Application of Analytics in Machine Vision using Big Data

Noha Elfiky

Department of Business Analytics
School of Economics and Business Administration
Saint Mary's College of California (Moraga, California, United States)
Email: nme5@stmarys-ca.edu

ABSTRACT— ***The Bag-of-Words (BoW) approach has been successfully applied in the context of category-level image classification. To incorporate spatial image information in the BoW model, Spatial Pyramids (SPs) are used. However, spatial pyramids are rigid in nature and are based on pre-defined grid configurations. As a consequence, they often fail to coincide with the underlying spatial structure of images from different categories which may negatively affect the classification accuracy.***

The aim of the paper is to use the 3D scene geometry to steer the layout of spatial pyramids for category-level image classification (object recognition). The proposed approach provides an image representation by inferring the constituent geometrical parts of a scene. As a result, the image representation retains the descriptive spatial information to yield a structural description of the image. From large scale experiments on the Pascal VOC2007 and Caltech101, it can be derived that SPs which are obtained by the proposed Generic SPs outperforms the standard SPs.

Keywords— Big Data Analytics, Machine Vision, Image Classification and Object Recognition Tasks, Bag of Words, Spatial Pyramids

1. INTRODUCTION

For category-level image classification and object recognition, the Bag-of-Words (BoW) approach has been successfully applied [1], [2], [3], [4]. The BoW is based on the occurrences of image features. Hence, it treats the image as an order-less collection of local features completely ignoring the spatial image layout. Hence, it treats the image as an order-less collection of local features completely ignoring the spatial image layout.

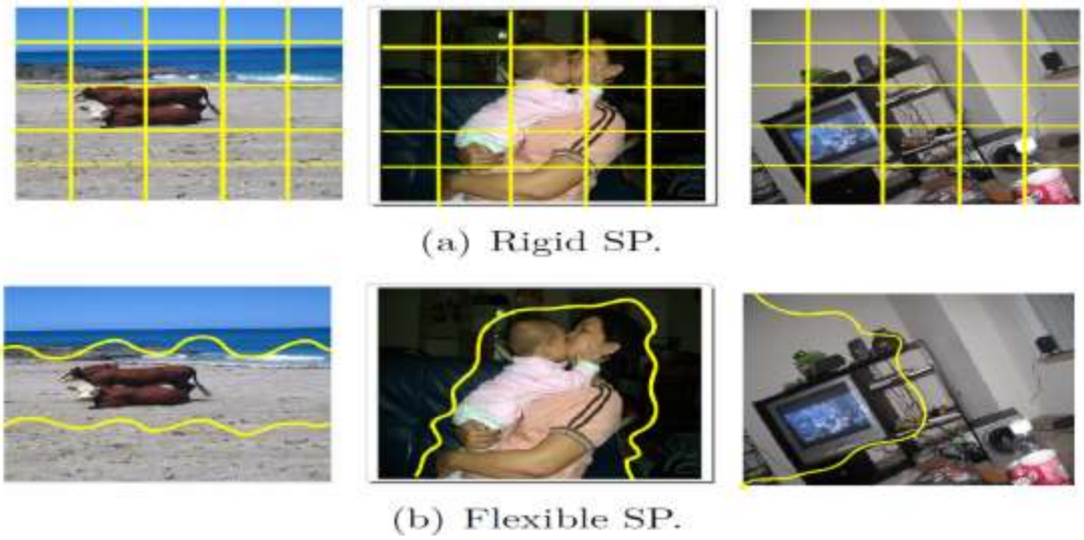
Extending the *BoW* with spatial information has therefore received considerable attention. Recently, several approaches consider the success of the Spatial Pyramid *(SP)* approach proposed by Lazebnik et al. [5]. It is shown that the use of *SP* outperforms the 1×1 image representation on challenging image classification tasks [5], due to the inclusion of image-to-image geometric correspondences. However, in general, SPs are based on rigid image subdivisions (e.g., grids). These rigid spatial configurations are not well suited for freely shaped objects and scenes. In Figure 1(a), some examples are shown taken from different image categories together with their standard SP sub-division. Sub-regions divide objects into two separate parts increasing the probability of dissimilar image features within cells and similar image features across cells. Hence, a rigid division may a negative equivalence-class configuration of image features.

Our aim is to use 3D scene geometry to steer the layout of spatial pyramids for category-level image classification. Images within a category usually share similar scene geometries. We exploit correspondences between categories by a scene geometry matching scheme. For example, Figure 1(b) shows the geometrical (depth) layers of some example images. The *cow example* corresponds to scene geometry style consisting of 3 segments: (1) ground, (2) background and (3) sky. The "ground" part depicts different objects than the background and sky. Each segment contains similar features and features across segments are more dissimilar. Therefore, the *BoW* should be applied separately to each geometrical scene sub-region.

In this paper, we propose a method to obtain a holistic image representation by inferring the constituent geometrical parts of a scene. The method steers the image layout on the basis of 3D scene geometry (i.e., "Stages") computed from a single image. We propose the Generic SP approach to obtain structural image representations from 3D scene geometries by exploiting the 3*D* scene geometry of images. The 13 stages of [7] are used as 3D priors. After the image scene geometry is estimated, the most appropriate stage per object category is selected as the spatial pyramid.

