Title
PixNet: A Localized Feature Representation for Classification and Visual Search

Permalink
https://escholarship.org/uc/item/8js5z36z

Journal
IEEE Transactions on Multimedia, 17(5)

ISSN
1520-9210

Authors
Pourian, Niloufar
Manjunath, BS

Publication Date
2015-05-01

DOI
10.1109/tmm.2015.2410734

Peer reviewed

PixNet: A Localized Feature Representation for Classification and Visual Search

Niloufar Pourian, *Student Member, IEEE*, and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—This paper presents a novel localized visual image feature motivated by image segmentation. The proposed feature embeds relative spatial information by learning different image parts while having a compact representation. First, an attributed graph representation of an image is created based on segmentation and localized image features. Subsequently, communities of image regions are discovered based on their spatial and visual characteristics over all images. The community detection problem is modeled as a spectral graph partitioning problem. This results in finding meaningful image part groupings. A histogram of communities forms a robust and spatially localized representation for each image in the database. Such a region-based representation enables one to search for queries that might not have been possible with global image representations. We apply this representation to image classification and search and retrieval tasks. Extensive experiments on three challenging datasets, including the large-scale ImageNet dataset, demonstrate that the proposed representation achieves promising results compared to the current state-of-the-art methods.**

***Index Terms*—Community detection, feature extraction, image classification, segmentation.**

I. Introduction

SEARCHING for images with a specific visual content has been a topic of intense research in recent years [1]. However, much of this recent work is focused on a global image representation. Searching for small regions of interest in larger images is still a challenging problem. An example of such a case is when one is looking for a car with a specific logo in the photos taken by people. In such a scenario, having an effective, localized feature representation is crucial. While localized methods have been well investigated in the context of detection and recognition of objects in a scene, they have not found wide applicability in scalable visual search. The primary contribution of this paper is developing a novel feature representation that is localized and compact.

Conventional methods usually represent an image based on low level global features. These include various global color and texture descriptors, SIFT and GIST features, and Bag of Word models [2]–[6]. While many of underlying features (e.g., SIFT) are well localized, the image representations are usually global (e.g., histogram of gradients). One can obviously build localized versions by either explicitly partitioning the image using segmentation methods or by imposing a pre-defined image grid. In such cases, additional steps will be needed to process the query as well as the search results.

There is a wealth of published literature that demonstrates the usefulness of spatial information in visual classification and search [7]–[9]. In [9]–[11], a large number of key-point based descriptors are computed and their relative spatial relationships are encoded. Also, [12] calculates the location offset of two matched features. The work in [7] incorporates spatial layout by introducing a Gaussian location model per visual word and encoding only the absolute spatial information. Utilizing localized grids into the feature representation is also another common approach to integrate spatial information [8], [13]–[15]. These methods often result in a high-dimensional representation and rely on a pre-defined partitioning of the image which is independent of its content. [16] models the spatial layout of images by sampling a large number of windows per image and weighting local features proportional to the number of windows that overlap them when computing a Fisher vector representation. While [16] does not rely on partitioning of the image, it still leads to an increase in the dimensionality of the Fisher vector.

In addition, [17] and [18] are based on fast approximate spatial verification. However, due to the high computational cost, these methods are only applied to the top ranked images in retrieval as a post-processing step.

Some researchers utilize part based models for object detection and image-level annotation [19], [20]. The work in [20] proposes a new class of object models by specifying a set of points on the target object boundary in training images with respect to a set of pre-defined parts (e.g., dog head, dog leg, etc.). Also, [19] involves training a part-based model from images labeled with bounding boxes around the object of interest. These approaches require a manual labeling that is costly to be scalable. However, in our work, we are only dealing with image level labels.

In addition, one can focus on introducing codebooks by encoding image parts [21]–[23]. For instance, the work in [23] focuses on a semantic-aware image retrieval by introducing ObjectWord defined as a collection of discriminative image patches annotated with the corresponding objects. To create each ObjectWord, [23] requires that each image be annotated by a single class. Also, [21] introduces a Part-Book for image parsing. The approaches in [21]–[23] rely on a pre-defined partitioning of the image which is independent of its content.

Manuscript received June 02, 2014; revised October 20, 2014 and February 16, 2015; accepted February 21, 2015. Date of publication March 05, 2015; date of current version April 15, 2015. This work was supported by the ONR under Grant N00014-12-1-0503 and by the UCSB Doctoral Scholar Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiao-Ping Zhang.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: npourian@ece.ucsb.edu; manj@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2410734

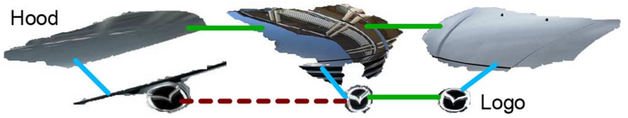

Fig. 1. Robustness to non-ideal segmentation by encoding the relative spatial information. The blue line represents the connection due to spatial adjacency, the green line represents the connection due to attribute similarity, and the red line demonstrates the correct identification of the over-segmented logo. Best viewed in color.

Alternatively, one can compute localized features in segmented image regions [24]–[33]. These methods adapt to the image content, and help in defining a limited number of spatial neighborhoods as the image dimensionality goes from millions of pixels to few tens of segments. The approaches of [24], [26] are sensitive to segmentation quality and thus they resort to manual segmentation in the training phase. [27] uses template matching between segmented regions while ignoring the relative spatial information existing among regions. In [25], the authors propose a graph based segmentation framework that is able to integrate cues from multi-layer superpixels simultaneously and shows improvement in segmentation. However, it does not tackle the classification problem. The works of [29], [32]–[34] focus on a semantic segmentation and propose a model to recover the pixel labels of the training images. However, the aforementioned methods require that every image in the training set to be labeled by all the classes that it contains. This is unrealistic for most scenarios of practical value and consequently is limited to much smaller datasets. In contrast, the proposed approach tackles image classification and retrieval and does not require a user to annotate training images with all of its associated image labels. Instead, it automatically groups related image parts across the training set using spectral clustering. [28] proposes an approach based on soft-matching tree-walks, however it requires that every image be segmented into equal number of regions.

