Autonomous Robots (2019) 43:1005–1022
https://doi.org/10.1007/s10514-018-9776-8

Viewpoint invariant semantic object and scene categorization with RGB-D sensors

Hasan F. M. Zaki[1] · **Faisal Shafait**[2] · **Ajmal Mian**[3]

Received: 28 November 2016 / Accepted: 13 June 2018 / Published online: 5 July 2018

Abstract
Understanding the semantics of objects and scenes using multi-modal RGB-D sensors serves many robotics applications. Key challenges for accurate RGB-D image recognition are the scarcity of training data, variations due to viewpoint changes and the heterogeneous nature of the data. We address these problems and propose a generic deep learning framework based on a pre-trained convolutional neural network, as a feature extractor for both the colour and depth channels. We propose a rich multi-scale feature representation, referred to as convolutional hypercube pyramid (HP-CNN), that is able to encode discriminative information from the convolutional tensors at different levels of detail. We also present a technique to fuse the proposed HP-CNN with the activations of fully connected neurons based on an extreme learning machine classifier in a late fusion scheme which leads to a highly discriminative and compact representation. To further improve performance, we devise HP-CNN-T which is a view-invariant descriptor extracted from a multi-view 3D object pose (M3DOP) model. M3DOP is learned from over 140,000 RGB-D images that are synthetically generated by rendering CAD models from different viewpoints. Extensive evaluations on four RGB-D object and scene recognition datasets demonstrate that our HP-CNN and HP-CNN-T consistently outperforms state-of-the-art methods for several recognition tasks by a significant margin.

Keywords Object categorization · Scene recognition · RGB-D image · Multi-modal deep learning

1 Introduction

In realizing long-term mobile robot autonomy, semantic scene and object understanding capabilities are crucial and have gained considerable research attention in the past decade (Asif et al. 2015b; Bo et al. 2011, 2012; Lai et al. 2011; Lowry et al. 2016; Schwarz et al. 2015; Zaki et al. 2016). Generally, developing a highly accurate and robust recognition system involves training the robotic vision in an off-line mode where the robot is given a set of labelled training data and is then asked to predict the semantics of novel instances in a complex environment. Performing scene and object recognition can be very challenging due to a range of factors; drastic illumination and viewpoint changes, heavy clutter, occlusions and the problem of perceptual aliasing where two or more scenes may look extremely similar, forcing the robot to define subtle discriminative features for recognition (Angeli et al. 2008; Lowry et al. 2016).

In developing a highly accurate and robust visual recognition system, certain design criteria and their contributing factors should be put into consideration. Firstly, the recognition system should acquire a good generalization capability for various domain types so that the robot can adapt to a novel environment which are significantly peculiar from the one that the robot has been trained on. To realize this criterion, the system must be presented with a set of labelled training data containing all possible variations. Furthermore, the learned system should be robust to the effect of intra- and inter-class variability that can be a challenging nuisance for recognition. Therefore, the feature representation must be descriptive of the visual elements being captured as well as discrimi-

✉ Hasan F. M. Zaki
hasan.mohdzaki@research.uwa.edu.au

Faisal Shafait
faisal.shafait@seeks.edu.pk

Ajmal Mian
ajmal.mian@uwa.edu.au

[1] Department of Mechatronics Engineering, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia

[2] National University of Sciences and Technology, Islamabad, Pakistan

[3] School of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia

native to differentiate between two different elements that may appear very similar. Moreover, the recognition algorithm must be efficient to perform real-time decisions which is essential for robotics applications. The advent of low-cost multi-modality sensors such as RGB-D cameras has opened up a number of possibilities to fulfil these design considerations.