The proposed method to generate SPs will be compared to existing rigid SP for object recognition tasks. To this end, two benchmark data sets are used in the experiments: Pascal VOC 2007 [6] and Caltech-101 [10]. Furthermore, a large data set is provided *(denoted as "stage data set")* to learn each stage.

This paper is organized as follows. First, in section 2 and 3, we give the motivation of our approach and discuss related work on rigid spatial pyramids. In section 4, the method is proposed to generate dynamic spatial pyramids. In section 5, the experimental setup is discussed and the results are given. Finally, we give the conclusion in Sec. 6.


(a) Rigid SP.

(b) Flexible SP.

Figure 1: Show example images from different categories. (a) shows the standard SP proposed by Lazebnik et al. [5]. (b) shows the proposed flexible spatial partitionings which best suit each category. Images are from the Pascal dataset [6].

2. MOTIVATION

The subdivision scheme should consider the trade-off between two important design properties *invariance* and *descriptiveness*. *Larger* subregions are preferred to gain invariance to viewpoint changes (translation, orientation and scale) and object occlusion. Sub-regions should cover the range of possible positions of occurring objects. For example, the entire image is invariant to all possible object positions. *Smaller* sub-regions are required to obtain more descriptive regions and spatial layout. Sub-regions should depict similar object/background augmenting the descriptive ability of the SP. Finally, regions should not be constrained in shape allowing for a natural division of the image into its constituent parts.

In this paper, we propose a strategy to divide the image into its constituent *scene geometry* parts to obtain an *invariant* and *descriptive* image representation. The aim is to split the image into sub-regions corresponding to generic scene (depth) layers. These layers provide a middle ground between low-level features and high-level object categories. A number of methods have been proposed to estimate the rough scene geometry from single images [11], [12], [13]. We use the scheme which derives scene information for a wider range of generic scene categories by using *stages* [7]. Stages are defined as a set of prototypes of often recurring scene configurations. They can be seen as discrete classes of scene geometries. Typical classes of discrete 3D *scene geometries* include single-side backgrounds (e.g. walls and buildings) or three sides (e.g. corridor and narrow streets). A number of stage models are shown in Figure 2. These models are dependent on the inherent geometrical structure of images. In this paper, 13 different stages are used excluding *noDepth or tab+pers+bkg,* as these stages are specific characteristics of the data set used in [7].

As shown in Figure 2, the scene structures of the stage models are shown in different colors. The stage models are used to determine how the image is divided in subregions. For instance, images of stage *sky+backgnd+gnd* are divided into three layers: sky (in blue), background (in yellow) and ground (in brown). In Fig. 3(a), it is shown that the example images from Fig. 1 are instantiations of the *"sky+backgnd+gnd"*, *"person"* and *"gnd+diagonal"* stages, respectively. Each scene (depth) layer is equivalent to an image segment. Hence, SPs are constructed based on 3D scene geometries in which each geometry layer (e.g., ground, background and sky) is represented by a different sub-region.

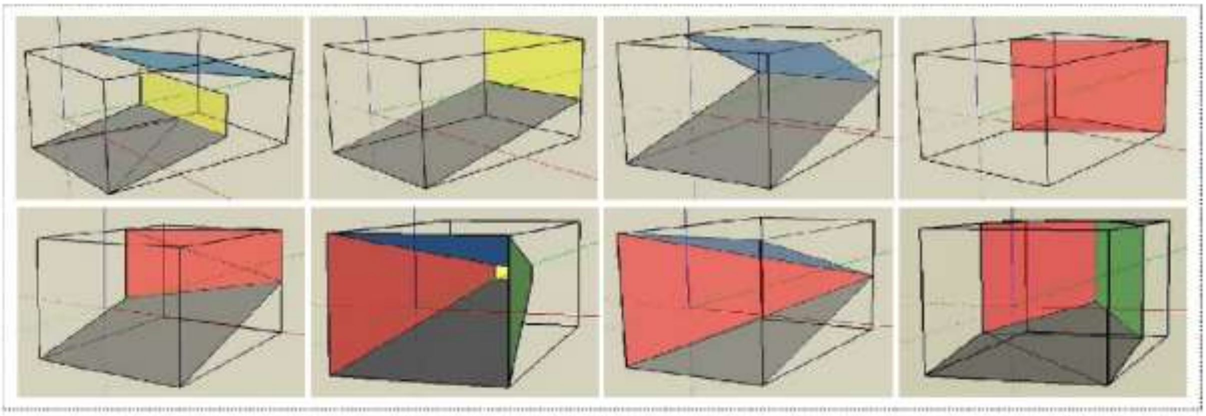
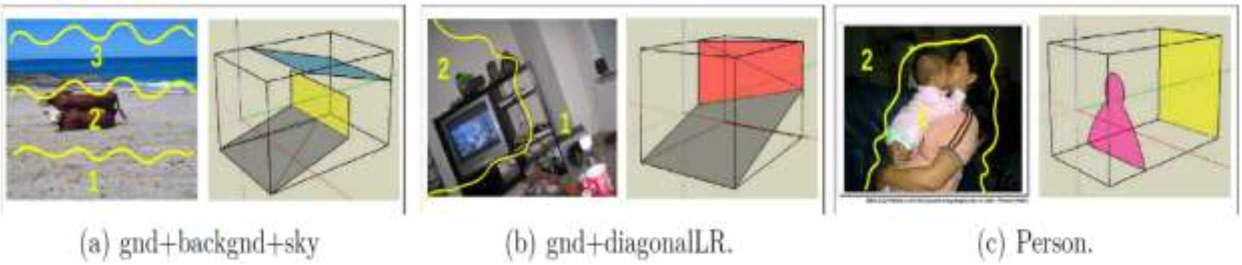


Figure 2: Stage models and their corresponding instantiations. Top row, from left to right: "sky+backgnd+gnd", "backgnd+gnd", "sky + gnd", "gnd + diagalBackgndLR". Bottom row: "diagalBackgndLR", "box", "1side-wallLR", "corner". This figure is taken from [14].