In the following, we introduce a spatially localized representation that captures the attribute similarity with the relative spatial information without encoding the spatial information at the pixel level or a strong dependence on segmentation. We do this by posing a problem of learning image parts that are composed of segmented regions. These regions are homogeneous in color and texture feature space. An attributed graph for each image based on segmented regions captures the relative spatial information. In addition, collective consideration of all the regions' attributes in the dataset allows us to determine their visual similarity. We combine this visual similarity with the localized spatial information by creating a network of segmented regions. Different communities of related regions are discovered based on spatial and visual characteristics. In addition to learning parts of an image in an adaptive way, this community detection compensates for variations in segmentation.

As an example, one can consider an image of a car with over-segmented logo. In our network representation, a logo is connected to a hood due to spatial adjacency. The hood is linked (based on visual similarity) to other hoods that are themselves adjacent to logos. This configuration enables the system to identify the logo correctly as demonstrated in Fig. 1.

In the proposed network representation, one can think of the communities of related regions of all images as codebooks with embedded spatial information. Each segmented region in an image is associated with a histogram representing the likelihood of belonging to different communities. An image can now be represented by combining the individual region histograms, resulting in a robust descriptor with embedded relative spatial information. We call this new visual image feature a *PixNet*. Representing an image by a PixNet visual image feature is analogous to learning a puzzle from its pieces.

In summary, the main contributions of this paper is to introduce a novel representation that:

— integrates spatially localized information;

— is robust to segmentation variations;

— has a considerably more compact representation than the other state of the art methods, making the classifier learning more efficiently and helping in scaling to larger datasets.

The remainder of this paper is organized as follows. In Section II, we describe the overall framework of PixNet feature representation. The applicability of the representation is illustrated in Section III through an image classification, and query retrieval problem on three challenging datasets. Finally, we conclude the paper with some final remarks and directions for future research in Section IV.

II. PixNet Feature Representation

Given a set of images, the goal is to find a compact representation that encodes the relative spatial information between object/image parts. The overall framework for the creation of this feature is illustrated in Fig. 2. A graph is constructed for every image based on segmented regions and each region is represented by a node in the graph. Two nodes are connected by an edge if the corresponding regions are adjacent. We combine the graphs of all images in a large network that represents different object parts among different images in the database. Furthermore, we find groups of similar/related nodes in the network by community detection. Each group (community) defines a visual codeword that is embedded with the relative spatial information. Then, each region is represented by a histogram of these codewords with each bin indicating the strength of association of the region to the corresponding codeword. Finally, each image is represented by sum pooling of its regions' histograms. We call this image representation a PixNet visual feature. In the following, we describe the different stages of processing leading to the construction of the PixNet feature.

A. Graph Representation

Let D denote the number of images in the dataset. For every image I in the dataset, $I \in \{1, \ldots, D\}$, we incorporate the relative spatial information using a graph structure. Let $G^{(I)} = (V^{(I)}, E^{(I)})$ be a graph corresponding to image I where $V^{(I)}$ and $E^{(I)}$ represent the nodes and edges of this image, respectively. Each node represents a segmented region and two nodes are connected by an edge if the corresponding regions are adjacent, i.e. $E_{ij}^{(I)}$ is 1 iff region i is adjacent to region j and 0 otherwise. An example of the graph representation of each image is shown in Fig. 3. To compute the segmentation, we

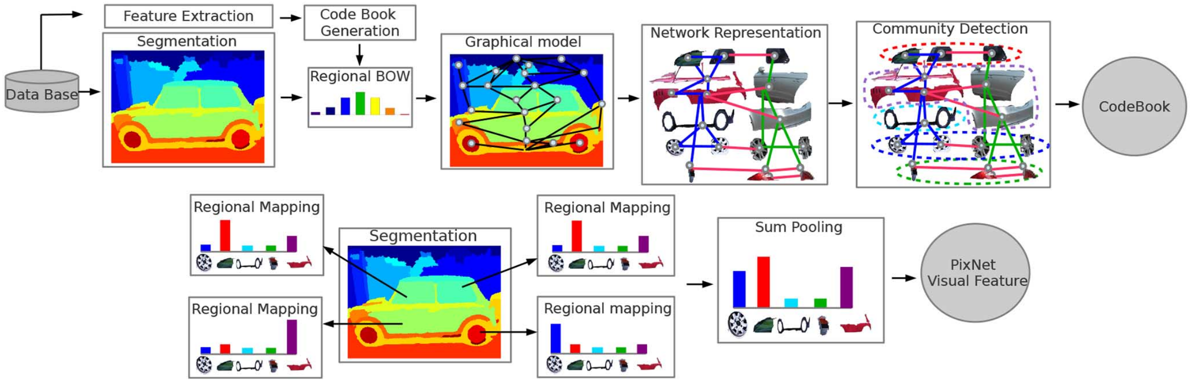

Fig. 2. PixNet visual features construction as described in Section II. This includes the following steps: segmentation, graph and network representations, community detection of the network, and the regional mappings.

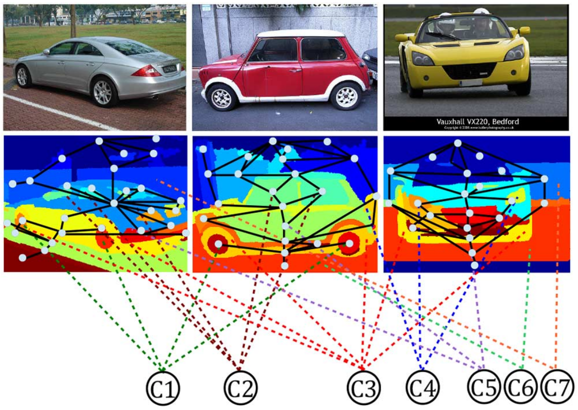

Fig. 3. Graph representations of different images. Each segmented region represents a node in the graph. Communities c_1–c_7 represent the new visual codewords embedded with localized spatial information. Regions belonging to each community are indicated by dotted lines. To avoid clutter, only the most common communities are represented. Image is best viewed in color.

