Multi-View Object Recognition using View-Point Invariant Shape Relations and Appearance Information

Wail Mustafa*, Nicolas Pugeault† and Norbert Krüger*
*Mærsk Mc-Kinney Møller Institutte University of Southern Denmark,
Campusvej 55, 5000 Odense C, Denmark
Email: wail@mmmi.sdu.dk
†Centre for Vision, Speech and Signal Processing, Faculty of Engineering & Physical Sciences, University of Surrey,
Guildford GU2 7XH, United Kingdom

***Abstract*—We present an object recognition system coding shape by view-point invariant geometric relations and appearance. In our intelligent work-cell, the system can observe the work space of the robot by 3 pairs of Kinect and stereo cameras allowing for reliable and complete object information. We show that in such a set-up we can achieve high performance already with a low number of training samples. We show this by training the system to classify 56 objects using Random Forest algorithm. This indicates that our approach can be used in contexts such as assembly manipulation which require high reliability of object recognition.**

I. Introduction

The task of object recognition in an industrial assembly set-up (as shown e.g., in Fig. 1) is fundamentally different from the 'general object recognition problem' as addressed, for instance, in the Pascal Challenge [1] as well as in 3D databases (such as [2]) which are in particular recently discussed with the availability of cheap RGBD sensors, such as the Kinect camera[1]. The main difference in an industrial set up is that the sensors as well as the number thereof can be chosen freely as well as the fact that illumination can be controlled to a large degree by, e.g., letting the illumination conditions be dominated by the light sources attached to the platform (see Fig. 1). A particular challenge is that the goal of performing large sequences of actions in assembly processes requires very reliable and also fast object recognition and localization and hence if performance is not close to 100%, any assembly process is severely affected.

In this paper, we address the task of object recognition in the well controlled scenario assuming objects occurring only in the rather restricted work space of the robot as shown in Fig. 1b—resembling an 'intelligent work-cell' in an advanced production scenario. The task at hand is to estimate likelihoods of the presence of objects in the working space covered by the three pairs of Kinect and stereo cameras. In contrast to the object recognition problem addressed by standard databases operating on 2D images, in our set-up we can operate on rather complete 3D data computed by 3 different views arranged in a triangle (see Fig. 1a). This likelihood is then supposed to be used to trigger other mechanisms such as pose estimation, grasping or manipulation actions (such as peg-in-hole or screwing actions) as well as monitoring such processes in the context of complex assembly operations.

In this paper 3D *texlets* (see Sect. IV), extracted to serve as visual representations of objects, are acquired by two different sensors—stereo and Kinect cameras—simultaneously. From these, view-point invariant representation based on appearance and geometric relations are computed. The use of 3D information is attractive since it allows us to extract view-point invariant (see Fig. 4) features in terms of geometric relations between 3D entities (such as distance or angle). The fact that we operate in a limited and controlled workspace leads to very reliable 3D and appearance information. In our representations, both aspects—shape and color—are represented separately, allowing us to investigate their relative importance. This space of feature relations (in the following also called 'relational space') can be expressed in (potentially high-dimensional) histograms providing unique and interpretable descriptors for specific objects (e.g., the distance between two parallel surfaces, see Sect. V) which, besides being view-point invariant, is also rather specific for a certain object. As we will show in this paper, this is useful for efficient learning since relatively few object recordings are required to learn representations for reliable object recognition.

As a classification algorithm, a multi-class Random Forest approach [3], [4] is applied in the context of this paper. Random Forests (RFs) have been found to be efficient since they combine the simplicity of decision trees with the stability of voting methods. In our context, it is of particular importance that by means of RFs, the relevant features for a task can be identified, i.e., that they inherently do feature selection. This allows for an interpretation of in-between stages of the algorithm, and for efficient classification. The algorithm is trained with a set of real objects represented with a combination of the their relations and appearance histograms (see Sect. V).

[1]http://www.xbox.com/kinect

The main achievements of our work can be summarized as follows:

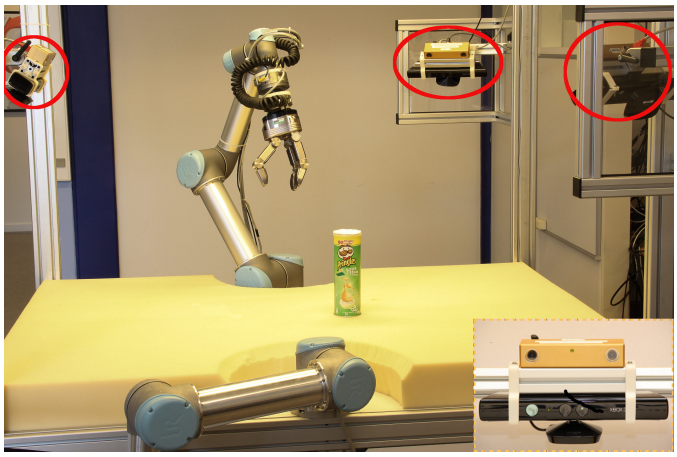
- we demonstrate the potential of applying 3D view-point invariant relations by achieving high-performance classification with very few training samples. The remaining miss-classifications are caused by the small size of some objects relative to sensor resolution that prevents the representation coding shape differences effectively.
- we can achieve a significant improvement in performance by using multiple cameras comparing to single—or even two—cameras. This is due to the fact that certain significant aspects of objects are expressed in our representation by relations which only manifest themselves with a rather complete 3D representation, only achievable by means of 3 views from rather different perspectives.
- we can show that our approach, when applied to Kinect sensor data, has a slightly better performance in comparison with the sensor data extracted by standard stereo cameras.
- we can show that, even in a very controlled environment as in our setup with color mainly produced by one light source, shape is a much stronger feature than color and that, at least in our object dataset, the combination of color and shape only slightly outperforms shape alone.