(a) gnd+backgnd+sky

(b) gnd+diagonalLR.

(c) Person.

Figure 3: Example depth images with their corresponding 3D geometries

3. PRELIMINARIES

In this section, we briefly discuss the method for obtaining standard spatial pyramids and for determining scene geometries (stages).

3.1 Rigid Spatial Pyramids

The spatial pyramid scheme proposed by [5] is a simple and computationally efficient extension of an order-less bag-of-words image representation. This approach represents an image by using weighted multi-resolution histograms which are obtained by repeatedly sub-dividing an image into increasingly finer sub-regions. Histograms are computed over the resulting sub-regions. For each resolution level, the image is subdivided into the cells of a grid. At resolution *l*, the grid has 2^{2l} cells. The number of points in each grid cell is then recorded.

Marszalek et al.[15] evaluate both regular and irregular grids. Further, they consider a broader set of coarse subdivisions for each dimension, such as a 1×1 grid corresponding to the standard representation of the bagof-words, a 2×2 grid (i.e. four blocks), a horizontal 3×1 grid as well as a vertical 1×3 one. They show that dividing the image plane in three horizontal (i.e. 3 × 1 grid) regions, provides the highest recognition performance. Further, this approach reduces the dimensionality of the conventional 4 × 4 (i.e. sixteen blocks) structure; from *vocabularysize* × 21 to *vocabularysize* × 8.

The use of various image layouts shows the influence of the image configuration and spatial image representation. However, unconstrained spatial image representations have not been studied [16]. Moreover, pyramids commonly applied to BoW are not designed for the specific task of categorization, due to the assumption of having fixed rigid grid representation that suits all the dataset categories. The proposed approach resolves the use of rigid spatial pyramids. The aim is to generate more natural spatial pyramids based on the underlying image geometry.

3.2 Image Segmentation to Obtain Depth Layers

For each scene geometry, the different image segments correspond to a scene part at a certain depth (layer). Each segment represents geometrical entities like walls, ground, and sky. The image divisions provided by the scene geometry models will be used to learn the best geometry that suits each category of concern. Segmentation is based on the occurrence probability in the training set. Ground truth is obtained by manual annotation, thereby dividing the training set according to the scene geometry patterns, and fitting the parameters of each geometry model (horizon, vanishing points) such as to visually best fit the underlying data. For this purpose, the stages data set described in section 5.1 will be used for obtaining the segmentation masks used to represent each scene geometry.

More precisely, suppose that an image belongs to stage S, which is composed of N layers, correspondingly there will be N mask maps. The mask map for the i^{th} partition T_i is obtained by taking the average of the mask maps for each image:

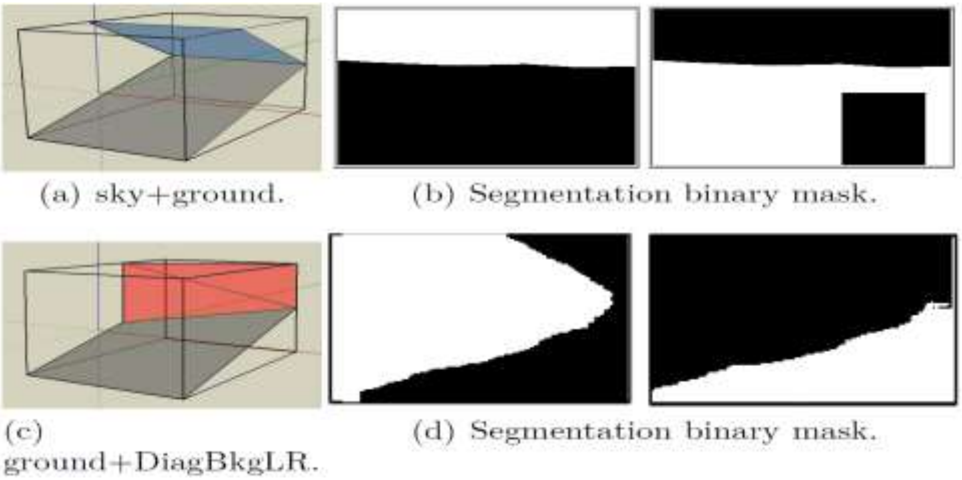
$$T_i(x) = \frac{\sum_{j=1}^{n} M_{j,i}(x)}{n}, \qquad (1)$$

where n is the total number of images in the training data set, and $M_{j,i}(x)$ is the mask map of the j^{th} image for i^{th} partition. Note that $M_{j,i}(x)$ is an indicator function: $M_{j,i}(x)$ =1, if $\boldsymbol{x}$ belongs to the i^{th} partition and 0 otherwise.