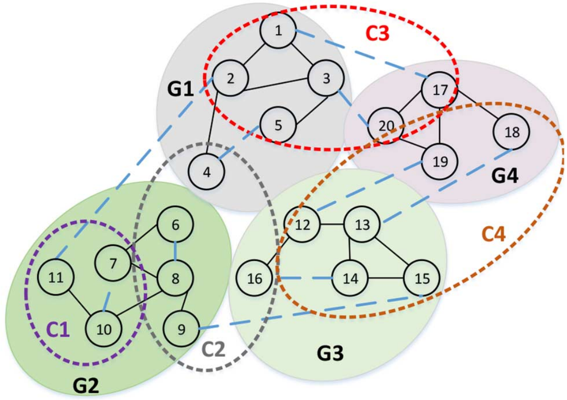

Fig. 4. Illustration of the network representation. This figure represents an example of a network created by using a database of four images. Each node represents a segmented region in this network. Black lines correspond to the connections due to spatial adjacency while dotted blue lines correspond to connections based on visual similarity. G_i, with $i \in \{1, \dots, 4\}$, denotes the graphical representation of image i in the database. Each dotted ellipse C_j, with $j \in \{1, \dots, 4\}$, is a detected community of related regions across that network.

use the method proposed in [35] since the software was publicly available and produces a reasonable number (10-100) of segmented regions per image. We did explore super-pixel segmentation methods but they typically resulted in a significant over-segmentation of the images. Since segmentation itself is not the focus of this work, we did not explore optimizing various segmentation criteria.

To represent regions (nodes), we extract densely sampled SIFT features [36] from each image, and map each 128 dimensional feature vector to a segment that they belong to. Each node is then represented by vector $h^{(i)}$ using the Bag of Words (BOW) model [6], or the Fisher Vectors (FV) model [37]. In the following, the appearance of node i is denoted by $h^{(i)}$.

B. Network Creation

Next, we create a network that is a collection of all nodes among all images. Until now, nodes are connected solely based on spatial adjacency. The idea of creating a network representation is to integrate the visual similarity between all regions in the network with the localized spatial information. We create a network of segmented regions $M = (V; A)$ where $V = \cup_{I=1}^{D} V^{(I)}$ represents the nodes in the network M, and A is the corresponding adjacency matrix. The (i, j)th element of A, denoted by $A_{i,j}$, is the weight indicating the strength of connectivity between nodes i and j. In creating this network, two nodes are connected if they are relevant either due to visual similarity or spatial adjacency. This will be done in the following two steps. An example of the network representation is shown in Fig. 4.

First, two nodes i and j are connected by a weighted edge equal to their attribute similarity (defined in equation (2)) if node i/j belongs to the set of T most similar nodes to node j/i. Second, we incorporate the spatial adjacency to this network. Two spatially adjacent nodes are connected with a weighted edge equal to the average of the weights of all edges connected to the corresponding nodes. This process is illustrated in Algorithm 1, and can be summarized in the following equation:

$$\begin{aligned} A_{i,j} = {} & \omega(i,j)\mathcal{I}\{i \in \mathcal{T}_j \quad or \quad j \in \mathcal{T}_i\} \qquad (1) \\ & + \frac{1}{2T}\left[\sum_{p \in \mathcal{T}_i} \omega(i,p) + \sum_{q \in \mathcal{T}_j} \omega(j,q)\right] \\ & \cdot \mathcal{I}\{i \notin \mathcal{T}_j \& j \notin \mathcal{T}_i \& i \in \mathcal{H}_j\} \end{aligned}$$

where $\omega(i, j)$ denotes the attribute similarity between two node i and j which will be discussed shortly, $\mathcal{T}_i$ is the set of T most similar nodes to node i, and $\mathcal{H}_i$ denotes the set of all nodes in spatial neighborhood of node i. The $\mathcal{I}\{x\}$ represents the indicator function and is equal to 1 if x holds true and zero otherwise.

The attribute similarity between two nodes i and j is given by the following:

$$\omega(i,j) = \underbrace{e^{-d(h^{(i)},h^{(j)})}}_{regional\ similarity} \underbrace{\gamma^{\mathcal{I}\{L(i)==L(j)\}}}_{label\ similarity} \tag{2}$$

where $d(h^{(i)}, h^{(j)})$ represents the distance between appearances of two nodes i and j, $L(i)$ denotes the label associated with the image that node i belongs to, and γ is a constant larger than 1. We set $\gamma > 1$ to give a higher weight to the visual similarity of two nodes that belong to images with the same label.

Algorithm 1 Creating Network M

```
Input: $G^{(I)} = (V^{(I)}, E^{(I)}) \quad \forall \quad I \in \{1, \ldots, D\}, \quad T$
Output: $V$, $N$, $A$
$V = \cup_{I=1}^{D} V^{(I)}$
$N = |V|$
$A \leftarrow N \times N$ zero vector

Comment: integrating attribute similarity into adjacency matrix of $M$
for $i = 1 \to N$ do
    for $j = 1 \to N$ do
        compute $\omega(i,j)$
    end for
end for
for $i = 1 \to N$ do
    $index \leftarrow sort(\omega(i,:), descend)$
    $\mathcal{T}_i \leftarrow index(1:T)$
    for $j = 1 \to N$ do
        if $j \in \mathcal{T}_i$ then
            $A_{i,j} \leftarrow \omega(i,j)$
        end if
    end for
end for

Comment: integrating the spatial context into adjacency matrix of $M$
for $i = 1 \to N$ do
    for $j = 1 \to N$ do
        if $i$ and $j$ are not connected due to attribute similarity then
            if $i$ and $j$ are connected due to adjacency then
                $A_{i,j} \leftarrow \frac{1}{2T}\left(\sum_{P \in \mathcal{T}_i} \omega(i,P) + \sum_{Q \in \mathcal{T}_j} \omega(j,Q)\right)$
            end if
        end if
    end for
end for
```

When using BOW model to represent nodes, we use the Hellinger metric [38] to compute the distance between $h^{(i)}$ and $h^{(j)}$. Hellinger distance has been shown to be a good metric for computing the distance between histograms in classification and retrieval problems [39]. For $\mathcal{L}_1$ normalized $h^{(i)}$ and $h^{(j)}$, distance $d(h^{(i)}, h^{(j)})$ is computed by

$$d(h^{(i)}, h^{(j)}) = \left(\sum_{k=1}^{K} \left(\sqrt{h_k^{(i)}} - \sqrt{h_k^{(j)}} \right)^2 \right)^{1/2} \tag{3}$$

with K denoting the size of the codebook for BOW (number of clusters found by approximate kmeans). In the case of Fisher Vector (FV) model, we choose $d(h^{(i)}, h^{(j)})$ to be the Euclidean distance between $h^{(i)}$ and $h^{(j)}$.

C. Community Detection

Our goal is to find similar/related regions in the network representation by which one can detect spatially localized codewords. We call each group a *community*. By partitioning the network into different communities, it is possible to learn similar parts of the objects in the database, for instance a group of logos in the examples of Fig. 1. In addition, such a community detection can compensate for variations in segmentation.