RGB based visual recognition algorithms have moved from hand-crafting features to learning features through deep neural networks. Particularly, convolutional neural networks (CNN) based methods have obtained unprecedented performance for RGB visual recognition (Krizhevsky et al. 2012), where the success is mainly credited to the availability of computational resources and large-scale labelled training datasets with diverse variations [e.g. ImageNet (Deng et al. 2009)]. However, labelled training data for RGB-D recognition is currently limited and manual annotation of images captured by low-cost depth sensors such as microsoft kinect camera is time consuming and expensive. Additionally, unlike the RGB based recognition which could benefit from high resolution data, RGB-D sensors have low resolution, capture noisy data and contain multi-modal heterogeneous data which pose challenges for the learning algorithms. To avoid these problems, recent works (Lai et al. 2011; Bo et al. 2011, 2012; Socher et al. 2012; Cheng et al. 2014; Asif et al. 2015b; Liu et al. 2015b; Jhuo et al. 2014; Zaki et al. 2015) have resorted to shallow networks which are more tractable to train using limited amount of data. Deep neural networks require large-scale annotated training data otherwise they tend to result in poor convergence and overfitting (Bengio et al. 2013).

Recent development has shown that CNN models that were trained on a large-scale datasets can be effectively used as a generic feature extractor for a wide range of other applications (Krizhevsky et al. 2012; Chatfield et al. 2014; Razavian et al. 2014; Gupta et al. 2014; Azizpour et al. 2016), even without re-training on the target tasks. The factors of transferability of the learned features in CNN can vary from the architecture of network and data distribution (Azizpour et al. 2016) but most techniques extracted the fully-connected layer activations before the classification layer as a feature representation, leaving the antecedent convolutional layers relatively unexplored. Although the former acquires high degree of semantically descriptive information, it does not preserve spatially relevant information of the input (Hariharan et al. 2015). Therefore, they are less effective in capturing subtle details (He et al. 2015). Expensive pre-processing steps such as data augmentation and segmentation (Chatfield et al. 2014; Razavian et al. 2014) are generally carried out as a complimentary factor for the recognition systems.

In this paper, we propose an effective recognition framework based on a deep CNN with a particular attention to address the above problems. Firstly, we formulate a technique for transferring knowledge of input from depth sensors to a model well-trained on a large-scale RGB data which allows a seamless integration between these modalities. Next, we propose a global feature representation which utilizes the activations of all convolutional layers that is able to encode coarse-to-fine information of the modalities. We term our proposed representation as convolutional hypercube pyramid (HP-CNN) which is used in conjunction with the semantically-descriptive features from the fully connected layer. The encoding of this representation is done by re-sampling the convolutional tensors into three pyramid levels. For each pyramid, multi-scale features are harvested by employing a spatial pyramid pooling method and the concatenation of the pooled features defines the representation for each pyramid. To average the behaviour of the features from these multi-scale pyramid, we then apply max pooling over these features to produce the final HP-CNN representation. Note that these encoding steps are done for all modalities. Finally, we propose a late feature fusion technique to combine the features of our HP-CNN and the fully connected layer from different modalities where we empirically show that it consistently increases the discriminative property and compactness of the feature representation. The overview of our HP-CNN encoding is depicted in Fig. 1.

This paper is a major extension of our previous work (Zaki et al. 2016) in a number of directions. Firstly, we perform extensive experiments on additional benchmark RGB-D visual scene/place recognition datasets including NYU v1 Indoor Scene (Silberman and Fergus 2011) and SUN RGB-D Scene (Song et al. 2015) besides the evaluation on object recognition datasets. Secondly, we propose a view-invariant version of HP-CNN, termed as HP-CNN-T which is extracted from a multi-view 3D object pose (M3DOP) model. M3DOP is a deep network learned end-to-end using RGB-D images of 3D CAD models rendered from multiple viewpoints. Finally, the experimental results and analysis based on the extended work (see Sect. 6) have reinforced our hypothesis in the previous work (Zaki et al. 2016) that the proposed method acquires a high degree of efficacy and generalization capability as a result of effective transfer learning and domain adaptation. In summary, our core contributions are as follow:

1. We present convolutional hypercube pyramid descriptor (HP-CNN) as a discriminative feature representation that encodes spatially-descriptive information for RGB-D image recognition.
2. We propose an effective encoding technique for depth sensor inputs to allow knowledge transfer to a model that was well-trained on RGB camera inputs (Sect. 3.2).
3. We combine the features from our HP-CNN and the fully connected layer activations using a feature fusion method based on extreme learning machines which not

