Active Classifier Selection for RGB-D Object Categorization using a Markov Random Field Ensemble Method

Maximilian Durner, Zoltán Márton, Ulrich Hillenbrand, Haider Ali and Martin Kleinsteuber‡
German Aerospace Center (DLR), Oberpfaffenhofen ‡ Technical University of Munich

ABSTRACT

In this work, a new ensemble method for the task of category recognition in different environments is presented. The focus is on service robotic perception in an open environment, where the robot's task is to recognize previously unseen objects of predefined categories, based on training on a public dataset. We propose an ensemble learning approach to be able to flexibly combine complementary sources of information (different state-of-the-art descriptors computed on color and depth images), based on a Markov Random Field (MRF). By exploiting its specific characteristics, the MRF ensemble method can also be executed as a Dynamic Classifier Selection (DCS) system. In the experiments, the committee- and topology-dependent performance boost of our ensemble is shown. Despite reduced computational costs and using less information, our strategy performs on the same level as common ensemble approaches. Finally, the impact of large differences between datasets is analyzed.

Keywords: ensemble learning, active classification, RGB-D object recognition

1. INTRODUCTION

Over the last two decades, a plethora of image and shape descriptors and classification techniques, each with many variants, has been proposed. Different settings were shown to stand out depending on data and problem considered. In order to overcome the lack of a universally superior recognition system, one strategy is to combine several classifiers, the so-called experts, into an ensemble or committee. While some of these approaches aim to exploit the strength of the individual experts task-wise, others fuse all or several experts to generate a final prediction. Individual expert predictions are typically combined either by some voting procedure or by another classifier trained on the experts' outputs, i.e. classifier stacking. Another line of work is the concatenation of several image descriptors to higher-dimensional vectors. In this variant, only one classifier is utilized, rather than one expert for each individual descriptor type. This descriptor combination is conceptually simple but inflexible, as any change induces a complete re-training of the classifier, which can be very slow for large numbers of dimensions. Nevertheless, the concatenation shows very good performance and is widely used [1, 2].

The present work considers object category recognition from segmented RGB-D data by combining different global descriptors, both shape- and texture-based, and different classifiers. Our main contribution is a new ensemble approach which utilizes a MRF at the combination layer, called MRF ensemble method (as illustrated in Figure 1). While more common combination techniques often use the experts' estimated class posterior probabilities, our approach fuses only their predicted class labels. Although this leads to some loss of information, the performance thus obtained is still comparable to diverse state-of-the-art combination methods we evaluate here. Offering the benefits of an undirected graphical model, our MRF ensemble method is implementing two inference types on a single network. The MRF ensemble method is able to execute with the same trained model the fusion of all ensemble experts as well as a dynamic expert selection. Thus, the next used experts can be selected based on expected benefit for each individual prediction. We show that this strategy not only results in an increased performance of the overall classifier, but also greatly reduces the average computational effort, due to waiving descriptor computation for some experts.

The performance of trained classifiers on test data critically depends on how well the training data resemble the testing domain. For a robot operating in a largely unconstrained environment, such as human homes or

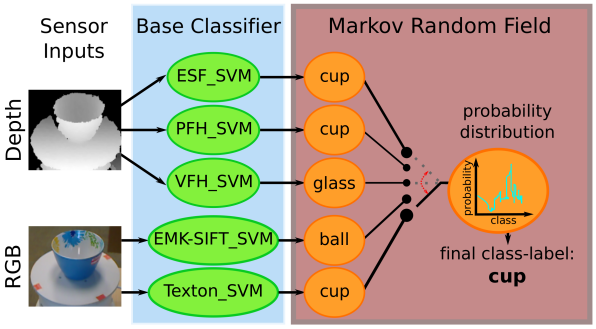

Figure 1: General pipeline of the MRF ensemble method (images of cup from [1]), where the line thicknesses at the switch indicate how often an expert is triggered during DCS.

public places, off-line training data will likely have some characteristic and distribution different from the target domain. Here we address this issue by using two sets of RGB-D data in our study, one taken from [1], the other from [3]. Although both sets are acquired with the same kind of sensor under similar conditions, classifier accuracy drops significantly when one is used for training and the other for testing [3]. However, the main relations between observed performances of ensemble types turn out consistent across all experiments.