For graph partitioning, we use the normalized cut method as described in [40]. In this algorithm, the quality of the partition (cut) is measured by the density of the weighted links inside communities as compared to the weighted links between communities. The objective is to maximize sum of the weighted links associated with a particular community while minimizing sum of the weighted links associated between this community and other communities.

Suppose C be the number of partitions (communities) in the network. This graph partitioning provides a codebook of size C that is integrated with spatial information. If we choose C to be small (fewer number of communities), the detected communities may be such that they include all parts of a particular object as a whole. While larger values of C result in a case that different parts of objects fall into different communities.

Once this community detection technique is applied to the images in the database, one can find communities of related segmented regions. Fig. 5 shows an example of how different parts of eight sample images belong to the detected communities.

D. Regional Mappings

Once we detect communities of similar regions, each region is represented by a histogram of these communities. The bins of each histogram indicates the strength of association of a region to the corresponding community. Let $\mathcal{H}_i$ denote the set of all nodes in the spatial neighborhood of node i, ϕ_c be the community c with $c \in \{1, \ldots, C\}$, and $\mathcal{T}'_i$ denote the set of all nodes that are in the top T' nearest neighbors of node i. The strength of association of a node i to a community ϕ_c is measured by two factors: first by the attribute similarity between node i and community ϕ_c, second by considering the attribute similarity between neighbors of node i and different communities in the network along with the relation between community ϕ_c and each of the communities in the network. Fig. 6 illustrates the mapping process of each node to the detected communities.

Let $g(i \in \phi_c)$ denote the attribute similarity between node i and community ϕ_c. The function $g(i \in \phi_c)$ is defined by the

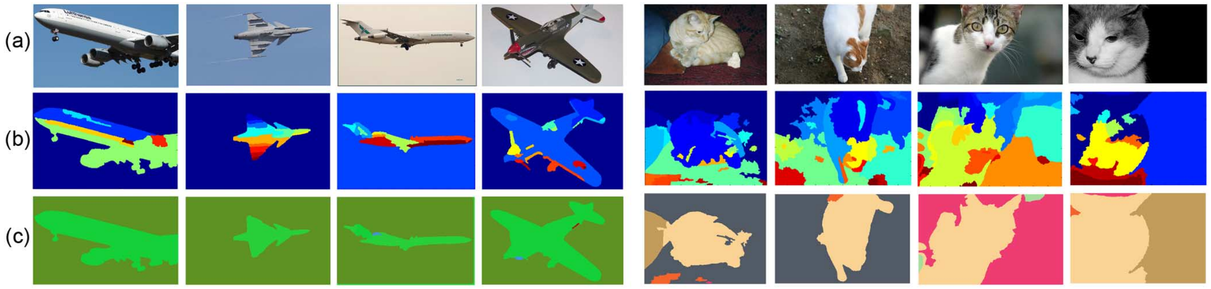

Fig. 5. Row (a) illustrates eight sample images from the VOC07 database. Row (b) represents the corresponding segmentations of the samples images. Each color denotes a segmented region. The segmentation is obtained by online software provided by authors in [35]. In row (c), each color denotes a community that the segmented regions of figures in row (b) belong to. These communities belong to the set of all detected communities over the entire database. The community detection algorithm is described in Section II-C.

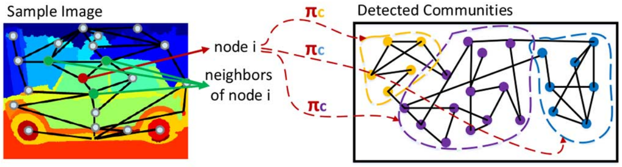

Fig. 6. Illustration of the mapping process of each node to the detected communities. Segmentation of a sample image is depicted (left) and the detected communities are represented by dotted closed-curves (right). The degree of association of each node i and each of the communities is denoted by π_c^i.

fraction of top T' nearest neighbors to node i that belong to community ϕ_c

$$g(i \in \phi_c) = \frac{\sum\limits_{j \in \mathcal{T}_i'} \mathcal{I}\{j \in \phi_c\}}{T'}. \tag{4}$$

Moreover, we define $f(\phi_c', \phi_c)$ to measure the relation between two communities ϕ_c' and ϕ_c

$$f(\phi_c', \phi_c) = \frac{\sum\limits_{i \in \phi_c'} \sum\limits_{j \in \phi_c} \mathcal{I}\{A_{i,j} > 0\}}{\sum\limits_{i \in \phi_c'} \sum\limits_{j=1}^{N} \mathcal{I}\{A_{i,j} > 0\}} \tag{5}$$

where $N = |V|$ denotes the total number of nodes in the network. In particular, $f(\phi_c', \phi_c)$ measures the number of links between the two communities ϕ_c' and ϕ_c divided by the total number of links between community ϕ_c' and all other communities. Thus, the strength of association of a node i to a community ϕ_c can be determined by π_c^i in (6), as shown at the bottom of the page. Now one can use π_c^i to represent each node by a histogram of length C. In such representation, each bin c is equal to π_c^i that is the likelihood of node i belonging to community c. Finally, each image is defined by summing the histograms that are associated with the segmented regions of the image.

Fig. 7. Sample images from VOC07 (top row), TREC (middle row), and ImageNet2010 (bottom row) databases with their corresponding labels.

III. Experimental Results

We evaluate our method in several different settings by using three challenging datasets: PASCAL VOC07, TREC, and ImageNet'10 (ILSVRC2010) dataset. We compare our method with 3 different baseline methods for image classification described in Section III-B. Moreover, it is shown how the performance

$$\begin{aligned} \pi_c^i &= \frac{\sum\limits_{j \in \mathcal{H}_i} \left[f(\phi_c, \phi_1) g(j \in \phi_1) + \ldots + f(\phi_c, \phi_C) g(j \in \phi_C) \right] g(i \in \phi_c)}{\sum\limits_{c''=1}^{C} \sum\limits_{j \in \mathcal{H}_i} \left[f(\phi_{c}'', \phi_1) g(j \in \phi_1) + \ldots + f(\phi_{c}'', \phi_C) g(j \in \phi_C) \right] g(i \in \phi_{c}'')} \\ &= \frac{\sum\limits_{j \in \mathcal{H}_i} \left[\sum\limits_{c'=1}^{C} f(\phi_c, \phi_c') g(j \in \phi_c') \right] g(i \in \phi_c)}{\sum\limits_{c''=1}^{C} \sum\limits_{j \in \mathcal{H}_i} \left[\sum\limits_{c'=1}^{C} f(\phi_{c}'', \phi_c') g(j \in \phi_c') \right] g(i \in \phi_{c}'')} \end{aligned} \tag{6}$$

