Evaluation of Feature Selection and Model Training Strategies for Object Category Recognition

Haider Ali* and Zoltan-Csaba Márton*

Abstract—Several methods for object category recognition in RGB-D images have been reported in literature. These methods are typically tested under the same conditions (which we can consider a “domain” in a restricted sense) such as viewing angles, distances to the object as well as lighting conditions on which they are trained. However, in practical applications one often has to deal with previously unseen domains.

In this paper, we investigate the effect of domain change on the performance of object category recognition methods. We use the public RGB-D Object Dataset from Lai et al. [1] for training, and for testing we introduce the DLR-RGB-D dataset, representing a similar, but different domain. The data present in both datasets holds various object instances grouped into general object categories. Object category detectors are trained using the objects of one domain and tested on the objects of the other domain. We then explored how do different 3D features perform when the model trained on the source domain is applied on the target domain, and evaluated two feature selection strategies.

In our experiments we show that a domain change can have significant impact on the model's accuracy, and present results for improving the results by increasing the variability of the objects in the training domain. Finally, we discuss the relevance of the descriptors and the properties they capture.

Index Terms—object categorization, cross-domain learning, feature selection, domain adaptation, RGBD object databases

I. INTRODUCTION

We consider the problem of domain change, as one of the major obstacles faced in practical adaptation of object categorization algorithms is that object instances used for testing are different than the training ones, and they might be captured under slightly different conditions than the ones used for building a training database. Domain in this context refers to properties of a dataset that are dependent on the data capturing process and conditions.

For RGB data, it has been shown that domain change severely affects recognizer performance, in spite of the fact that feature detectors and descriptors aim to be as invariant as possible. Therefore, we are continuing our previous work on categorizing high-variance RGB-D data [2], focusing here on increasing the robustness of the used classifiers by selecting the right features and augmenting the training dataset.

In order to evaluate the effect of relying on an online RGB-D object database during category recognition performed in a different environment, we have used the largest available benchmark RGB-D dataset for training, and a new dataset (DLR-RGB-D) for testing. Our database contains 21 object categories recorded on a table with Microsoft Kinect [3], and it is publicly available for the community, in order to support other benchmarking experiments. Using the 21 object categories that are common between the two sources, we train our classifier using the categories from one domain and see how that classifier performs on samples from the other domain. The main objective is to see how much robustness is provided by depth information against domain change.

We have used combinations of well known 3D features (VFH, ESF, PFH as summarized in [4]), and trained SVM classifiers. For feature selection we have used the minimum Redundancy Maximum Relevance (mRMR) and Maximal Relevance (MaxRel) feature selection methods [5]. The mRMR feature selection algorithm selects features with minimizing redundancy and takes into account their highest relevance to the target class. The MaxRel feature selection algorithm selects features with the consideration of highest relevance of features to the target class. We have tested our classifier on the RGB-D Dataset using 5-fold cross validation as well as on the DLR dataset, without adaptation, to highlight the difficulties encountered during cross-domain categorization. During our evaluations we investigated various combination of features as well as the amount of data used to train the classifier for optimizing performance as reported in the experimental section of the paper.

We report on the following contributions in this work:

- quantifying the generalizing power of object category recognition between datasets;
- evaluating the VFH, ESF and PFH features and their combinations (full and partial concatenation);
- analyzing the results given by features selection methods (mRMR and MaxRel);
- quantifying the effect of domain change from the RGB-D (original) to the DLR (target) domain;
- improving the categorization accuracy obtained through adapting the training set.

In the following we will give an overview of the related approaches, then describe our work in sections III-V, present the experimental results in Section VI and discuss our findings in Section VII before finally concluding and discussing further research directions in VIII.

II. RELATED WORK

Object recognition has been an active area of research in computer vision. Object category recognition extends this concept to recognize classes of objects (chair, car, ...) instead of individual object instance detection. In a typical scenario for object recognition, training is performed on a
subset of a Dataset and tested on the remaining subset. Recent challenges are to improve the robustness of the object recognition systems as well as their detection accuracy.

