# Towards Extensible Automatic Image Annotation with the Bag-of-Words Approach

Robert Nagy
Chair for Data Management
University of Erlangen-Nürnberg
Erlangen, Germany
robert.nagy@cs.fau.de

Klaus Meyer-Wegener
Chair for Data Management
University of Erlangen-Nürnberg
Erlangen, Germany
klaus.meyer-wegener@cs.fau.de

## ABSTRACT

Visual-word-based image categorization has proven to be very effective in several publications and contests. Recently, various approaches have been proposed to address the need for scalability and computational performance of classification based on Bag of Words. Despite these efforts, extensibility still remains an issue. Classifiers and histograms of visual words are both heavily dependent on an immutable general visual vocabulary created during the training step based on training images. Adding a new category that is insufficiently represented by the visual words in the vocabulary requires recreation of the visual vocabulary, complete recomputation of histograms and retraining of classifiers. When adding a new category, current approaches need to fully rebuild the whole recognition system. We address the problem of extensibility by combining class-specific vocabularies with outlier visual words. Classification is achieved by computing a scoring function for each class-specific vocabulary and selecting the highest score value. We show first results of our highly parallelizable and distributable approach on the Caltech 256 dataset.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Image databases*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Experimentation

## Keywords

object recognition, bag of words, extensibility

## 1. INTRODUCTION

In recent years, significant steps have been taken to address the problem of classifying a huge number of categories. [2] estimates that humans are able to readily distinguish 30,000 objects (assuming 10 types for each of the estimated 3,000 categories). Current object recognition approaches still do not scale well enough and are far away from achieving good classification accuracy for such a high number of categories. Besides scalability and computational performance another issue remains open: extensibility.

Many current approaches for object recognition like [3, 18, 28, 31] are based on the popular and very successful Bag-of-Words model. First, points of interest are extracted from training images using sparse or dense sampling and then described by local descriptors like SIFT [16]. For dimensionality reduction, all or a sampled part of the descriptors extracted from all the training images are clustered by $k$-means into a visual vocabulary. Each cluster center resembles a visual word. The resulting vocabulary is used for computing the histogram of visual words for each training image by assigning each descriptor to the closest visual word in the vocabulary. Various discriminative and generative strategies have been proposed for learning the class-label relationships. The most successful approaches use k-Nearest Neighbors ($k$NN) or Support Vector Machines (SVMs) combined with specialized kernels and diverse strategies for estimating the optimal weights.

Our ultimate aim in the Pixtract[1] project is to build an image annotation and search framework as depicted in figure 1. We have evaluated several user studies regarding image search for art [4], history [5], press [1, 7, 25] and web [9, 12, 13] and have concluded that image search is and in the future still will be text-based. Consequently, our idea is to separate image annotation and image search. The latter is implemented based on text annotations and established text indexing methods. For the text-based image search to work well, good text annotations must be provided. From the previously mentioned user studies we collected following requirements for text annotations of images: *people*, *objects*, *locations*, *events* and *actions* present in the image should always be annotated, because they are often searched for. However, *colors*, *shape*, *texture* and *abstract concepts* like emotions or impressions are less often searched for and can be neglected. For annotating objects we use object recognition based on the Bag-of-Words model. Categories are learnt from user-selected groups of images depicting one single category and stored as object identifiers. In the course of time, the annotation system is more than likely to grow. Our current effort in this paper is to establish an easily extensible concept for the object identifiers.

