International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 5, Issue 1, Dec- Jan 2016, 19-26

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

BAG OF VISUAL WORDS

PRERNA RAI[1] & SANJOY GHATAK[2]

[1]Department of Computer Science Engineering, SIKKIM, India

[2]Assistant Professor, Department of Computer Science Engineering, SMIT, SIKKIM, India

ABSTRACT

In this document, the image classification techniques and their steps have been put up using Bag of words which is extended to an image, making it bag of features or bag of visual words. It also focuses on

the different techniques under Bag of Visual words. A comparative study also been done on Bag of visual words and its two different techniques. Spatial Pyramid Matching Scheme which encodes local spatial information(SPM) and geometry preserving visual phrase(GVP) which encode local as well as long range spatial information. The comparative study has found that Spatial pyramid Matching (SPM) and Geometric Preserving Visual Phrase techniques are introduced to improvise upon the basic BoV representation by incorporating local as well as long range spatial information.

KEYWORDS: Mean average, Transformation, Invariance, Image

INTRODUCTION

An **image[9]** is an artifact that depicts or records visual perception, for example a two-dimensional **picture**, that has a similar appearance to some subject—usually a physical object. Inorder to classify or categorize images a technology named as Bag of Visual words is being used in today's day. The Bag of visual words has been taken from the concept "Bag of words" which has its subsequent origin as texture recognition. Texture is characterized by the repetition of basic elements also called as textons, in terms of words and pixel or picture element in terms of visual word.

The **Bag-of-words model** finds its application in natural language processing and context based information retrieval (CIR). In this model, a text (such as a sentence or a document) is represented as the bag of its words, where the grammar and word order are ignored but keeping multiplicity. It is used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

Bag-of-Words model (BoW model) can be applied to image classification, by treating image features as words and thus can be renamed as Bags of Visual Words (BoV).

Bag of Visual words [3][20]approach uses the local appearance information from the image and discards all spatial and geometry information that is available in the images. This information can be available using two other techniques Spatial Pyramid Matching scheme and Geometry preserving visual phrase.

Therefore, this representation of the image is advantageous in computational complexity and invariance within category.

RELATED APPROACH

To categorize an image using Bag of visual words model, an image can be treated as a document. Similarly,

"words" in images need to be defined too.

To achieve this, it usually includes following three steps:

- Feature Detection,
- Feature Description, and
- Codebook generation.

Feature Detection

The first step to image categorization is feature detection. It is the methods for finding parts of an image which is relevant to classifying image.

Feature Representation

After feature detection, each image parts are represented by several local patches. Feature representation method deals with the representation of the patches as numerical vectors. These numerical vectors are called feature descriptors. A good descriptor should have the ability to handle intensity, rotation, scale and affine variations to some extent. One of the most famous descriptors is Scale-invariant feature transform (SIFT).

Codebook/Visualword Generation

The final step for the Bag of visual words model is to convert vector represented patches or feature descriptors to "codewords", which further produces a "codebook" or visualword. A visualword can be considered as a representative of many similar patches. One simple method is performing k-means clustering over all the vectors. Visualwords are then defined as the centers of the learned clusters.

Thus, each patch in an image is mapped to a certain visualword through the clustering process and the hence the histogram is computed using visual words.

Advantages

- BoV is orderless, as it is not affected by position and orientation of object in image.
- It has fixed length vector irrespective of number of detections.
- It is Simple and Efficiency.

Disadvantages

- No explicit use of configuration of visual word positions
- It cannot localize objects within an image
- Does any consider any geometrical information.
- Does not consider spatial layout of the features in the image.

To overcome the problems associated with Bag of Visual words there are many other Techniques used for object Recognition using Bag of Visual Words.

They are Spatial Pyramid Matching, Geometry Pyramid Matching, Clustering Algorithm, Support Vector Machine, etc.

Here in our study we would be focusing on Spatial Pyramid Matching and Geometry Preserving Visual Phrase.

SPATIAL PYRAMID MATCHING

Bov model as discussed above does not include any spatial layout information and thus cannot take advantage of the regularities in image composition and the spatial arrangement of the features, which can make very powerful method for scene classification task.

