ACTIVE OBJECT CATEGORIZATION ON A HUMANOID ROBOT

Keywords: Active vision; humanoid robot; object categorization.

Abstract: We present a Bag of words-based active object categorization technique implemented and tested on a humanoid robot. The robot is trained to categorize objects that are handed to it by a human operator. The robot uses hand and head motions to actively acquire a number of different views. A view planning scheme using entropy minimization reduces the number of views needed to achieve a valid decision. Categorization results are significantly improved by active elimination of background features using robot arm motion. Our experiments cover both, categorization when the object is handed to the robot in a fixed pose at training and testing, and object pose independent categorization. Results on a 4-class object database demonstrate the classification efficiency, a significant gain from multi-view compared to single-view classification, and the advantage of view planning. We conclude that humanoid robotic systems can be successfully applied to actively categorize objects - a task with many potential applications ranging from edutainment to active surveillance.

1 INTRODUCTION

Hand-held object manipulation by a humanoid robot is particularly challenging, because we cannot expect precise, reproducible hand trajectories, or perfect pose for a grasped object. This is in contrast to industrial robots with much higher precision, and also more computing power available. Furthermore, when we compare object recognition tasks in industrial environments in view of a limited variety of specific objects to the much more general task of object category recognition by a humanoid, it is evident that we need algorithms that can tolerate imprecise manipulations as well as significant ambiguity due to intra-class variability and inter-class similarity (cf. (Pinz, 2006)). We present a novel solution on a small humanoid robot (Nao), extending known concepts of active recognition and view planning of specific objects to the much harder task of active object category recognition.

The benefit of active multiple view-based techniques has been demonstrated for the recognition of specific objects from an object database a decade ago (Schiele and Crowley, 1998; Borotschnig et al., 1998). (Deinzer et al., 1998) focus on view point selection to reduce the number of steps needed through effective view planning (see also (Roy et al., 2004)). The accumulation of data from multiple views of an object is also commonly used in 3D object recognition schemes (Bustos et al., 2005). These systems are usually trained with all possible views of an object and provide a pose hypothesis along with the object hypothesis, when given a test image. The pose hypothesis is further used to plan the next best view to be chosen for effective recognition.

Object categorization is still a focus of current vision research (Pinz, 2006; Dickinson et al., 2009). But we do not propose a novel categorization algorithm in this paper. Our innovation is in the combination of the well known Bag of words (Sivic and Zisserman, 2003) concept for object categorization from individual object views with active acquisition of additional views, view planning, and fusion of individual view classification results using a Bayesian scheme. We adopt a view planning approach based on entropy reduction, similar to (Borotschnig et al., 1998). The entire algorithm is implemented on the slim computational platform of the Nao robot, and it can tolerate the robot's limited arm motion and grasp-

ing capabilities. An active approach for foreground feature separation makes the system robust to background clutter. Finally, we extend this scheme towards active categorization in the case of unknown pose of the grasped object.

2 THE NAO ROBOT

The Nao robot is a medium sized humanoid capable of grasping small objects and imitating human limb movements. Nao has an $X86$ AMD GEODE 500 MHz CPU with a 256 MB SDRAM and a functional LINUX Operating system. The robot is equipped with two CMOS 640×480 cameras. One camera is placed at its forehead and the other at its mouth level. The cameras facilitate visual observation of the object handed to the robot. The 5 DoF arm motion coupled with the 2 DoF head motion help manipulating the object view. The Nao hands are shaped like a pincer with three fingers, unlike a human hand. This pincer-like shape constrains the poses in which a rigid object can be held by the robot. The present work does not focus on autonomous grasping of objects by the robot. A human operator selects an object to be categorized and hands it to the robot. The object should be handed in a proper pose to maintain a firm grip during the arm motion. Once the object is held firmly by the robot, the arm and head are moved to pre-programmed positions to capture the desired views. An image of the Nao robot inspecting a toy horse is shown in fig. 1(a) and the robot's view is depicted by fig. 1(b). An image of the 36 objects of the 4-class toy object database used in the experiments is shown in fig. 1(c).

3 ACTIVE CATEGORIZATION

An object database may be composed of object classes sharing similar views. Active object categorization aims at disambiguating such object classes by considering different views of the same object. The movement to different view points is carried out in a planned manner to reduce the steps needed to attain the final result. In this paper, an active object categorization scheme based on the Bag of Words (BoW) approach is proposed for the Nao humanoid platform. This scheme further demonstrates the use of robotic arm movement to separate the foreground information from the image, thus removing the need for object segmentation.