[1]Picture Annotation Extraction

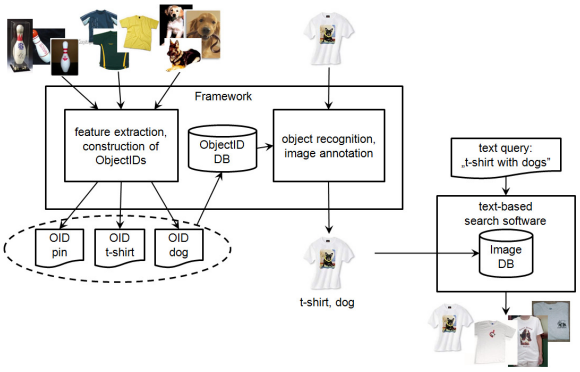



**Figure 1: Image annotation and search framework.**

The major problem regarding the extensibility of Bag-of-Words approaches is the immutability of the visual vocabulary. After creating the vocabulary in the training step, it remains unchanged for ever. Also the histograms of visual words, the classifiers and therefore the classification accuracy are heavily dependent on the suitability of the visual vocabulary. If a new category has to be introduced to the recognition system, the vocabulary needs to be extended. In current approaches, this results in a complete recomputation of the visual vocabulary, the histograms of visual words for all training images and the training of classifiers. Adding a new category currently requires the rebuilding of the whole recognition system.

Recently, hierarchical approaches were proposed for making Bag of Words more scalable and more efficiently computable. These approaches are still not easily extensible, but they guided us to the idea to use several vocabularies instead of one general vocabulary. We address the problem of the extensibility of the Bag-of-Words model by using class-specific vocabularies as introduced in [8, 27]. Instead of merging class-specific vocabularies into one single general vocabulary and computing histograms of visual words and classifiers, we keep our class-specific vocabularies independent and propose a scoring function for computing category membership for images. The proposed method is easily extensible, highly parallelizable and distributable.

The rest of the paper is organized as follows: In the next section we discuss the major drawbacks of the Bag-of-Words model regarding the static vocabulary when dealing with extensible image categorization. In section 3 we give an overview of related work. In section 4 we describe our proposed method in detail. We show first results and discuss our achievements in section 5. We address some further steps that need to be taken towards an extensible image annotation system in section 6.

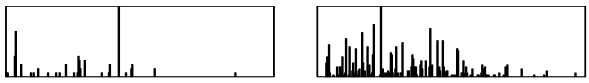
# 2. PROBLEMS WITH BAG OF WORDS

In this section we point at two major problems regarding the visual vocabulary: the optimal choice for the number of visual words ($k$) and the issue of extensibility. Afterwards we discuss possible solutions for the extensibility question.

## 2.1 Issues with the Visual Vocabulary

One of the main problems of Bag-of-Words approaches – addressed in [14, 29] – is the accurate choice of the value $k$ in the creation of the visual vocabulary. Besides reducing the dimensionality, the visual vocabulary also has the task of grouping together noisy versions of the same or very similar descriptors. The bigger the value $k$, the smaller the clusters providing a good precision for the descriptor, but lowering the probability that a noisy version of a descriptor is assigned to the correct visual word. With smaller $k$ values the clusters get bigger assuring that almost all noisy versions of a descriptor are assigned to the same visual word, while reducing the discriminative power of single visual words. Choosing the right $k$ is a compromise between the quantization noise and the descriptor noise.

Another major problem of commonly used Bag-of-Words approaches is the immutability of the vocabulary. The visual vocabulary is usually generated only once during training based on the training data and is fixed for all future images. Imagine now the situation of adding a new category to the already existing and learnt ones. The previously learnt and fixed vocabulary will likely not contain enough visual words that would accurately describe the points of the new category. For example, take a vocabulary that was created based on fruit images and try adding the category motorbike. Figure 2 shows histograms of visual words for images depicting a grape and a motorbike using a fruit vocabulary constructed with $k = 300$ from apple, banana and grape images. Clearly, the descriptors of the motorbike image cannot be assigned well to the visual words of the fruit vocabulary. Obviously, the visual words in the vocabulary are not discriminative enough for new categories.



(a) grape (b) motorbike

**Figure 2: Histograms of visual words for objects sufficiently (grape) and insufficiently (motorbike) represented in the fruit visual vocabulary.**

## 2.2 Extensibility of the Visual Vocabulary

We concluded that the visual vocabulary needs to be adjusted to contain words also from new categories. Following options can be considered as a solution:

1. use a general-purpose visual vocabulary,
2. recompute the general visual vocabulary
   (a) every time a new category has been added,
   (b) after the discriminative power of the vocabulary falls below a predefined margin, or
3. use class-specific vocabularies instead of one static general vocabulary.