TABLE I
COMPARISON OF MAP BETWEEN SPATIAL PYRAMIDS (SPM) WITH ℓ LEVELS, SPATIAL FISHER VECTORS (SFV), AND PIXNET VISUAL FEATURES (PIXNET) ON VOC07 (TOP) AND TREC (BOTTOM) DATABASES. BAG-OF-WORDS (BOW) AND FISHER-VECTORS (FV) ARE USED FOR CODING REGIONAL APPEARANCES. RESULTS ARE REPORTED FOR $C = 100$

Database: VOC07							
K / Method		SPM			SFV	PixNet	
		ℓ=0	ℓ=1	ℓ=2		Coarse	Fine
BOW	50	29.1	37.0	41.4	35.7	45.3	48.0
	200	38.1	44.1	47.1	42.4	50.6	53.1
	1000	45.9	50.1	51.5	50.3	54.2	56.1
FV	50	54.1	55.8	56.4	55.4	59.0	61.1
	100	55.0	56.5	56.9	56.2	62.5	64.5

Database: TREC							
K / Method		SPM			SFV	PixNet	
		ℓ=0	ℓ=1	ℓ=2		Coarse	Fine
BOW	50	21.7	27.8	30.7	25.3	32.6	35.7
	200	33.1	39.1	41.3	38.2	43.9	45.5
	1000	42.5	48.3	49.5	47.9	50.9	51.8
FV	50	44.0	46.2	46.4	46.1	47.2	48.1
	100	45.1	46.7	46.8	46.1	48.5	50.3

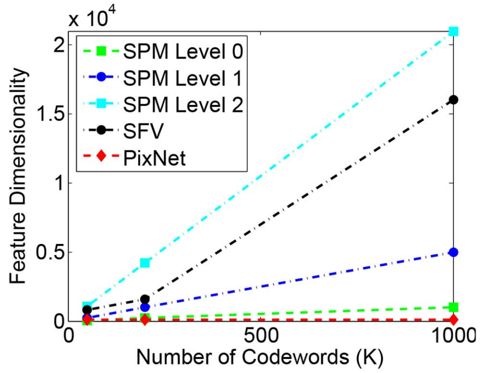

Fig. 8. Comparison of the dimensionality of SPM, SFV, and PixNet features.

TABLE II
COMPARISON OF THE ERROR RATE OF PIXNET VISUAL FEATURES ON IMAGENET 2010 DATASET WITH THE STATE OF THE ART REPORTED RESULTS. PIXNET ACHIEVES A SLIGHTLY HIGHER ERROR RATE THAN CNN AT A MUCH LOWER COMPUTATIONAL COST. PIXNET RESULTS ARE REPORTED FOR $C = 5,000$

Database: ImageNet 2010	
Method	Error Rate
Sparse coding [44]	47.1
SIFT + FV [45]	45.7
CNN [46]	37.5
PixNet	40.6

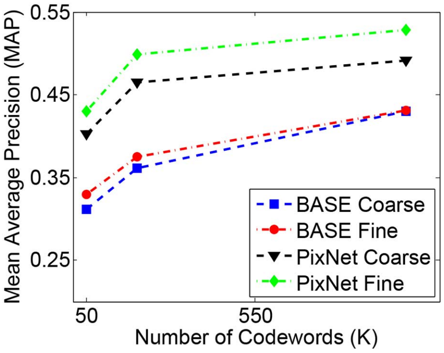

Fig. 9. Comparison between the performance of BASE and PixNet method. Results are reported for PASCAL VOC07 database.

of our method varies as the level of segmentation changes. By setting the parameters of [35], we achieve two levels of segmentation with an average number of 20 and 50 segments per image which are referred to as “Coarse” and “Fine”, respectively. We further evaluate the performance of our approach as a function of the number of detected communities. In addition, we show the sensitivity of our method on parameter T, number of nearest neighbors based on similarity in the creation of the network. Finally, we illustrate the applicability of our method in a query retrieval setting where the query image contains only the object of interest.

A. Datasets

We evaluated the performance of the PixNet visual features on three datasets: ImageNet 2010 (ILSVRC2010) [41], PASCAL VOC07,[1] and TREC. ImageNet 2010 is a publicly available dataset containing 1,000 object classes and 1.2 million training images, 50,000 validation images and 150,000 test images. PASCAL VOC07 is another publicly available dataset containing 20 object classes, 5, 011 training images and 4,952 test images. To evaluate the performance of different methods in identifying the object of interest when it occupies only a small portion of the image in a cluttered background, we have collected a set of images by extracting frames from TRECVID 2012 instant search (INS) dataset (consists of about 70,000 short video files) which we refer to as TREC dataset. Since the groundtruth was only published for a subset of the data, we have only considered classes that have sufficient numbers of true positives. TREC dataset contains 10 object classes (Mercedes logo, Brooklyn bridge tower, Eiffel tower, Golden Gate Bridge, London Underground logo, Coca cola logo, Stonehenge, US Capital exterior, Hoover Dam exterior, One World Trade center building), 5, 083 training images, and 5, 206 test images. Fig. 7 illustrates examples of images in each of the databases.

B. Baseline Methods

We compare our method with Spatial Pyramid method (SPM) [8] which encodes the global positions of features in the image, and also with the Spatial Fisher Vectors (SFV) [7] which incorporates spatial layout by introducing a Gaussian location model per visual word. Since the code for SFV was not publicly available, we re-implemented the algorithm based on [7].

To show the importance of encoding the spatial adjacency of segmented regions, we have also tested the performance by creating the network solely based on attribute similarity of the segmented regions in the database while ignoring the spatial neighborhood of the regions. We refer to this method as the baseline (BASE) method. We show how the presence and absence of the relative spatial neighborhood information affects the performance of the system.

