Towards Robust Object Categorization for Mobile Robots with Combination of Classifiers

Christian A. Mueller, Nico Hochgeschwender, and Paul G. Ploeger

Bonn-Rhein-Sieg University of Applied Sciences, Germany
{christian.mueller@smail.inf.,nico.hochgeschwender@,paul.ploeger@}h-brs.de

Abstract. An efficient object perception is a crucial component of a mobile service robot. In this work we present a solution for visual categorization of objects. We developed a prototypic categorization system which classifies unknown objects based on their visual properties to a corresponding category of predefined domestic object categories. The system uses the Bag of Features approach which does not rely on global geometric object information. A major contribution of our work is the enhancement of the categorization accuracy and robustness through a selected combination of a set of supervised machine learners which are trained with visual information from object instances. Experimental results are provided which benchmark the behavior and verify the performance regarding the accuracy and robustness of the proposed system. The system is integrated on a mobile service robot to enhance its perceptual capabilities, hence computational cost and robot dependent properties are considered as essential design criteria.

Keywords: object categorization, Bag of Features, feature extraction, clustering, machine learning, classifier combination.

1 Introduction

Due to the social phenomenon of a greying society it can be expected that elderly people will be assisted by service robots in their everyday activities at home more and more. In many cases objects play a central role in these activities and thus a service robot must be able to detect and classify known and unknown objects in such domestic environments. Most of the current approaches for object recognition have two characteristics, namely firstly the perception is instance based and secondly the instances of the objects have to be known in advance via some teach-in process. Yet the identification of a particular object instance is not necessarily required or sometimes even not feasible, e.g. serving and delivering drinks require the detection of glasses as such, the recognition of a particular glass is often not needed. It would be very wasteful or even infeasible to teach this large set of individual glass instances in the individual home when just any element from the category glass will do the job. This shift of focus may also be illustrated in the development of the current rule set in the *RoboCup@Home* competition,

T. Röfer et al. (Eds.): RoboCup 2011, LNCS 7416, pp. 137–148, 2012.

which is a well-established international benchmark for service robots in domestic environments [9]. For instance, in the recently established *supermarket* scenario, the service robot should fetch e.g. a cup – as a missing item of a shopping list – from a shelve of the supermarket. Here a object categorizing capability is needed since any cup will resolve the problem.

The presented work develops an object perception system that categorizes unknown domestic object instances like cups, glasses, bottles or cell-phones in their respective category. Our system addresses robustness and reliability in several ways. It can cope with object categorization challenges like perspective variations due to object-rotation-angle and robot-object-distance variations or intra category variations i.e. deformations of object instances of a common category. The system extracts expressive visual properties of example objects and *generalizes* these visual properties according to their respective object category in order to categorize *unknown* objects. In the remaining paper we discuss the related work and our contribution, followed by the requirements and assumptions we made. Further on the components of the system are explained followed by the evaluation. Finally, a discussion and conclusion are given.

2 Related Work and Contribution

The presented work is grounded on 2D image information. Two common approaches are available, geometric- and geometric-free based approaches. *Geometric-based* approaches work on the global geometric appearance of objects. These approaches rely e.g. on shape models or shape descriptors. They show robustness to strong shape deformations. Hence, they provide the capability to make a reliable decision about a corresponding shape class; however a reliable decision about the shape class of an object is not sufficient for object categorization purposes, since objects from *different* object categories with *similar shapes* might fall into the same shape class and thus become indistinguishable. Geometric-based approaches demand a precise image segmentation algorithm, since precisely extracted object boundaries are required as input for an optimal performance. Sophisticated image segmentation algorithms are often computationally expensive ($\gg$*1sec*), and therefore unattractive to be applied to a mobile service robot where a short-response-time is required.

Consequently, our work relies on a *geometric-free* approach (see Fig. 1): called *Bag of Features*(*BoF*) [1,6,8]. This approach has shown its reliability and robustness to object occlusions, illumination changes and especially to geometric deformations of objects which belong to a common category, since the *BoF* approach does not rely on global geometric information; instead it relies on the extraction of local invariant features. Categories which are hardly distinguishable by shape-based approaches, become distinguishable (e.g. cups and glasses) since the extracted features provide information about the structure and texture of objects. The *BoF* approach is based on the assumption that each object category is distinguishable by its individual independent statistical appearance of salient-invariant-local features which are extracted from images. The idea is

by comparing the features of a query object to the distribution of those features from previously analyzed objects, to infer the category of the query object. A so-called *visual dictionary* of generalized features is generated based on extracted features from a training set. These generalized features are expressive features which provide a high discriminability regarding the categories. The extracted features of a query object are mapped to these generalized features; the generalized features can be seen as visual words in the visual dictionary whose presence or absence lead to a decision about the actual category.