The first option has the advantage that the vocabulary never changes, histograms don't need to be recomputed and class-label relationships don't need to be relearned. But several questions arise: How to create a general-purpose vocabulary? How many and which visual words should the general-purpose vocabulary contain to be able to distinguish 3,000 categories? In [29] it was shown that the optimal size of the vocabulary (regarding classification accuracy) clearly depends on the training image set. Obviously the previously

enumerated questions cannot be answered unless the future categories, that are likely to be added to the system are known.

Another option would be to regularly recompute the general visual vocabulary. This could be done every time after adding a new category or after reaching a defined margin measuring the loss of discriminative power of the visual words in the vocabulary. The main benefit of this solution is that the general vocabulary always adapts to the categories in the system and can also be extended during its lifetime. Another positive effect would be, that after learning a huge amount of categories the vocabulary would turn into a general-purpose vocabulary which never ever has to be changed. For this adaptability one has to pay with high and steadily increasing computational costs. Every time the vocabulary changes, all the histograms of visual words have to be recomputed for all categories and the class-label relationships need to be relearned from scratch. Obviously, this solution is not practicable.

In the third option separate class-specific vocabularies are computed for each class instead of a general-purpose visual vocabulary. With this approach, histograms of visual words are not required any more. Other benefits are, that the system is easily extensible with new categories and no recomputation of any already learnt categories is needed. This approach can also be easily parallelized, distributed and used in a hierarchy. The only disadvantage is, that instead of training a classifier with established methods like SVMs a good scoring function is needed to compare different categories for a given image.

# 3. RELATED WORK

In this section we first present approaches using class-specific vocabularies; in a second part we discuss methods similar to class-specific vocabularies – that is hierarchical vocabularies – and finally we briefly introduce hierarchical approaches for classification.

## 3.1 Class-Specific Vocabularies

Class-specific vocabularies were not really popular in the recent past. Most Bag-of-Words approaches like [3, 18, 28, 31] use a single vocabulary generated with $k$-means using keypoints extracted from all training images. [8] introduce the idea of building class-specific vocabularies and merging these into a single general vocabulary. The main benefit of computing class-specific vocabularies first is that class labels are incorporated in the clustering of keypoints, while this discriminative information would have been lost during all-at-once clustering. Following this idea, [27] first creates a general vocabulary based on all training samples. The general vocabulary is then adapted using class-specific training samples which results in class-specific vocabularies. For each category both the general and the class-specific vocabularies are then joined as bipartite histograms and used as input for one-vs-all SVMs. Neither approach is extensible because after adding a new category the whole training and learning has to be repeated.

## 3.2 Hierarchical Vocabularies

Hierarchical vocabularies or vocabulary trees are related to class-specific vocabularies because single or groups of leaf nodes might represent separate classes. Due to the scalability issue, recently several approaches addressed the problem of organizing visual words in a hierarchical vocabulary tree. [24] adapted the kD-tree approach of [16]. Their tree is heavily dependent on the training set and needs to be reconstructed from scratch if new categories are added. [23] proposed a hierarchical top-down $k$-means algorithm for organizing large numbers of visual words in a vocabulary tree. Leaf nodes represent single visual words and use a scoring function for creating an inverted list of images containing the given visual word. The vocabulary tree is created only once during the training step and remains unchanged even if new images are added. New images are only added to the inverted lists in the leaf nodes, so adding several new categories will sooner or later require the recomputation of the whole vocabulary tree.

## 3.3 Hierarchies for Classification

Many Bag-of-Words approaches use SVMs for classification. For multi-class classification, several binary SVMs have to be trained. Traditionally, multi-class SVMs are implemented as one-vs-one (voting, 1:1), one-vs-all (competition, 1:N), or gradual exclusion via a directed acyclic graph (DAG). The latter two have linear, while 1:1 has quadratic complexity depending on the number of classes. With an increasing number of categories in the field of object recognition, several approaches addressed the problem of building classification hierarchies to classify images with polylogarithmic, logarithmic or even constant complexity.