In addition, to investigate the performance of PixNet features on a large dataset with large number of classes, the error rate

[1] “The Pascal Visual Object Classes Challenge 2007 (VOC2007) Results,” [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/

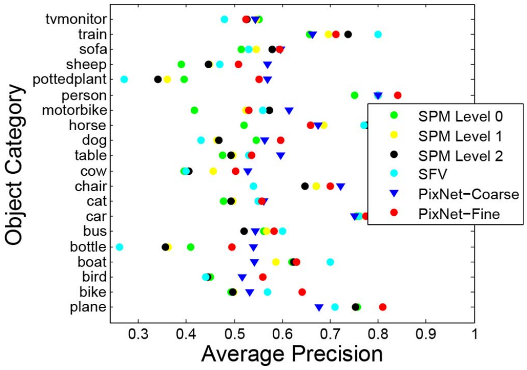

Fig. 10. Per-class classification accuracy of spatial pyramids (SPM) with ℓ levels, spatial fisher vectors (SFV), and PixNet visual features (PixNet) with "Coarse" and "Fine" levels. Fisher-vectors (FV) are used for coding regional appearances. Results are reported on the PASCAL (VOC07) dataset with K = 100.

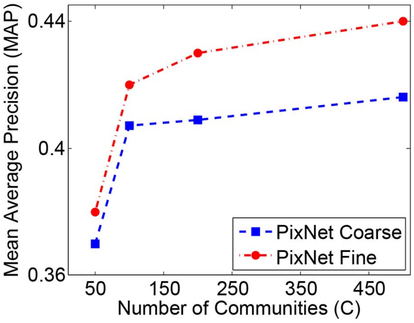

Fig. 11. Effect of number of communities (C) on the classification accuracy. Results are reported on the PASCAL (VOC07) dataset with K = 50.

of the proposed approach is compared with the state of the art results achieved on the challenging ImageNet dataset.

Finally, one should note that we did not compare with other segmentation based approaches [24], [26] as they require ground truth segmentation in the training phase.

C. Evaluation

Classification: For image classification, we learn a binary SVM classifier per class for all image representations, and evaluate the performance using Mean-Average-Precision (MAP). The Average Precision (AP) is computed by the interpolated average precision [42]. MAP is chosen in order to be comparable with the results reported on PASCAL VOC database.

Table I shows the MAP scores for different vocabulary sizes using BOW and FV as appearance models for segmented regions applied to VOC07 and TREC datasets. As shown, our approach achieves higher classification accuracy compared to the SPM and SFV methods while having a considerably more compact representation. It can be seen that Fine PixNet achieves a higher classification accuracy compared to the Coarse PixNet. This is due to the fact that larger number of segmented regions results in encoding more relative spatial information. Fig. 8 shows the comparison between the dimensionality of different features. The PixNet feature dimentionality does not change as the number of codewords (K) varies and is less than the dimension of all other methods. Together with the results in Table I, it is evident that among the methods considered here,

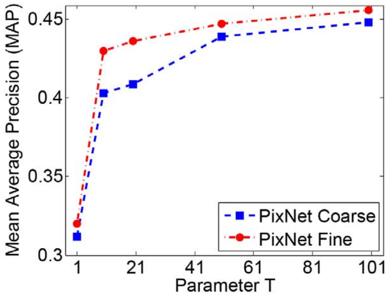

Fig. 12. Effect of parameter T on the classification accuracy. Results are reported on the PASCAL (VOC07) dataset with K = 50.

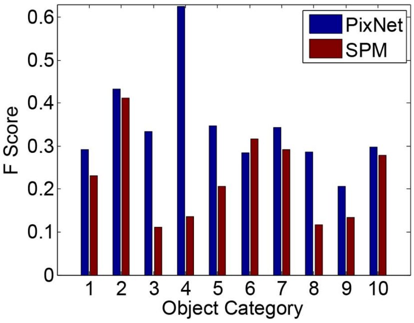

Fig. 13. Comparison between the performance of SPM with $\ell = 0$ and PixNet method in a query retrieval setting. Results are reported for TREC database with K = 50.

PixNet achieves the highest performance with the smallest feature dimension.

To compare with the state of the art results reported on ImageNet 2010, the performance of PixNet is measured by error rates. Table II illustrates such a comparison between PixNet and the state of the art [43]–[45]. It is worth noting that CNN [45] has a slightly lower error rate than PixNet. The work in [45] involves training a convolutional neural networks (CNN). The network consists of eight layers with weights; first five are convolutional and the remaining three are fully connected. The output of the last fully connected layer is fed to a softmax which produces a distribution over the class labels. Their network maximizes the multinomial logistic regression objective. To achieve their accuracy and train their large network, they found that they had to increase the size of the training data. Hence they used different data augmentation strategies such as random cropping of sub-images and random perturbations of the illumination. However, in the proposed work we did not need to do that. In general, CNN exploits rectangular image patches to learn high order local features using multilayer architectures. The learning capacity of these networks depends on the number and size of the kernels in each layer and the number of kernel combinations between layers. The model in [45] included 650,000 neurons, 60 million parameters, and 630 million connections and is trained with stochastic gradient descent on two GPUs for about a week. This is much more computationas computational complexity is discussed in III-D. This result emphasizes the suitability of PixNet for large scale datasets.

Fig. 9 compares the performance of our approach and the BASE method. One can see that PixNet method has a higher accuracy than the BASE method, thus highlighting the importance

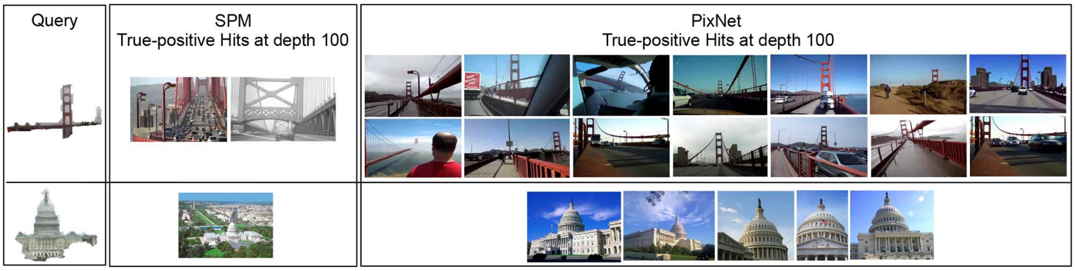
Query
SPM
True-positive Hits at depth 100
PixNet
True-positive Hits at depth 100

Fig. 14. Sample queries of Golden Gate Bridge (top) and U.S. State Capital Building (bottom) for TREC database with their true positive hits at depth 100 in a query retrieval problem for SPM and PixNet visual features.

of encoding the relative spatial information between segmented regions in an image.

The per-class classification accuracy of PixNet and baseline methods are shown in Fig. 10. It can be seen that PixNet achieves the highest boost in classification accuracy for "pottedplant" class compared to other methods.