[1] Lazebnik, et al. introduced Spatial-Pyramid Matching (SPM) which encodes spatial information based on a modification of pyramid match kernels [3](Grauman and Darrell,

2005). [1] This method, follows "subdivide and disorder" strategy. It works by repeatedly subdividing an image into levels and computing histograms of image features over the resulting subregions and hence carry out histogram matching.

The method had initially been created for recognizing scenes such as highway, office, street, forest etc.

[7]The SPM model is inspired by the intuition that people can recognize scenes while overlooking various details and thus perceive scenes in a holistic way. Thus scenes may be recognized or classified based on the spatial layout of the image while neglecting the details.

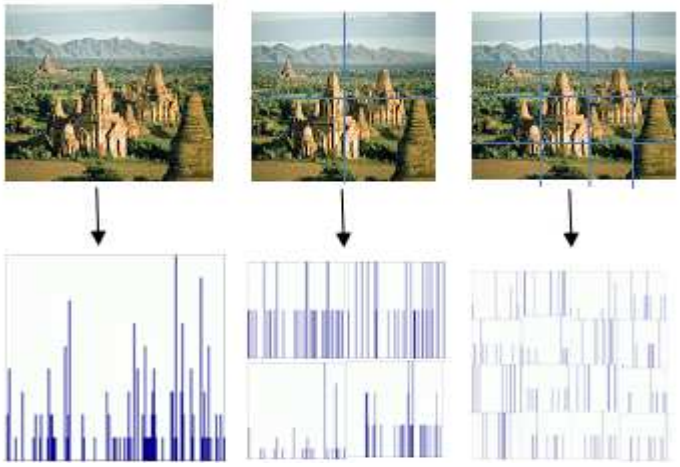
Spatial pyramid matching scheme finds its application in efficient scene recognition in large datasets, as well as for capturing contextual information.

[1]This strategy has been practiced numerous times in computer vision, for global image description.

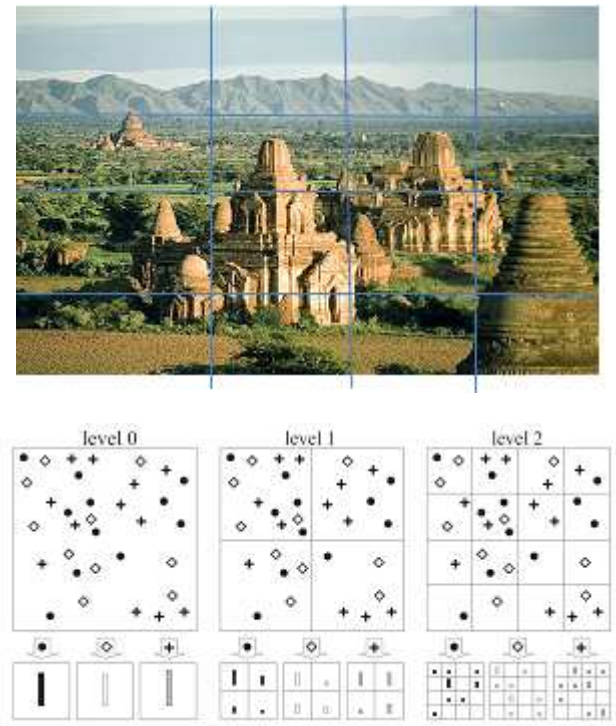
As per the study in [1][8]Spatial pyramid formation:

- Partition the image recursively.
- Accumulate visual word counts separately

By performing spatial partitioning and taking histograms for each of the level.



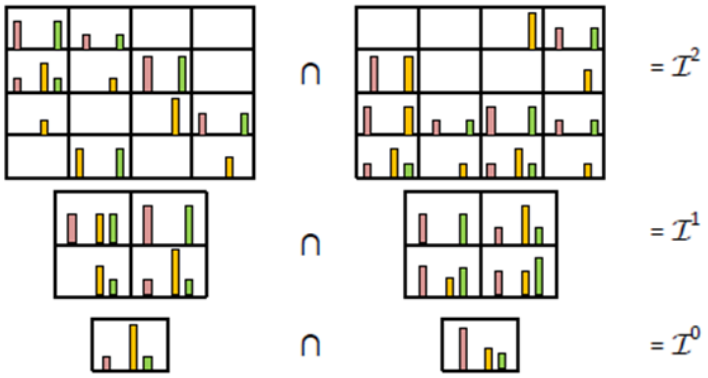
[1][8]Figure 1: Subdivide and Disorder



[1]**Figure 2: Subdivide and Disorder**

Once spatial pyramids have been formed, matching can be performed following a strategy inspired by pyramid matching kernel [Grauman &Darrell].[3]

In this case at each level of the pyramid the number of elements is the same.