The image of an object obtained by the robot at a particular view point may vary with the pose of the object held in the robotic hand. Hence, the image I of an object o_i held at a pose ϕ_k, obtained by the camera at a view point v_j can be represented as $I = f(o_i, v_j, \phi_k)$. However, in this section, the object pose is assumed to be a constant ϕ_c for all experiments, i.e., $I = f(o_i, v_j)|_{\phi_k = \phi_c}$. The case of varying object poses is discussed in Section 4.

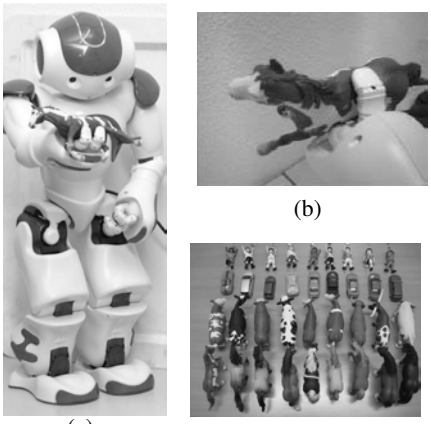

(a) (b) (c)

Figure 1: (a) Nao inspecting a toy horse, (b) Horse image obtained by Nao, and (c) Object database with 4 classes (toy horses, cows, cars, and soccer players).

3.1 The Bag of Words Scheme

The BoW model considers an image to be a collection of local features describing key patches in an image. In the proposed scheme, we use SIFT (Lowe, 2004) features to represent the key patches in an image. The SIFT feature represents each local patch as a 128-dimension vector. The patch can then be described by the descriptor and the position of the patch. The features obtained in an image are vector quantized to form visual words which describe the image. This quantization is achieved by clustering all the features obtained in the database into K clusters. The SIFT descriptors obtained in any image are then assigned to one of these clusters. The vector quantization in our scheme is carried out through hierarchal K-means clustering as suggested in (Zhang et al., 2009). The image is finally represented by a frequency vector denoting the frequency of different visual words in the image. The frequency vectors are in turn used to train and test suitable classifiers for object classification.

3.2 Elimination of Background Features

In object categorization systems, a strong segmentation scheme to separate the object from the back-

(a) (b) (c)

Figure 2: Foreground feature extraction: (a) SIFT features for object image with background, (b) SIFT features for background image, (c) SIFT features retained on the object image after foreground feature extraction.

ground preceding the categorization step is known to be advantageous. Especially the BoW is known to be a "model free" representation, unable to localize and separate the object (Pinz, 2006). In cluttered scenes, where many words might be found in the background and confusing the BoW scheme, foreground object separation gains more importance. This segmentation step can be circumvented, if the background information of the image is known a priori. In the proposed scheme, we exploit the advantage of the robotic arm movement to extract the background information from the image.

Once the object is grasped and the robot attains the desired view point, the object image I_{obj} is captured. Subsequently, the object is moved out of the camera's range by moving the robotic arm, keeping the camera position fixed. This enables the robot to capture the background image I_b without the object in the scene. SIFT features are extracted from both the images. A SIFT feature pertaining to I_{obj} is eliminated, if a similar feature is found within a 5 pixel radius in I_b. The similarity measure used in the proposed scheme is the Euclidian distance between SIFT features. For a given SIFT feature in I_{obj}, the best three matches in I_b are considered as similar features. This method is effective, if the background image remains static during the time taken by the robot to capture both images. The time taken to switch from I_{obj} to I_b is $2-3secs$, hence it is a reasonable assumption for most practical situations. Figure 2 shows an example illustrating the effectiveness of the scheme.

3.3 Detection from Multiple Views

The 5 DoF arm movement and 2 DoF head movement help the robot capture multiple views of the same object held in the hand. In each step of active object categorization, the robot attains a different view point. The robot moves on to the next view point after processing the data obtained in the current step. Let N_c be the number of classes in the database. The preprogrammed view points attainable by the robot are given by the set $V = \{v_i | i = 1,2,\ldots,N_v\}$, where N_v is the total number of views. In the proposed scheme, N_v different classifiers given by the set $C = \{C_i | v_i \in V\}$ are trained corresponding to each view. The classifier C_i is trained with a database of images obtained by the robot in the viewpoint v_i. Random forest classifiers (Bosch et al., 2007) are used in the experiments demonstrated in the paper.