Fig. 11 illustrates the effect of different number of communities on the classification accuracy. For PASCAL database, the performance of our approach barely changes as the number of communities varies for values larger than 100.

Furthermore, we investigated the sensitivity of our approach on parameter T, number of nearest neighbors based on similarity in the creation of the network. Fig. 12 shows that our method is robust in T over the range of 10 to 100. Very large values of T (due to the connection of unrelated nodes) adversely affect the performance.

Retrieval: Since the proposed method has the advantage of breaking the image into object parts, we also investigated the applicability of PixNet features in a query retrieval setting where the query image contains only the segmented region indicating the object of interest. The similarity score is computed by $\exp(-d(h^{(I)}, h^{(Q)}))$ where d is the Hellinger metric defined in equation (3), and $h^{(I)}$, $h^{(Q)}$ are the image representations of images I and Q. The performance is evaluated using F measure.

Fig. 13 illustrates that PixNet features achieve a higher retrieval score compared to the SPM at level 0. In addition, Fig. 14 shows results for a sample query along with its true positive hits at depth 100. The PixNet approach returns higher number of true positives. We have not reported results for SPM at higher levels or SFV as these methods work on rigid shapes and are not relevant in this case.

D. Scalability

The compact representation of the PixNet visual features makes it suitable for large scale datasets. A moderate computational complexity is important when considering scaling to thousands of images and hundreds of categories. When training our method the largest computational cost is in finding the top T nearest neighbors to each node as part of the algorithm for constructing the network. The simplest solution is to compute the distance from each node to all the other nodes in the network. This approach has a running time of $O(Nd)$ where N is the number of nodes in the network and d is the dimensionality of the data. We have greatly reduced this cost by performing the search using a space partitioning data structure (k-d tree) with running time of $O(\log N)$. We note that the original k-d tree might exhibit poor performance [46]–[52], and several remedies have been proposed [53]–[57]. We used the implementation of k-d tree in opencv library [58] and achieved good enough of a performance for all our experiments (see Table I). As the size of the database increases, the network can be constructed using a subset of the images from the training data. Consequently, one would not need to increase the cost associated with the network construction. This is analogous to the clustering of features using kmeans for codebook creation. At the test time, the largest bottleneck is in mapping each region of the test image to the communities of the network again by finding the top nearest neighbors of each region (and its surrounding regions) in each of the communities. At this stage, we take the similar approach as the training phase. In our experiments the overhead cost for VOC07 at training time for 5, 011 images is less than 5 minutes and at test time for 4,952 images is less than 3 minutes which is an average of 0.03 seconds/image using a quad core computer with 3.0 GHz processors. This cost is negligible compared to the gained compact regional representation.

E. Remark

The localized information corresponding to image-parts that occupy a relatively small portion of the image is typically lost in the majority of the image representations. Most such representations are global in nature, even when the features are computed locally. In contrast, the PixNet representation enables one to partition the image and learn its salient parts. In the proposed representation a segmented region is mapped to communities in PixNet and each community is given equal weight independent of number of pixels belonging to it. This is illustrated through the qualitative results shown in Fig. 14. As shown, the query occupies a small portion of many of the retrieved images, however the PixNet features are able to retrieve them. Furthermore, the quantitative results show that PixNet achieves a higher retrieval accuracy than the baseline methods. Another important aspect of the proposed PixNet features is that the detected communities map back to localized regions in the image, as illustrated in Fig. 5. Further, localized descriptors have a higher discriminative power as shown in applications of image classification (see Section III-C1), or content based image retrieval (see Section III-C2).

IV. Conclusion

This paper presented a compact spatially localized visual image feature by combining the visual similarity of image parts with the localized relative spatial information through a network of segmented regions. A graph partitioning algorithm is employed to discover groups (communities) of related segmented regions across the database. A histogram of communities forms a robust and spatially localized representation for each image in the database.

We illustrated the applicability of PixNet visual features in an image classification problem and presented results in a query retrieval example. Further, we showed that the proposed approach is robust to segmentation variations, and can achieve a more compact representations than the state of the art approaches. We show that even with sum pooling of the localized features for a classification task, PixNet is able to achieve a good performance. Sum pooling allows one to have a vector representation for an image as it might be more desirable in a classification task. This is because one can use the vector representation in utilizing many of the machine learning techniques. Depending on the application, one may choose to use a vector representation of an image, or choose to use the updated localized representation of image-parts. Our future work focuses on ways of better representing the localized feature representations (histogram of communities) and integrating them in a formulation to retrieve images with specific spatial configuration of objects.

The proposed method also has the potential for object class localization by highlighting regions that contribute to a specific object. A possible direction for future research is to explore image compression using PixNet visual features. Another direction for future research is in-video product annotation [59] and video analysis [60]. We also plan to investigate the applicability of PixNet visual features in multi-modal retrieval problems.

Acknowledgment

The authors would like to thank Dr. C. Lang and Dr. S. Karthikeyan for their helpful discussions.

References

[1] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, “The devil is in the details: An evaluation of recent feature encoding methods,” in *Proc. BMVC*, 2011, pp. 1–12.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *Proc. Workshop Statist. Learn. Comput. Vis.*, 2004, pp. 1–2.

[3] E. Kasutani and A. Yamada, “The MPEG-7 color layout descriptor: A compact image feature description for high-speed image/video segment retrieval,” in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2001, vol. 1, pp. 674–677.

[4] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 18, no. 8, pp. 837 –842, Aug. 1996.

[5] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[6] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1470–1477.

[7] J. Krapac, J. Verbeek, and F. Jurie, “Modeling spatial layout with fisher vectors for image categorization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1487–1494.

[8] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2169–2178.

[9] N. Morioka and S. Satoh, “Building compact local pairwise codebook with joint feature space clustering,” in *Proc. ECCV*, 2010, pp. 692–705.

[10] D. Liu, G. Hua, P. Viola, and T. Chen, “Integrated feature selection and higher-order spatial feature extraction for object categorization,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[11] T. Weyand, T. Deselaers, and H. Ney, “Log-linear mixtures for object class recognition,” in *Proc. BMVC*, 2009, pp. 1–11.

[12] Y. Zhang, Z. Jia, and T. Chen, “Image retrieval with geometry-preserving visual phrases,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 809–816.

[13] A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel,” in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.

[14] A. Hegerath, T. Deselaers, and H. Ney, “Patch-based object recognition using discriminatively trained gaussian mixtures,” in *Proc. BMVC*, 2006, pp. 519–528.