[8]**Figure 3: Pyramid Formation**

At each level spatial configuration detail importance is increased. Matches are only counted once and Level 0 is similar to standard Bag-of-Words

This approach has been refined as follows:

quantize all feature vectors into *M* discrete types, and make the simplifying assumption that only features of the same type can be matched to one another. Each channel *m* gives us two sets of two-dimensional vectors, *Xm* and *Ym*, representing the coordinates of features of type *m* found in the respective images. The final kernel is then the sum of the separate channel kernels:

$$K^L(X, Y) = \sum_{m=1}^{M} k^L(Xm, Ym) \tag{1}$$

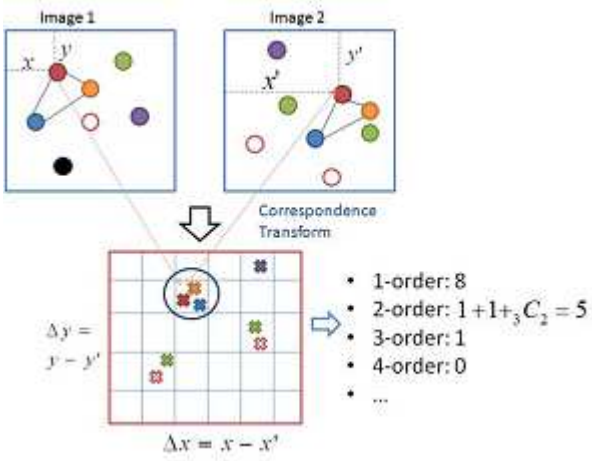
Where K^L is a single histogram intersection of long vectors formed by concatenating the appropriately weighted histograms of all channels at all resoulution for L levels and M channels.

$$M\sum_{l=0}^{L} 4^l a^{n-k} = M\frac{1}{3}(4^L + 1 - 1) \tag{2}$$

The practical success of spatial pyramids observed in experiments[1] states that locally orderless matching may be a powerful mechanism for estimating overall perceptual similarity between images.

Geometry Preserving Visual Phrase

The other BoV technique used to classify object is Geometry preserving visual phrase. This model finds the co-occurrences of visual words to form visual phrase along with capturing the local and long-range spatial layouts of the words.



[9]Figure 4: Representing Visual Phrase

Each circle in the top two images are represented as visual word (local feature). Different colors represents different words. Two images are transformed to the offset space (bottom image) in order to find the co-occurrence of high order features. Each cross in the offset space is created by a pair of same words (same color) form the two input images known as visual phrase. The main idea is that when n points have the same location in the offset space, we have a particular co-occurring n order feature.

As cited in [4] A geometry-preserving visual phrase (GVP)of length k is defined as k visual words in a certain spatial layout. Different words and different spatial layouts both define different phrases. An image will be represented as a vector defined with the GVP.

Similar to the BoV model the vector representation V k(I) of an image I is defined as: the histogram of GVP of length k, with the ith component representing the frequency (tf) of phrase pi.

It is proven that the dot product of such vectors of two images equals the total number of co-occurring GVP in

these images.

The algorithm proposed in [ ] is to identify the co-occurring GVP in two images. The algorithm is illustrated in figure 4.1 For each pair of the same word j in images I and I', we calculate their offset (Δxj ;Δyj), which is the location of the word in image I' substracts that in image I. Then a vote is generated on the offset space at (Δxj ;Δyj). On the offset space, k votes locating at the same place correspond to a co-occurring GVP of N length k.