One of the simplest approaches is to build a binary or $k$-nary tree of subsequent SVMs like proposed in [15, 30]. [11] derives a visual taxonomy represented by a binary tree of SVMs on the Caltech 256 dataset. For the top-down approach, a spectral clustering algorithm is applied, while the bottom-up approach subsequently combines pairs of categories with highest pairwise confusion. The resulting visual taxonomy proves the hypothesis that visual and lexical similarities between objects can be divergent. [19] implements a DAG-SVM based on the Caltech 256 dataset which appears to scale better than other hierarchical SVM approaches.

Other approaches try to incorporate external knowledge about relationships between objects into the hierarchy of SVM classifiers. Sources for object relationships are either home-built taxonomies or empirical ontologies like WordNet [22]. [32] uses four different object hierarchies for the combination of object classifiers and achieves both higher precision and higher recall. [17] extends the labels assigned to training images with lexical relationship for words extracted from WordNet and learns an extended hierarchical model with SVMs. Although the evaluation showed no performance increase, the complexity was reduced to sublinear. Despite the efforts of incorporating lexical relationships to the task of object recognition the question whether lexical and visual relationships are parallel or orthogonal remains open.

All approaches in this section present steps towards a more efficient and scalable object recognition, but all still rely on a previously computed static vocabulary and therefore are not dynamically extensible with new categories.

# 4. PROPOSED METHOD

As discussed in section 2, a static vocabulary has drawbacks when handling the problem of extending the number of learnt categories. As a first step, we propose a new method inspired by common image retrieval algorithms, hierarchical approaches, clustering and $k$NN classification. In the follow-

ing sections, we describe the main idea as well as the steps for training and classification using our method in detail.

## 4.1 Main Idea

Our proposed method addresses the problem of adding a new category to an object recognition system based on the Bag-of-Words model. Classic Bag-of-Words approaches use a fixed visual vocabulary created during the training step. Although the visual words in the vocabulary describe images well that belong to categories used for training, they are inaccurate for the description of images of previously unseen categories.

From a top-down perspective, visual vocabularies can be separated for different category groups getting as many histograms of visual words for each image as vocabularies are used. This could be seen as a variation of the idea proposed in [26], where bipartite histograms are used to distinguish whether an image belongs to a given class or to other classes. The current scenario results in n-partite histograms. Following this idea, visual vocabularies could be separated further on deeper levels, too (like apple, banana, grape, orange vocabularies for fruit categories etc., see figure 3), arriving at the point where for each category a separate vocabulary is used. So, the number of histograms of visual words to be computed equals the number of different categories on the current branch of the hierarchy. During the training a classifier also need to be trained for each node in the hierarchy.

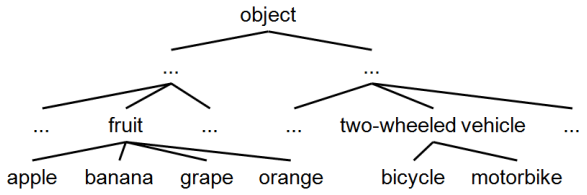



**Figure 3: Example object hierarchy.**

From a bottom-up perspective, a class-specific vocabulary could be computed first for each category. The class-specific vocabularies could then be merged together into grouped class vocabularies. This merging step could be repeated several times, resulting in a hierarchy where the root node contains a general-purpose visual vocabulary like in [8, 27]. In the end still as many histograms of visual words need to be computed for each image as many different categories there are on the current branch of the hierarchy. Another problem is, that with the addition of a new leaf category, all vocabularies of inner nodes along the path from the new category to the root node would need to be recomputed - along with the computation of all histograms of visual words for each training image and the retraining of classifiers.

