Pattern Discovery for Object Categorization

Edmond Zhang, Michael Mayo

Department of Computer Science, The University of Waikato,
Hamilton, New Zealand.
Email: ez1@cs.waikato.ac.nz

Abstract

This paper presents a new approach for the object categorization problem. Our model is based on the successful 'bag of words' approach. However, unlike the original model, image features (keypoints) are not seen as independent and orderless. Instead, our model attempts to discover intermediate representations for each object class. This approach works by partitioning the image into smaller regions then computing the spatial relationships between all of the informative image keypoints in the region. The results show that the inclusion of spatial relationships leads to a measurable increase in performance for two of the most challenging datasets.

Keywords: Image Processing, Keypoints, Recognition, Categorization

1 Introduction

Generic object recognition is a challenging problem in computer vision. In particular, our research focuses on the task of object class categorization – a task that is so natural and effortless for the human visual system, but proven to be difficult for current computer vision algorithms. This is mainly due to variability, and the need to generalize across variations in the appearance of objects belonging to the same class [2][4].

Recently, many approaches using machine learning techniques have been proposed, such as [5]. Broadly speaking, the machine learning approach requires conversion of an image to feature vector. This task of generating feature vectors from images can be done either with global image features or local patch features. The global approach is easy to implement and inexpensive to compute. However, one inherent disadvantage is that it is susceptible to local and global variation (e.g. changes in viewpoints or illumination). Local features on the other hand provide a better foundation to handle local and global variability such as various forms of transformation (e.g. affine, scale, rotation). However, one of the fundamental drawbacks of the local approach is that in order to include spatial relationships between local features, they must be modelled explicitly [7][8]. Spatial relationships between image features are important in the sense that they provide a kind of 'linkage' information between independent object parts. This information informs us how object parts are related to each other, and enables classifiers to better discriminate object categories from each other.

This paper focuses on the problem of finding frequently occurring keypoints and keypoint patterns from images. Our two main contributions are: 1) We argue that spatial relationships between keypoints are worth analysing because their inclusion increases accuracy, compared to when they are not used. 2) Our model enables frequent keypoints and keypoint patterns to be visualised and interpreted, see Figure 1 for some examples.

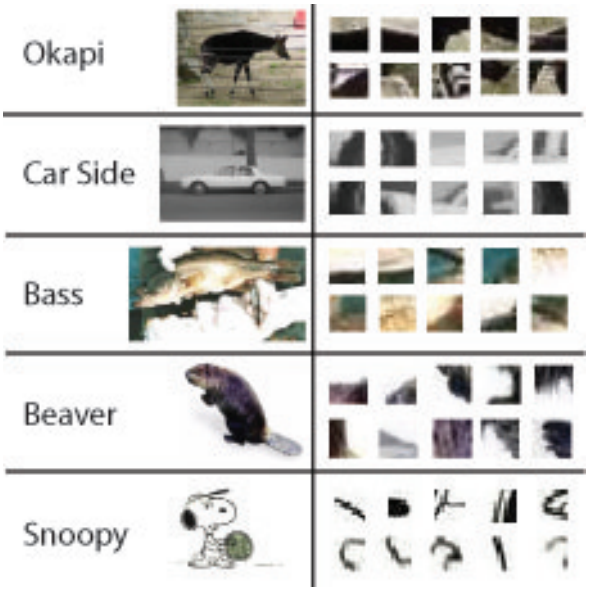

Figure 1: The top 10 patches for their respective classes for the Caltech101 dataset.

In Section 2, we will discuss some of the similar cur-

978-1-4244-2582-2/08/$25.00

rent approaches. Section 3 describes our approach and the reasoning behind it. We then show the results for two of the datasets used for our experiments in Section 4. Results from the experiments will be explained in Section 5. Finally, we finish with conclusion and future work in Section 6.

2 Related work

The 'bag of words' (BOW) approaches have made a great progress in object categorization [4][10][11]. The pioneering BOW approach [2][4] works by representing an image as orderless collection of local features (SIFT keypoints [13]) without any intermediate representation. Intermediate representation has been seen as a 'bridge' in reducing the gap between low-level and high-level image processing, and therefore to better match the object model with human perception.