LITERATURE SURVEY

Comparative study on the above mentioned techniques is done based on the following parameters:

- Mean Average Precision: Average Precision (AP), averaged over accurate output of all queries and reported as a single score.
- Performance: This parameter is measured on mean average precision value. Higher the mean average precision value better is the performance.
- Computation Time: This is the amount of time computed for retrieval of classifying image in a large dataset.
- Spatial Information: This is number of pixel values per area.
- Geometric Information: Align and scaling image
- Database Preprocessing: Processing of datastore before analysis.
- Tolerance for Transformation: [9]A geometric transformation is any bijection of a set having some geometric structure to itself or another such set. Specifically, a geometric transformation is a function whose domain and range are sets of points. Tolerance for transformation determines how much an image can tolerate this transformation in the geometry of any image.

TECHNIQUES FOR COMPARISION

Techniques that are being compared are:

- Bag of Visual words
- Spatial Pyramid Matching
- Geometry Pyramid Matching

Here in our study we would be focusing on comparative study on Bag of visual words, Spatial Pyramid Matching and Geometry Preserving Visual Phrase based on matrices mentioned above.

COMPARATIVE STUDY ON BoV, SPM AND GVP

Comparison undertaken as per the Experiments in [2][4][6] conducted for image classification using techniques , Bag of Visual word, Spatial pyramid matching and Geometry preserving visual phrase.

The dataset used for carrying out Experiment are:

- WWW-40K : 40,000 images

- NUS-WIDE
- Oxford 5K :5062 images with more than 16M features
- Flicker 1M :contains 1M images with more than 2 billion features.

Outcome of the study undertaken based on the Experiment in [2][4][6] are as follows:

Table 1: Comparative Outcome

S no	Metrices	BoV	SPM	GVP
1	Spatial Information	It considers individual pixel value	Encodes local range Spatial information	Encodes local and long range spatial information
2	Geometrical Information	It lacks geometrical information	It does not encode geometrical information	It is based on geometrical information.
3	Mean Average precision[2][4][6]	Worst(0.634)	Significant improvement(0.651)	It outperforms both the scheme(0.696)
4	Performance	Worst performance	Significant improvement	Outperforms both
5	Tolerance for Transformation	Variant to transformation	Variant to transformation	Invariant to transformation.
6	Computation time[2][6]	Shorter computing time(73%faster)	Shorter computing time	Longer computation time(73% slower)
7	Database processing[2]	Very Less preprocessing of database	Less preprocessing of database	More preprocessing of database

CONCLUSIONS

The concept “Bag of Visual Words” and the techniques Spatial pyramid matching and geometry preserving visual phrase which when compared considering Experiments as in[2][4][6], it is found that Geometry Preserving Visual Phrase (GVP) algorithm consistently outperforms SPM and Bag of Visual words(BoV). BoV have the lowest precision on average for the dataset and GVP have higher precision. GVP have high performance with respect to retrieval accuracy.

REFERENCES

1. Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories."*Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2006.
2. Qi Zhang Frank Anemaet Leiden University Institute for Advanced Computer science.URL:http://press.liacs.nl/publications/InvariantBagofWordsImageRetrieval.pdf.Dt:3.09.15
3. Grauman, Kristen, and Trevor Darrell. "The pyramid match kernel: Discriminative classification with sets of image features." *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. Vol. 2. IEEE, 2005.
4. Zhang, Yimeng, and Tsuhan Chen. "Efficient kernels for identifying unbounded-order spatial features." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.
5. Semantics-Preserving Bag-of-Words Models and Applications Lei Wu, Steven C. H. Hoi, *Member, IEEE*, and Nenghai Yu IEEE transactions on image processing.

6. Zhang, Yimeng, Zhaoyin Jia, and Tsuhan Chen. "Image retrieval with geometry-preserving visual phrases." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.
7. Li, Xinchao, Martha Larson, and Alan Hanjalic. "Pairwise Geometric Matching for Large-scale Object Retrieval." *IJCV* 60.2 (2004): 91-110.
8. Semantics-Preserving Bag-of-Words Models and Applications Lei Wu, Steven C. H. Hoi, *Member, IEEE*, and Nenghai Yu IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 19, NO. 7, JULY 2010
9. https://en.wikipedia.org/wiki/**Bag-of words**_model_in_computer_vision/dt:9.9.15
10. Extending bag of words with locality information:www.micc.unifi.it/.../wp.../A56SpatialPyramidmatch.pdf.Dt:10.9.15.
11. www.wikipedia.com/dt:3.9.15
12. www.wikipedia.com/dt:10.9.15