Based on this line of thought, our idea is to extract independent class-specific vocabularies for each category separately. Because of variations in the object representation and background changes we assume that only a part of all visual words found in images belonging to a category are truly relevant for the description of the given category. Therefore only those visual words are kept in the class-specific vocabulary, that appear most frequently in images belonging to the given category. This also results in several "outlier" visual words, which don't belong to the top visual words for a given category. Our hypothesis is that later in the classification step, images of objects belonging to their class should have less outliers than images of objects belonging to other classes. To leverage the computation of histograms of visual words, the vocabulary and the histogram of visual words are merged into a single class description. As a result, each class-specific vocabulary contains an additional value for the frequency of a visual word in an "average" image belonging to the given category. Instead of the repeated training of classifiers we propose a scoring function inspired by classic image retrieval approaches.

## 4.2 Learning Class Descriptions

Class-specific vocabularies are computed instead of a general vocabulary. First, each image is scaled to 128x128 pixels, converted to HSV color space, then points of interest are extracted using the Hessian Affine detector from the V channel and SIFT descriptors are computed using the software provided by [20, 21][2]. This way we get about 275 SIFT descriptors on average per image using the Caltech 256 dataset provided by [10]. In the next step all SIFT descriptors of all training images of each class are clustered with $k$-means using the L2 distance. After clustering the frequency of the points for each cluster are computed and the clusters are sorted according to their population. The top $n$ clusters are kept for each class ($n < k$). Next, the means for each of the $n$ clusters (class-specific visual words) and the maximum distance between a point belonging to a given cluster and its mean (cluster radius) are computed. For each class the values of the top $n$ visual words, the average number of points per visual word and per image belonging to the cluster and the cluster radius are stored as a class description (object identifier). This way we get for each class a vocabulary containing $n$ visual words along with the information how much noise is allowed for a descriptor to be assigned to the corresponding visual word and how often these visual words appear on average in a single image. In other words: we merged the visual vocabulary with the histogram of visual words.

## 4.3 Scoring and Classification

The first steps of classifying previously unseen images are similar to the training (resizing, detection of points of interest, SIFT descriptors). For measuring the similarity between an image $I$ and a class description $C$ a good scoring function is required. The idea is to assign each descriptor of image $I$ to the $n$ cluster centers of the class description $C$ using the L2 distance and the allowed noise (cluster radius). Descriptors having a bigger distance to their closest cluster centers than the radius of the cluster allows are considered as "outliers" and are assigned to a special outlier cluster $n+1$. We assume that the number of outlier descriptors will be lower for images from the same class than for images from different classes. Next for all clusters the number ($freq(I_i)$) and the average distance ($avgdist(I_i)$) of the descriptors of the given image $I$ are computed. $freq(C_i)$ is the average number of descriptors for cluster $i$ in class description $C$. For the outlier cluster $freq(C_{n+1})$ is set to 0 and for $avgdist(I_{n+1})$ the average distance of all outlier descriptors is computed. The average distance should already be quite high because the outliers don't belong to any visual word in the class description and the more outliers there are, the higher this average outlier distance is weighted by $freq(I_{n+1})$. This way we have already included a penalty for outliers in our scoring function. Based

[2]http://www.robots.ox.ac.uk/∼vgg/research/affine/

on these values the score for image $I$ and class description $C$ is computed using the following scoring function:

$$score(I,C) = \frac{1}{\frac{1}{n+1}\sum_{i=1}^{n+1} |freq(C_i) - freq(I_i)| \cdot avgdist(I_i)} \quad (1)$$

The closer an image gets to the class description, the higher the value of $score(I,C)$, so, intuitively, image $I$ is assigned to the class $C$ with the highest score value. There are several parameters in our proposed method that need further exploration. We will show our first results in the following section.

# 5. EXPERIMENTS AND DISCUSSION

In our experiments, we try to find the optimal values for the initial number of clusters ($k$) and the number of most populated visual words to be kept as class descriptions $n$. For all tests we used the scoring function from equation 1, the Caltech 256 dataset provided by [10] and 20 training images per category. All tests have been repeated for 100 different combinations of 3, 4, 5, 10 and 20 categories and the classification precision has been averaged to get an impression how general the approach is.