Segmentation: mask maps are used to automatically divide the images. Assuming that the images of a stage can be partitioned into *N* layers, there exist *N* mask maps corresponding to the partitions in the training data set. Then, the binary mask map is defined as follows:

$$\mathrm{T'}_i(\mathbf{x}) = \begin{cases} 1, & \mathrm{T}_i(\mathbf{x}) = \max_{\mathbf{j}=1}^{\mathbf{N}} \mathrm{T}_{\mathbf{j}}(\mathbf{x}), \\ 0, & otherwise. \end{cases} \qquad (2)$$

As a consequence, the values in the mask map are either 0 or 1, as shown in Figure 4. In the next section, scene geometry (i.e. scene depth) maps will be used in order to achieve a proper selection of the spatial partitionings that suit each object category of concern.



(a) sky+ground.
(b) Segmentation binary mask.
(c) ground+DiagBkgLR.
(d) Segmentation binary mask.

Figure 4: An example of segmentation binary mask maps. Top row image belongs to *sky+ground* scene geometry. Bottom row image belongs to *ground+DiagBkgLR* scene geometry. The mask maps are of the same size as the original image.

4. SPATIAL LAYOUTS DERIVED FROM 3D SCENE GEOMETRY

The SP scheme proposed by [5] represents an image by using weighted multi-resolution histograms which are obtained by repeatedly sub-dividing an image into increasingly finer sub-regions, where the spatial pyramid at level $l \in \{0, \dots, L\}$ has $R(l) = 2^{2l}$ sub-regions. For image X, all features are assigned to their best visual word v selected from a vocabulary V. The frequency of v inside sub-region i of image X is given by the histogram bin $H_X^i(v)$. The similarity or matching rate between images X and Y at level l, is given by the histogram intersection function [17]:

$$I^l(X,Y) = \sum_{i=1}^{R(l)} \sum_{v=1}^{|V|} \min(H_X^i(v), H_Y^i(v)). \qquad (3)$$

Matches found at finer resolutions are closer to each other in the image space and are therefore more heavily weighted. To accomplish this, each level is weighted to $^1/_2\, L - l$ which results in the final SP:

$$\kappa^L(X,Y) = \frac{1}{2^L} I^0(X,Y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l(X,Y). \qquad (4)$$

For the geometry-driven pyramid we use the same approach, only for geometric equivalent classes. Formally, for each scene geometry or stage s of n stage types, let $R(s)$ denote the number of sub-regions of s. Instead of using a fixed pyramid, we propose that a spatial stage pyramid is created by computing the similarity between images X and Y for stage s by

$$I^s(X,Y) = \sum_{i_s=1}^{R(s)} \sum_{v=1}^{|V|} \min(H_X^{i_s}(v), H_Y^{i_s}(v)). \qquad (5)$$

Where the different sub-regions i_s for stage s correspond to a scene part at a certain depth (layer). For each stage s there are $R(s)$ sub-regions.

We propose two alternative approaches for selecting the appropriate spatial image representation for each category. This is achieved by learning a proper class-specific spatial model. These models encode the proper spatial partitioning for each category; which is used further to obtain class-specific spatial models. To this end, we first exploit the use of the standard 3D geometry model as a prior for learning the best candidate template for each category. Then, we introduce an adaptive approach to generate spatial image representation that suits each category based on the ground-truth (GT) information of its training images. Lastly, we propose a learning approach for learning the most suitable category-model based on information theory.

4.1 Generic Spatial Pyramids sets

In this section, the 13 different scene geometries $\{S1, \dots, S13\}$ proposed by [7] are used. It has been shown that these geometries cover most of the image partitionings encountered in real-world scenarios. Hence, 3D geometry structures are used to determine how the image should be divided; where each geometry depth corresponds to a pyramid region i_s. Therefore, the 13 prior stages are considered as generic models. Images of each category are classified into one of these stages in order to select the most appropriate spatial representation. More formally, the proposed method consists of the following steps: first, training images are spatially represented according to each of the 13 binary mask maps. We use the stage data set to obtain the binary mask maps. The training set is manually annotated and divided into scene geometries as in [18]. The parameters of each geometry (horizon, vanishing points) are computed to fit the underlying data. Image segmentation is based on the occurrence probability in the training set, see [18] for details. For each category, we then train 13 geometry models and learn on the validation set which geometry or even the combination of geometries that best suits each category. The whole process is demonstrated by the block diagram in Figure 5.

