A Fast Object Recognition and Categorization Technique for Robot Grasping Using the Visual Bag of Words

Mohamed Hannat, Nabila Zrira, Younès Raoui and El Houssine Bouyakhf
LIMIARF, FSR, Mohammed V University, B.P.1014 RP Rabat, Morocco
mohamedhannat@gmail.com, nabilazrira@gmail.com, raoui@fsr.ac.ma, bouyakhf@mtds.com

***Abstract*—We present in this paper a real time method for visual categorization to do robot grasping. We describe an object database with SURF feature points which we quantify with the Kmean clustering algorithm to make visual words. Then, we train a Support Vector Machine classifier having as entries the distribution of the bag of features extracted earlier. Next we do object recognition using samely the SVM algorithm. The real time implementation is done with the OpenCV GPU. The application we show consists to pick an object and drop it using our robot manipulator equipped with a camera using our visual system. Finally we present the results of our experiments of the object recognition which average of recognition is between 95% and 100%.**

Keywords*—*Visual categorization, Bag of Words (BoW), OpenCV GPU, SURF, Object recognition.

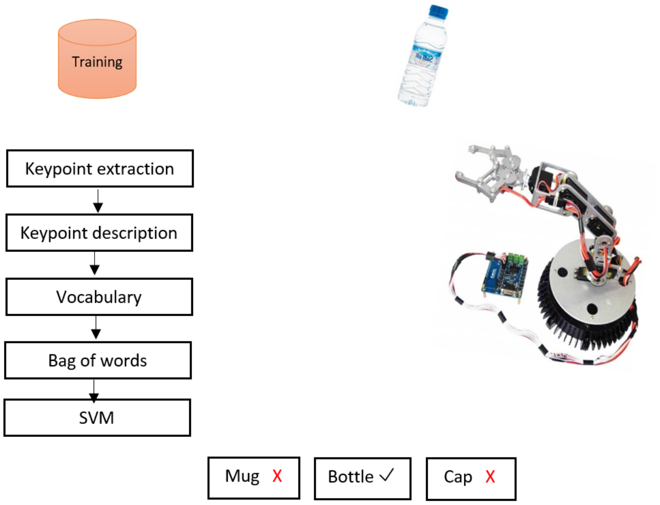

Figure 1. Overview of our approach for object recognition and manipulation using Bag of Words.

I. Introduction

The visual categorization is an important task in computer vision, and it is widely used for object recognition, content based image retrieval and robot grasping. In the appearance based image description approach, objects should be described with visual feature points with methods such as SIFT or SURF (0) because of their invariance to the scale and the orientation and almost to the illumination changes. The visual categorization has a close relationship with the learning with both Bayesian or bio inspired methods. Fei Fei lee has introduced in (?) the use of the Bayesian approach in the visual categorization and creates a model for large objects database through representing the visual cues with parameters of the appearance and the shape. Methods such us bag of features (?) and spatial pyramids (?) have been used to represent the likelihood of each feature inside the view. For building a model, we should train a classifier on the extracted visual features or words with methods like SVM or recurrent neural networks. Next, the recognition is done by maximizing the likelihood giving the probability if a feature is present in an image seen by the robot and the learnt model. In this work, we present a new method for visual categorization using SURF as visual features and SVM for the training and the recognition. Additionally, we implement our algorithm in a real time robot grasping's application using the CUDA technology. We summarize our method in these steps:

- Extraction of the keypoints and descriptors using SURF detector/descriptor;
- Extraction of the vocabulary using Kmeans algorithm;
- Computation of the Bag of Words;
- Training the SVM classifier on the SURF visual words;
- Determination of the object category with the LibSVM classifier;
- Optimization of our approach using GPU device;
- Pick up and drop the target object with Dagu robotic arm (see fig. 3).

We present our paper as following, in the section II we present an overview of our method, in section III, we present related works, then we describe in the section IV our proposed algorithm, and in section V we present experiments showing the advantages of using CUDA and giving an overview about the performances of our method. Finally we conclude and we show some of our future works.

II. Method Overview

In this paper, we suggest a new approach for fast object recognition and manipulation for mobile robotic applications. Figure 1 summarizes the main steps of our proposed approach.

Training set: represents a set of data (images) used on our experiment. Training means, creating a dataset with all objects we want to recognize.

Keypoint extraction: is the first step of our approach. It consists of extracting keypoints (interest points) from data. They reduce the computational complexity by identifying particularly those regions of images, which are important for descriptors, in terms of high information density.