In the first step of the *BoF*-based object categorization process, an *extraction of invariant features* from images is exploited to transform the visual image information into a compact representation, which provides rich recallable information of the image, i.e. if the image content is transformed by scale, shift or rotation, similar information are extracted. Commonly Scale-Invariant-Feature-Transform (*SIFT*) has been often successfully applied [5]; however our experiments have shown that Speeded-Up-Robust-Features (*SURF*) performs a better feature extraction, due to its feature recallability and computational cost. Next, the *visual dictionary* is created, which analyzes the feature frequencies for a set of images that have passed the feature extraction process. Therein, the features are grouped by similarity, in order to generate clusters of similar features. Based on a cluster, a generalized feature is constructed which represents a *visual word*. Mostly k-means-based algorithms are applied for clustering due to its simplicity and low computational cost [5,8,1]. Other contributions group the features e.g. by randomized cluster trees [7] or mean-shift clustering [6]. After the dictionary is generated, the extracted features of a query image are assigned to the nearest visual words by, e.g. nearest-neighbor-search. The comparison between the visual word frequencies, i.e. distribution of the visual words, of a query image and of labeled example images leads to a decision about the corresponding category of the query image. Often supervised machine learning approaches like Support Vector Machines(*SVM*) are applied [1,8], since they have shown an enhanced robustness to discriminate sets of categories. The learners are trained with the visual word frequencies of examples objects to generate a *prediction model*.

In our work we do not rely on the decision of a single classifier, since a single classifier provides a certain accuracy and also a high risk of a misclassification bias for specific categories. To enhance the accuracy and to reduce the influences of those biases, a set of classifiers is trained and their outcomes are *combined* to make a more *robust* and *reliable* decision about a category. Additionally the performance of each classifier is improved by different *feature-selection algorithms*. Moreover our approach does not completely neglect the object shape information, since it provides a useful indication about a corresponding category. We combine the set of feature-based classifiers with an additional *shape-based classifier* in order to support an appropriate final decision. An appropriate number of clusters (*dictionary size*) is a crucial factor which influences the categorization performance. The discriminability is decreased if a too small or too large dictionary is used; in both cases the efficiency of the dictionary is negatively influenced. Most approaches heuristically examine the dictionary size or they set

the dictionary size to a fixed number [8,7]. In contrast, we systematically analyze the dictionary size by the *examination of structure and relation* of the generated visual words to each other, in order to generate a discriminative dictionary; also the *importance* and *relevance* of each visual word is analyzed with respect to the object categories. In many approaches[1,8,5] the entire image is examined as a single entity. Image distortions like cluttered backgrounds, which do not contribute meaningful information, influence the categorization result; moreover multiple object occurrences are generally not considered. We show a basic, but sufficient approach which allows to *detect multiple objects* in an image by an image segmentation algorithm; afterwards the detected objects are classified by the feature- and shape-based classifiers. Most object categorization approaches are stationary and less concerned about the response time. Our system is integrated on a mobile service robot, hence issues like the computational cost or the robustness towards the large variety of perspective variations to objects have to be considered in the development of the system. Further a monocular camera is applied, rather than a cost intensive Time-of-Flight- or Stereo-Vision-Camera.

3 Requirements and Assumptions

The object categorization system acquires 2D images from the camera as input. The camera is mounted on the top of the robot and provides images from a typical service-robot-height of about *100-120 cm* in a resolution of *640×480* pixels. The system is trained for a robot-object distance of about *30-40 cm*. The robot camera has to be focused on a plane surface, as for instance a table. The system is supposed to perceive the objects on the table top and to classify them to a supported object category; up to four different object categories (cup, cell-phone, bottle, glass) are supposed to be classifiable. We verify the system performance and behavior using these four categories in certain constellations; however our system can easily be *extended with additional categories*. Due to the current object detection approach (see section 4.4), the background is assumed to be less cluttered and the objects are mainly present on a uniformly colored surface. The objects are positioned – completely visible and not occluded – in a reasonable distance from the camera and to each other. The system is applied under artificial light conditions. Further on, an *object database* is required in order to supply the system with sufficient object-related-information for an efficient prediction-model generation. It consists of a set of images of each category which is used as training set. Each image contains a single object, thereby different images are taken from different randomly chosen positions and orientations, and on varying uniform backgrounds. Also a test set of images for each category is provided. Both sets are mutually exclusive in terms of particular object instances.

