obtained at different retrieval number of images for 3 different cases (Fig. 4) with 64 and 128 Gaussians to model the GMM. The best results were obtained with 128 Gaussians and contour partitioning. However, 64 Gaussians and contour partitioning gives better results than 128 Gaussians and rectangular partitioning. Thus, the impact of contour partitioning is higher than the impact of increasing Gaussians in the GMM. The retrieval results are lower without spatial partitioning as shown by the MAP obtained for baseline approach with 128 and 64 Gaussians in the GMM.

Table 1: MAP in % on synthetic dataset with multiple objects.

No. of retrieved images	5	20	60	100
Contour partitioning, 128	94.93	81.65	66.82	60.48
Contour partitioning, 64	94.72	79.92	65.45	59.26
Rectangular partitioning, 128	93.38	76.85	59.99	53.45
Rectangular partitioning, 64	91.39	74.87	58.90	52.40
Baseline approach, 128	92.41	74.09	57.33	50.55
Baseline approach, 64	90.35	72.54	56.48	49.94

Table 2: MAP in % on real dataset with multiple objects.

No. of retrieved images	3	7	11	15
Contour partitioning, 128	26.04	27.34	25.48	24.98
Baseline approach, 128	12.36	14.24	14.44	13.90

For conducting experiments in real dataset, Giuseppe toys dataset [17] was considered. Fig. 5 shows the spatially different regions identified in some of the images. As expected the partitioning is better with a good saliency map. We added the images from the clutter-257 category from the Caltech dataset [15] which contains 827 images along with the 52 multi-toy images making a total of 879 images in the first database. This was done to have more images in the database to measure the retrieval performance. By using contour partitioning, 343 spatially partitioned regions are identified which resulted in a second database with $827 + 343 = 1170$ images. The FVs are formed for all the images using a GMM with 128 Gaussians. The database has 124 images with a single toy which belong to 31 different categories making 124 queries. Table 2 shows the retrieval results on the above 2 databases. We could observe that spatial partitioning improved the retrieval performance.

(a) Query image (b) Without partitioning (c) Rectangular partitioning (d) Contour partitioning

Fig. 4: Top 4 retrieval results in synthetic dataset with multi-object images for the query image with single object in (a) for 3 different cases. Images in black boxes represent the wrong retrievals.

(a) (b)

Fig. 5: Contours and the corresponding spatial regions identified for the toy images with multiple toys. (a) Images with proper partitioning. (b) Images with wrong detections because of heavy background clutter and overlap between objects.

5. CONCLUSIONS

We proposed a method to solve the problem of loss of spatial information while forming global descriptors by forming a BoFVs representation. A saliency model was used to identify the spatial locations of the objects. Two types of spatial partitioning were evaluated. Even though, contour partitioning tends to trim the object regions in some cases, it was found better than using rectangular partitioning which adds background information. Furthermore, it was observed that the contour partitioning is having a higher impact in increasing the retrieval accuracy than the increase in number of Gaussians in the GMM. An interesting research direction will be to compress these BoFVs to form a compact representation as now we require more memory to store FVs corresponding to each spatial region.

6. REFERENCES

[1] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," *ECCV Workshop on Statistical Learning in Computer Vision*, vol. 1, no. 1-22, pp. 1–2, 2004.

[3] Hervé Jégou, Florent Perronnin, Matthijs Douze, Javier Sanchez, Pablo Perez, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[4] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[5] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2169–2178, 2006.

[6] Rene Grzeszick, Leonard Rothacker, Glenn Fink, et al., "Bag-of-features representations using spatial visual vocabularies for object classification," in *Proceedings of the 20th IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 2867–2871.

[7] Edmond Zhang and Michael Mayo, "Improving bag-of-words model with spatial information," in *Proceedings of the 25th IEEE International Conference of Image and Vision Computing New Zealand (IVCNZ)*, 2010, pp. 1–8.

[8] Jorge Sánchez, Florent Perronnin, and TeóFilo De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2216–2223, 2012.

[9] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1155–1162.

[10] Tommi Jaakkola, David Haussler, et al., "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, 1999, pp. 487–493.

[11] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[12] Rafael C Gonzalez, Richard Eugene Woods, and Steven L Eddins, *Digital image processing using MATLAB*, 2004.

[13] Benjamin W Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 4, 2007.

[14] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[15] Gregory Griffin, Alex Holub, and Pietro Perona, *Caltech-256 object category dataset*, 2007.

[16] M-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1447–1454, 2006.

[17] "Giuseppe toys dataset," `www.vision.caltech.edu/pmoreels/Datasets/Giuseppe_Toys_03/`, [Online; accessed 18-Aug-2015].