First we wanted to see how the classification accuracy behaves for different numbers of clusters $k$. We fixed the ratio of $n/k$ at 0.5, so we always kept half of the clusters as class descriptions. We experimented with $k = 40$, 60, 80 and 100. We didn't try higher values for $k$ because the average number of descriptors per image is 275 for 128x128 pixel images using Hessian Affine detector and SIFT descriptor on the V channel. The results are shown in figure 4(a). Similarly to [23] we found that – even for the class-specific vocabulary case – the bigger the vocabulary, the better the classification accuracy. Further investigation is required for determining the optimal ratio between the average number of descriptors per image and the number of visual words $k$ for the vocabulary computation.

Next we wanted to know how much we can trim the class-specific vocabularies on average without losing classification accuracy. Because in the previous step we got the best results using $k = 100$, we experimented with keeping the top $n = 5$ to $n = 90$ visual words. The results are shown in figure 4(c). On average about 20% of visual words for each class-specific vocabulary are sufficient for achieving same precision as with a higher number of visual words. We also repeated our tests using the Harris Affine detector and the SIFT descriptor. The results depicted in figure 4(d) show a similar limit around 20-25%. The question how general the optimal ratio $n = k/5$ is, needs further investigation using different detectors and vocabulary sizes. Figure 4(e) shows a ROC curve for $k = 100$ and $n = 5$ to $n = 50$. The current area under curve is around 0.72, which is good but needs further improvement.

We also implemented tests to verify our hypothesis that images of objects belonging to a given class have less outliers than images of objects belonging to other classes. Figure 4(b) shows the relation of the number of class outliers compared to the number of non-class outliers. Our assumption that images of objects belonging to a given class have less outliers than images of objects belonging to other classes was proven right in all scenarios. In further improvement of our scoring function we plan to integrate this knowledge and penalize outliers more heavily to improve classification accuracy.

# 6. CONCLUSION

We propose an extensible approach for object recognition based on the Bag-of-Words model. The main idea is to compute independent class-specific vocabularies instead of one static general vocabulary. Without a general vocabulary, we are unable to compute fixed length histograms of visual words for SVMs, so we merged the histograms and the class-specific vocabularies and introduced a scoring function which calculates the closest category based on the class-specific vocabularies. Our proposed method is clearly easily extensible by new categories without the need of heavy recomputation. The evaluation of our approach on the Caltech 256 dataset shows promising results, but scalability still remains an issue.

In the future, we will evaluate our approach using color channels, various distance measures, incorporating spatial information, different features and their combinations. This also might require adjustments in our scoring function which we plan to evolve, too. Other future work is the extensible hierarchical organization of our independent class-specific vocabularies according to visual and lexical similarities.

Our approach is highly parallelizable and distributable. Both the training and the classification steps are currently optimized for multi-core CPUs. Further steps are to implement our approach with MapReduce [6] and boost computing time by using large clusters.

# 7. REFERENCES


[1] L. H. Armitage and P. G. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[2] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE PAMI*, 30(4):712–727, April 2008.

[4] H.-l. Chen. An analysis of image queries in the field of art history. *JASIST*, 52(3):260–273, 2001.

[5] Y. Choi and E. M. Rasmussen. Searching for images: The analysis of users' queries for image retrieval in american history. *JASIST*, 54(6):498–511, February 2003.

[6] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Communciations of the ACM*, volume 51, pages 107–113, January 2008.

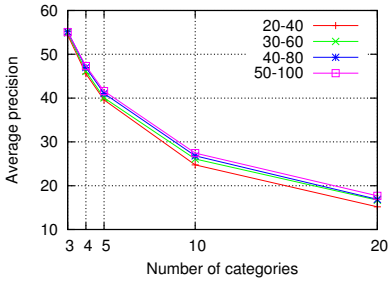
[7] P. G. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52, 1993.

[8] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation: Generative models and pdf-kernels. Technical report, Department of Electronics and Computer Science, University of Southampton, 2005.