2. RELATED WORK

The effect of ensemble methods was investigated in various fields [4, 5], finding an empirical boost in the recognition rate. Even greater performance increase can be obtained by combining information derived from different sensors and modalities (e.g. RGB and depth) [1, 2, 6, 7].

Since a large number of discriminative classifying systems exist, each of them having different competences, one could also dynamically select the classifier based on the application. The selection of the most suitable classifier should rely on a competence measurement on the current setup. In the literature, a large amount of different selection criteria for this purpose can be found [8, 9, 10]. A good overview of such approaches is given by the survey of Britto et al. [11]. An issue of combining a large amount of classifiers is the higher risk of redundant information. Fusing of different classifiers is only gainful, if the committee shows a high diversity among themselves [12, 13, 14, 15]. Hence, when given a large committee, not all predictions are necessary. In order to overcome this issue, several DCS systems were proposed [4, 11, 15]. Similar to ours is the one by Gao and Koller [16]. A benefit of our MRF ensemble method is the possibility to generate a lookup table for the selection process offline, leading to a more efficient DCS.

3. MARKOV RANDOM FIELD ENSEMBLE METHOD

The main idea of the MRF ensemble method is to model the joint probabilities between the expert predictions and the category of queried object. Therefore the predictions of the experts and the ground truth label are treated as random variables in a MRF. The extracted sets of features are classified by pre-trained classifiers, employed at the base level. In the next stage, the combinational level, the resulting predictions are forwarded into the MRF. In the MRF network we defined two types of nodes: the *expert nodes* (X_i^e) represent the predictions of the associated classifier, while the *ensemble node(s)* (X_j^c) contain the information about the final prediction.

3.1 Network Topology

Based on the presented theory, two basic architectures were investigated, which are appropriate in terms of computational complexity. While both architectures consist of one vertex for each expert, they differ in the representation of the final class prediction. The first structure (here denoted as Label_*) contains one single ensemble node. The domain of each node is equal to the predictable classes of the associated expert, respectively the ground truth classes. In the other basic structure (here denoted as Binary_*), each target category is represented as a binary node with the domain $\mathcal{L}_{X_j^c} = \{yes, no\}$. Inspired by the *Multi Labeling* approach presented by Shahbandi and Lucidarme [17], the advantage of this structure is the individual prediction of every class. Besides the essential connection between each expert node and the ensemble node(s), the topology evaluation showed a positive influence by connecting the expert nodes among themselves. To include these combinational dependencies a so-called *pairwise MRF* is used. Several works [18, 19] utilize the pairwise MRFs for reasons of lower complexity. These networks simply consider cliques involving a single or a pair of nodes. Although, a pairwise MRF does not strictly satisfy the *Hammersley-Clifford theorem*, it is a good approximation in order to shrink the complexity. At the end we decided to use the best performing topologies for both basic architectures in the experiments below. For both structures this have been the pairwise MRF connecting also the expert nodes among themselves (here denoted as *_pairCliques*).

3.2 Inference Techniques

With all utilized topologies presented, another main part of the MRF ensemble method is the inference. Firstly the common inference is presented. This type of inference does not exploit the whole power of an undirected graphical model. Since the characteristic of undirected edges leads to a higher computational complexity, for such inferences normally a Conditional Random Field (CRF) is utilized. This sub-group of MRFs is directly modeling the standard prediction problem $\mathbf{p}\left(X_j^c|\mathbf{mb}\left(X_j^c\right)\right)$. This inference restriction leads to more accurate probabilities. Secondly, an active classification approach is described, which presents several benefits.

3.2.1 Forward Inference

For the *Forward Inference* (*FwInf*) technique, all expert nodes X_i^e are set to their associated class prediction. For the topologies Label_pairCliques the probabilities of each label $\in \mathcal{L}_{X_1^e}$ are computed. Inferring the Binary_pairCliques graphs, requires the computation of the probabilities of each ensemble node being exclusively "yes". Both cases result in a Probability Density Function (PDF) which is then used to determine the highest class probability p_{final} and the final class prediction $class_{final}$.