Several methods proposed in the literature for object recognition are based on the available textural information in RGB images data. The focus have been to propose different features based on the available textural information. For instance, [6] has proposed a real time multi-resolution object detection framework using Histograms of Oriented Gradient (HOG) features. A hierarchical region based object detection framework based on coherent probabilistic model has been proposed in [7]. Similarly, A region based object segmentation method using salient information (holistic properties of object shapes, geometric relationship of object boundaries) has been introduced by [8]. Another work [9] has proposed a learning framework of object detection and classification by concatenation of feature and context information using Support Vector Machine (Context-SVM).

In robotics and vision community, a major challenge is the real time object recognition and pose estimation for manipulation tasks on RGB-D (color+depth) data using sensors like Microsoft Kinect [3]. Lai et. al [10] has presented an object recognition framework for based on the image-level depth features using hierarchical kernel descriptors. Burrus et. al [11] have recently proposed a model based 3D shape reconstruction approach using classical histogram comparison for the task of pose estimation. Therefore, different methods have been proposed to adapt classifiers trained on one domain to the other domain.

A visual object recognition framework with domain adaptation using web data has been introduced by Lai et. al [12]. They introduce a probabilistic exemplar-based method using SVM on Google 3D Warehouse and local Datasets. Bo et. al [13] have presented an unsupervised object recognition framework using hierarchical feature representations from RGB-D data. They have introduced dictionary learning mechanism to generate RGB-D depth and color image features. They have reported accuracy for category and instance recognition in comparison to their existing work and previously available approach (Convolutional K-Means descriptor [14]).

III. RGB-D DLR OBJECT DATASET

The RGB-D DLR Dataset is collected using by Microsoft Kinect [3]. The acquisition of each frame is performed by the Point Cloud Library (PCL) [15], using the OpenNI Grabber Framework. The Kinect is placed about one meter from the table where the objects to be sampled are placed. The data was taken with the Kinect mounted at 45° above the horizontal. An example frame is shown in Figure 1, on which standard 3D object segmentation methods from PCL are used to extract the objects.

One revolution of each object was recorded, the rotation of the object was done manually rotating it around 5° in each frame, and in some cases, depending on the shape of the object, the object was rotated in more than one axis. This procedure gives a total of 7893 RGB-Depth frames in the DLR RGB-D Object Dataset. The DLR RGB-D Object Dataset contains 70 different objects in 21 categories, these categories are also included in RGB-D Dataset from Lai [1]. Figure 2 shows the categories and objects collected.

Once we have the cropped point cloud, To find the table and object, we assume that all the sampled objects are lying on the planar surface, then we perform RANSAC plane fitting to find the table plane and take points that lie above it to be the object. To remove the outlier in segmented objects, we have applied Radius Outlier Removal filter PCL [15]. The segmentation results are shown in Figure 3.

The resulting DLR dataset can be found at:  

IV. FEATURE EXTRACTION

As we found in our previous work [16] that geometric features perform better for categorization of previously unknown object instances, we focused here on three promising global 3D object descriptors that were evaluated and described in [4]: the Point Feature Histograms (PFH), the Viewpoint Feature Histogram (VFH), and the Ensemble of Shape Functions (ESF).
All three capture various 3D surface properties, and have a high number of dimensions: 125 for PFH, 308 for VFH and 640 for ESF. We used the tools available in PCL to process the input point clouds, namely to estimate surface normals, subsample the points s.t. the resolution is around 1 cm, and to compute the feature descriptors for each input file. The feature matrices were concatenated, thus creating a descriptor of length 1073, as such concatenations provide a good way to fuse the information from different sources [17], [18]. However, some tests were performed using the individual features separately as well.

V. Multi-Class Object Classification

We used multi-class Support Vector Machine (SVM) classifiers with a linear kernel [19]. A linear kernel was chosen because it is faster to train and it gives comparable results to RBF Kernel for such large number of features/samples.

In order to investigate which positions in the high dimensional feature descriptors hold the most discriminative power, we used the mRMR and MaxRel algorithms [5] for the identification of best 500 features. We have then selected the best 50, 100, 150, ..., and 500 as well as all the features for training different models.

For different combination of features descriptors (VFH, ESF and PFH) using mRMR and MaxRel, we have repeated the given sequence of feature selection separately. In addition to that we also combine the top scoring of each of the feature types (VFH, ESF, PFH) Top 50, Top 100 and compare with the globally best 150 and 300 selected by mRMR and MaxRel.