II. State of the art

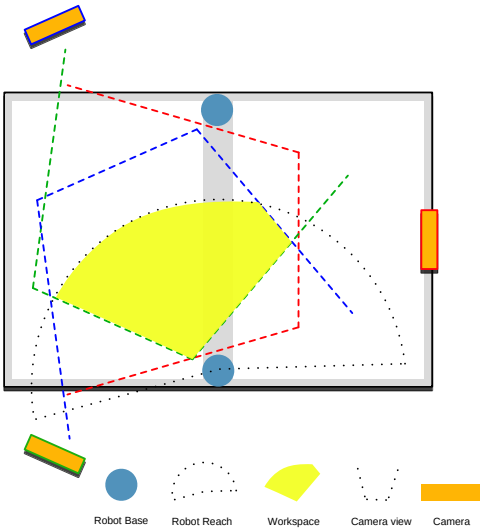
Object recognition in industrial setups: Object recognition has been used in industrial production set-ups mainly for the identification for a small set of objects (mostly one). Such approaches are nowadays part of standard vision packages such as HALCON[2] and have mostly been applied to 2D images. Recently, approaches using 3D data have entered the market [5]. The novelty of our approach lies in the explicit use of multiple simultaneously recorded views, utilizing view-point invariant relations that can only be generated based on the combination of all three views. By that, we show that we can achieve object recognition in controlled—but, from a point of view of industrial production, realistic—environments with a large number of objects attaining close to 100% performance with a few training samples. This allows systems to perform object recognition for assembly processes with some complexity, based on visual information.

ECV contour and surface representation: The visual representation used for object recognition is a subset of the Early Cognitive Vision (ECV) system as described in [6]. While the ECV system provides a deep hierarchy [7] of edge and surface features at different level of granularity, in the work described in this paper we only make use of local surface descriptors (which we call 'texlets') which are quite similar to descriptors such as 'patchlets' developed by other authors (see [8]). A distinguishing feature of our representation is the use of view-point invariant descriptors based on *global 3D relations* between local entities. We want to stress that a more extensive use of the different levels of the hierarchy in the ECV system is expected to lead to even better performance of the system.

[2]http://www.mvtec.com/halcon/



(a)



(b)

Fig. 1. The setup. (a) overview of the setup showing the robot arms and the camera pairs (the close-up view shows one pair of Kinect and stereo). (b) a top-view sketch for the setup depicting the workspace.

Shape relations: Relative shape information as visual descriptors has been applied under the concept of shape context introduced in [9]. The shape context of a point encodes the relative distribution of other points on the shape. It has been used as such to perform point-to-point matching in 2D. In [10], the shape context was extended to 3D, and defined for a local neighborhood. The difference to our approach is the global scope of our relations, and that we use supervised learning for classification.

Object recognition and classification learning, Random Forest: The problem of object recognition and classification has been intensively studied over the last decades, as evidenced by the annual PASCAL challenges e.g. [1], [11], that promoted rigorous evaluation and comparison of object recognition algorithms. Despite this, some criticism has been raised that

the typical visual class recognition may learn pose and context-specific features rather than the object itself, notably Nicolas Pinto and colleagues showed that a simple model of the V1 cortical area of the human brain could perform well on a typical natural image benchmark [12]. We can distinguish between two classes of classification algorithms: the first class of methods effectively performs image *retrieval* and is based on nearest-neighbour matching (eg., [13]); the second makes use of discriminative classification algorithms (such as Support Vector Machines [14] or Boosting [15]). Generally, discriminative approaches lead to higher classification performance, but can suffer from poor generalization when using weak visual features or when the variety of the training data is too limited. This work differs in many ways from traditional approaches to object recognition: First, it is based on a multi-view setup that is specific to an industrial scenario, aiming at high level recognition performance; Second, this set-up allows us to develop a feature describing the objects' 3D-shape in a pose-invariant fashion, allowing the robust use of discriminative classification methods.

III. The Setup (Intelligent Work-Cell)

The environment in which we want to solve the object recognition task is an intelligent robot work-cell (used e.g., in industrial assembly processes). The work-cell consists of two robot arms performing manipulation tasks with a variety of objects. For vision, 3 pairs of Kinect and Stereo cameras are mounted in a close to equilateral triangular configuration. By having such a configuration, a rather complete representation of the objects' 3D shape is obtained by combining the three views. Fig. 1a shows an overview of the setup and the camera pairs in use.

Fig. 1b is a sketch (plan view) of the setup, showing the field of view of each camera and the area of reach of the main robot arm. The yellow-shaded area depicts the workspace in which our system operates. The workspace is defined by the intersection of the three fields of view and the reach area. The requirement that all cameras cover the area is strictly limiting the useable workspace. On the other hand, for complicated manipulation tasks, such as the ones supposed to be executed in this setup, high performance of object recognition is needed. As we show in this paper, having multiple views can enhance the performance significantly by providing a rather complete 3D representation allowing encoding of a rich set of relations which are not available in single views (e.g., opposite surfaces). Furthermore, such a multi-view approach should also increases robustness against occlusion.