4 Object Categorization with Bag of Features

The system is divided into two phases, namely training phase and evaluation phase. These phases and the involved components with their parameters are depicted in Fig. 1 and discussed in the remaining section.

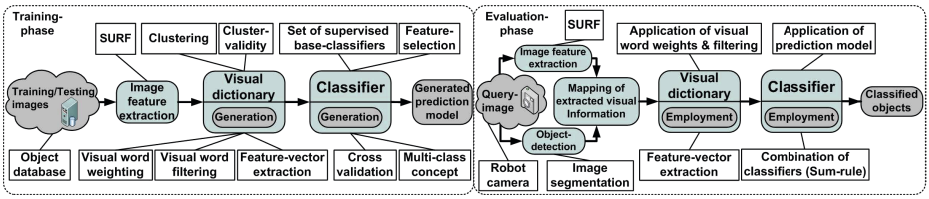


Fig. 1. Illustration of the system design. The components shown, are involved during training- and evaluation phase.

4.1 Image Feature Extraction

Firstly, during training phase features are extracted from images of the provided object database. Note that, only object-related features are extracted since a single object is positioned on a uniform background. During evaluation phase the features are extracted of a query image which is acquired of the robot camera. In this case, object detection has to be applied to find features related to each presented object in the query image (see section 4.4).

The extraction (detection and description) of the features is performed by *SURF*: in the given conditions of our application *SURF* shows a up to *4%* lower classification error and it tends in average to a faster feature extraction than *SIFT*. Each (*SURF*-)feature is described by a vector of *64* elements, moreover up to *120* features per object are extracted since they have proofed to provide sufficient object-related-information for further processing purposes. After the extraction of the features from each object of the object database and the query image, the extracted feature set of each object and query is called *BoF*.

4.2 Visual Dictionary Generation and Employment

During training phase the visual dictionary is generated from the set of *BoFs* of the *training set*. The visual dictionary provides a representation which is able to describe - in a compact and efficient way - the visual information. The dictionary contains a set of visual words, which are generated through grouping the features from the *BoFs* by similarity. The fast k-means clustering algorithm is applied to group and find similarities in the extracted features. In succession of the grouping of the features, the center of each group represents a *visual word*. The goal is to find a discriminative dictionary: the appropriate k(*dictionary size*) is determined by the *Dunn-validity-Index*[3]. Thereby in our experiments the dictionary size is varied from *100*(min) to *1000*(max) visual words with an increment of *10*; the *validity value* and the *classification accuracy* of each dictionary size is examined. An indication of an appropriate dictionary size which leads to an enhanced classification performance, is found by the identifications of local maxima of the global *validity values*. That *local maximum* whose corresponding dictionary size leads to the lowest classification error is selected. In addition, a modified *soft-assignment weighting scheme*[5] based on the examination of the feature frequency related to particular visual words is applied, in order to give

particular visual words less or more importance in the categorization process. Also a *filtering scheme* of less informative visual words is applied by first ranking ascendingly the visual words according to their frequency proportions to each object category, and later neglecting the lower ranked words.

After the generation, the dictionary is employed to the *BoF* of objects from the object database (*training and test set*) and of *query objects*. Through the dictionary employment, each *BoF* is represented as a composition of the previously generated visual words in a histogram which is also denoted as *feature-vector*. Thereby the visual word occurrences in each *BoF* are examined by the nearest-neighbor-search: the nearest visual word to each feature is determined.

4.3 Supervised Classifier Generation and Employment

Through the representation of each object as a *feature-vector*, the categorization problem has been converted to a pattern matching problem which we handle as a machine learning problem. Hence, efficient and powerful techniques from the machine learning field are exploited.