Because the BOW approach disregards all information about the geometrical layout of the image features, it has limited descriptive abilities. In particular, the BOW models are incapable of separating the object of interest from background clutter and image noise [12]. Recently, in [10] and [11], researchers showed that the inclusion of intermediate representations/themes could improve recognition accuracy significantly. However, the task of discovering intermediate representation has proven to be quite challenging because the recognition system must take into account image noise, background clutter, viewpoint variation, occlusion and image scale changes.

There are two main inherent weaknesses of current BOW approaches. Firstly, the number of features (codebook) extracted from images is often very large – thousands of high dimensional features (128 dimensions for SIFT keypoints) extracted from a single 640 by 480 pixel image is not uncommon. The BOW model also requires a large number of features because they are the 'parts' that made up the object. Too few features will not be sufficient to represent the object, while too many features will introduce too much background noise and image clutter. It is also computationally expensive to compare large numbers of high dimensional features.

The second weakness of the BOW model is that it does not take into account that object parts are dependent on each other. Instead, physical objects are more than just the sum of its parts. For example, the current BOW approach might wrongly classify a motorbike object to a bike object because both objects contain wheel parts. However, in theory, this problem might be managed better if geometrical information that describes how features are related is computed and learned.

3 Our approach

Our work is based on the BOW model. We differ from the previous work in the following areas. 1) A large portion of keypoints extracted from images are not useful, and these mainly consist of background clutter and image noise. In order to filter out these keypoints, we developed a frequent keypoint selection technique, based on frequent itemset mining [1], for discovering frequent and informative keypoints. 2) In order to discover spatially related keypoints to make objects more discriminative, we developed a pattern discovery technique that discovers patterns between frequently occurring keypoints. 3) Finally, we then developed a fast and efficient method of generating low dimensional feature vectors from high dimensional keypoints.

3.1 Frequent keypoint selection

In the original BOW approach [4], the k-means clustering technique was used to identify informative keypoints - that was done by only keeping the largest clusters and discarding smaller ones. In practice, we found that the performance and accuracy tradeoff is not very good. For example, if the number of initial clusters(C) is small (i.e. C = 100), too many unrelated features are grouped into the same cluster. However, if the initial cluster number is large (i.e. C = 10000), the amount of time this technique requires for computing the centroid of the clusters is impractical. That is because at every iteration, the centroid of each cluster must be computed. This is computationally expensive because in the majority of cases, we are working with a large number of high dimensional keypoints.

Our frequent keypoint selection uses a frequent item set mining approach, which is significantly faster because we do not need to cluster the keypoints. Algorithm 1 describes our approach.

Algorithm 1:

1. For each object category, detect and extract keypoints from all training images.
2. Traverse through one keypoint at a time, generating a new frequent itemset for that keypoint if there is no existing frequent itemset for that keypoint
3. If there exist a frequent itemset for that keypoint, increase the weight counter on that frequent itemset.
4. After traversing through all of the keypoints, rank all of the frequent itemsets based on their weight counter and select only the top N number of frequent itemsets.

Our model selects only the top N number (N = 50) of frequent itemsets because frequent keypoints are the most distinctive and informative about a particular object category. Lower ranked keypoints are ignored as they are more likely to be background clutter or image noise. This is because they do not frequently appear through out all the training images. See Figure 2 for an example.

Figure 2: All keypoints are selected for the image on the top. Only frequent keypoints are selected for the image on the bottom.

Because keypoints are high dimensional (128 attributes per keypoint), we cannot simply apply direct comparisons between them. Consequently, two keypoints are treated as the same if they are deemed 'close'. We have tried several different high dimensional number comparison algorithms, and we found that the X^2 distance [3] measure is the fastest and also has the highest recall rate (see equation (1)).

$$d_{X^2}(H,P) = \sum_i \frac{(h_i - m_i)^2}{m_i} \qquad (1)$$

where:

$$m_i = \frac{h_i + p_i}{2}$$

H and P represent keypoint 1 and 2, while h_i and p_i are the bin index for each of the keypoints respectively. If the two features are identical, then the X^2 distance between them is 0. However, the chance of finding two identical keypoint matches is extremely low. Initial testing determined the best value for our datasets at 400. Therefore, if the X^2 distance between two keypoints is under 400, then they are considered to be the same, and are therefore treated as co-occurrences of the same frequently occurring keypoint.