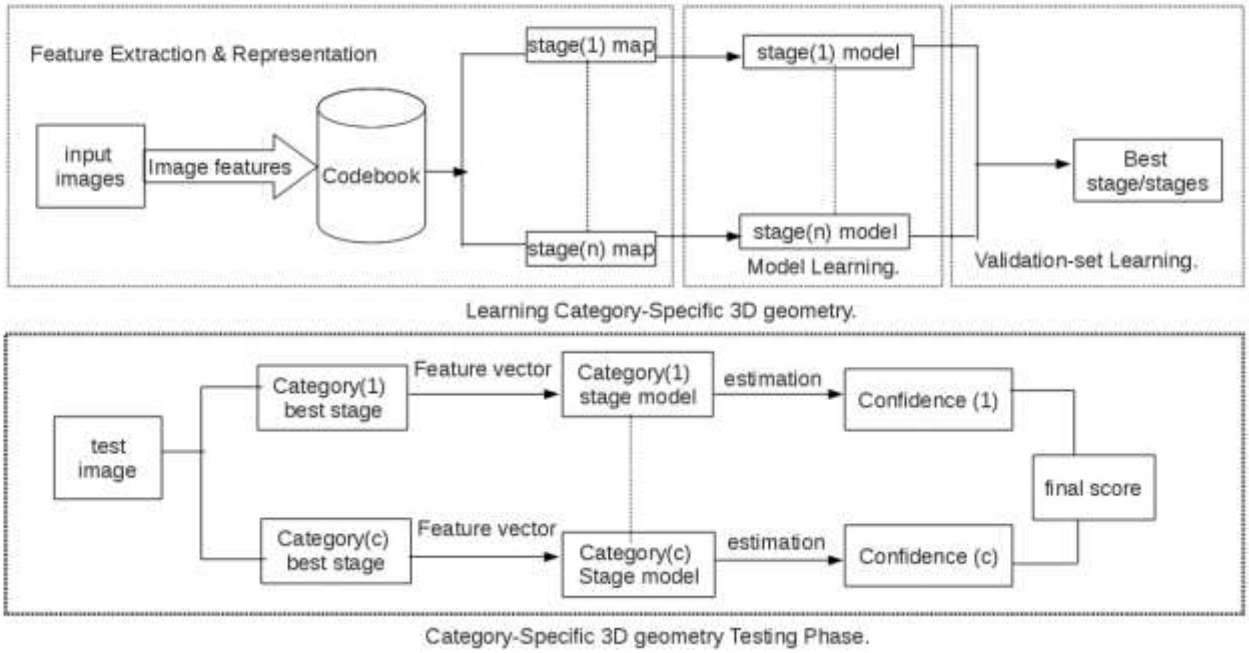
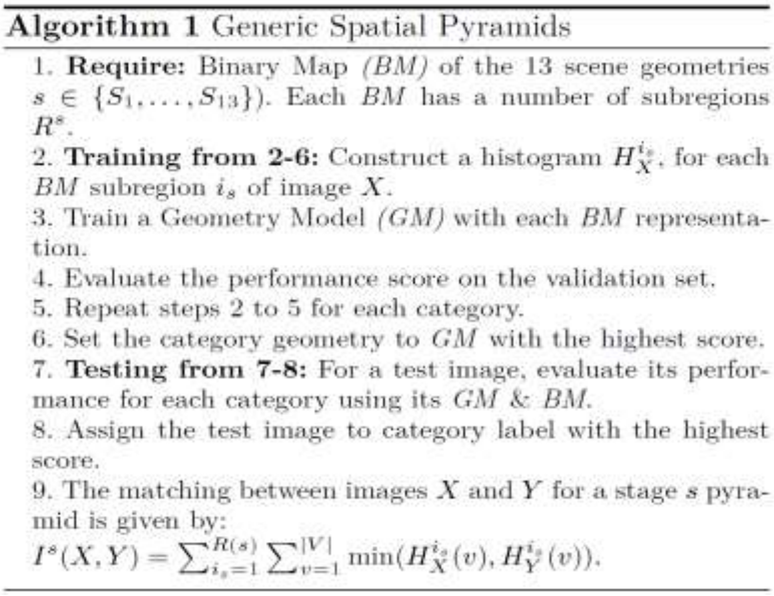


Figure 5: Outline of the spatial representation using 3D scene geometry. Note that the codebook models and the stage models are obtained off-line. For each category, the proper stage model is obtained.

For a new image, it is represented using the 13 binary maps (off-line step). We evaluate the learned geometry model of each category w.r.t. its appropriate representation. Consequently, the test image will have a score towards each category and is assigned to the category with the highest score. We summarize the whole procedure in Algorithm 1.



```
Algorithm 1 Generic Spatial Pyramids
1. Require: Binary Map (BM) of the 13 scene geometries $s \in \{S_1, \ldots, S_{13}\}$). Each BM has a number of subregions $R^s$.
2. Training from 2-6: Construct a histogram $H_X^{i_s}$, for each BM subregion $i_s$ of image $X$.
3. Train a Geometry Model (GM) with each BM representation.
4. Evaluate the performance score on the validation set.
5. Repeat steps 2 to 5 for each category.
6. Set the category geometry to GM with the highest score.
7. Testing from 7-8: For a test image, evaluate its performance for each category using its GM & BM.
8. Assign the test image to category label with the highest score.
9. The matching between images X and Y for a stage s pyramid is given by:
$I^s(X,Y) = \sum_{i_s=1}^{R(s)} \sum_{v=1}^{|V|} \min(H_X^{i_s}(v), H_Y^{i_s}(v))$.
```

.

5. EXPERIMENTS

In this section, the proposed methods to generate flexible spatial pyramids will be compared to the existing state-of art rigid based pyramids in the context of object recognition. In Sec. 5.1, the data sets used in all experiments are given. The experimental setup used is shown in Sec. 5.2.