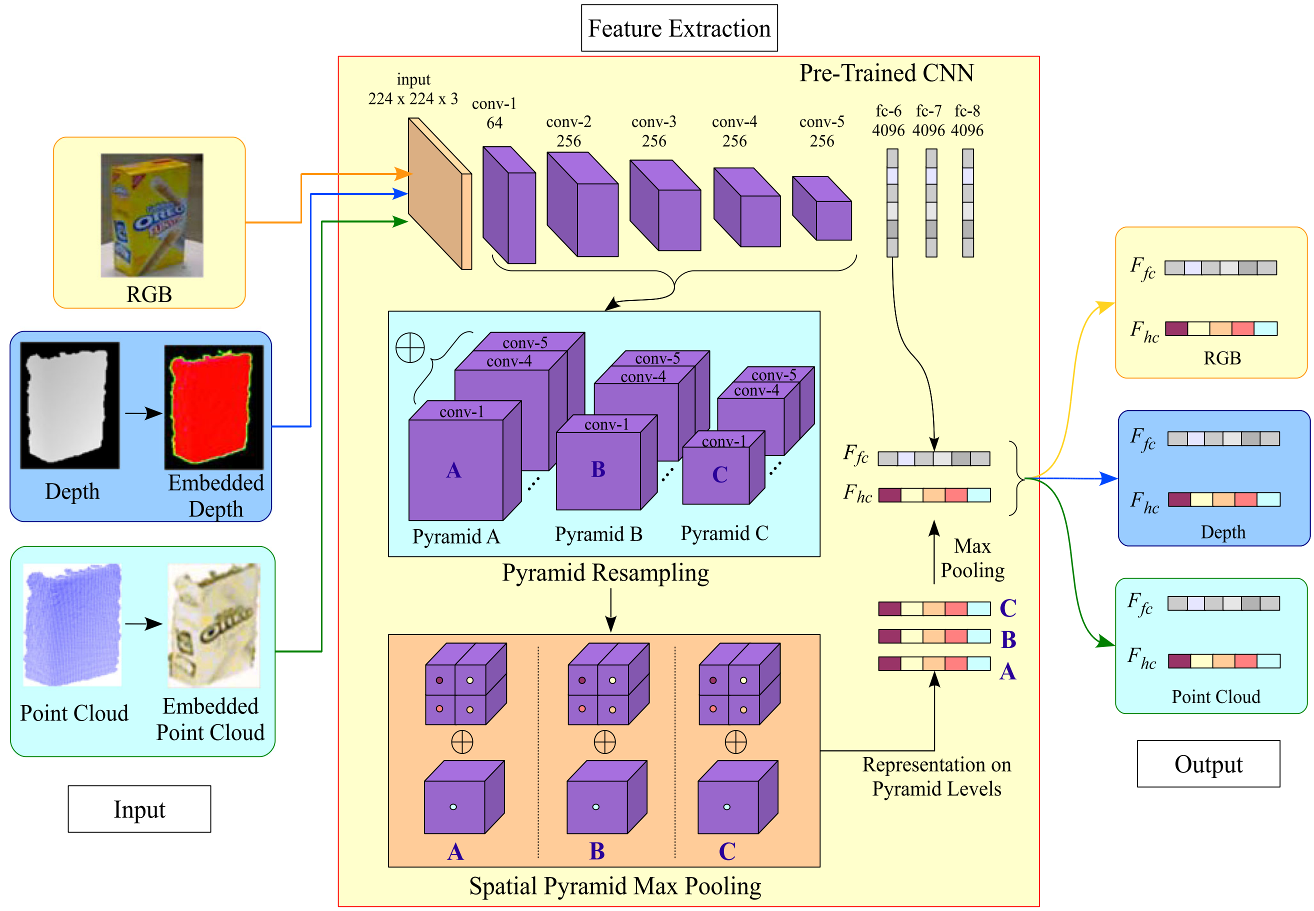

Fig. 1 Illustration of the proposed convolutional hypercube pyramid feature extraction. The feature representation is extracted in a coarse-to-fine manner separately for each RGB, depth and point cloud image. This is done by resampling the convolutional feature maps into three pyramid levels *A*, *B*, and *C* and concatenate all feature maps at each pyramid level separately (refer text for detailed description). The symbol ⊕ denotes concatenation

only reduces the dimensionality but also increases the discriminative properties of the features (Sect. 4).