As a preprocessing step to enhance the quality of the feature-vectors, *feature-selection algorithms*[4] are applied to identify discriminative features of the vectors during *training phase* and filter out those features during *evaluation phase*. Three filters are applied, based on *Principle Component Analysis* (*PCA*), *Entropy* and *Iterative Adaptive Feature Selection* (*IAFS*); *IAFS* iteratively adds a feature from a ranked set of features[1] and trains accordingly a classifier; if an improvement is achieved, a new feature from the set is selected, else the last added feature is removed, and replaced with a new feature.

We experimented with two popular supervised machine learning techniques, namely *SVM* and *AdaBoost* which we combined with the feature-selection algorithms in order to generate a pool of classifiers. Six (base-)classifiers[2] were defined which are *independently* trained during the training phase by the generated feature-vectors of the training set in order to learn a *prediction model*. The accuracy is verified by a *10-fold cross validation* and the test set. Since the categorization problem is a multi-class problem, we decided to apply the majority-voting-strategy-based *One-vs.-One* multi-class concept, due to its misclassification recovery property. We also experimented with one extra *shape-based* base-classifier to support the final category decision. It is trained with feature-vectors which are based on shape descriptor results[10] of extracted contours[3] of objects from the object database: to each object contour descriptors are applied as statistical moments, Hu-moments, direction, eccentricity, normalized central moments, spatial moments, contour area, contour length, and Fourier descriptor.

During the evaluation phase, a set of the trained base-classifiers is employed to a feature-vector of a query image; the constellation of the set is evaluated regarding its classification accuracy (see section 5). The classification outcomes of the

[1] The features are ranked by their discriminability through feature-selection algorithms like *PCA* or *Entropy*.

[2] In Table 1 of section 5 the actual set of base-classifiers is listed.

[3] The contours are extracted during the object detection process (see section 4.4).

base-classifiers are combined with an extended *sum-rule*[2], in order to classify reliably the feature-vector to an appropriate category and to be robust to misclassification of base-classifiers. The *sum-rule* is extended with four *weighting-factors* which are applied to the outcomes: firstly, the outcome of each base-classifier is weighted by the base-classifier's *classification accuracy*(*self-confidence*), which is gathered during training phase. Secondly, each outcome is weighted by the *confidence in the actual outcome* of the base-classifier. Thirdly, biases of correct classifications and misclassifications towards particular categories of each base-classifier are considered. Hence, a *penalty to each outcome* is applied according to the biases of the base-classifier. Fourthly, *classification majorities of the categories* from the outcomes are weighted in order to overcome misclassifications by base-classifiers. The sum-rule involves two factors: an implicit *voting* to the most probable outcome and giving a higher importance - by *weighting* - to accurate base-classifiers and less importance to less accurate base-classifiers.

4.4 Multiple Object Detection

During the evaluation phase a query image of the robot camera is acquired. The system has to classify single or multiple occurrences of object instances. Initially the features of the entire image are extracted. However the entire set of features cannot be classified like a single object as during the training phase, because of the probable presence of cluttered background or multiple object occurrences. Hence, the detection of potential objects in the given image and the correspondence of features to the respective object is needed to be determined. An accurate image segmentation algorithm is not the scope of our work and is not necessary, since only regions of interests of potential objects are required to be found. A sufficient segmentation algorithm based on contour extraction is

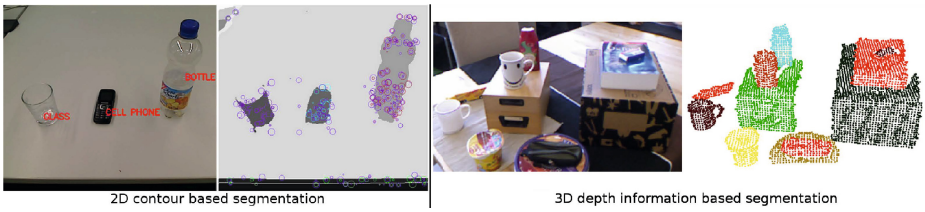


Fig. 2. Left: object detection result is shown at a distance of $\approx 30cm$ to the robot with the extracted object boundaries and detected features. Right: 3D depth-based detection result is shown (point cloud is randomly colored for each detected object).

applied to gather the boundaries of potential objects in the image and to map the corresponding features to the object boundaries as shown in Fig. 2(left). Alternatively, a detection based on 3D depth information is under development. This approach segments objects on different planes in cluttered and occluded environments with a higher reliability compared to the contour-based approach, see

Fig. 2(right). The single plane extraction is based on surface normals extraction and *RANSAC*(*Random Sample Consensus*) plane fitting. Afterwards, a hierarchy of different heights of overlapping planes is created. Potential objects are extracted via grouping points above each plane by euclidean clustering. Due to the infancy of this approach the following results are based on the contour-based approach.