5.1 Data sets

Three independent data sets are used in the experiments. The first data set is a large dataset consisting of 3589 images classified as 15 different categories representing the standard generic scene geometries. 151 “sky+ background + ground”, 333 “background + ground”, 81 “sky + ground”, 212 “ground”, 139 “ground + Diag-BkgLR”, 132 “ground + DiagBkgRL”, 75 “diagBkgLR”, 71 “diagBkgRL”, 84 “box”, 57 “1sidewallLR”, 69 “1sidewallRL”, 266 “corner”, 960 “persBkg”, 833 “noDepth”, and 126 “tabPersBkg”. Images are take are under large variety of lighting conditions and imaging conditions (including indoor, outdoor, desert, cityscape, and other settings). We refer to this data set as “stages data set”. This data set is used to generate the binary mask maps (Sec. 3.2) used by the *Generic Spatial Pyramids* approach. Some example of images that are in this dataset are shown in Figure 6.

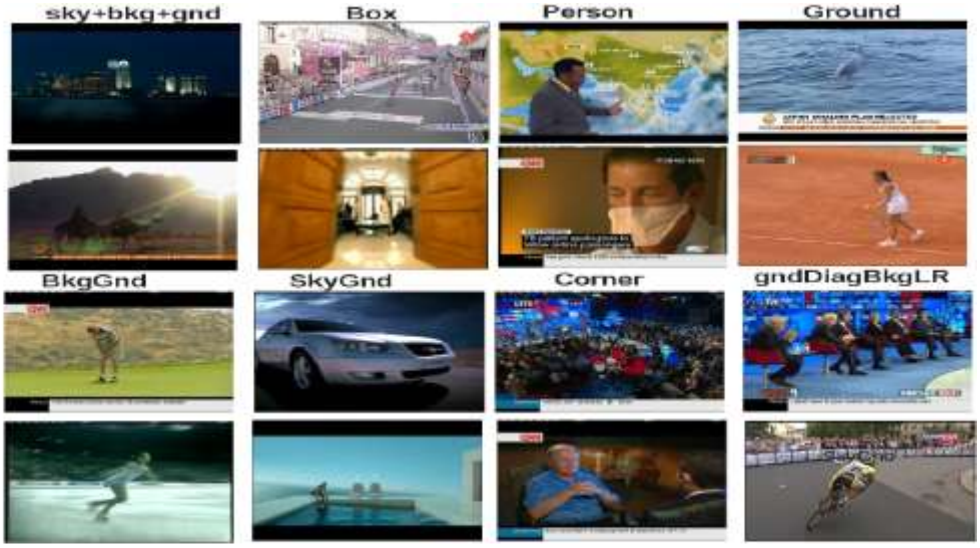


Figure 6: Example images of stages data set

We also use Caltech-101 [10] and Pascal 2007 [6] datasets as benchmark data sets for evaluating our approach. The Pascal VOC 2007 data set [6] which consists of 9963 images of 20 different classes with 5011training images and 4952 testing images. The Caltech-101 data set which contains 9144 images of 102 different categories. Some example of images that are in Caltech and Pascal data sets are shown in Figure 7(a) and Figure 7(b), respectively.

Figure 7: Example images of Caltech and Pascal data sets

5.2 Experimental Setup

To compare the different spatial pyramids, a standard BoW image classification approach is used. SIFT features [20] of 16 × 16 pixel patches are used.

For Caltech-101, we use 30 images per category for the training and 50 for testing. The general architecture follows [5]. The SIFT descriptors are extracted on a dense grid rather than interest points, as this procedure has been shown to yield superior performance for scene classification [3]. We use a codebook of size 300. Experiments are conducted over 10 random splits of the data, and the average per-class recognition rates are recorded for each run. The final result is reported as the mean accuracy and its standard deviation from the individual runs.

For Pascal 2007, a standard multi-scale grid detector is taken together with a Harris-Laplace point detector [2], and a blob detector. SIFT descriptors are computed for all regions in the feature descriptor step, which are then quantized to a codebook of size 1000 visual-words. We used the standard average precision (AP) criteria to validate the performance on the object recognition task. The average precision is equivalent to the area under a precision-recall curve. Mean average precision (MAP) is used to evaluate the performance of the features over all the data set categories, which is obtained by averaging the AP over all categories. In our experiment, we compare our method with the standard three-levels Spatial Pyramid SP proposed by Lazebnik et al.[5] as the baseline.

5.3 Generic Spatial Pyramids (Generic SP)

In this section, the “Generic SP” approach proposed in Sec. 4.1 is evaluated using 3D scene geometries. For each category, the geometry with the highest score is selected for representing it. In table 1, we show the obtained scores compared with the standard SP. It is demonstrated that the Generic SP improves the results by 8.5% and 9.0% (relative to the baseline) on the Pascal and Caltech data sets, respectively. We attribute this to, (i) the generic 3D geometries contains a wide range of spatial partitionings which cover most of the “real-world” object categories; (ii) the obtained representations of our approach, are tailored for each category, and therefore can efficiently capture the variabilities that exists within each category.

Table 1: Results obtained on Pascal (MAP score) and Caltech (Average per-class recognition rates) data sets using Generic SP. The proposed approach improves the scores over the standard SP proposed by Lazebnik et al.[5]. The best geometric split-up learned over multiple kernels ***MKL*** improves the overall performance significantly (see text).