VI. Experiments and Results

In this section we will present the results of our evaluations. All experiments, where only a subset of the feature descriptor vectors’ dimensions were used, were performed twice, once using the feature ranking provided by mRMR and then using the ones by MaxRel. The two methods select the same top 500 features, but order them differently, as shown in Figure VII. The top 118 dimensions according to MaxRel are scored in the same order by mRMR, but the remaining ones are mingled.

First, we checked the 5-fold cross-validation results we obtained on the RGB-D dataset separately. The SVM parameters were selected such that they maximize this cross-validation accuracy, and very good results were obtained, as shown in Figure 5.

Interestingly, contrary to our expectations based on [5], the feature vector dimensions scored higher by MaxRel performed in general much better than the ones scored highly by mRMR. This finding was confirmed in successive experiments, as we will show, suggesting that for the task of identifying the most important feature dimensions (and indirectly the object properties) the MaxRel method is more useful.

Looking at the features individually (Figure 6), we can see that a small subset of the ESF and PFH features already captures most of the variance between the object categories, while in the case of VFH the accuracy increases steadily as more and more dimensions are added. This means that a large portion of the information that PFH and ESF capture is redundant, or not too relevant for this categorization task. On the other hand, a small portion of them (between 100 and 150 dimensions) already has the same discriminative power as the full 308 dimensions of VFH.

We also evaluated the effect of selecting the top scoring 50 or 100 features from each descriptor and concatenating
Fig. 6. Linear SVM 5-fold cross-validation accuracies on the RGB-D Object Dataset using the three descriptors separately. The top scoring feature dimensions according to mRMR and MaxRel were tested in steps of 50, and compared to the case when every feature dimension is kept.

Fig. 7. Linear SVM 5-fold cross-validation accuracies on the RGB-D Object Dataset using the 50 (top) or 100 (bottom) highest scoring dimensions of the three descriptors separately, compared to their concatenation and the globally best 150 or 300 dimensions, respectively.

Fig. 8. Linear SVM prediction accuracies on the DLR dataset using the concatenation of all three descriptors. The top scoring feature dimensions according to mRMR and MaxRel were tested in steps of 50 up to 500, and compared to the case when every feature dimension is kept.

Fig. 9. Linear SVM prediction accuracies on the DLR dataset using the three descriptors separately. The top scoring feature dimensions according to mRMR and MaxRel were tested in steps of 50, and compared to the case when every feature dimension is kept.

The results highlight the difficulties posed by domain change, due to the limited generalization power of the training data and features, reaching an accuracy of approximately 60%. This is in strong contrast to the randomized cross validation results presented above, as shown in Figures 8, 9 and 10.

Thus, we can safely conclude that some sort of domain adaptation is required. Instead of a weighting approach performed in [12], we left one object instance per category out of our testing dataset, and included them during training. To compensate for the added training examples, one of the original training instances was left out of training. We then compared the results on the remaining objects obtained using the original and the updated model, shown in Figure 11.

The adaptation step was performed multiple times in a randomized way in a form of jackknifing, and the minimal and maximal accuracies shown by the error bars. The baseline of 60.62% using the original model is based on a single classification of the DLR objects, where one object per category was removed (which produced a negligible improvement with respect to Figure 8). However, based on the randomized results for the updated model (and those that
will be presented in Figure 12), we don’t expect a deviation larger than a few percents for this case either.

Even this simple mixing of domains resulted in an improvement to an average of 70% in the categorization accuracy on the new domain’s objects. In total around 15000 RGB-D point clouds from the original training domain were replaced by around 2500 from the new one, however, the original clouds had large overlaps (the authors proposed using every fifth scan only). As there are on average around 5-6 object instances per category, the data from the new domain constitutes only a small portion of the updated dataset, but it was sufficient for improving the results considerably. The parameters obtained by cross-validation on the original dataset were reused, so the re-training lasted only 4.5 minutes.