IV. Visual Representation

The visual representation for the object recognition system proposed in this paper is acquired through the Early Cognitive Vision (ECV) system [6]. The ECV system produces a hierarchy of visual entities in both 2D and 3D spaces with a multi-modal description for each entity. This description contains geometric attributes, appearance attributes and uncertainty estimates (for details see [6]). The ECV hierarchy works in two domains; edge domain and surface domain where entities lie in different levels, from low-level features (such as line segments and texlets) to high-level ones (such as contour and surflings) [7]. With respect to sensors, ECV supports both the classical stereo vision (in edge and surface domains) as well as Kinect cameras [16] (in the surface domain). The proposed method in this paper considers only the low-level aspects of the ECV system, namely 3D texlets, as a visual representation for objects extracted by stereo and Kinect cameras. Fig. 2 shows the 3D texlets extraction hierarchy. When operated in GPU, 3D texlets extraction with Kinect can be achieved with approximately 5 Hz [16].

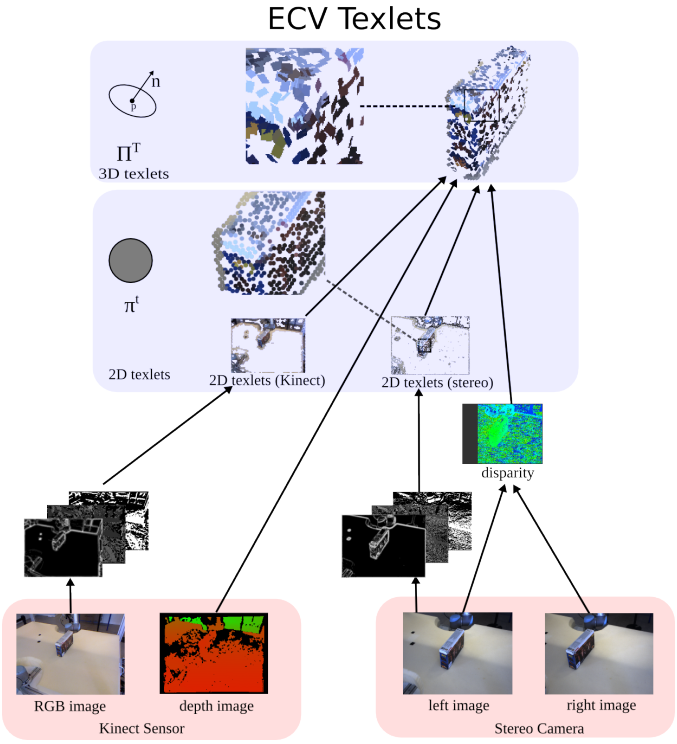


Fig. 2. The hierarchical representation of the ECV texlets. Example images from Kinect and stereo cameras are shown at the bottom. In the middle, 2D texlets are extracted after filtering operations. On top are the extracted 3D texlets from different cameras. This figure is best viewed in color.

Based on the extracted 3D texlets, our method transfers the absolute geometric information (3D position and orientation) into a relational space. To do so, we define a set of geometric relations to encode the shape information of the object. Using relations in this way provides an explicit and view-point invariant representation (see Fig. 4). More specifically, for a pair of 3D texlets (Π_i^T and Π_j^T) we compute: an angle relation $R_a(\Pi_i^T, \Pi_j^T)$, an Euclidean distance relation $R_d(\Pi_i^T, \Pi_j^T)$ and a normal distance relation $R_{nd}(\Pi_i^T, \Pi_j^T)$. Fig. 3a depicts how those relations are defined.

The overall set of all relations is used for representing the object. For example, if we want to represent an object O with 3D texlets angle relations, we define a matrix representation O_a^T as:

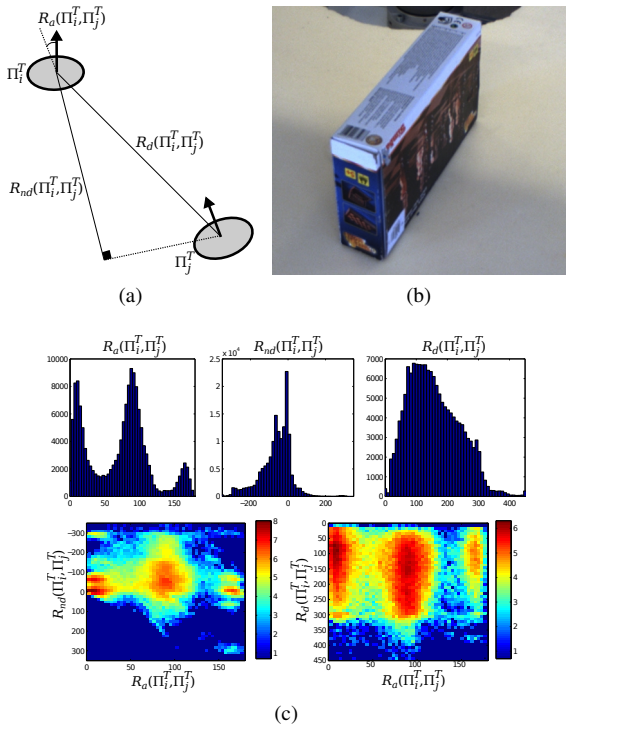


Fig. 3. Texlet relations. a) definition of shape relations between texlet Π_i^T and texlet Π_j^T; Euclidean distance $R_d(\Pi_i^T,\Pi_j^T)$, angle $R_a(\Pi_i^T,\Pi_j^T)$, and normal distance $R_{nd}(\Pi_i^T,\Pi_j^T)$. c) relation histograms of all pairs of texlets extracted from object shown in (b).