Method	Pascal	Caltech
Lazebnik et al. [5], linear kernel	48.3	51.4 ± 0.9
Lazebnik et al. [5], intersection kernel	51.5	64.6 ± 0.8
Generic SP, linear kernel	52.5 (+8.7%)	55.5 ± 1.6 (+8.0%)
Generic SP, intersection kernel	**55.9** (+8.5%)	$\mathbf{70.5 \pm 1.3}$ (+9.0%)
Generic SP + MKL	**59.7** (+16.0%)	$\mathbf{77.3 \pm 1.1}$ (+19.7%)

In Figure 8, we show some example images for various data set categories together with their most appropriate 3D scene geometries. These quantitative results illustrate that different 3D scene geometries are selected for representing the various data set categories. For instance, the plane category is instantiated from the sky+gnd scene geometry. While, the bird category is instantiated from the box geometry. Hence, using 3D scene geometries is important for efficiently capturing the spatial layout per category.

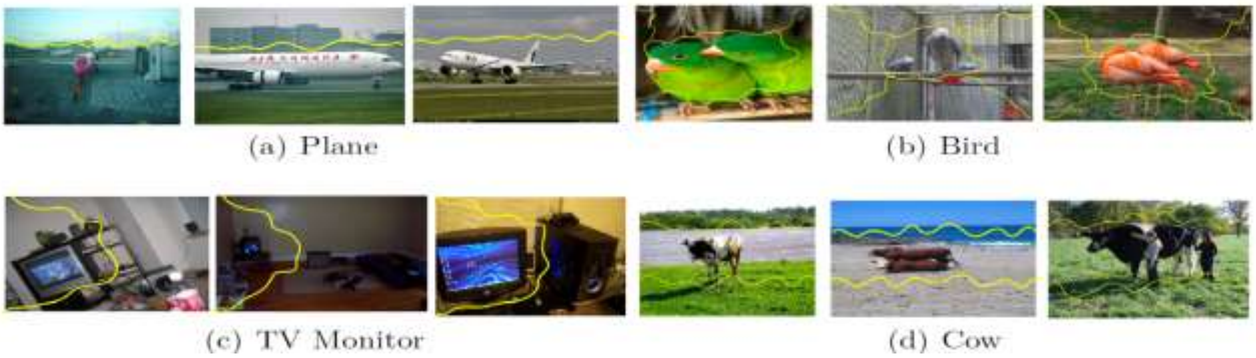

(a) Plane
(b) Bird
(c) TV Monitor
(d) Cow

Figure 8: Pascal data set examples with their learned geometries. Plane instantiated from sky+gnd, Bird from *box* geometry, TV monitor from *gnd+DiagBkgLR* and Cow from *sky+backgnd+gnd*

Another advantage of the "Generic SP" scheme is its ability of reducing the dimensionality of the generated histograms. The maximum number of partitions that exist for representing a category is 6 (i.e., "box geometry"+ BoW). This leads to a final representation of size 6×|V |, where |V | is the size of the vocabulary. On the other hand, the number of partitions of the standard "SP" is 21. This leads to a final representation of size 21 × |V |. Finally, we investigate the use of "Multiple Kernels Learning (MKL)" proposed by Gehler et al. [21], for the selection of the most appropriate geometry or the combination of geometries per category among multiple kernels. The results, in table 1, demonstrate the importance of using multiple kernels for our approach. This improves the performance significantly by 16.0% and 19.7% (relative to the baseline) on Pascal and Caltech data sets, respectively.

5.3 Comparison with State of the Art

In Figure 9, we compare the performance of our approach to the state-of-art SP using various vocabulary sizes (i.e., 1k, 2k, 4k and 6k) on the Pascal data set. The results show that our approach outperforms the standard state-of-art SP over all the examined vocabularies. Similar results is obtained on Caltech datasets under various sizes. Moreover, our results confirm the experimental findings of the work of [21], where the use of MKL improves the overall performance for various vocabulary sizes.

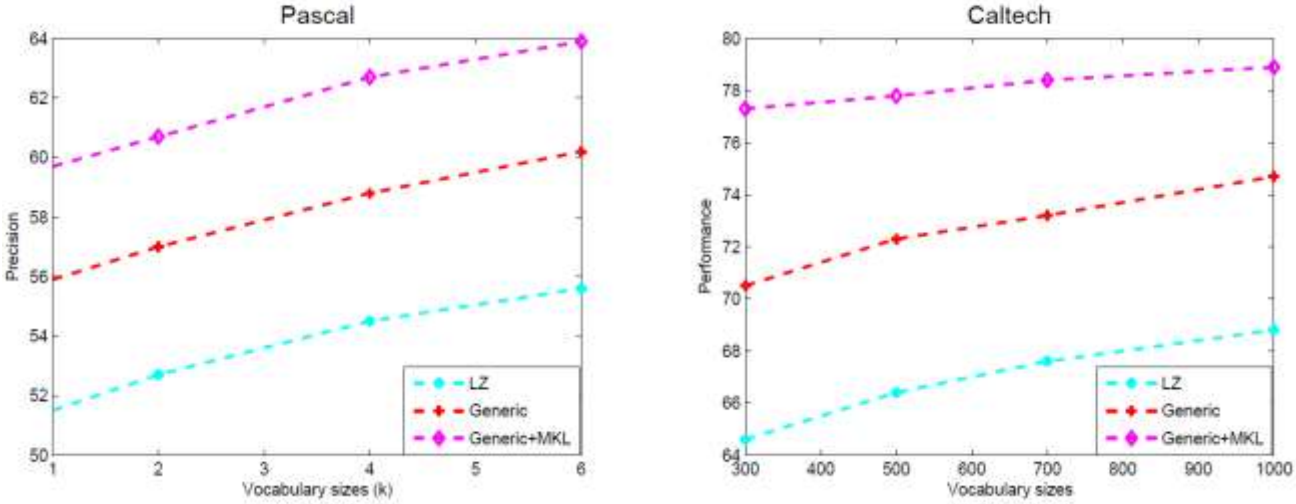


Figure 9: Comparison between our *Generic SP* and *Generic SP + MKL* approaches with standard SP (denoted as *LZ*) using the SIFT descriptor under different vocabulary sizes on Pascal (left) and Caltech (right) data sets, (see text).