[15] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *Proc. ECCV*, 2010, pp. 143–156.

[16] J. Sánchez, F. Perronnin, and T. d. Campos, “Modeling the spatial layout of images beyond spatial pyramids,” *Pattern Recog. Lett.*, vol. 33, no. 6, pp. 2216–2223, 2012.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[18] G. Tolias and Y. Avrithis, “Speeded-up, relaxed spatial matching,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1653–1660.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[20] L. Zhu, Y. Chen, and A. Yuille, “Learning a hierarchical deformable template for rapid deformable object parsing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1029–1043, Jun. 2010.

[21] K. Yang, L. Zhang, Y. Rui, and H.-J. Zhang, “Partbook for image parsing,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 17–24.

[22] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, “Objectpatchnet: Towards scalable and semantic image annotation and retrieval,” in *Proc. Comput. Vis. Image Understand.*, 2014, pp. 16–29.

[23] S. Zhang, Q. Tian, Q. Huang, and W. Gao, “Objectbook construction for large-scale semantic-aware image retrieval,” in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2011, pp. 1–6.

[24] B. Fulkerson, A. Vedaldi, and S. Soatto, “Class segmentation and object localization with superpixel neighborhoods,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Sep.–Oct. 2009, pp. 670–677.

[25] Z. Li, X.-M. Wu, and S.-F. Chang, “Segmentation using superpixels: A bipartite graph partitioning approach,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 789–796.

[26] S. Savarese, J. Winn, and A. Criminisi, “Discriminative object class models of appearance and shape by correlatons,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2033–2040.

[27] J. R. Smith and C.-S. Li, “Image classification and querying using composite region templates,” in *Proc. Comput. Vis. Image Understand.*, 1999, pp. 165–174.

[28] Z. Harchaoui and F. Bach, “Image classification with segmentation graph kernels,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[29] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, “Weakly supervised semantic segmentation with a multi-image model,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 643–650.

[30] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, “Weakly supervised structured output learning for semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 845–852.

[31] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, “Active learning for semantic segmentation with expected change,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3162–3169.

[32] J. Verbeek and B. Triggs, “Region classification with Markov field aspect models,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[33] A. Vezhnevets and J. M. Buhmann, “Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3249–3256.

[34] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, “Weakly-supervised dual clustering for image semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2075–2082.

[35] Y. Deng and B. Manjunath, “Unsupervised segmentation of color-texture regions in images and video,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[36] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 1999, vol. 2, pp. 1150–1157.

[37] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[38] M. S. Nikulin, “E. of mathematics,” in *Hellinger Distance*. New York, NY, USA: Springer, 2001.

[39] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.

[40] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[41] J. Deng *et al.*, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.

[42] G. Salton and M. J. McGill*, Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.

[43] A. Berg, J. Deng, and L. Fei-Fei*, Large scale visual recognition challenge 2010*, 2010 [Online]. Available: http://image-net.org/challenges/LSVRC/2010/index

[44] J. Sánchez and F. Perronnin, “High-dimensional signature compression for large-scale image classification,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1665 –1672.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012, pp. 1097–1105.

[46] A. W. Moore*, An Intoductory Tutorial on kd-Trees*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 1991.

[47] V. Gaede and O. Günther, “Multidimensional access methods,” *ACM Comput. Surveys*, vol. 30, no. 2, pp. 170–231, 1998.

[48] J. L. Bentley and J. H. Friedman, “Data structures for range searching,” *ACM Comput. Surveys*, vol. 11, no. 4, pp. 397–409, 1979.

[49] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[50] J. H. Friedman, F. Baskett, and L. J. Shustek, “An algorithm for finding nearest neighbors,” *IEEE Trans. Comput.*, vol. C-24, no. 10, pp. 1000–1006, Oct. 1975.

[51] J. H. Friedman, J. L. Bentley, and R. A. Finkel, “An algorithm for finding best matches in logarithmic expected time,” *ACM Trans. Math. Softw.*, vol. 24, no. 10, pp. 1000–1006, 1977.

[52] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani*, The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, vol. 2, 1.

[53] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, “Locality-sensitive hashing scheme based on p-stable distributions,” in *Proc. 20th Annu. Symp. Comput. Geom.*, 2004, pp. 253–262.

[54] P. Indyk and R. Motwani, “Approximate nearest neighbors: Towards removing the curse of dimensionality,” in *Proc. 30th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.

[55] H. Jagadish, B. C. Ooi, K.-L. Tan, C. Yu, and R. Zhang, “idistance: An adaptive b+-tree based indexing method for nearest neighbor search,” *ACM Trans. Database Syst.*, vol. 30, no. 2, pp. 364–397, 2005.

[56] H. Xu *et al.*, “Complementary hashing for approximate nearest neighbor search,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1631–1638.

[57] C. Silpa-Anan and R. Hartley, “Optimised kd-trees for fast image descriptor matching,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[58] G. Bradski, “The OpenCV Library,” *Doctor Dobbs J.*, vol. 25, no. 11, pp. 120–126, 2000.

[59] G. Li, M. Wang, Z. Lu, R. Hong, and T.-S. Chua, “In-video product annotation with web information mining,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 8, no. 4, p. 55, 2012.

[60] M. Wang *et al.*, “Unified video annotation via multigraph learning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

Niloufar Pourian (S’13) received the B.S. and M.S. degrees in electrical and computer engineering from the University of California, Santa Barbara, CA, USA, where she is currently pursuing the Ph.D. degree with the Computer Vision Lab.

Her research interests include feature extraction, image classification, and multimedia retrieval problems.

Ms. Pourian is the recipient of the UC Regents Fellowship and Dean’s Doctoral Scholar Award in 2007 and 2010, respectively.

B. S. Manjunath (S’88–M’91–SM’01–F’05) received the B.E. degree from Bangalore University, Bangalore, India, the M.E. degree from the Indian Institute of Science, Bengaluru, India, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991.

He is currently a Professor of Electrical and Computer Engineering and Director of the Center for Bio-Image Informatics with the University of California, Santa Barbara, Santa Barbara, CA, USA. He has authored or coauthored over 280 peer-reviewed articles and is a co-inventor on 24 U.S. patents. His research interests include large scale image/video databases, camera networks, information security, and bio-image informatics.

Dr. Manjunath has served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and the IEEE SIGNAL PROCESSING MAGAZINE. He was a co-author of the paper that was the recipient of the Best Paper Award from the IEEE TRANSACTIONS ON MULTIMEDIA in 2013.