4. We propose HP-CNN-T, a view-invariant descriptor extracted using multi-view 3D object pose (M3DOP) model which is trained on a newly generated synthetic dataset (Sect. 5) to extract 2D and 3D features. We show that the proposed HP-CNN not only generalizes well for cross-modality purposes, but also across different sets of applications. We will make the M3DOP model publicly available.

2 Related literature

2.1 RGB-D visual recognition

Channel-specific hand-crafted features are among the most prevalent methods used in prior works for colour and 3D visual recognition. Generally, a recognition system requires to densely compute descriptors such as SIFT (Lowe 2004) and spin images (Lai et al. 2011) that describe rapidly changing elements in localized regions of the image. The encoding of the global representation is done using a bag-of-visual-words (BoVW) method which seeks to quantize the densely sampled features into pre-determined visual vocabularies. However, these methods require the knowledge of the input distribution beforehand, which is normally not available in most real-time robotic applications (e.g. long-term navigation, object grasping).

Recent works have proposed feature learning methods to mitigate the need to encode channel-specific features where the features of multiple heterogeneous modalities can be readily learned using the same algorithm. These include the previously proposed convolutional K-Means (CKM; Blum et al. 2012), convolutional-recursive neural networks (CNN-RNN; Socher et al. 2012), hierarchical matching pursuit (HMP; Bo et al. 2012), deep regularized reconstruction independent component analysis (R^{2}ICA; Jhuo et al. 2014), cascaded random forests (CaRFs; Asif et al. 2015b), localized deep ELM (LDELM; Zaki et al. 2015) and discriminative feature learning (Asif et al. 2015a). However, these methods perform independent learning of the modality features and need to repeat the same process for each modality. Besides increasing the execution time, the methods do not address the input heterogeneity problem in designing the representation. Additionally, the models are trained from relatively limited training datasets which often leads to sub-optimal performance.

2.2 Deep learning based visual representation

Inspired by the remarkable performance of the new generation CNN for RGB visual recognition, many methods exploit the discriminative properties of the internal CNN layers for other cross-domain tasks. In a common scenario, a deep CNN with an abundance of non-linear computation units is trained from scratch on a large-scale dataset and the feed-forward activations at a certain layer are extracted as a feature representations for another dataset (Chatfield et al. 2014; Razavian et al. 2014; Azizpour et al. 2016; He et al. 2015; Zeiler and Fergus 2014). In most of these recognition tasks, the activations of the fully connected layers are used as the feature representation. This approach has shown promising performance when transferred to a vast array of target recognition tasks. However, recent works have demonstrated that the earlier convolutional layers in CNN also contain a degree of semantically meaningful features. Moreover, in contrast to the fully connected layers, the activations in the convolutional layers preserve the spatial locations of the input pixels. Therefore, these layers can be used to extract multi-scale features (Liu et al. 2015a; Yang and Ramanan 2015) and more precise localizations of the elements of a scene in the whole image (Hariharan et al. 2015). For example, Hariharan concatenated several consecutive convolutional layers and extracted activations in local windows as a fine-grained pixel representation. Liu extracted sub-arrays of a convolutional layer at the regions detected by the previous convolutional layers in a guided cross-layer pooling technique. In this paper, instead of using convolutional layers activations as local features, we devise a technique to encode a global representation of the input based on all convolutional layers as a unified feature tensor which leads to a compact yet powerful representation for RGB-D recognition tasks.