$$O_a^T = \begin{pmatrix} R_a(\Pi_1^T,\Pi_1^T) & R_a(\Pi_1^T,\Pi_2^T) & \cdots & R_a(\Pi_1^T,\Pi_n^T) \\ R_a(\Pi_2^T,\Pi_1^T) & R_a(\Pi_2^T,\Pi_2^T) & \cdots & R_a(\Pi_2^T,\Pi_n^T) \\ \vdots & \vdots & \ddots & \vdots \\ R_a(\Pi_n^T,\Pi_1^T) & R_a(\Pi_n^T,\Pi_1^T) & \cdots & R_a(\Pi_n^T,\Pi_n^T) \end{pmatrix}$$

Where n is the total number of the 3D texlets extracted from O.

Similarly, we can get O_d^T and O_{nd}^T for Euclidean and normal distance relations, respectively. Using all or a subset of these three relations we get a visual shape representation for the object. Fig. 3c shows the relational representation of the object in Fig. 3b with histograms.

The appearance information is represented directly by the color components associated to the 3D Texlets. We use the HSV space for color encoding since it helps providing a rather stable representation in different lighting conditions. If we want, for example, to represent O with the H component of the 3D texlets, we define O_h^T by:

$$O_h^T = \begin{pmatrix} h_1^T & h_2^T & \cdots & h_n^T \end{pmatrix}$$

Where h_i^T for a 3D Texlet i is the H color component. The same applies to the rest of the color components (S and V).

V. Object Recognition

A. 2D histograms

The 3D texlets' shape relations and color information described above are used as an input for the object recognition system. The size of the shape relations of an object (e.g., O_a^T) is variable and determined by the number of feature extracted which depends on the object properties and the extraction parameters. The first step of the object recognition method proposed here is to compute histograms of the 3D texlets' shape relations and color. Traditionally, histograms are used as an estimate of the probability distribution for any data. Data granularity, in this case, is controlled through the histogram bin size. This also allows us to have a generic and fixed-length representation required by supervised learning.

In this paper, we propose using two-dimensional histograms incorporating two shape relations (e.g., with O_a^T and O_{nd}^T to get $Hist_{a,nd}^T$) or two appearance (color) components (e.g., O_h^T and O_v^T, to obtain $Hist_{h,v}^T$). By applying 2D histograms, we are able to acquire a more distinctive representation since it reveals the correlation across different dimensions within the data. As a result, the performance of the classification algorithm will be enhanced (see Fig. 8).

Fig. 4 shows both the 1D histograms and 2D histograms for 4 objects. The presented objects are two instances of a rectangular box (see Fig. 4a and 4b), one instance of a similar box which is, however, open at one side (Fig. 4c), and one instance of a cylindrical object (Fig. 4d). The first two objects are closed boxes with the same color but with different pose. The third object is an open box (i.e, the same as the first box but with a missing side) where the inside faces have a different color. Fig. 4 shows histograms corresponding to appearance information; hue and saturation in 1D histograms (i.e., $Hist_h^T$ and $Hist_s^T$, and in 2D histogram ($Hist_{h,s}^T$). The figure also shows histograms corresponding to shape: angle and normal distance in 1D ($Hist_a^T$ and $Hist_{nd}^T$), and in 2D histograms ($Hist_{h,s}^T$). All features in this figure are extracted from three combined views by three kinect sensors.

Looking at the appearance histograms (at the top right of each block in Fig. 4) of the first and the second objects, we can notice that they are very similar. Analogously, the histograms representing the shape relations (at the bottom right of each block) of the first and and the second objects are very similar although a significant pose change has occurred. This exemplifies the view-point invariance of the shape relations.

The variation in color visible in Fig. 4c (in which the brown inside faces are visible) with respect to the first and the second objects is reflected in the appearance histograms being different (an additional peak value at hue 0.9 and saturation 0.1 appears). The shape histograms as well the appearance histograms for the cylindric object (Fig. 4d) are significantly different from the box like objects.