3.2.2 Active Classification

A main component of the MRF is the property of undirected edges which allows the implementation of a DCS. The two main ideas behind this inference are the reduction of processing costs and the generation of a more accurate PDF. Although the ensemble of classifiers empirically shows a performance gain in most of the literature, a general theory is still missing [11]. There does not exist a general rule for how many experts should be fused nor for which applications combining experts leads to a boost [20]. It might be the case that the prediction of one expert is already enough to obtain the right prediction with a sufficiently high likelihood. But, on the other hand, the committee could be confronted with a queried instance for which the class probabilities of only one classifier are not enough. Thus, the DCS procedure leads to a final prediction which is based on the highest expertise – fused or single – available. The other main advantage is the reduced computation of only required classifiers, because the selection is applied right in the classification phase. This can lead to reduced processing costs, if the selection process is less expensive than generating the expert predictions. The decrease of processing effort is especially important for autonomous and mobile robotic platforms.

Furthermore, our approach enables the system to react to already known predictions. This is a major advantage over most of the state-of-the-art DCS which execute the classifier selection off-line. The basic concept of the *Active Classification* (*ActClass*) is shown in Algorithm 1.

Algorithm 1: Active Classification (*ActClass*) algorithm used by the MRF. Notations were simplified for easier readability.

Require: entropy threshold e_{th}; state number threshold s_{th}

$p_{prior}(X^c) \leftarrow uniform$
$\mathbf{experts}_{rem} \leftarrow X^e_{1..n}$
repeat
  initialize **votes**
  inference() $\leftarrow$ possible states $< s_{th}$? *Exhaustive* : *MCMC*
  for all labels $l \in \mathcal{L}_{X^c}$ **do**
    set X^c to l
    for all $E \in \mathbf{experts}_{rem}$ **do**
      apply inference() on $E \rightarrow p(E)$
      compute $H_E\,(p(E))$
    end for
    votes $(\mathrm{argmin}_E\, H_E)$ += $p_{prior}(X^c = l)$
  end for
  $E^* = \mathrm{argmax}_E$ **votes** (E)
  compute E^* corresponding expert and set E^* to sample-label
  remove E^* from $\mathbf{experts}_{rem}$
  apply inference() on $X^c \rightarrow p(X^c)$
  $p_{prior}(X^c) = p(X^c)$
until $H\,(p(X^c)) > e_{th}$ **or** $\mathbf{experts}_{rem} \neq 0$

Ensure: $class_{final} = \mathrm{argmax}_l\, p(X^c = l)$

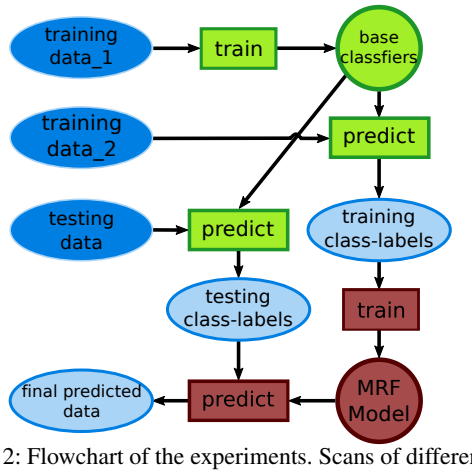

Figure 2: Flowchart of the experiments. Scans of different objects are used for the training and the testing data, in some experiments even coming from independently acquired datasets.

The *ActClass* inference predicts the probable outcome of aggregating a specific expert to the current expert ensemble (starting with the best performing expert during training since no information is given). For the evaluation criteria the entropy H, which is a commonly used representation for the information in a PDF, is used. The currently best expert is derived based on the entropies H_E of the experts' PDFs for each target label l. Therefore, the ensemble node(s) is/are set to evidence and the PDF for the domain of each unset node is computed. With each iteration an expert is selected and its computed class prediction set to evidence. Hence the selection pool $\mathbf{experts}_{rem}$ and with it the number of possible states decreases with each iteration. Instead of setting each expert node individually to evidence and infer the ensemble node, this direction of inference was chosen to enable the combination of experts with different domains. Following [16] we evaluated as stopping criteria the information gain, however, the entropy of the current PDF gave better results.

Since not all expert nodes are filled, inference in a partially filled network is required. We applied Markov chain Monte Carlo (MCMC) which lead to reliable convergence after 5000 iterations in our experiments. However, we also observed an increasing rate of misclassifications if the number of iterations is far too high (relative to the number of actual permutations). Since the number of possible permutations decreases with an increasing number of iterations, the possibility of misclassifications due to a improper amount of iterations is given. Therefore, if the number of permutations is too low a so-called *Exhaustive Inference* is executed. Hereby, every possible permutation of the absent nodes is processed and the average over the generated PDFs results in the final class probabilities. Besides the prevention of misclassifications, this distinction also leads to reduced

computational costs. Applying the *Markov blanket* property of the network, the number of iterations, hence, the computational costs could be reduced even more, because only nodes in the Markov blanket are considered.