3.2 Spatially related feature discovery

Recall that the original BOW model does not take into account how keypoints are related to each other spatially [4]. We argue that single independent keypoints are not always unique to any object classes. However, keypoint pairs (K = 2) or triplets (K = 3) in theory, are more distinctive and informative to object classes. The value of K determines the size of the keypoint pattern. For our model, K is limited to 1, 2 or 3.

Since we have already produced a set of frequent keypoints from the previous Algorithm, it is natural to generate spatial relationships from those frequent keypoints only. This approach speeded up the process significantly by avoiding generating patterns for background clutter and image noise (lower ranked frequent keypoints). Algorithm 2 shows our approach in generating frequent keypoint patterns.

Algorithm 2:

1. Traverse through all frequent keypoints extracted from each training image.
2. Determine all the frequent keypoints that are within a predefined radius of the currently selected frequent keypoint
3. Generate all unique pairs or triplets (depending on our choice of K) of keypoint patterns from the set of selected keypoints from step 2
4. Keypoint patterns are compared and ranked similar to the single keypoint approach described in Algorithm 1
5. Select only the top N number of patterns from each object class

Once again, we select only the top N number (N = 50) of most frequent patterns from each object class because they are the most distinctive and informative.

We experimented with various different radius sizes ranging from 10 pixels to the entire image; we

found that the radius of 50 pixels is the best for both the datasets.

3.3 Binary feature vector generation

Our method for feature vector generation differs from the original BOW approach [4]. In the original approach, feature vectors are constructed by appending all of the attributes from selected keypoints. This resulted in enormous feature vectors because thousands of keypoints are included and each keypoints contains 128 attributes. The original approach also does not support spatial feature patterns required for our model.

In our model, for each of the object classes, we take the top N number (N = 100) of single keypoints (see Algorithm 1) and the top N number of keypoint patterns (see Algorithm 2), as they are in theory the most informative features that describe that object class. A table is then formed by combining all of the single keypoints and patterns from all of the object classes (see Table 1). Essentially, the table contains all the frequent keypoints and keypoint patterns for every object class in the dataset.

Table 1: Example of binary feature vector generation. Where K stands for keypoint and P stands for pattern.

Image	K1	K2	K3	...	P1	P2	Class
Image 1	45	2	0	...	2	0	Bike
Image 2	1	0	12	...	1	20	Car
Image 3	0	0	14	...	0	13	Car

Once the table is constructed, all of the keypoints from the training images are compared to the table. If a match is found, then the counter for that keypoint of that image on the table is incremented by 1. The same applies to patterns. This is done by first finding a match for one of the keypoint patterns, and then within the predefined radius, our system will try to find the remaining matches for the other keypoints in the pattern. The counter of that pattern of that image will also increment if a match for that pattern is found.

By representing high dimensional keypoints and keypoint patterns with a single number, our approach not only reduces the size of feature vectors, but most importantly, it provides a simple and intuitive way in representing keypoint patterns.

4 Experiments and results

We have experimented with our model on two of the most popular datasets: 15 scenes [12] and Caltech101 [14]. We report the experiment setup and results in this section. All experiments are repeated 10 times with randomly selected training and test images. Multi-class classification is done with the Polynomial Support Vector Machine classifier with default parameters as specified in WEKA V. 3.5.5 [9], except the exponent value was set to 0.5.

4.1 Caltech101

The first dataset we experimented with was the Caltech101 dataset. This is probably one of the most diverse datasets in the research community. Each object class contains between 31 and 800 images. The resolution for most of the images is about 300 by 300 pixels. See Figure 3 for some examples.

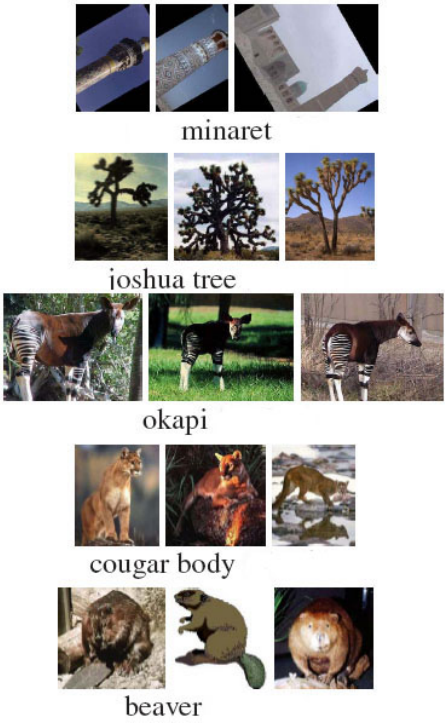

Figure 3: The Caltech101 dataset.