Regarding the shape relations, we can see that the 2D histograms reveal certain peak values at different combinations of angle and normal distance. These peaks, for Fig. 4a, 4b and 4c, correspond to the dimensions of the box, i.e., the distances

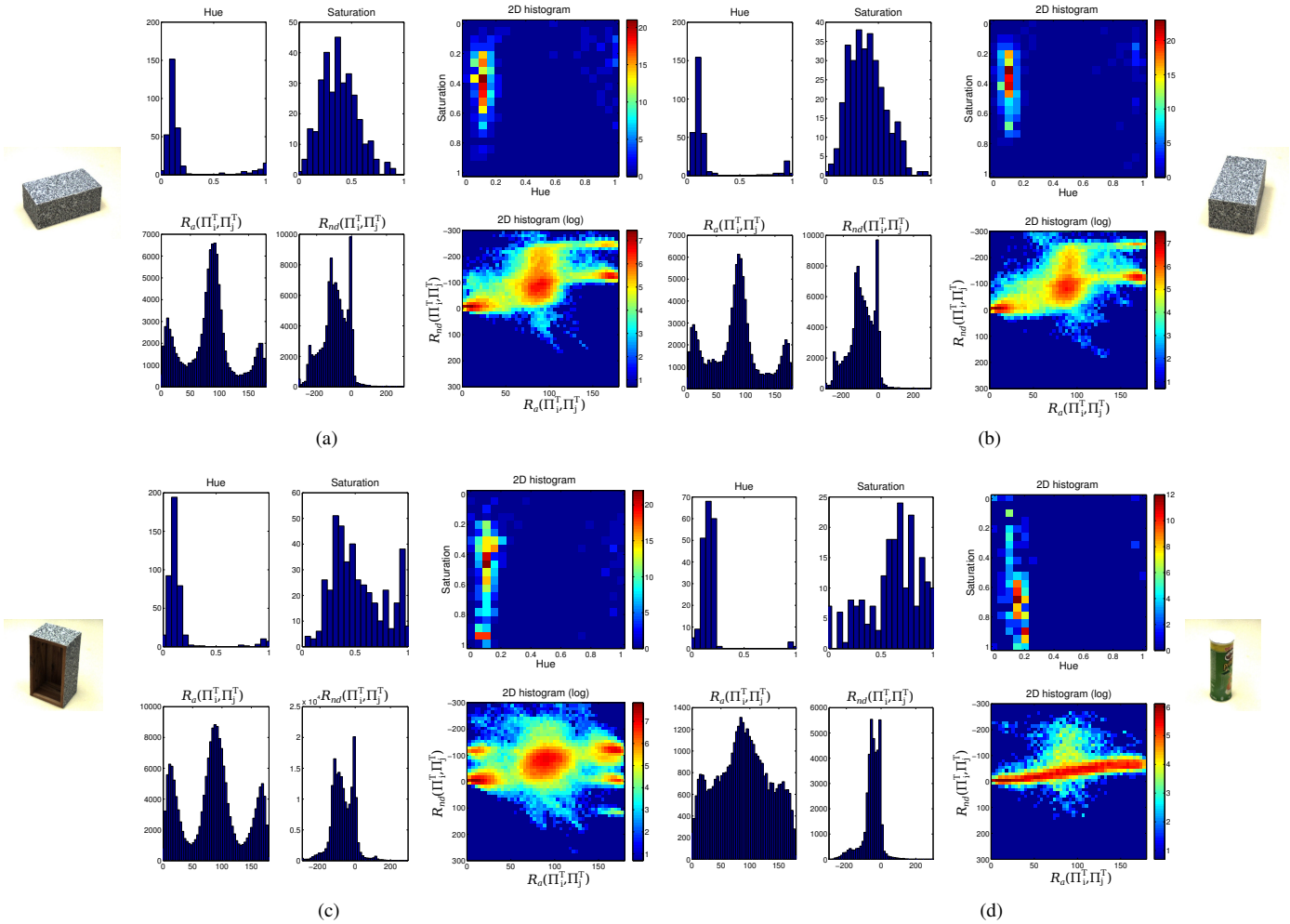


Fig. 4. Four different scene configurations and corresponding histograms. The histogram blocks for each scene consists of the following components: (top row) Two appearance histograms representing the hue (H) and the saturation color information of the extracted texlets, while the right most histogram shows the second order hue vs. saturation histogram. (bottom row, left) shows two aspects of the geometric information: (1) The histogram corresponding to the angles $R_a(\Pi^T_i, \Pi^T_j)$ of pairs of texlets, (2) their normal distance $R_{nd}(\Pi^T_i, \Pi^T_j)$. (bottom row, right) The two dimensional histogram spanned by angle $R_a(\Pi^T_i, \Pi^T_j)$ and normal distance $R_{nd}(\Pi^T_i, \Pi^T_j)$ for all possible pairs of texlets.

of the parallel planes. Note that negative value indicates an outward direction. Peaks at a specific normal distance with angle of about $0°$ represent parallel surfaces pointing in the same direction (i.e., the table and the top surface) and of about $180°$ degrees when they are pointing in opposite directions (the four side planes) indicating the occurrence of additional surfaces.

For $Hist^T_{a,nd}$ of object Fig. 4c, we can see a peak value at about $180°/120\,mm$ that does not exist in the other box objects. The openness of the box in this case allows for the extraction of texlets from the inside faces. As a result of that, we obtain texlets pointing in opposite directions. In the case of the fourth object Fig. 4d, we can see again that we obtain a completely different shape histogram making a clear distinction from the other cases.

In order to show that 2D histograms are more distinctive than two 1D histograms combined, we refer to $Hist^T_{a,nd}$ in Fig. 4c (2D histogram at bottom right). The histogram shows some object-specific peak values, among those the one at about $0°/-120\,mm$. When observing the 1D histogram $Hist^T_a$, we can notice a peak at about $0°$. This peak value exist in all objects and is by no means distinctive. So, using two 1D histograms will result in not exploiting $0°$ angle relations to classify the object while the 2D histogram will make it possible. This angle also accounts for the peak of $Hist^T_{a,nd}$ at about $0°/0\,mm$ which is a redundant component.