It is worth mentioning that the aforementioned methods performed the knowledge transfer of CNN model to different applications but still operated in the same modality space. Knowledge transfer across modalities (e.g. RGB to depth) is much more challenging as there exist an ambiguity of the relationship between heterogeneous modalities that represent different type of information (i.e. RGB pixels represents the colour while depth represents the distance from the sensor). Recent works have obtained encouraging results for cross-modality recognition by augmenting the target modality (i.e. depth) to be in a close resemblance to the source modality (i.e. RGB; Gupta et al. 2014; Schwarz et al. 2015). We devise two approaches to address this problem. Firstly, we propose the encoding of the depth maps and the point cloud so that they closely resemble the typical input of a pre-trained CNN model. Secondly, we learn deep models end-to-end specifically for RGB-D recognition tasks. Since RGB-D training data are expensive to acquire, we propose to learn these models using synthetic training data which are generated by rendering 3D models from multiple viewpoints. By doing so, we show that the recognition performance can significantly improve for real RGB-D data.

3 Convolutional hypercube pyramid

Objects and scenes captured in the real-world environment typically lie on non-linear manifolds, especially when dealing with various viewpoint changes and occlusion. However, most existing methods (Bo et al. 2012; Socher et al. 2012; Blum et al. 2012) for RGB-D image recognition extract features based on linear based encoding. Moreover, these methods need to be heavily tuned on specific datasets and applications. Inevitably, this will lead to huge computational burden as the model learning process repeats for each dataset and thus not scalable for new incoming data which is crucial for real-time robotics application. To overcome these drawbacks, we employ a deep CNN model, VGG-f (Chatfield et al. 2014) which has been pre-trained and optimized on a large-scale image dataset, ImageNet (Krizhevsky et al. 2012) for feature extraction on all datasets.

3.1 Feature extraction and encoding

Let us assume a CNN model which consists of consecutive modules of convolutional layer $L(k, f, s, p)$, max-pooling $MP(k, s)$, local contrast normalization LCN, fully connected layers $FC(n)$ and rectified linear unit (ReLU) RL, where $k \times k$ is the receptive field size, f is the number of filters, s denotes the stride of the convolution and p indicates the spatial padding. The architecture of the model is given by: $L(11, 64, 4, 0) \rightarrow RL \rightarrow LCN \rightarrow MP(3, 2) \rightarrow L(5, 256, 1, 2) \rightarrow RL \rightarrow LCN \rightarrow MP(3, 2) \rightarrow L(3, 256, 1, 1) \rightarrow RL \rightarrow L(3, 256, 1, 1) \rightarrow RL \rightarrow L(3, 256, 1, 1) \rightarrow RL \rightarrow MP(3, 2) \rightarrow FC(4096) \rightarrow RL \rightarrow FC(4096) \rightarrow RLFC(1000)$. Many methods (Razavian et al. 2014; Gupta et al. 2014; Schwarz et al. 2015) only consider the fully connected CNN layers as features. While these layers contain rich semantic information, they do not preserve the localized image information and ignore the convolutional layers. On the other hand, the convolutional layers carry locally-activated features (Hariharan et al. 2015; Liu et al. 2015a). We formulate an effective framework that extracts features from all convolutional layers hence preserving the spatially-relevant information in the feature representation. These features are then used alongside the holistic fully connected layer features to obtain a global and local visual representation.

Note that we perform the same feature extraction procedure for all modalities (RGB and Depth). In this section, we present the formulation of our proposed HP-CNN for only one modality (e.g. RGB). Therefore, we drop the modality-specific notation from our description. The feature maps activation can be visualized as an $i \times j \times n$ convolutional tensor for each convolutional layer $l^{(i)} = \{l^{(1)}, \ldots, l^{(L)}\}$. Each convolutional node can be expressed mathematically as

$$a^{(l)}_{i,j,n^{(l)}} = \sigma\left(\sum_{w,h,c} k_{w,h,c,n^{(1)}} * a^{(l-1)}_{i-w,j-h,c} + b^{(l)}_{i,j}\right), \quad (1)$$

where b is a bias term and $\sigma(.)$ denotes the non-linear function ReLU (rectified linear unit). The three-dimensional w-by-h-by-c learned filter kernels are indicated by k such that it convolves the c feature maps at previous layer $(l-1)$ to produce n feature maps with dimension i-by-j at the current layer l. The number of feature maps (i.e. the convolutional layer depth) in each convolutional layer is $n^{(l)} = \{64, 256, 256, 256, 256\}$ giving $N = 1088$ feature maps as shown in Fig. 1.