Let us denote by I_i and $u_i \in V$, the image obtained and the view point attained by the robot in the i^{th} step respectively. Let g_i denote the frequency vector of foreground visual words in I_i . The probability that the object held by the robot belongs to the class o_j is given by $P(o_j|I_i)$. Let $C_k(o_j,g_i)$ denote the ratio of votes obtained in favor of o_j in the trees to the total number of trees in the classifier C_k, when tested with the frequency vector g_j. $P(o_j|I_i)$ is then given by Eq. 1. The overall probability $P(o_i|I_1,I_2,\ldots,I_n)$ after n steps is given as shown in Eq. 2.

$$P(o_j|I_i) = C_k(o_j,g_i),\ u_i = v_k \in V \quad (1)$$

$$P(o_i|I_1,\ldots,I_n) \propto P(o_i|I_1,\ldots,I_{n-1})P(o_i|I_n) \quad (2)$$

The view planning step determines the next best view point for object detection. View planning is carried out in the lines of (Borotschnig et al., 1998), by choosing the view point that is expected to result in the highest entropy change. The term $H(o_i|I_1,\ldots,I_n)$ denoting the entropy of the object class, when provided with images from the first n steps is defined in Eq. 3. If o_i is assumed to be the correct object hypothesis, the loss in entropy ΔH on choosing a view point v_r after the n^{th} step can be estimated as shown in Eq. 4.

$$H(o_i|I_1,\ldots,I_n) = -\sum_{o_i} P(o_i|I_1,\ldots,I_n) log(P(o_i|I_1,\ldots,I_n)) \quad (3)$$

$$\Delta H(v_r,o_i,I_1,\ldots,I_n) = H(o_i|I_1,\ldots,I_n) - \int_G P(g|o_i,v_r)H(o_i|I_1,\ldots,I_n,g,v_r)dg \quad (4)$$

Hence, for $v_r \in V$, the estimated entropy change $s_n(v_r)$ at the n^{th} step is given by Eq. 5. The view resulting in the maximum value of s_n is chosen as the next view.

$$s_n(v_r) = \sum_{o_i} P(o_i|I_1,\ldots,I_n)\Delta H(v_r,o_i,I_1,\ldots,I_n), \quad (5)$$

The integration carried out in Eq. 4 is over the entire space of feature vectors G. However, in order to speed up computation, the integration is limited to the

summation over training feature vector space. The probability and entropy terms under the integration in Eq. 4 can be further broken down as shown in Eq. 7 and Eq. 6 respectively.

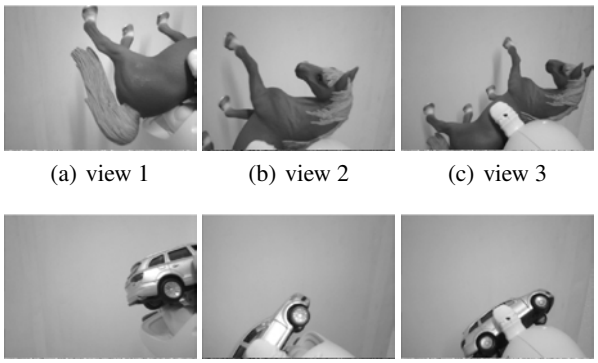
$$H(o_i|I_1,\ldots,I_n,g,v_r) = \sum_{o_k}\big[P(o_k|g,v_r)P(o_k|I_1,\ldots,I_n) \; log(P(o_k|g,v_r)P(o_k|I_1,\ldots,I_n))\big] \quad (6)$$

$$P(g|o_i,v_r) = P(o_i|g,v_r)\frac{P(g|v_r)}{P(o_i|v_r)} \propto C_r(o_i,g) \quad (7)$$

$P(o_k|g,v_r)$ terms are calculated offline for the training feature vectors to reduce computation time. In order to avoid moving to a view point more than once, the term $s_n(v_r)$ is set to zero for all previously visited view points. The object categorization algorithm is terminated when the highest probability for an object class exceeds a certain threshold or when all view points have been exhausted. The first view point is chosen randomly for all the experiments.