Keypoint description: once keypoints are extracted, descriptors are computed on the obtained keypoints and these form a description that is used to represent the images.

Vocabulary: after the extraction of descriptors, the approach uses the vector quantization technique to cluster descriptors in their feature space. Each cluster is considered as "visual word vocabulary" that represents the specific local pattern shared by the keypoints in this cluster.

Bag of words: is a vector containing the (weighted) count or occurrence of each visual word in the image which is used as feature vector in the recognition and classification tasks.

Support Vector Machine (SVM): all images in training set are represented by their Bag of Words vectors which represent the input of SVM classifier. Our approach can predict the class of real-world objects, then, the arm picks up the target object.

III. Related Work

Recently, the approaches that were based on a Bag of Words (BoW), also known as Bag of features produced the promising results on several applications, such as object and scene recognition (0) (0), localization and mapping for mobile robots (0), video retrieval (0), text classification (0), and language modeling for image classification and retrieval (0) (0) (0).

Sivic et al. (0) use Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis (pLSA) in order to compute latent concepts in images from the cooccurrences of visual words. The authors aim to generate a consistent vocabulary of visual words that is insensitive to viewpoint changes and illumination. For this reason, they use vector quantized SIFT descriptors which are invariant to translation, rotations and re-scaling of the image.

Csurka et al. (0) developed a generic visual categorization approach in order to identify the object content of natural images. In the first step, their approach detects and describes image patches which are clustered with a vector quantization algorithm to generate a vocabulary. The second step constructs a bag of keypoints that counts the number of patches assigned to each cluster. Finally, the authors use Naive Bayes and SVM to determine image categories.

Fergus et al. (0) suggested an object class recognition method that learn and recognize object class models from unlabeled and unsegmented cluttered scenes in a scale invariant manner. The approach exploits a probabilistic model that combines shape, appearance, occlusion and relative scale, as well as an entropy-based feature detector to select regions and their scale within image.

Philbin et al. (0) proposed a large-scale object retrieval system with large vocabularies and fast spatial matching. The authors, extract features on each image in some high-dimensional descriptor space which are quantized or clustered to map every feature to a "visual word" that is used to index the images for the search engine.

Wu et al. (0) proposed a new scheme to learn optimized bag of words models called Semantics Preserving Bag of Words (SPBoW) that aims to map semantically related features to the same visual words. SPBoW computes distance between identical features as a measurement of the semantic gap and tries to learn a codebook by minimizing this semantic gap.

Khan et al. (0) suggested a new approach to integrate spatial information in the bag of visual words. The approach model the global spatial distribution of visual words that consider the interaction among visual words regardless of their spatial distances. The first step consists on computing pair of identical visual words (PIW) that saves all the pairs of visual words of the same type. The second step represents a spatial distribution of words as histogram of orientations of the segments formed by PIW.

Larlus et al. (0) combined a bag of words recognition component with spatial regularization based on a random field and a Dirichlet process mixture in order to insure category level object segmentation. The random field (RF) component assures short-range spatial contiguity of the segmentation, while a Dirichlet process component assures mid-range spatial contiguity by modeling the image as a composition of blobs. Finally, the bag of words component allows strong intra-class imaging variations and appearance.

Vigo et al. (0) exploited color information in order to improve the bag of words technique. Their approach chooses highly informative color-based regions for feature detection. Then, feature description, focuses on shape, and can be improved with a color description of the local patches. The experiments show that color information should be used both in the feature detection as well as the feature extraction stages.

IV. The proposed method

A. Visual categorization and object recognition

The visual categorization and object recognition is very important in robot grasping with the use of cameras. We present in this section the method we use so that we describe and categorize a class of objects through classifier's training. Next, we show that we could recognize an object of a certain using the SVM method

1) The visual categorization with the bag of word distribution:

- An overview of the Speeded Up Robust Transform detector descriptor:
    1) Keypoint extraction Before applying the detector of SURF, we divide the image into small sub-images with the integral of integral images. Given an image I(x,y), com compute several images which sum is the image I:

$$S(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j) \tag{1}$$

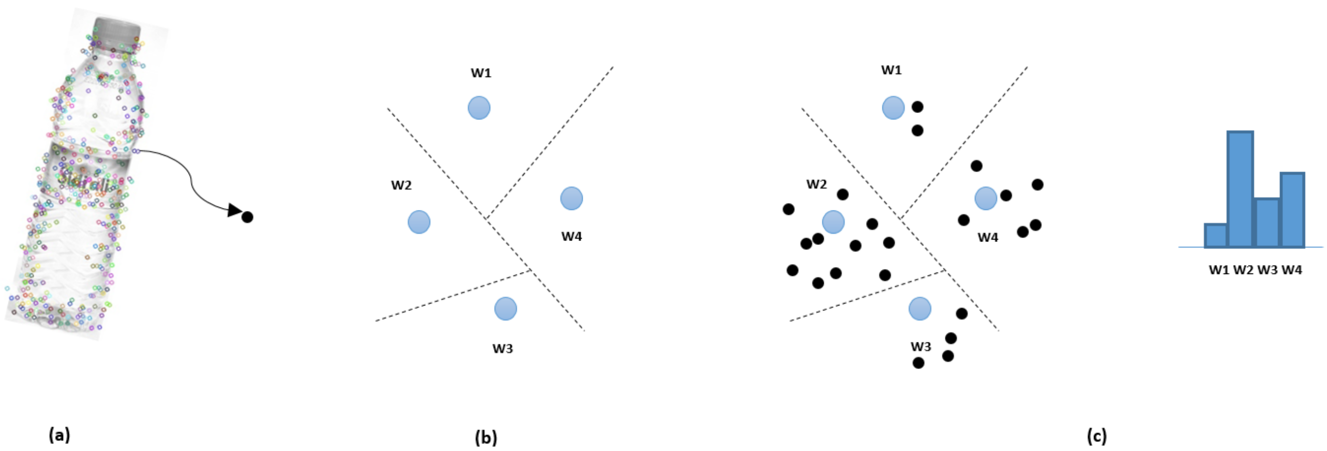

Figure 2. The schematic illustrates visual vocabulary construction and word assignment. (a) the black dot represents SURF keypoint, the object contains in total 240 SURF keypoints. Next, the approach computes SURF descriptor on each keypoint. (b) Visual words (W1, W2, W3 and W4) denote cluster centers. (c) The sampled features are clustered in order to optimize the space into a discrete number of visual words. A bag of visual words histogram can be used to summarize the entire image. It counts the occurrence of each visual word in the image.

Then the goal is to compute a set of feature points with their characteristic scales and orientations. The points of the SURF detector are computed with the determinant of the Hessian matrix measuring the local changes around the point. We have to maximize this determinant so that the pixels where it is computed are salient points. Additionally this determinant is used as well to determine the scale σ. The Hessian is given with :

$$H(p,\sigma) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{xy}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix} \tag{2}$$

Where p(x,y) is an image and :

$$L_{xx} = S * G_{xx}(\sigma) \tag{3}$$

$$L_{xy} = S * G_{xy}(\sigma) \tag{4}$$

$$L_{yy} = S * G_{yy}(\sigma) \tag{5}$$

Furthermore, the Gaussian G is approximated with a box filter of size 9*9 at the scale $\sigma = 1.2$.

Because of the importance of the invariance to the rotation, the orientation of each feature point needs to be detected. Around each feature point, we compute the Haar wavelet responses in the x and y directions in a circle of size 6σ where σ is the scale. Next, we convolve this circle with a Gaussian kernel giving us a matrix of values in the horizontal and vertical axes. The orientation is calculated by summing all values of this matrix with a sliding orientation window of size $\pi/4$. While the horizontal and vertical responses within the window are summed and the two summed responses then gives a local orientation vector.

2) Keypoint description:
We take a window region of size 20s oriented along the orientation computed before . Next we split it up into smaller 4*4 square sub-region, and the Haar wavelet response is extracted at 5*5 regularly spaced points. Additionally, these responses are weighted witha Gaussian to improve the robustness for deformation noise and translations.

- Visual categorization:
After describing each of the images inside a class with SURF transform, we should construct a model that represents all the images which are not of the same object but within the same class of a particular object (see fig. 2). That is why we need to make a visual categorization using the probabilistic approach. The method we use consists to apply a quantization operation with kmean clustering, constructs visual words with the well known method of bag of features, and finally classifies the words using the support vector machine.

Kmean clustering The Kmean clustering algorithm is done by following these steps:

1) Select initial centroids at random;
2) Assign each keypoint to the cluster with the nearest centroid;
3) Compute each centroid as the mean of the objects assigned to it;
4) Repeat previous 2 steps until no change.