The above examples showed that, for all similar objects, very similar histograms of color and shape are obtained despite their differences in pose. Also, it shows that the shape relations code object properties in an easily identifiable way as individual peaks in histograms. They also show the potential benefit of using 2D histogram in order to obtain a more distinctive representation for better selection of features in the classification algorithm.

B. Random forests

The quality and invariance properties of the histogram representation presented in the previous section makes it attractive for the purpose of object recognition. Supervised classification is a field that is well explored in machine learning (eg, [17], [18]). In this work, we make use of Random Forest classification [3], [4]. The reasons for this choice are multiple: first, RF can be trained efficiently and are very fast at classification time, even for large input dimensions; second, RF are intrinsically multi-class allowing for an efficient learning in contrast to 1vs. all approaches; finally, RF have shown to reach very high level of performance on a variety of tasks (notably [19], [20]); finally, RFs effectively perform a form of dimension selection and the models are interpretable. The Random Forests are learnt using the concept of *Bagging* to learn a population of randomized decision trees. If we consider a dataset $D = (x_j, y_j)_{j\in[1..|D|]}$, where x_j is an observation (here it is a 2D histogram e.g., $Hist^T_{a,nd}$) and $y_j \in [1..C]$ is a class label and $|D|$ denotes the number of samples in D, then Bagging splits the dataset into M subsets $D_i \subset D$, $\forall i \in [1..M]$, and training a population of classifiers $F = T_{i i\in[1..M]}$ such that T_i is trained from the subset D_i. Typically, the subsets D_i are drawn randomly such that $|D_i| = \sigma|D|$ (we used a common value of $\sigma = 0.5$). From each subset D_i, we train a Randomized Classification Tree (RCT). RCT are binary trees, where each node n splits the input space (and thereby the dataset such that $D_l \cup D_r = D_n$) recursively in order to maximize class purity in all partitions, and sending the samples that fall on each side of the partition to each child node. The recursion stops when a node receives too few samples to split ($|D_n| < 5$ here) or reaches a maximum depth ($\text{depth}(n) > 10$ here)—such nodes are called leaf nodes, and label the corresponding region according to the majority label in the available samples. The split operation is done along a hyperplane, by applying a threshold operation to one input dimension. The learning procedure selects a random set of input dimensions, and finds the optimal dimension and threshold amongst them, by minimizing all partitions' class impurity, using the so–called Gini coefficient $G(D_n)$:

$$G(D_n) = 1 - \sum_{k=1}^{C} \left(\sum_{j=1}^{|D_n|} \frac{\mathbf{I}_k(y_j)}{|D_n|} \right)^2 , \qquad (1)$$

where $\mathbf{I}_k(y_j)$ is an indicator function that returns 1 if $y_j = k$, and 0 otherwise.

Finally, the RF response $F(x)$ for an input vector is obtained by calculating the class with the largest amount of votes amongst all RCTs T_i.

$$F(x) = \arg \max_{k\in[1..C]} \sum_{i\in[1..M]} \mathbf{I}_k(T_i(x_i)) \qquad (2)$$

The hierachical greedy search for splits allows for a high performance classification, while the randomization and redundancy provided by the bagging reduces the model's overfitting, increasing generalization and robustness.

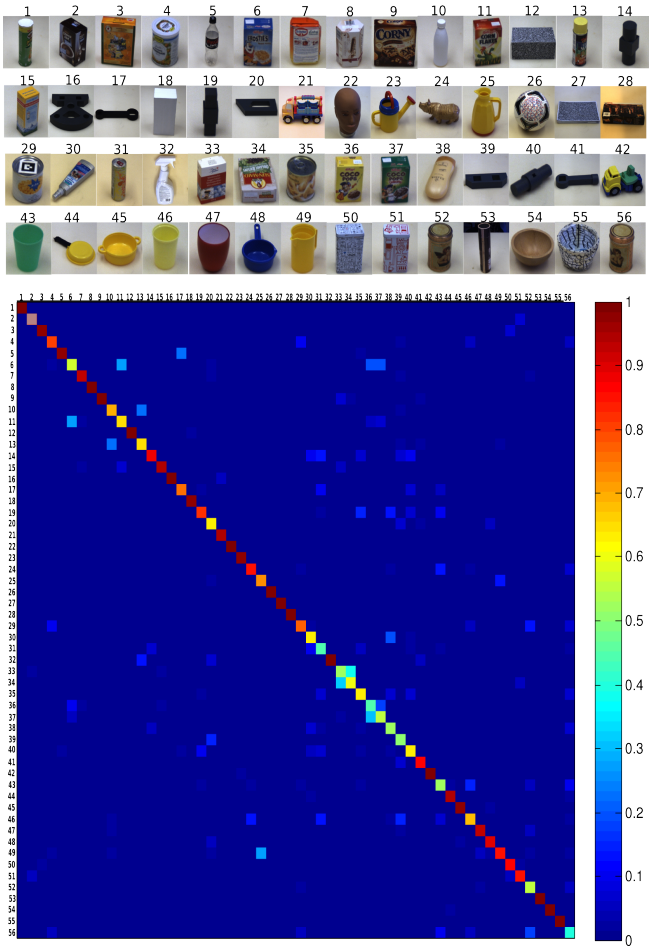


Fig. 5. Classification confusion matrix of 56 objects (shown above). It is obtained from a testing subset of 8 samples for each object while 12 samples are used for training. The visual representation is extracted from 3 Kinect cameras