3.4 Multiple View Experiments

The system presented in the paper cannot be evaluated on a pre-existing image database since the evaluation is an online process requiring the objects to be handed to the Nao robot. Hence, our scheme is tested on a database developed in the lab containing 4 object classes: toy 'horses', 'cows', 'cars', and 'soccer players'. The object database has 9 objects per class, amounting to 36 objects in total as seen in Fig. 1(c). The results in this section are presented by considering a maximum of 5 different pre-programmed views for an object. Sample images of objects belonging to the classes 'horse' and 'car' are shown for three different views in Fig. 3. The database is composed of object classes sharing similar appearances along certain views, like the 'horse' and 'cow' classes. As seen in fig. 4, the objects belonging to classes 'cow' and 'horse' may have a very similar torso and appearance. The two classes are hence separated by a small inter-class distance. Also, objects within the same class may exhibit considerable diversity in appearance, as evident from the two 'horse' images in Fig. 4. All the experiments are performed on grayscale images. The experiments in this section deal with objects grasped at a fixed pose during both the training and testing phase. The training in all the experiments is performed offline, using the images obtained by the Nao robot for different views of the object database. A visual vocabulary of 500 visual words is used in the experiments.



(a) view 1 (b) view 2 (c) view 3

(d) view 1 (e) view 2 (f) view 3

Figure 3: Sample images from different views for the object classes 'horse' and 'car'.

Figure 4: Sample objects from the classes 'horse' and 'cow'. This example shows the difficulty of handling rather small inter-class distances (cow vs. the horse in the middle) and high intra-class variability (the two horses).

The effectiveness of the object classification system is demonstrated by using the 'leave one out' cross validation scheme. The training and testing is carried out 36 times, where each time a different sample in the object database is left out from the training set and is used to validate the system. Owing to the high computational expense of the 'leave one out' scheme, the experiments are carried out on an external Intel(R) Core(TM) *i*7, 2.67 GHz processor. However, the proposed scheme was implemented on the Nao robot and tested for individual object samples, with identical, but slower behavior. The percentage accuracy for 5-view object categorization was found to be 78.88% using the 'leave one out' validation scheme, compared to only 50.00% for single-view categorization. The change in processor is not expected to cause any difference in terms of classification performance. The average time taken by the robot (running on its original *x*86 AMD GEODE 500 MHz processor) if made to run through all 5 view points is found to be 119.17 secs. On average, 23.83 secs are spent for analyzing a view point. A major chunk of the time is spent in computing the SIFT features. The robot outputs intermediate results after analyzing a given view point. The intermediate outputs are the probability distribution for the best three classes at every stage. This is done by voice output after analysing images obtained from each view; for example the robot says

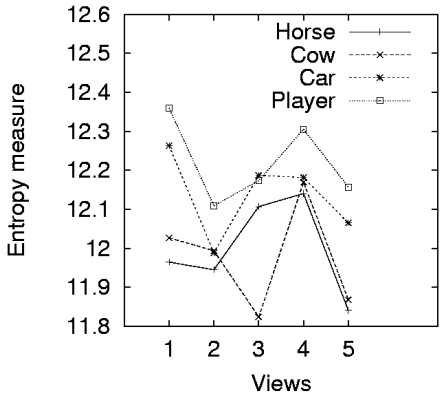


Figure 5: Entropies for different views of each object class.

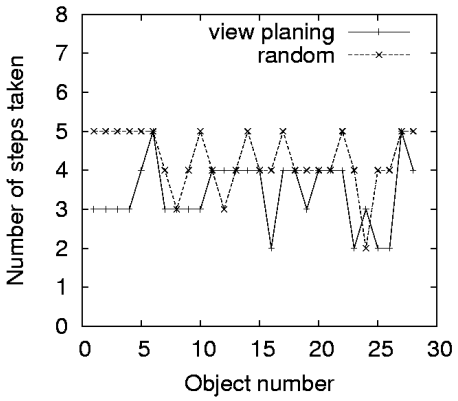


Figure 6: A comparison of the number of steps taken for different objects, between a random view and planned view approach. The view planning approach produces the result in fewer steps for most objects, hence indicating a faster performance.

"45% horse, 30% cow, 15% car".

Figure 5 shows a plot of the average entropy for each view and for all the object classes. It can be seen that some views are more informative than the others for a given class of objects; thus, justifying the need for a multi view approach. The effectiveness of a view planning approach in reducing the number of steps taken by the robot to reach the final result is demonstrated in Fig. 6. The average number of steps taken for each object is plotted for the view planning approach and a random view approach (where the next best view at any step is chosen randomly). A probability threshold of 60% was chosen for the experiments. The view planning based approach is seen to produce the results in fewer number of steps for a large number of objects.