Visual Bag Of Words Instead of considering each feature point a visual word, we consider thanks to the quantization that each of the clusters' center represent a word. The bag of words algorithm consists to compute the number of occurrences of each word in the model database. It is like a probability of the number of words inside the class of objects. Subsequently, it is a step towards computing a codebook or, in other words, a dictionary of several classes of object. Moreover, it will be useful for grasping tasks because it will give a unified model for perception-grasping.

2) Training a classifier and object recognition with SVM: Support Vector Machines (SVMs) (0) are a useful method for data classification. They are based on the concept of decision planes which separate between a set of objects that have different class memberships.

Let's consider a training set of instance-label pairs (x_i, y_i) where $i = 1,...,l$, $x_i \in R^n$, and $y \in (-1,1)^l$, the optimization problem is defined as:

$$min_{w,b,\xi} \qquad \frac{1}{2}W^TW + C\sum_{i=1}^{l}\xi_i \tag{6}$$

subject to

$$y_i(W^T\phi(x_i)+b) \succeq 1-\xi_i \tag{7}$$

with $\xi_i \succeq 0$ and $C \succ 0$.

x_i are mapped into a a higher dimensional space by the function ξ. SVMs try to find a linear separating hyperplane with the maximal margin in this higher dimensional space. C is the penalty parameter of the error term.

In our work, we are interested in multi-class classification. For this purpose, we use C-Support Vector Classification (C-SVC) for two-class and multi-class classification.

Let's consider a training vectors (x_i) where $i = 1,...,l$, $x_i \in R^n$, in two classes, and an indicator vector $y \in (-1,1)^l$, the optimization problem is defined as:

$$min_{w,b,\xi} \qquad \frac{1}{2}W^TW + C\sum_{i=1}^{l}\xi_i \tag{8}$$

subject to

$$y_i(W^T\phi(x_i)+b) \succeq 1-\xi_i \tag{9}$$

with $\xi_i \succeq 0$ and $i = 1,...,l$.

$\phi(x_i)$ maps x_i into a higher-dimensional space. Due to the possible high dimensionality of the vector variable W, usually we solve the following dual problem.

$$min_{\alpha} \qquad \frac{1}{2}\alpha^TQ\alpha - e^T\alpha \tag{10}$$

subject to

$$y^T\alpha = 0 \tag{11}$$

with $0 \preceq \alpha_i \preceq C$ and $i = 1,...,l$.

Where $e = [1....1]^T$ is the vector of all ones, Q is an l by l positive semi-definite matrix,$Q_{ij} \equiv y_iy_jK(x_i,x_j)$ as well as $K(x_i,x_j) \equiv \phi(x_i)^T\phi(x_j)$ denotes the kernel function. After problem (6) is solved, using the primal-dual relationship, the optimal W should satisfy the following equation.

$$W = \sum_{i=1}^{l} y_i\alpha_i\xi_i \tag{12}$$

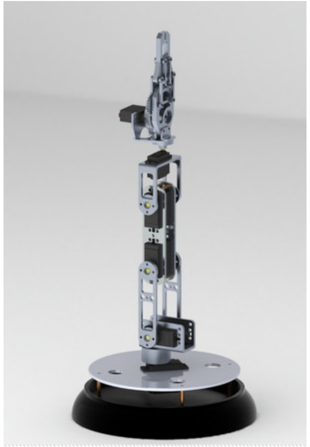

Figure 3. Dagu robotic arm with 6 degrees of freedom used in our experiment.

V. The experiments

A. The Experimental setup

We used the Amsterdam Library of Object Images (ALOI) (0) dataset which contains 1200 small objects of the same size and belonging to several classes (see fig. 4). First, we extracted many words, reaching 1500, from the images following a quantization step to cluster the SURF visual features. Second, we train a classifier with the LibSVM C++ library. We chose the RBF's kernel, a non linear one, mapping the words to the labels even when relationship between them is non linear. We used our robot hand to make experiments for our visual categorization method.