Afterwards the visual dictionary is employed to the extracted features from each object boundary. So, the content of each object is projected to a feature-vector which is classified to an appropriate category by the employment of the trained classifiers whose outcomes are combined to a final decision.

5 Experimental Evaluation

The following first part of the evaluation is based on the training- and test set of the object database. Our experiments have shown that *190* images as training set and *55* images as test set, deliver sufficient object category related information for an efficient prediction model generation.

The identification of an appropriate dictionary size leads to an enhanced classification accuracy: Fig. 3 shows the visual word dedications with respect to the categories, if an appropriate dictionary size is chosen. The top-ranked *20%*

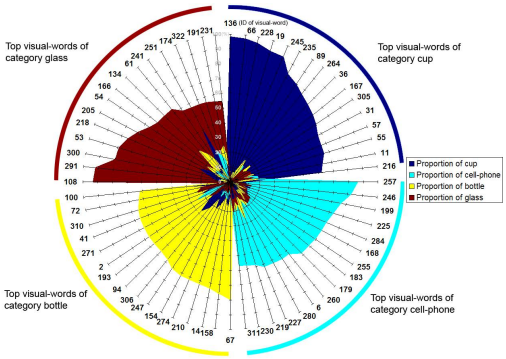


Fig. 3. The top discriminative visual words with their proportions of frequency regarding each category (*16* selected visual words of each category). The visual words are sorted in ascending order of their proportion in the respective dedicated category. E.g. in case of the cup category: the most discriminative visual word is the one with *ID-136* which has a proportion of *97.7%* in the cup category, *0.8%* in the cell-phone category and *1.4%* in the bottle category.

visual words (*16* selected visual words of each category) are shown of a dictionary which contains in total *325* words and supports the four categories (cup, cell-phone, bottle, glass). The selected set of *16* visual words for each category

does not intersect with the sets for the other categories; also the top *20%* visual words are strongly dedicated to a specific category. Even lower ranked visual words from the top *20%* show a dedication of ≈*50%* to a certain category, which is still discriminative, since this proportion has the majority; in that case not a single proportion of another category reaches a level of more than *30%* – with one exception: visual word with ID-*251*.

Further on, Table 1(green) shows that dictionary sizes indicated by the local maxima of the Dunn-validity-index values provides in average a *lower classification error* than randomly chosen dictionary sizes: an appropriate dictionary size with the lowest classification error has been found for *2*-categories with *270* words, respectively for *3*-categories *400* words and for *4*-categories *325* words. These dictionary sizes are chosen as the base-classifiers return the lowest

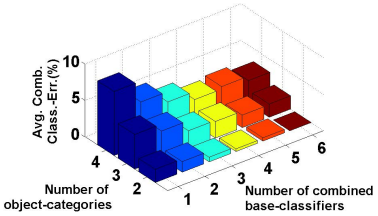
Table 1. The average classification error (%) regarding the test set is shown of each base-classifier which is trained with randomly chosen and Dunn-validity-index indicated dictionary sizes. The classification error in the brackets shows the error if an appropriate dictionary size is chosen: 2-cat.=*270* words, 3-cat.=*400* words, 4-cat.=*325* words.

Base-classifier approach	Number of supported object categories					
	2		3		4	
	Rand.	Dunn.	Rand.	Dunn.	Rand.	Dunn.
SVM	1.91	0.23(0)	5.14	2.04(2.4)	8.01	6.65(5.9)
SVM+Entropy	1.71	0.45(0)	4.84	2.29(1.2)	7.38	6.28(2.7)
SVM+PCA	1.91	0.67(0.9)	3.93	2.78(1.2)	6.73	5.87(4)
AdaBoost	5.34	5.65(3.6)	7.78	7.50(9)	15.1	13.17(12.7)
AdaBoost+PCA	3.62	2.49(2.7)	7.27	7.59(7.8)	10.98	9.30(9)
AdaBoost+PCA+IAFS	4.23	3.17(2.7)	6.66	6.18(6)	10.37	9.85(9.5)

classification error compared to other dictionary sizes of their respective number of supported categories. Note that, the dictionary size *does not increase proportionally* with the increase of supported categories: the discriminability between categories is a decisive factor that affects the dictionary size.