Since the weights are fixed after training, the expert selection ranking for all permutations can also be computed off-line once. Thus, in contrast to the approach presented in [16] and other publications, the expert selection does not negatively influence the computational effort of the inference phase.

4. EXPERIMENTAL SETUP AND DATASETS

In our experiments, we focus on two goals. i) Investigate the performance of the MRF ensemble method with regard to the different topologies and inference methods, and ii) study the effects of domain differences between train and test data. For the latter, a comparison is made to a range of alternative ways of combining base classifiers, through variants of voting and stacking, and also to concatenation of the image descriptors. We here describe the datasets and their usage in our study, and then turn to the results on MRF and cross-domain performance in the following sections.

The MRF performance is evaluated on the RGB-D dataset from [1], here denoted by D_1. As proposed by the authors, only every 5^{th} object view is included so as to reduce redundancy and prevent over-fitting, and 10-fold cross-validation on the object-instance level is applied. Since two training phases are required for the two levels of the ensemble classifiers, the training sets have to be split again, s.t. in all training and testing phases the system is confronted with novel object instances; see Figure 2 for an illustration of the data flow. Let $D_1 = \cup_c D_1^c$ be the partitioning of D_1 into subsets D_1^c of object class c, and $D_1^c(i)$ the subset of all views of instance i from class c. We hence have for each fold

$$\text{training\_data\_1} = D_1 \backslash [\cup_c (D_1^c(i^c) \cup D_1^c(j^c))] \;, \quad \text{training\_data\_2} = \cup_c D_1^c(i^c) \;, \quad \text{testing\_data} = \cup_c D_1^c(j^c) \;, \tag{1}$$

where the left-out instances i^c and j^c are selected randomly ($i^c \neq j^c$). Moreover, only 21 object categories out of the 51 available are included, in order to match the range covered by the other dataset taken from [3], here denoted by D_2, representing the second domain in our study.

The object images D_2 were acquired with the same sensor (Microsoft Kinect [21]) and from similar distance and viewpoints as images in D_1. Nonetheless, there is a marked drop of performance when one set is used for training, the other for testing, relative to cross-validation [3]. This setting hence mimics a training on a domain slightly different from the target domain, as may be expected, e.g., for robots operating in unconstrained human living environments.

For the cross-domain studies, the training data are taken from D_1, the testing data from D_2. We have run experiments without domain adaptation, meaning in detail

$$\text{training\_data\_1} = D_1 \backslash [\cup_c D_1^c(i^c)] \;, \quad \text{training\_data\_2} = \cup_c D_1^c(i^c) \;, \quad \text{testing\_data} = D_2 \backslash [\cup_c D_2^c(k^c)] \;; \tag{2}$$

with adaptation at the base level,

$$\begin{aligned} &\text{training\_data\_1} = D_1 \cup [\cup_c D_2^c(k^c)] \backslash [\cup_c (D_1^c(i^c) \cup D_1^c(j^c))] \;, \\ &\text{training\_data\_2} = \cup_c D_1^c(j^c) \;, \quad \text{testing\_data} = D_2 \backslash [\cup_c D_2^c(k^c)] \;; \end{aligned} \tag{3}$$

and with adaptation at the ensemble level,

$$\text{training\_data\_1} = D_1 \backslash [\cup_c D_1^c(i^c)] \;, \quad \text{training\_data\_2} = \cup_c D_2^c(k^c) \;, \quad \text{testing\_data} = D_2 \backslash [\cup_c D_2^c(k^c)] \;. \tag{4}$$

5. EVALUATION OF MRF ENSEMBLE METHOD

To arrive at more general conclusions about the MRF performance, different committees are evaluated. Some of the experts were based on the RGB information (2D-based), trained on the Texton (539D) [22] or the EMK-SIFT (1500D) descriptor [23]. The other experts used the depth information (3D-based), trained on Point Feature Histogram (PFH) (125D), Viewpoint Feature Histogram (VFH), or Ensemble of Shape Functions (ESF) (640D) [3, 24, 25]. In all experiments of this section, the extracted descriptor sets were classified by a linear Support Vector Machine (SVM) as base classifier, for which the library presented in [26] was used. The expert committees studied were 2D-based, 3D-based, and (2D+3D)-based.