B. The OpenCV GPU library

The problem was to make our application running in real time, and the solution that we chose it to use CUDA toolkit that increases dramatically the performances of computing exploiting the graphics processing units (GPUs). Indeed it is another option of OpenCV with a compiler for NVIDIA GPUs. Furthermore, we used SURF CUDA storing features in both CPU and GPU memory, and LibSVM CUDA reducing the computational cost for training data.
Table II illustrates the difference between the training vectors with LibSVM CUDA and original LibSVM. We can conclude that the use of GPU makes the training stage very fast 0.919S against 1.573s when using CPU. This result is very useful especially when the number of the training data is important. The figure 6 shows the computing time of SURF per the number of images for both CPU using INTEL and CPU using NVIDIA technologies. For the CPU, the curve roses sharply from 24s for 200 images to 150s for 1200 images. On the contrary for the GPU, it is less expensive in time computation as the curve roses steadily from 7s for 200 images to 28s for 1200 images. This remarkable difference is obvious, since GPU contains more efficient cores designed for handling multiple operations simultaneously. For this reason, we use the CUDA technology in implementation of the SURF algorithm.

C. Object grasping

For object manipulation task (see fig. 1), we use Dagu robotic arm that is equipped with ATMega Arduino card and serial port communication (see fig.3). Once the object is captured by the camera using OpenCV libraries, our approach extracts the Bag of Words vector of this image, then predicts the object class. Dagu picks up the target object and drops it on the precise area.

D. Results for the object recognition

The table I shows the accuracy of the recognition of a particular object. By fixing the vocabulary size in SVM parameters in the LibSVM CUDA to 1500 words, we obtain an accuracy of 100%. The figure 5 shows the accuracy of the recognition per the vocabulary size of the bag of words. The curve roses sharply from 20% for a size of 100 words to 92% for a size of 1000 words. Then it roses steadily from 92% to 100% for a vocabulary size of 1500 words. In consequence, the more the number of words is high, the more the accuracy of the object recognition is good.

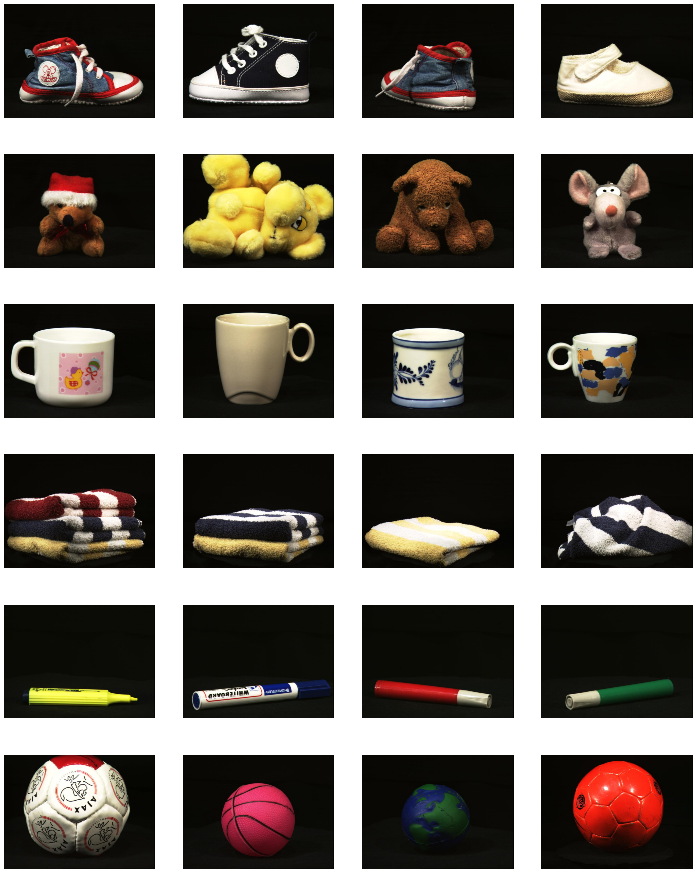

Figure 4. The sample images extracted from Amsterdam Library of Object Images (ALOI) dataset.

Kernel	Type	Gamma	C	Accuracy	Vocabulary size
RBF	C-SVC	0.50625	312.50	**100%**	1500

Table I. LibSVM characteristics used in our experiments. The obtained accuracy is 100% with vocabulary size equal to 1500 visual words.