VI. Data Set and Experiment

The dataset we use in this experiment consists of 56 objects. Objects are different in shape, size and color. 20 different samples are obtained from each object by placing them within the workspace (see Fig. 1b) covered by all the cameras in the setup. For generality, different poses per object were captured during the sampling. In all the following experiments, a subset of the samples is used for training and another for testing, both subsets are randomly selected. Also note that all results presented here are generated from the test subset. Regarding histogram calculations, we set the number of bins to 20 in all the following tests. For Random Forest, we train 100 trees with a bagging ratio of 0.5.

Fig. 5 shows the confusion matrix of classification for all the 56 objects with 12 training samples and 8 testing samples each. The figure shows the confusion matrix when $Hist^T_{a,nd}$ (see Sect. V) is used as feature vector for the Random Forest. The diagonal of the matrix reflects the accuracy of the classification. We can see that the confusion is generally low for most of the objects. Furthermore, the matrix shows that

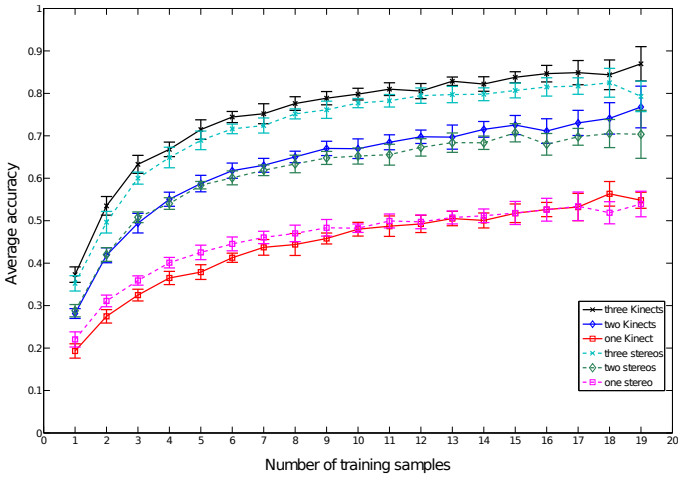


Fig. 6. Performance comparison of classification on single, double and triple views taken from both Kinect and Stereo cameras.

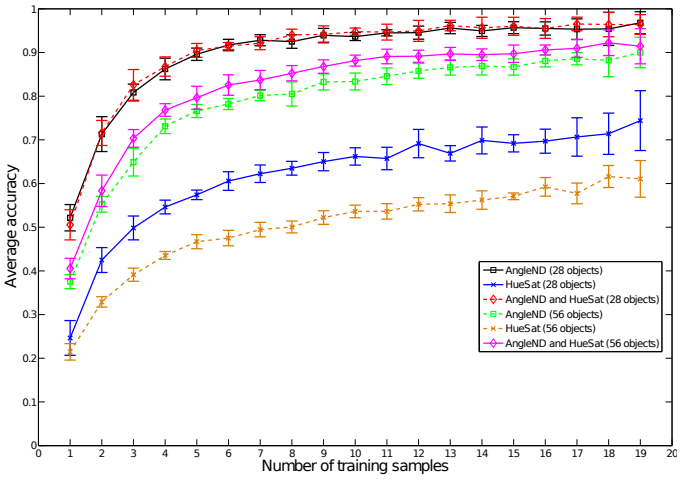


Fig. 7. Performance comparison for shape alone, color alone, and the combination of both. Shown for all objects in the dataset (56 objects) and for only half of them (28 objects).

objects with special shape or size obtained high classification accuracy while object with similar shape and size get low accuracy (e.g., objects 6,11,36 and 37).

In the subsequent results, we consider the trace mean of the confusion matrix as a single accuracy metric of classification for the whole set of objects. Then, its average on 10 different runs as well as the standard deviation (which is depicted as a bar in figures) are computed. For all figures, the average accuracy is drawn against different numbers of training samples.

Fig. 6 shows the classification average accuracy over different number of views (one, two and three) for Kinect as well as for stereo. The classification here is based only on shape features (we use $Hist^T_{a,nd}$). We are interested to make a performance comparison when different number of views are used, and a comparison between Kinect and stereo sensors. First to notice that even with a small number of training samples, we get a rather high classification accuracy (for three views acquired by Kinect, we get above 80% of accuracy with 13 samples only). The figure also shows the performance enhancement achieved by increasing the number of views. Furthermore, it shows that the Kinect sensor has a slightly better performance compared to stereo.

In Fig. 7, we show the performance when only shape information ($Hist^T_{a,nd}$) is used, when only color information ($Hist^T_{h,s}$) is used, and when both are combined ($Hist^T_{a,nd}$+$Hist^T_{h,s}$). This is also presented for half of the dataset (objects in the first two rows in Fig. 5) to show the effect of the dataset size. The figure shows that using color alone (in contrast to shape) results in poor performance (worse for bigger dataset). This is consistent with the fact that we cannot distinguish objects (at least in our dataset) based only on their appearance. However, adding color information does not make a significant improvement comparing to the shape alone case. Even though we do expect, generally speaking, high improvement when we add color to shape, this result can be explained in the case of our dataset because few objects share the same shape while differing in color.