The main reason for combining various experts to a committee is an expected performance boost over the single experts. However, as can be seen in Figure 3, the *FwInf* technique applied on the 3D-based committee slightly decreases the accuracy compared to the single best expert (ESF_{SVM}). We obtained the same result for the 2D-based expert pool against the single best expert ($Texton_{SVM}$, not shown). Conversely, the ensemble of all experts (2D+3D-based) with *FwInf* does show an increase of performance over all single experts. Further investigation suggests that the low performance of the 2D- and 3D-based ensembles under *FwInf* comes from

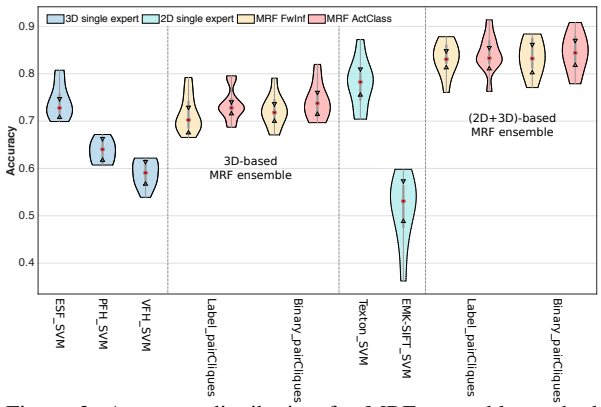

Figure 3: Accuracy distribution for MRF ensemble method on 3D-based and (2D+3D)-based experts inferred by ActClass and FwInf compared to the single experts.

	2D		3D		2D+3D	
no. of experts	Label_pairCliques	Binary_pairCliques	Label_pairCliques	Binary_pairCliques	Label_pairCliques	Binary_pairCliques
1	**90.3**	**92.1**	**48.3**	**52.8**	7.5	10.4
2	9.7	7.9	35.6	27.9	**33.5**	23.1
3	–	–	16.1	19.3	24.1	**26.4**
4	–	–	–	–	15.6	14.7
5	–	–	–	–	19.4	25.4

Table 1: Percentage of test images (10-fold cross-validation average) involving different numbers of experts by ActClass; **bold** indicates the highest percentage value.

a low diversity of the expert pool, as quantified by the double-fault measure [27] and correlation of predictions (not shown). In fact, these two ensembles have essentially failed, in that one of their experts performed better than the ensemble. The 2D+3D-based ensemble, on the other hand, contained more diversity and was a success.

The other observation from Figure 3 is the performance boost of the *ActClass* technique over the *FwInf*. The intuitive explanation is that, based on the PDF of the first selected classifier, the dynamic inference approach queries only further experts, if the predicted distribution is not distinctive enough. This way, strong experts like $Texton_{SVM}$ and ESF_{SVM} are often the only ones queried, hence class confusion decreases. The improvement is more evident for the sub-optimal 3D-based committee than for the better arranged 2D+3D-based committee.

Importantly, the MRF ensemble method inferred by the *ActClass* performs at least as good as the single best expert in *all* committees. This particular ensemble method hence provides a kind of robustness to the choice of experts in the committee, making a failure less likely to occur even with sub-optimal expert combination. Since the effect of a particular combination can hardly be predicted without thorough evaluation, the MRF ensemble method with *ActClass* is highly relevant for applications where a prior evaluation on the target domain is not reliably possible, e.g., for robots deployed in largely unconstrained environments.

Another advantage of MRF inference by *ActClass* over inference by *FwInf*, and over most other ensemble methods, is the reduction of computational costs, as image descriptors needed by non-queried experts are not computed. Table 1 shows the percentage of test images involving different numbers of experts for inference in the 2D-, 3D-, and 2D+3D-based ensembles. The low percentage for predicting the queried object based on only one expert in the 2D+3D case is due to the fact that a bigger and more complex network is generated. Since during training in the considered topologies the MRF is always confronted with all experts, the influence of a single expert aggregated with more experts changes. Still, in all three expert pools, the *ActClass* inference method leads to shorter computations through reduced expert recruitment.

6. DATASETS WITH DIFFERING CHARACTERISTICS

Robotic applications are commonly challenged with the issue of varying environmental circumstances during training and testing. Furthermore, the instances a classifier is trained on necessarily represent just a small subset of an indefinite amount of instances for each category. In order to simulate such conditions, in the following we utilize the additional dataset D_2 as a substantially differing test domain. Unfortunately, this dataset does not include RGB information, so only a reduced committee with the 3D experts can be used.