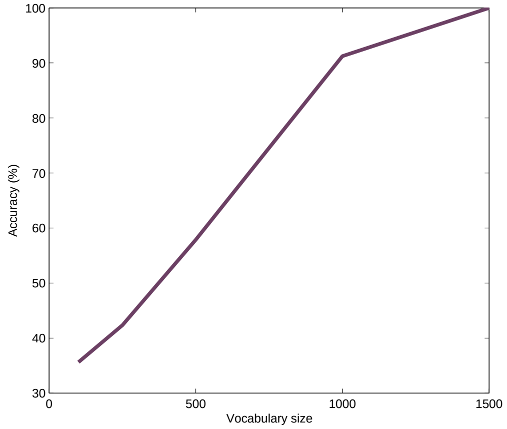

Figure 5. The classification performance at different vocabulary sizes.

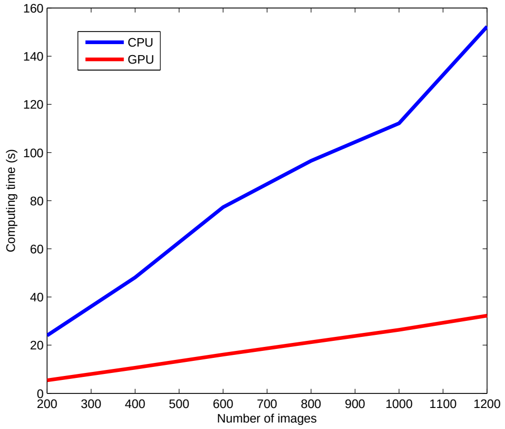

Figure 6. CPU and GPU computing time for SURF keypoints and SURF descriptors extraction.

Number of training vectors	CPU	GPU
1200	1.573 s	**0.919 s**

Table II. The computational cost difference between 900 training vectors on LibSVM classifier using CPU and GPU devices.

VI. Conclusion

In this paper, we have developed a visual system for visual categorization and object recognition with the SURF transform because of its fast computation and its distinctiveness. The invariance of rotation of SURF allows to get features that we transform with the quantization to have visual words. Next we use the bag of feature method to compute the occurrences of each word in the view which we utilize to train a classifier with SVM algorithm preceding the object recognition with SVM also. The obtained results are so promising because the average of the recognition is higher that 95%. For the future work, we will develop an approach for 3D object categorization and manipulation using 3D point clouds and PCL descriptors. For classification task, we will utilize deep learning methods such as Deep Belief Networks and Convolutional Neural Networks.

References

J. Bai, J.-Y. Nie, and F. Paradis, “Using language models for text classification,” in *Proceedings of the Asia Information*

Retrieval Symposium, Beijing, China, 2004.

A. Bolovinou, I. Pratikakis, and S. Perantonis, "Bag of spatio-visual words for context inference in scene classification," *Pattern Recognition*, vol. 46, no. 3, pp. 1039–1053, 2013.

T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 4, pp. 381–392, 2011.

M. Li, W.-Y. Ma, Z. Li, and L. Wu, "Visual language modeling for image classification," Feb. 28 2012. US Patent 8,126,274.

K. R. Mc Donald, *Discrete language models for video retrieval*. PhD thesis, Dublin City University, 2005.

L. Zhu, A. B. Rao, and A. Zhang, "Theory of keyblock-based image retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 2, pp. 224–257, 2002.

J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1470–1477, IEEE, 2003.

D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 3921–3926, IEEE, 2007.

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," 2005.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, pp. 1–2, Prague, 2004.

R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–264, IEEE, 2003.

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Computer Vision–ECCV 2006*, pp. 490–503, Springer, 2006.

L. Wu, S. C. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and applications," *Image Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1908–1920, 2010.

R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial orientations of visual word pairs to improve bag-of-visual-words model," in *Proceedings of the British Machine Vision Conference*, pp. 89–1, BMVA Press, 2012.

D. Larlus, J. Verbeek, and F. Jurie, "Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 238–253, 2010.

D. A. R. Vigo, F. S. Khan, J. Van de Weijer, and T. Gevers, "The impact of color on bag-of-words based object recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 1549–1553, IEEE, 2010.

H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*, pp. 404–417, Springer, 2006.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

C. Harris and M. Stephens, "A combined corner and edge detector.," in *Alvey vision conference*, vol. 15, p. 50, Citeseer, 1988.

W. Förstner, T. Dickscheid, and F. Schindler, "Detecting interpretable and accurate scale-invariant keypoints," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2256–2263, IEEE, 2009.

K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Computer Vision—ECCV 2002*, pp. 128–142, Springer, 2002.

K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA., 1967.

C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

J.-M. Geusebroek, G. J. Burghouts, and A. W. Smeulders, "The amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM, 1992.