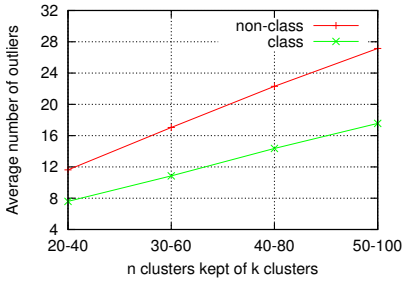
[9] A. Goodrum and A. Spink. Image searching on the excite web search engine. *IPM*, 37(2):295–311, March 2001.

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, March 2007.

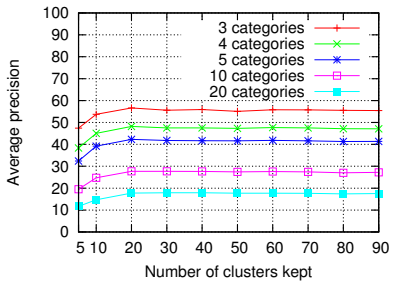
[11] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *IEEE CVPR*, pages 1–8, June 2008.



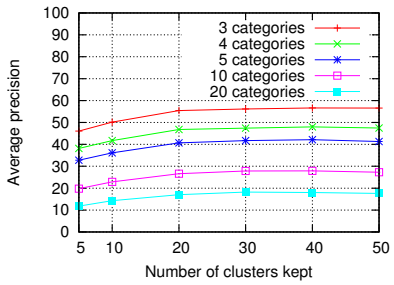

(a) $k = \{40, 60, 80, 100\}$, $n = k/2$



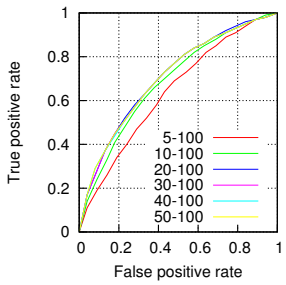

(b) Class vs. nonclass outliers.




(c) $k = 100$, $n = 5 \ldots 90$, Hessian+SIFT




(d) $k = 100$, $n = 5 \ldots 50$, Harris+SIFT




(e) ROC curves for $k = 100$, $n = 5 \ldots 50$, Hessian+SIFT

**Figure 4: Classification accuracy for different combinations of $k$ clusters and $n$ clusters kept.**


[12] L. Hollink, A. T. Schreiber, B. J. Wielinga, and M. Worring. Classification of user image descriptions. *IJHCS*, 61(5):601–626, November 2004.

[13] B. J. Jansen, A. Spink, and J. Pedersen. An analysis of multimedia searching on altavista. In *ACM MIR*, pages 186–192, November 2003.

[14] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008.

[15] S. Liu, H. Yi, L.-T. Chia, and D. Rajan. Adaptive hierarchical multi-class svm classifier for texture-based image classification. In *IEEE ICME*, July 2005.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[17] M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. In *IEEE CVPR*, pages 1–7, June 2007.

[18] M. Marszałek, C. Schmid, H. Harzallah, and J. van der Weijer. Learning object representations for visual object class recognition. In *Pascal Visual Recognition Challenge Workshop in Conjunction with ICCV*, 2007.

[19] J. Martinet and S. Satoh. A study of intra-modal association rules for visual modality representation. In *IEEE CBMI Workshop*, pages 344–350, June 2008.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 27(10):1615–1630, October 2005.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, November 2005.

[22] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, November 1995.

[23] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE CVPR*, volume 2, pages 2161–2168, June 2006.

[24] S. Obdrzálek and J. Matas. Sub-linear indexing for large-scale object recognition. In *BMVC*, volume 1, pages 1–10, 2005.

[25] S. Ornager. The newspaper image database: empirical supported analysis of users' typology and word association clusters. In *ACM SIGIR*, pages 212–218, 1995.

[26] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE CVPR*, pages 1–8, June 2007.

[27] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, pages 464–475, 2006.

[28] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE ICCV*, volume 2, pages 1470–1477, October 2003.

[29] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM MIR*, pages 197–206, 2007.

[30] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, and S. Li. Automatic video genre categorization using hierarchical svm. In *IEEE ICIP*, pages 2905–2908, October 2006.

[31] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, June 2007.

[32] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *IEEE ICCV*, pages 1–8, October 2007.