We compared our method to various common ensemble methods, with different complexities. First, so-called *naive* ensemble methods. While the *maxProb* strategy predicts the category with the highest class probability of all experts, *maxSumProb* derives its final prediction by summing up the probabilities class-wise over all experts and selecting the highest sum value. The *simple voting* considers each expert as equal class-voter achieving the final prediction by taking the most voted class. Similarly, the *precision weighted voting* rates the voter by the class-wise precision evaluated on a training data. The second ensemble group is the stacking approach [28] which processes the concatenated PDFs of the experts.

In order to analyze the ensemble behavior more precisely, two additional classifiers, namely k-Nearest Neighbor (k-NN) and Random Forrest (RF), are applied as base and combinational classifier (i.e. base level and combinational level, or expert and ensemble classifier), based on their implementation in Mathematica. All possible combinations of ensemble method and classifiers were applied. However, due to shortage of space, only the best performing methods from each of the three mentioned ensemble groups for all three base-classifiers are presented. Concerning the *naive* strategies, for all three base-classifiers the *maxSumProb* showed the highest

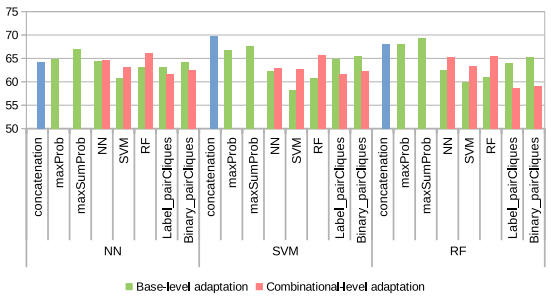

Figure 4: Averaged accuracies [%] of stacking approaches with different classifiers/fusion methods for base-level and combinational-level adaptation.

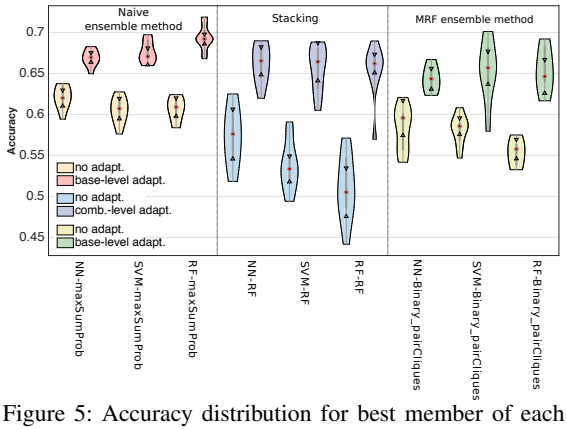

Figure 5: Accuracy distribution for best member of each ensemble method group for the three base-classifiers; MRF is inferred by *ActClass*; annotation: *classifier-ensemble*.

accuracy values. In the second ensemble group, the *stacking* approaches, the RF as combinational-classifier performed best. As could be already seen in the previous experiments the best performing MRF topology is the more complex Binary_pairCliques inferred by the *ActClass* technique.

Our first experiment in this section refers to the data according to eq. 2. As Figure 4 (no adaptation cases) shows, testing on the different dataset results in unsatisfying low performances. Thus, the slight differences in the recording setup already leads to a measurable negative effect on the accuracy. Another observable trend is the higher recognition rate of the *maxSumProb* over the stacking and MRF ensemble methods. This supports the theory proposed by Lam and Suen [29] that simpler fusion rules perform better on unknown datasets, since the complexity of a fusion method is correlated to the adaptation to the training data which results in worse results on new data. However, the MRF ensemble method still shows acceptable results compared to the common stacking approach for each base-classifier.

The low results obtained by the previous experiment lead to the requirement of an adaptation process. In [3], similarly to [30], this was achieved by including in the training phase the left out objects (one instance per category) of D_2. Because of the two training phases, the trained ensembles have two different ways of data adaptation, both using the same overall amount of training instances. The first adaptation (eq. 3) is applied on base level, by replacing one instance per category the training_data_1 with one instance of the D_2 dataset. In theory, this step leads to a higher generalizability with the consequence of better single experts. The second approach (eq. 4) switches in the D_2 instances at the combination level. This approach addresses the adaptation ability of the fusion method. By utilizing one instance per category of the D_2 data for the training_data_2, the combinational classifier is directly confronted with the target domain and can analyze the expected prediction behavior of the single experts.