For this dataset, we follow the experimental setup of J. Zhang et al. [15]. Specifically, 5, 15 or 30 images per class are used for training and the rest are flagged as test images.

Figure 4 shows some of the published results for this dataset.

J. Zhang et al. achieved the state-of-the-art for this dataset, which is at 66.2% for 30 training images. Our model obtained 54.5% accuracy with 30 training images. See table 2.

Table 2: Results for the Caltech101 dataset

J. Zhang	K = 1	K = 2	K = 3
66.2*	50.9±1.6	54.5±1.7*	52.3±1.6

Essentially, their method is also based on the BOW approach. Area of interest are detected and de-

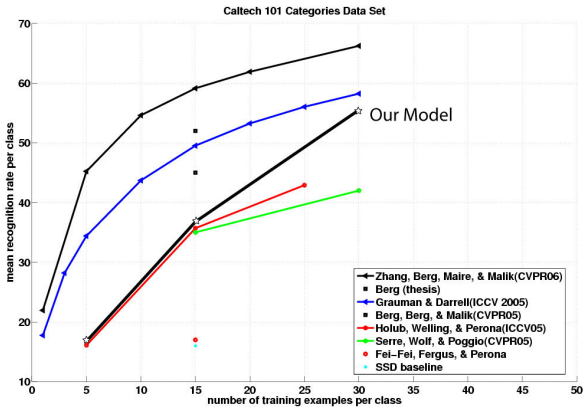

Figure 4: Some of the published results for the Caltech101 dataset [6].

scribed from images using keypoints, while the distribution of local features are used to represent images.

4.2 15 Scene Categories

The 15 scenes dataset contains fifteen categories. Each category contains 200 to 400 images with the average size about 300 by 250 pixels. See Figure 5 for examples.

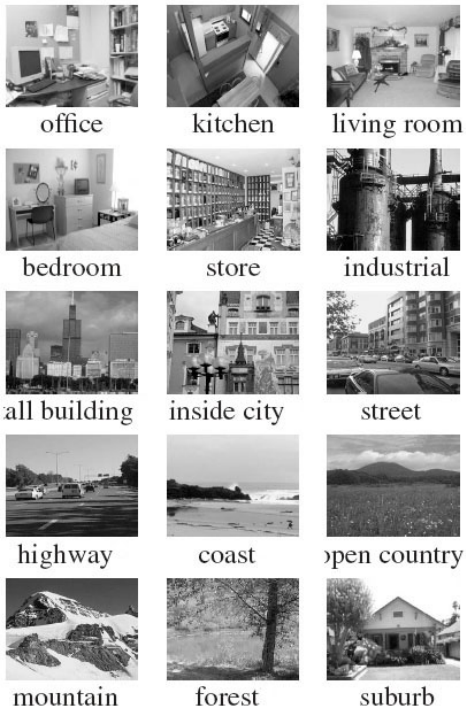

Figure 5: The 15 Scene dataset.

For this dataset, we followed the experimental setup of Lazebnik et al. [12]. That is, for each of the categories, 100 images are randomly selected for training and the remaining images are flagged as test images. Lazebnik et al. achieved the state-of-the-art accuracy for this dataset of 81.4%. The best result we obtained using K = 2 is 59.4%. See table 3.

Table 3: Results for the 15 Scene dataset

Lazebnik	K = 1	K = 2	K = 3
81.4*	52.5±3.0	59.1±2.9*	58.7±2.9

Their approach is based on the spatial pyramid method, which is an effective method that captures both global and local variability. It repeatedly subdivide an image, computing all features repeatedly for all progressively smaller sub-images. The primary advantage of spatial pyramids is that they capture the spatial distribution of features at the finer resolution, while also maintaining the global features that are in themselves highly effective features for classification.