In case all the original objects were kept for the updated training dataset, the accuracy was 69.41%, thus the accuracy increase was not due to leaving bad data out of the training set. For an in-depth analysis of the variability in object categories that is captured by the RGB-D Object Dataset, we performed another set of randomized experiments (10 runs each), as shown in Figure 12. First, in the left part, we can see that reducing/increasing the number of objects in the source dataset does not affect the general accuracy, but there are clearly objects that are more similar to testing ones than others. As the number of objects increase, their effect diminishes, and the variance between different splits decreases considerably. Second, on the right, the effect of adding an object from the new domain is made apparent: when random objects are added to the training set there is a consistent performance increase (in each run we test the same original training objects and testing objects, once without and then with adaptation).

Similarly to the left part of Figure 12, a reduction in variance can be seen in Figure 13, where the two feature selection methods are evaluated. Their performance follows the same general path than when using 100% of the object instances, with 66% of the data already showing relatively stable results, suggesting that there might be some redun-
in the ordering, and finally, the VFH dimensions’ positions.

ESF dimensions’ ordering, the second the positions of the PFH dimensions
MaxRel, shown for the three descriptors consecutively. First part shows the
two thirds of the object instances are used.

The usefulness of these normal angles is further high-
lighted by the results obtained by VFH. This descriptor has
two parts, one based on the fast variant of PFH (FPFH), and
another that is viewpoint dependent. Overall, VFH performs
very well, both for pose estimation and categorization [4],
but the highest ranking dimensions come all from the PFH-
based part of it. This makes sense, because we use PFH as a global descriptor,
capturing the variability of estimated surface normal angles
to distinguish objects of different categories. In such a task,
极端 values of angles correspond to normals pointing
in the same or opposite directions. The latter is quite rare
in RGB-D images, and the former does not have much
discriminative value. The use of PFH as a global descriptor
corresponds to its original design, as conceived at DLR [21],
only with one of the constituent features (point-to-point
distance) left out.

The ESF dimensions that made it to the top 500 are all
from the range [537, 640], with only dimension 539 missing,
and 537 being the one not part of the main cluster, at
position 403. Because ESF is a concatenation of ten 64-bin
histograms, this range corresponds to the ones for the shape
functions 8 (second half) to 10 as being most important for
object categorization. These are all point pair (line) features,
describing the lines which cross unoccupied space in the
point cloud. A similar approach to quantify the amount of
free space through point-to-point line traces is also the idea
behind GFPFH, the global feature that integrates besides this
information also local shape classification using FPFH [4].

While it is clear that MaxRel produces a better feature
ordering for our object categorization task, its absolute cor-
rectness is difficult to assess. Since SVM are not affected by
the Hughes phenomenon, the occasional dips in performance
as the number of dimensions increases suggests that some
of the features are wrongly selected. Figure 15 shows the
features’ performance in groups of 50, where a downward
trend would be expected if they are sorted correctly. These
combinations consider correlations between features as well,
so high values are not so surprising also using lower ranked
features. However, mRMR has a clearly inverted trend, and
considering the relation between the mRMR and MaxRel
ordering in Figure as well, it is safe to conclude that
MaxRel’s ranking is more reliable. Moreover, the selected
feature ordering seems to generalize well to a new dataset,
as there is a very high correlation in feature performance
on the source and target domain (0.99 for mRMR and 0.97
for MaxRel). On average around 60% of the accuracy is
preserved by the feature combinations when dealing with the new object instances.

VIII. CONCLUSION

In our evaluations we have quantified and discussed the effect of training object category recognizers on one dataset and testing it on another one, captured in a different environment, and slightly different conditions. The results show that even the largest RGB-D training database available online does not capture a sufficiently high variation in common object categories, and that some form of domain adaptation is needed. Even a simple approach proved to be highly effective, and could be improved further by methods presented in [22] for avoiding the costly re-training step, by updating existing models with new data.

Based on the comparison of features, and the useful insights gained by MaxRel, we can conclude that a well-chosen subset of features is already capable of capturing most of the information necessary for the presented categorization task involving 21 object categories. However, as we saw, the real-world variance in object categories can be quite difficult to capture in current datasets, and there are thousands of object categories of potential relevance [23]. Nonetheless, identifying the most relevant parts of long feature vectors, at least for a given limited application, can aid in optimizing performance, and also to guide future descriptor designs to focus on the most descriptive properties.

IX. ACKNOWLEDGEMENTS

The authors would like to thank Jorge Nuricumbo Morales for his help with the dataset, and the three anonymous reviewers for their helpful comments.

REFERENCES