The results in Figure 4 show that the combination-level adaptation performs better for all common stacking approaches, the opposite is valid for the *naive* and MRF ensemble methods. This can be mainly attributed to the different fusing concepts. The MRF (and naive) approaches determine the final prediction based on the fusion of the separated outcomes of experts. The class-specific expertise of each classifier is weighted based on its correlation to the associated ground truth. Thus, a higher quality of the experts leads to an increasing distinctiveness of the weights. In contrast, the common stacking methods generate one new feature vector by merging the classifiers' predictions. Thus, the combinational classifier does not reason based on the relations between the experts, but rather on the correlation among all obtained values from the concatenated PDFs. As a result, the influence of the individual expert performances is less important than the correlation between training and testing predictions (at the combinational level stage).

For the MRF ensemble, a higher quality of the experts (base-level adapt.) leads to a greater reinforcement of the weights for the individual predictions during training. Since the *ActClass* is based on the entropy of the final PDF, the stopping criteria are met earlier and fewer experts have to be inferred, as seen in Table 2. One disadvantage which occurs for the base-level adaptation, is the need to retrain all experts, if the system should be adapted to another target domain. Hence, the modularity of the MRF ensemble method decreases, the runtime reduction for three quarters of objects on average is of benefit for mobile robotic applications.

Focusing on the better performing ensemble methods, one can see the *maxSumProb* outperforming the others both before and after adaptation (Figure 5). This leads to the conclusion that the adaptation with one instance per category is not enough. The testing dataset still shows too much disparity against the training dataset. This is also implied by the large variance of the stacking and MRF, possibly caused by how informative the instances used for adaptation happen to be. Initial tests with two objects per category used for adaptation showed that the difference between *maxSumProb* and the MRF (and other stacking methods) is diminishing quickly.

		Label_pairCliques [%]			Binary_pairCliques [%]		
Number of experts:		1	2	3	1	2	3
Adaptation type:	no adaptation	15.1	**49.3**	35.6	15.7	**42.3**	42
	base-level	36.1	**36.3**	27.6	**40.4**	36.3	23.3
	combinational-level	6.3	43.9	**49.8**	6.8	38.6	**54.6**

Table 2: Percentage of test images (ten fold average) involving different numbers of experts by ActClass for no, base-level and combinational-level adaptation; **bold** indicates the highest percentage value.

Comparing the MRF ensemble method and the best stacking combinations, a significant difference cannot be observed (based on the notches, and more formally Mood's median tests, $p > 30\%$), although the MRF utilizes less information w.r.t. computed experts and the input information (a label instead of a PDF over the labels signifying confidences). In future work, the experts' prediction confidences should be included in the MRF by considering it as a node potential. A further increase in ensemble accuracy can thus be expected.

7. CONCLUSION

In this paper we have presented a new ensemble method for object category recognition. The proposed approach can be seen as a type of stacking in which the combinational classifier is replaced by a Markov Random Field. Therefore, the experts as well as the final prediction are defined as random variables and represent vertices in the graph, while dependencies between the experts and/or the final prediction are represented as undirected edges. This unique characteristic of the probabilistic graphical model is then used to develop a dynamic inference method besides the normal inference, leading to Dynamic Classifier Selection in the ensemble. The dynamic or Active Classification inference outperforms the usual Forward Inference.

The main advantage of Dynamic Classifier Selection during the Active Classification mode of inference over other ensemble methods and descriptor concatenation is the much lower execution time, due to fewer data descriptors being required. This is achieved without sacrificing performance, as shown in a large set of experiments with same and different characteristics of train and test data. Markov Random Field ensemble method with Active Classification has also shown a kind of robustness to sub-optimal experts in the ensemble. Moreover, a natural way of dealing with missing expert predictions is offered by the architecture.

Acknowledgments: This work has partly been supported by the European Commission under contract number H2020-ICT-645403-RobDREAM. The authors thank Dr. Rudolph Triebel for helpful discussions.

References

[1] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824, May 2011.

[2] Max Schwarz, Hannes Schulz, and Sven Behnke. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1329–1335, 2015.

[3] Haider Ali and Zoltan-Csaba Marton. Evaluation of feature selection and model training strategies for object category recognition. In *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, pages 5036–5042, Sept. 2014.