5 Discussion

Both Table 2 and 3 showed the benefits in combining keypoint patterns (i.e. K = 2, 3) compared to standard single keypoint (i.e. K = 1) approach. Results obtained from our current approach are not better than the state-of-the-art results for both datasets. However, we believe it was nevertheless a good first attempt with our spatially related keypoint approach, especially for the Caltech101 dataset.

Our model did not perform well for the 15 Scene Categories dataset, this is because for scene images, global image features are more important in representing the 'gist' of the image, than local keypoints. We believe that by combining both global image features and spatial patterns of keypoints, we can approach the state-of-the-art in the future.

Most importantly, we believe we can improve the accuracy of our model by including more image attributes such as colour information, local binary patterns and statistical information such as mean, median and standard deviation. One advantage with our model is that we know where informative patches are located in images. Therefore instead of extracting colour information and statistically information from the entire image, our approach allows us to extract that useful colour information around informative patches only.

We believe we can improve on the current accuracy because both the state-of-the-art models for the two datasets are using colour information and other important statistical values extensively. Whereas for our current approach, we are only concerned about whether certain keypoints or patterns exist in images.

6 Conclusion and future work

In this paper, we proposed a new BOW approach for image categorization. Our approach is different to the original model in the sense that we introduced a faster way to find the most informative patches from image. We also developed a spatially related keypoint pattern discovery and matching technique. Finally, we developed an efficient method in constructing the feature vector that also supports spatial keypoint patterns.

This paper focuses on the problem of finding frequently occurring keypoints and keypoint patterns from images. Our two main contributions are: 1) We argue that spatial relationships between keypoints are worth analysing because they increase accuracy, compared to when they are not used. 2) Our model enables frequent keypoints and keypoint patterns to be visualised and interpreted.

Results obtained so far are promising, especially for the challenging Caltech101 dataset. Most importantly, we believe we can improve on our current approach by including better base level features that encode colour and texture and get results better than state-of-the-art.

References

[1] Zou, Q., Chu, W., Johnson, D., and Chiu, H. 2002. A pattern decomposition algorithm for data mining of frequent patterns. Knowl. Inf. Syst. 4, 4 (Oct. 2002), 466-482. DOI= http://dx.doi.org/10.1007/s101150200016

[2] Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. Intl. J. of Comp. Vision 43 (2001) 29–44

[3] Rubner, Y., Tomasi, C., and Guibas, L. J. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. Int. J. Comput. Vision 40, 2 (Nov. 2000), 99-121. DOI= http://dx.doi.org/10.1023/A:1026543900054

[4] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In ECCV International Workshop on Statistical Learning in Computer Vision, 2004. 4

[5] Ilkay Ulusoy and Christopher M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 258–265, Washington, DC, USA, 2005. IEEE Computer Society.

[6] SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. Hao Zhang, Alex Berg, Michael Maire, Jitendra Malik. CVPR, 2006. 2126–2136

[7] R. Fergus and P. Perona and A. Zisserman and Ox Pj U. K. Object class recognition by unsupervised scale-invariant learning, In CVPR, 2003. 264–271

[8] D. Keysers, C. Gollan, and H. Ney. Classification of medical images using non-linear distortion models, 2004.

[9] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[10] J. Sivic, B. Rusell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. Technical report, CSAIL, MIT, 2005.

[11] Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning natural scene categories. In CVPR, 2005.

[12] Lazebnik, S., Schmid, C., and Ponce, J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (June 17 - 22, 2006). CVPR. IEEE Computer Society, Washington, DC, 2169-2178. DOI= http://dx.doi.org/10.1109/CVPR.2006.68

[13] Lowe, D. G. Object recognition from local scale-invariant features. Proceedings of the International Conference on Computer Vision (ICVPR'99), 1999.

[14] Wang, G., Zhang, Y., and Fei-Fei, L. 2006. Using Dependent Regions for Object Categorization in a Generative Framework. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (June 17 - 22, 2006). CVPR. IEEE Computer Society, Washington, DC, 1597-1604. DOI= http://dx.doi.org/10.1109/CVPR.2006.324

[15] Zhang, J., Marszaek, M., Lazebnik, S., and Schmid, C. 2007. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. Int. J. Comput. Vision 73, 2 (Jun. 2007), 213-238. DOI= http://dx.doi.org/10.1007/s11263-006-9794-4