6. CONCLUSION

Spatial Pyramids have been proposed which are steered by the 3D scene geometry. The geometry of a scene is measured based on image statistics taken from a single image. After the estimation of the scene geometry, the corresponding SP is selected as the geometrical representation. From large scale experiments on the Pascal VOC2007 and Caltech101, it can be derived that Generic SPs outperforms the standard SPs with 8.5% and 9.0% for Pascal VOC 2007 and Caltech101 respectively.

For future work, the proposed system will be extended to automatically learn a hierarchal class-specific adaptive shape model, where the highest levels will incorporate the important localization and/or segmentation knowledge for efficiently capturing the *ROI*, for restricting the objects spatial location to work with.

REFERENCES

[1] Dance, L. Fan, J. Willamowski, C. Bray.,Visual categorization with bags of keypoints., in: ECCV Workshop on Statistical Learning in Computer Vision.,2004.

[2] K. Mikolajczyk, C. Schmid., A performance evaluation of local descriptors, TPAMI 27 (10) (2005) 1615-1630.

[3] L. Fei-Fei, P. Perona., A bayesian hierarchical model for learning natural scene categories, in: CVPR, 2005.

[4] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: An in-depth study. A comprehensive study, IJCV 73 (2) (2007) 213–218.

[5] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006.

[6] M. Everingham, L. V. Gool, C. K. I.Williams, J.Winn,A. Zisserman, The pascal visual object classes challenge 2007 results. (2007).

[7] V. Nedovic, A. W. M. Smeulders, A. Redert, J.-M. Geusebroek.,Stages as models of scene geometry., IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 32 (9) (2010) 1673–1687.

[8] N. Slonim, N. Tishby, Agglomerative information bottleneck,in: NIPS, 1999.

[9] B. Fulkerson, A. Vedaldi, S. Soatto, Localizing objects with smart dictionaries, in: ECCV, 2008.

[10] L. Fei-Fei, R. Fergus, P. Perona., Learning generative visual models from few training examples., in: CVPR Workshop GMBV, 2004.

[11] D. Hoiem, A. A. Efros, M. Hebert., Geometric context from a single image., in: ICCV, 2005, pp. 654–661.

[12] E. Delage, H. Lee, A. Y. Ng., A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image., in: CVPR, 2006, pp. 2418–2428.

[13] E. Sudderth,Torralba, W. Freeman, A. Willsky., Depth from familiar objects: A hierarchical model for 3d scenes.,in: CVPR, 2006., pp. 2410–2417.

[14] V. Nedovic, A. Smeulders, A. Redert, J.-M. Geusebroek., Depth information by stage classification., in: ICCV, 2007.

[15] M. Marszalek, C. Schmid, H. Harzallah, J. van de Weijer, Learning object representation for visual object class recognition, in: Visual recognition Challenge Workshop, ICCV, 2007.

[16] J. van Gemert, Exploiting photographic style for category-level image classification by generalizing the spatial pyramid., in: ICMR, 2011.

[17] K. Grauman, T. Darrell., The pyramid match kernel: Discriminative classification with sets of image features., in: ICCV, 2005.

[18] L. Rui, A. Gijsenij, T. Gevers, V. Nedovic, X. De, J. Geusebroek., Color constancy using 3d scene geometry., in: ICCV, 2009.

[19] A. P. Moore, S. J. D. Prince, J. Warrell, U. Mohammed, G. Jones, Superpixel lattices, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[20] D. Lowe., Distinctive image features from scale invariant keypoints, IJCV 60 (2) (2004) 91–110.

[21] P. V. Gehler, S. Nowozin., On feature combination for multiclass object classification, in: ICCV, 2009.

[22] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, Largescale scene recognition from abbey to zoo, in: CVPR, 2010.

[23] F. Khan, J. van de weijer, M. Vanrell, Top-down color attention for object recognition, in: ICCV, 2009.

[24] Y. Su, F. Jurie, Visual word disambiguation by semantic contexts, in: ICCV, 2011.

[25] G. Sharma, F.Jurie, Learning discriminative spatial representation for image classification, in: British Machine Vision Conference (BMVC), 2011.

[26] Y. Boureau, F. Bach, Y. LeCun, J. Ponce., Learning midlevel features for recognition., in: CVPR, 2010.

[27] H. Zhang, A. C. Berg, M. Maire, J. Malik., Svm-knn: Discriminative nearest neighbor classification for visual category recognition., in: CVPR, 2006.

[28] O. Boiman, I. Rehovot, E. Shechtman, M. Irani., In defense of nearest-neighbor based image classification., in: CVPR, 2008.

[29] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: CVPR, 2009.

[30] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating color descriptors for object and scene recognition, TPAMI 32 (9) (2010) 1582–1596.