The system was also tested with object classes other than toys. An object database containing forks and pens (containing 9 objects each) in addition to the above mentioned 4 toy categories was also used to test the system. The classification accuracy for the 'leave one out' cross validation scheme was found to be 61.11% among 54 objects (five-view object categorization), and only 27.78% (single-view).

These results provide significant quantitative evidence for the benefit of the fusion of multiple views, as well as for the effectiveness of the view planning.

4 TOWARDS UNKNOWN POSE

Many active object recognition systems (Deinzer et al., 2003; Schiele and Crowley, 1998; Borotschnig et al., 1998) provide a pose hypothesis along with an object hypothesis. In these systems, the different object poses can be easily quantized and the system can be trained with any number of object views. However, in a humanoid unlike an industrial robot, the allowed poses for a rigid object vary depending on the object shape due to the pincer-like hand constraint. Hence, the task of precise pose estimation for an object held by a humanoid may not be very fruitful. In this section, we try to overcome the fixed pose constraint by a probabilistic approach where, given an object image, a probability distribution for the possible views is obtained. Note that the system is still trained with a fixed set of object views with object pose remaining consistent throughout the training phase.

4.1 Probability Distributions for Views

In order to obtain a probability distribution for the possible views of a given image, we take into account the positional information of the SIFT features as well. Let γ represent the combined information of the positional information along with the visual word frequency vector g for an image I. The probability $P(o_i|I)$ is then given as shown in Eq. 8.

$$P(o_i|I) = P(o_i|\gamma) \quad (8)$$

$$= \sum_{v_j \in V} P(o_i, v_j|\gamma) \quad (9)$$

The term $P(o_i, v_j|\gamma)$ can be represented as shown in Eq. 10. This term represents the probability that the newly obtained object image is similar to the view of an object belonging to class o_k seen from the view point v_j when it is held in the same pose as that used for training.

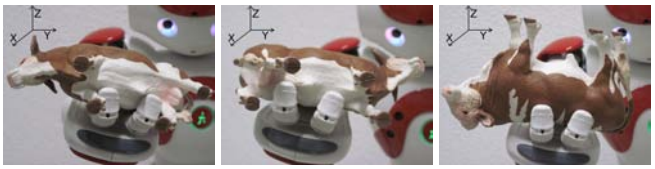
$$P(o_i, v_j|\gamma) = P(o_i|\gamma, v_j)P(v_j|\gamma) \approx P(o_i|g, v_j)\sum_{o_k} P(v_j, o_k|\gamma) \quad (10)$$

The term $P(o_i|g, v_j)$ can be obtained from the random forest classifier trained for the view point v_j as $P(o_i|g, v_j) = C_j(o_i, g)$. In order to estimate the second term in Eq. 10, we use an approach similar to the Implicit Shape Model (ISM) scheme (Leibe et al., 2004). We assume that the center of the object always coincides with the center of the robotic hand. This assumption is reasonable, since the object has to be held firmly during the robotic arm motion. The position of the hand is pre-programmed for a given view point and is fixed for a view point. Hence, the position of a reference point ("the hand of the robot") is known in every image, since the view point of an image is known to us. This enables us to find the distance of the various visual words in any image from the center of the object (assumed to be the hand position) in the image. During the training phase, we store the distance of individual visual words in an image from the reference point in the image. These values are stored in a codebook for each class of objects for every view point $v_j \in V$. Hence, we have N_c codebooks for every classifier/view point. Let us denote by CB_{o_i,v_j}, the codebook corresponding to the class o_i and view point v_j. The distance values corresponding to a visual word in the code book are clustered to decrease the memory storage required on the robot.

When a new image is obtained by the robot, the distance of the visual words from the reference point in the image is calculated as before. These distances are compared with the previously stored distances in the codebook CB_{o_i,v_j} . For every visual word in the image, a vote is issued in favor of the class-view point combination $\{o_i, v_j\}$, if the distance is within a certain range Δd of any of the stored distances for that visual word in CB_{o_i,v_j}. Let us denote by $\psi(o_i, v_j|\gamma)$ the number of votes issued in favor of a class-view point combination $\{o_i, v_j\}$ for the image I. $P(v_j, o_k|\gamma)$ can then be deduced as shown in Eq. 11.

$$P(v_j, o_k|\gamma) \propto \psi(o_k, v_j|\gamma) \quad (11)$$

The above described scheme effectively provides a probability distribution based on the hypothesis that similar views of objects belonging to the same class would have similar spatial distribution of visual words with reference to an object center. This hypothesis has been verified in the ISM scheme as shown in (Leibe et al., 2004). This scheme does not provide a view planning step since a definite pose hypothesis is not acquired for a given image.