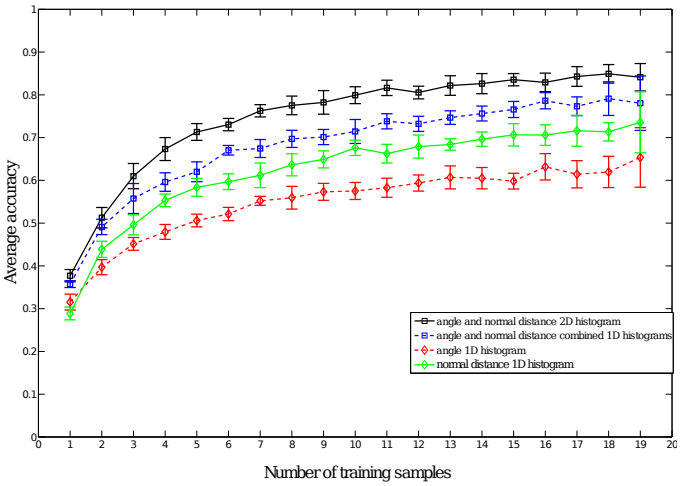


Fig. 8. 2D histogram of two features compared to combined 1D histograms, and when using only one feature at a time.

The performance of using a single 2D histogram in comparison with combining two 1D histograms is shown in Fig. 8. The figure shows that about 10% increase in performance is achieved through using $Hist^T_{a,nd}$instead of representing them with two combined 1D histograms ($Hist^T_a$ + $Hist^T_{nd}$). It also shows the performance when $Hist^T_a$ and $Hist^T_{nd}$ are individually used.

VII. Conclusion

We presented an object recognition system for an intelligent work-cell. Our system represents objects with view-point invariant 3D shape relations along with appearance information. Using multi-view representation (from three cameras) allows for rather complete 3D information. For object classification, the system is trained using Random Forest with a dataset of 56 objects.

The results show that our system can achieve high performance (in terms of classification accuracy) with a few training

samples. We have also demonstrated that by using multiple-view, a significant improvement in performance, comparing to a single-view, can be achieved. This is due to the fact that, with more complete 3D information, more shape relations can be detected and learned by our system. Comparing Kinect sensor to stereo vision, our results indicate that with Kinect we can obtain a slightly better performance. Finally, we showed that shape is a much stronger feature than appearance when compared separately, and that (at least on our dataset) the combination of both outperforms slightly shape alone.

As future work, we plan further investigations on using higher-order shape relations (relations between more than two features). This might lead to an even more distinctive representation that would enhance the performance of our system further.

ACKNOWLEDGMENT

This work has been funded by the EU project Xperience (FP7-ICT-270273).

REFERENCES

[1] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009)," Summary presentation at the 2009 PASCAL VOC workshop, 10 2009.

[2] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2011, pp. 1817–1824.

[3] Y. Amit and D. G. Y, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, pp. 1545–1588, 1997.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] A. M. Pinto, L. F. Rocha, and A. P. Moreira, "Object recognition using laser range finder and machine learning techniques," *Robotics and Computer-Integrated Manufacturing*, vol. 29, no. 1, pp. 12 – 22, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0736584512000798

[6] N. Pugeault, F. Wörgötter, and N. Krüger, "Visual primitives: Local, condensed, and semantically rich visual descriptors and their applications in robotics," *International Journal of Humanoid Robotics (Special Issue on Cognitive Humanoid Vision)*, vol. 7, no. 3, pp. 379–405, 2010.

[7] G. Kootstra, M. Popovic, J. Jørgensen, K. Kuklinski, K. Miatliuk, D. Kragic, and N. Kruger, "Enabling grasping of unknown objects through a synergistic use of edge and surface information," *The International Journal of Robotics Research*, vol. 31, no. 10, pp. 1190–1213, 2012. [Online]. Available: http://ijr.sagepub.com/content/31/10/1190.abstract

[8] D. Murray and J. Little, "Patchlets: Representing stereo vision data with surface elements," in *ars in: Seventh IEEE Workshops on Application of Computer Vision. WACV*, vol. 1, 2005, pp. 192–199.

[9] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 509–522, Apr. 2002. [Online]. Available: http://dx.doi.org/10.1109/34.993558

[10] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, May 2004.

[11] M. Everingham, A. Zisserman, C. K. I. Williams, L. van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. S. Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang, "The 2005 PASCAL Visual Object Classes Challenge," in *Pascal Challenges Workshop*, ser. LNAI. Springer, 2006, vol. 3944, pp. 117–176.

[12] N. Pinto, J. J. DiCarlo, and D. D. Cox, "How far can you get with a modern face recognition test set using only simple features?" 2009.

[13] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150 –1157 vol.2.

[14] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[15] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 994–1000. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.254

[16] S. M. Olesen, S. Lyder, D. Kraft, N. Krüger, and J. B. Jessen, "Real-time extraction of surface patches with associated uncertainties by means of kinect cameras," *Journal of Real-Time Image Processing*, pp. 1–14, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11554-012-0261-x

[17] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[19] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[20] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011, pp. 1297 –1304.