In the following, the *combination* of the feature-based base-classifiers is evaluated. In the experiment, all 63^4 combinations of the six base-classifiers are analyzed (see Fig. 4). The plot in Fig. 4 illustrates that the addition of base-classifiers *generally improves* the classification accuracy, in other words the average classification error of four supported categories has dropped from *8.7%* (single classifier) to *3.1%* (six combined classifiers). Further, in this case of a 4-category trained system, the combination of three base-classifiers, namely *SVM*, *SVM+Entropy* and the *AdaBoost+PCA* leads to the lowest classification error of *2.2%* (see in Fig. 4 right-side), i.e. the application of a *particular sub-set* of the six base-classifiers leads to a *-0.9%* decreased classification error compared to the classification error if *all* six base-classifiers are combined. Also the combination of those three classifiers shows a lower classification error than the *most accurate single base-classifier* (*SVM+Entropy=2.7%* – see Table 1 in brackets). The same behavior is observed for 2- and 3-category trained systems. However in case of the

[4] In total *63* combinations of *6* base-classifiers: summed by the number of combinations of one(*6*), two(*15*), three(*20*), four(*15*), five(*6*) and six(*1*) applied classifiers.



Num. of supported categories	Base-classifier comb. with the lowest classification error	Error(All classifiers combined)
2	SVM, S.+Entropy, S.+PCA, AdaBoost, A.+PCA+IAFS	0(0)
3	S.+Entropy, S.+PCA, AdaBoost	0.6(1.8)
4	SVM, S.+Entropy A.+PCA	2.2(3.1)

Fig. 4. Left: average classification error according to the number of base-classifiers which are combined and the number of supported categories. Right: combinations of base-classifiers which result to the lowest classification error(%) of test set.

2-category trained system the combination of all six classifiers has been chosen, which lead to the same accuracy than the most accurate single classifier(0%); this combination has been chosen, due to the robustness against misclassifications. It is worth to mention, that the single *AdaBoost* base-classifier which has the *lowest accuracy*, still contributes to the combination with the *lowest combined classification error* in case of 3-categories (see in Fig. 4 right-side). These results show that our method of combining sets of particular base-classifiers, effectively reduce the classification error.

The shape-based classifier has shown reasonable results for distinctive categories regarding the object shape, e.g. in system configurations where cups and cell-phones (*5%*) or cups, cell-phones, and bottles (*10%*) are categorized; when additionally glasses are involved (4-categories), the classification error increases (*27.5%*) due to shape similarities between instances of the categories like of cups and glasses. We conclude, that this classifier is helpful to support the combined classification but it is a matter of future work to enhance it.

In the following evaluation, the trained system is treated as a black-box: images of the robot camera are acquired and evaluated. The Table 2 presents

Table 2. The classification accuracy (%) regarding the four categories. The system is trained to support 2-, 3-, and 4-categories. 10 objects of each category are involved in this experiment. (x=category is not applied in respective system configuration).

Actual Category	Number of supported object categories											
	2				3				4			
	Cup	Cell-ph.	Bot.	Gla.	Cup	Cell-ph.	Bot.	Gla.	Cup	Cell-ph.	Bot.	Gla.
Cup	95	5	x	x	92.2	5.6	2.2	x	96.7	1.7	0.8	0.8
Cell-phone	3.3	96.7	x	x	4.4	95.6	0	x	3.4	95.8	0.8	0
Bottle	x	x	x	x	7.8	0	92.2	x	8.3	0	91.7	0
Glass	x	x	x	x	x	x	x	x	2.5	7.5	0	90

the classification accuracy regarding the four categories: certain misclassification biases for particular categories are observed. Further we investigated the categorization behavior depending on the *robot-object-distance* and *object-rotation-angle* – see Fig. 5. The classification accuracy among the supported categories behaves differently under the same experimental setup i.e. *object-robot-distance* and *object-rotation-angle*. Several factors are responsible for this behavior:

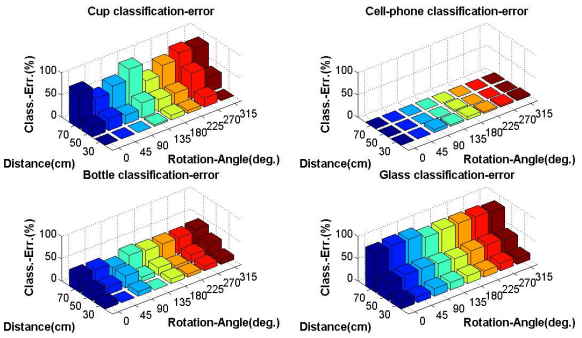


Fig. 5. The classification error results of each category with respect to object-robot-distance and object-rotation-angle. *10* objects of each category are involved in this experiment. The system is trained to support *4* categories.

robot-object-distance implies that the farther the object is positioned the fewer descriptive features are extracted which lead to an increase of the classification error; the *object-material* with respect to the contrast between the object and the background, can increase the classification error because of a weak object detection in farther distance like in the case of glasses (partial transparency). Also the *object-rotation-sensitivity* of object influences the classification accuracy. A sensitivity to the absence of descriptive features due to the object-rotation-angle is observed for cups (e.g. cup-handle is not visible). Bottles and glasses show less rotation sensitivity due to the symmetry of the extracted features from different rotation angles. In case of cell-phones mostly all possibly available features are extracted – regardless the object-rotation-angle – due to its general flat shape and the upper robot-camera-perspective.

6 Discussion and Conclusion

The evaluation has shown that the number of extracted features and especially the *presence* and *absence* of descriptive features due to the distance and viewing angle to the objects from the robot have an important impact on the classification result. Also it was observed that categories have *biases* for being misclassified to particular categories. The construction of an efficient visual dictionary is a crucial factor for the classification performance: the determination of an appropriate number of visual words (dictionary size) plays an important role in order to generate discriminative visual words which show a bias for certain categories. The evaluation of an appropriate dictionary size by exploiting the *Dunn-validity-index* as an indicator for the size plus *visual word weighting* and *filtering* have shown in the experiments reasonable results. The choice of an appropriate machine learning technique and feature-selection algorithm is important: it is observed that a more accurate classification is achieved if additionally a feature-selection algorithm (e.g. *PCA* or *Entropy*) is applied than the application of basic *AdaBoost* or *SVM* learning approaches. Moreover the combination

of a *certain number of base-classifiers* for a combined classification has shown in the evaluation reasonable improvements compared to the application of a single classifier. Also the evaluation has shown that a combination has to be determined of *particular base-classifiers*, rather than to combine the top-n base-classifiers; even *less accurate* base-classifiers (i.e. *AdaBoost*) can contribute to an efficient combination of base-classifiers. The object detection based on an basic image segmentation has shown a satisfying trade-off between computational cost and accuracy of the extracted object boundaries. However, the detection has shown its limitation in the evaluation. As mentioned previously we are working on an enhanced detection based on 3D depth information which can provide more reliable indications of object candidates compared to purely image intensity based approaches; we focus on the detection of objects on multiple planes in varying constellations as for instance on shelves in a supermarket. Additionally in the future work, we focus on to increase the number of supported object categories.

Further experiments have shown an average execution time of $\approx$*2.6s* for the categorization of six concurrently present objects ($\approx$*802ms* for a single object). This execution time provides the feasibility of the system to be applied on service tasks, which require *occasional* to *frequent* categorization of objects.

In this paper a prototypic object categorization system has been described which is based on *BoF*. The presented evaluation has shown the behavior and the competitive categorization performance. This system equips a service robot with an ability which supports the application of advanced object-related tasks.

References

1. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints (2004)
2. Duin, R.P.W.: The combining classifier: To train or not to train? In: International Conference on Pattern Recognition, vol. 2 (2002)
3. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Cybernetics and Systems 3, 32–57 (1973)
4. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
5. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval (2007)
6. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV 2005: Tenth IEEE Intl. Conf. on Computer Vision, vol. 1, pp. 604–610 (2005)
7. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems 19, pp. 985–992 (2006)
8. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
9. Wisspeintner, T., van der Zant, T., Iocchi, L., Schiffer, S.: Robocup@home 2008: Analysis of results. Tech. rep. (2008)
10. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition 37(1), 1–19 (2004)