(a) (b) (c)

Figure 7: The three different poses used in the experiment are shown for a toy cow (view point remains unchanged for all the images). (a) $Pose_{train}$, (b)$Pose_{180}$ and (c) $Pose_{90}$

Object Pose	% of objects correctly categorized
$Pose_{train}$	72.22%
$Pose_{180}$	61.11%
$Pose_{90}$	55.56%

Table 1: The percentage of correctly classified images using the proposed scheme for 3 different poses including 2 previously unseen poses.

4.2 Unknown Object Pose Experiments

In this section, experiments are presented to demonstrate the effectiveness of the proposed technique for active object categorization with unknown object pose. The scheme is tested on the 4-class object database containing cows, horses, cars, and soccer players. Results are shown for three different object poses including two poses which are different from the pose used for training the robot. Due to limitations on the poses in which certain rigid objects can be held by the robot, results are shown only for poses applicable to all object categories. Two object poses $Pose_{90}$ and $Pose_{180}$ which are different from the training pose $Pose_{train}$ are obtained by rotating the object 90° about the $X-axis$ and 180° about the $Z-axis$ respectively, starting from the original training pose as shown in Fig. 7. The results for the 'leave one out' cross validation experiment are shown for the database with 36 objects in Table. 1. It is to be noted that the effectiveness of the scheme would reduce, if the pose of the object is such that the images obtained from some view points are vastly different from any of the training images of objects in its category. This effect can be seen in the case of $Pose_{90}$, where the images obtained show more variation from the training images than $Pose_{180}$.

5 CONCLUSION

We have presented a study on active object categorization on a humanoid robotic platform. The nov-

elty of our approach is twofold: First, we demonstrate the viability of active recognition of categories by generalizing previous concepts on active recognition of specific, individual objects. Second, we use the particular abilities of the humanoid arm to efficiently eliminate background features. Compared to categorization from a single view, our experimental results clearly demonstrate a significant gain in correctly categorized test objects by this 'active' approach. Furthermore, view planning can reduce the number of active steps needed to produce the final categorization result. Finally, the overall computational complexity of object categorization is significantly reduced by the integration of various views so that we can hope to see applications on rather slim computing platforms like Nao in the near future. There certainly is an extremely high potential for this kind of hand-held, active inspection of objects by a humanoid robot, including human-robot interaction, home and service robotics, edutainment, active inspection, and active surveillance.

REFERENCES

Borotschnig, H., Paletta, L., Prantl, M., and Pinz, A. (1998). Active object recognition in parametric eigenspace. In *British MAchine Vision Conference*.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classification using random forests and ferns. In *International Conference on Computer Vision*.

Bustos, B., Kein, D., Saupe, D., Schreck, T., and Vranic, D. (2005). Feature-based similarity search in 3d object databases. *ACM Computing Surveys (CSUR)*, 37(4):345–387.

Deinzer, F., Denzler, J., Derichs, C., and Niemann, H. (1998). Integrated viewpoint fusion and viewpoint selection for optimal object recognition. In *British MAchine Vision Conference*.

Deinzer, F., Denzler, J., and Niemann, H. (2003). Viewpoint selection - planning optimal sequences of views for object recognition. In *Computer Analysis of Images and Patterns*, pages 64–73. Springer Berlin / Heidelberg.

Dickinson, S., Leonardis, A., Schiele, B., and Tarr, M., editors (2009). *Object Categorization*. Cambridge University Press.

González, E., Adán, A., Batlle, V. F., and Sánchez, L. (2008). Active object recognition based on fourier descriptors clustering. *Pattern Recognition Letters*, 29(8):1060–1071.

Leibe, B., Leonardis, A., and Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Pinz, A. (2006). Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4):255–353.

Roy, S., Chaudhury, S., and Banerjee, S. (2004). Active recognition through next view planning: A survey. *Pattern Recognition*, 37:429 – 446.

Schiele, B. and Crowley, J. L. (1998). Transinformation for active object recognition. In *International Conference on Computer Vision*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477.

Zhang, S., Tian, Q., Hua, G., Huang, Q., and Li, S. (2009). Descriptive visual words and visual phrases for image applications. In *Proc. ACM Int. Conf. on Multimedia*, pages 75–84.