[4] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 2013 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 847–855, 2013.

[5] Veronika Cheplygina, David M.J. Tax, and Marco Loog. Dissimilarity-based ensembles for multiple instance learning. *Transactions on Neural Networks and Learning Systems*, PP(99), 2015.

[6] Manuel Blum, Jost Tobias Springenberg, Jan Wülfing, and Martin Riedmiller. On the applicability of unsupervised feature learning for object recognition in rgb-d data. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[7] Zoltan-Csaba Marton, Florian Seidel, Ferenc Balint-Benczedi, and Michael Beetz. Ensembles of Strong Learners for Multi-cue Classification. *Pattern Recognition Letters (PRL), Special Issue on Scene Understandings and Behaviours Analysis*, 2012.

[8] Giorgio Giacinto and Fabio Roli. Methods for dynamic classifier selection. In *Proceedings of the 1999 International Conference on Image Analysis and Processing*, pages 659–664, 1999.

[9] Marianna Madry, Dan Song, and Danica Kragic. 2D/3D object categorization for task based grasping. In *European Robotics Forum: RGB-D Workshop on 3D Perception in Robotics*, Apr. 2011. extended abstract.

[10] M. Sabourin, A. Mitiche, D. Thomas, and G. Nagy. Classifier combination for hand-printed digit recognition. In *Proceedings of the 1993 Second International Conference on Document Analysis and Recognition*, pages 163–166, Oct. 1993.

[11] Alceu S. Jr. Britto, Robert Sabourin, and Luiz E.S. Oliveira. Dynamic selection of classifiers - a comprehensive review. *Pattern Recognition*, 47(11):3665–3680, 2014.

[12] G. Giacinto, F. Roli, and G. Fumera. Design of effective multiple classifier systems by clustering of classifiers. In *Proceedings of the 2000 15th International Conference on Pattern Recognition*, volume 2, pages 160–163 vol.2, 2000.

[13] Giorgio Giancinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing Journal*, 19:699–707, 2001.

[14] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[15] A. Santana, R.G.F. Soares, A.M.P. Canuto, and Marcilio C P de Souto. A dynamic classifier selection method to build ensembles using accuracy and diversity. In *2006 Ninth Brazilian Symposium on Neural Networks*, pages 36–41, Oct. 2006.

[16] Tianshi Gao and Daphne Koller. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems 24*, pages 1062–1070. 2011.

[17] S.G. Shahbandi and P. Lucidarme. Object recognition based on radial basis function neural networks: Experiments with rgb-d camera embedded on mobile robots. In *Proceedings of the 2012 1st International Conference on Systems and Computer Science (ICSCS)*, pages 1–6, Aug. 2012.

[18] Ben Taskar, Vassil Chatalbashev, and Daphne Koller. Learning associative markov networks. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 102, 2004.

[19] Haider Ali, Faisal Shafait, Eirini Giannakidou, Athena Vakali, Nadia Figueroa, Theodoros Varvadoukas, and Nikolaos Mavridis. Contextual object category recognition for rgb-d scene labeling. *Robotics and Autonomous Systems*, 62(2):241–256, Feb. 2014.

[20] Michał Woźniaka, Manuel Grañab, and Emilio Corchadoc. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.

[21] Zhengyou Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, Feb. 2012.

[22] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal Computer Vision*, 43:29–44, June 2001.

[23] Liefeng Bo and Cristian Sminchisescu. Efficient match kernel between sets of features for visual recognition. In *Advances in Neural Information Processing Systems 22*, pages 135–143. 2009.

[24] Radu B. Rusu, Zoltan-Csaba Marton, Nico Blodow, and Michael Beetz. Persistent point feature histograms for 3d point clouds. In *Proceedings of the 10th International Conference on Intelligent Autonomous Systems (IAS)*, 2008.

[25] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R.B. Rusu, S. Gedikli, and M. Vincze. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. In *Robotics Automation Magazine, IEEE*, volume 19, pages 80–91, Sept. 2012.

[26] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[27] Giorgio Giacinto and Fabio Roli. An approach to the automatic design of multiple classifier systems. *Pattern Recognition Letters*, 22(1):25–33, 2001. Machine Learning and Data Mining in Pattern Recognition.

[28] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[29] Louisa Lam and Ching Y. Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16(9):945–954, 1995.

[30] Yuyin Sun and Dieter Fox. NEOL: Towards never-ending object learning for robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.