We convert the feature maps into the HP-CNN representation so that it encodes multi-scale information. To do this, each convolutional feature map is first sub-sampled using bilinear interpolation into three pyramid levels by sub-sampling the spatial dimension (i, j) of each feature map in all convolutional layers into $p^{(1)} = m \times m$, $p^{(2)} = 2m \times 2m$ and $p^{(3)} = 0.5m \times 0.5m$ respectively. The sub-sampling captures distinctive features of the convolutional layers at multiple scales (Lowe 2004). The sub-samples are then concatenated along the depth dimension at each pyramid level to produce a pyramid of Hypercube descriptors (see Fig. 1 for illustration). The hypercube at each pyramid level P is given by

$$H_P = \left[a^{(1)}_{p,n^{(1)}}, a^{(2)}_{p,n^{(2)}}, \ldots a^{(L)}_{p,n^{(L)}}\right], \quad (2)$$

with $H_k \in \mathbb{R}^{p^{(k)} \times N}$ where $k = 1, \ldots, P$.

This operation produces three different size Hypercubes. To enhance the discriminative properties of the descriptors and reduce the hypercube dimensionality spatial pyramid max pooling (SPM; Bo et al. 2012) is performed. The hypercube at each pyramid level is divided into two SPM levels. The complete hypercube is used as one cell for SPM level one, whereas each hypercube is partitioned into four cells of equal size for the second SPM level. The pooled feature vectors for each cell are then calculated through component-wise maximum over all feature maps within that cell. Note that the features extracted from each cell has dimension equal to the depth (N) of the respective hypercube. Thus, five equal-dimensional feature vectors are generated and concatenated to create a single vector per pyramid level. Finally, max pooling is done to combine the feature vectors of the three pyramid levels and produce a compact discriminative representation $F_{hc} \in \mathbb{R}^{5N}$ of the pyramidal hypercube.

3.2 Depth map and point cloud encoding

RGB-D sensors capture two channels with complementary and incongruous information. The objective of depth encoding is to render the depth information as RGB in a domain adaptive manner allowing knowledge transfer using the pre-trained CNN model. We use the depth image and point cloud representation together to embed richer depth information and render two independent RGB images. Starting from a single depth map channel $d(u)$, where $u = (x, y)$ and d denotes pixel-wise depth value at the x-y location, we calculate the vertical and horizontal derivatives

$$\begin{aligned} G_y &= K_y * d(u) \\ G_x &= K_x * d(u), \end{aligned} \quad (3)$$

where K_y and K_x are the vertical and horizontal Prewitt kernels respectively, and $*$ is a 2D convolution operator. Next, we compute the gradient magnitude and direction

$$\begin{aligned} G_m &= \sqrt{G_y^2 + G_x^2} \\ G_\theta &= \arctan(G_y, G_x). \end{aligned} \quad (4)$$

A three-channel depth map is constructed by combining the original depth map with the gradient magnitude and direction as $D(u) = \left[d(u), G_m, G_\theta\right]$. The motivation of this technique is to encode the shape of the object with gradient direction and sharp edges and boundaries through gradient magnitude. Figure 2a shows the result of this encoding. We can see that the combined channels capture richer shape information.

For point cloud $p^{(i)} = \{a^{(i)}, b^{(i)}, z^{(i)}\}$, $i \in 1, \ldots, P$ encoding, we first project it onto a canonical view[1] and then apply a gray colour map along the depth axis. This is followed by a colour transfer algorithm (Welsh et al. 2002) which is applied with the corresponding colour image to approximate the RGB values at each pixel. This technique transfers the chromatic information from the source image (RGB) to the target image (gray-scale) by matching their luminance and

[1] In practice, we define the canonical view as the $-27.5°$ and $20°$ off the azimuth and elevation angles.

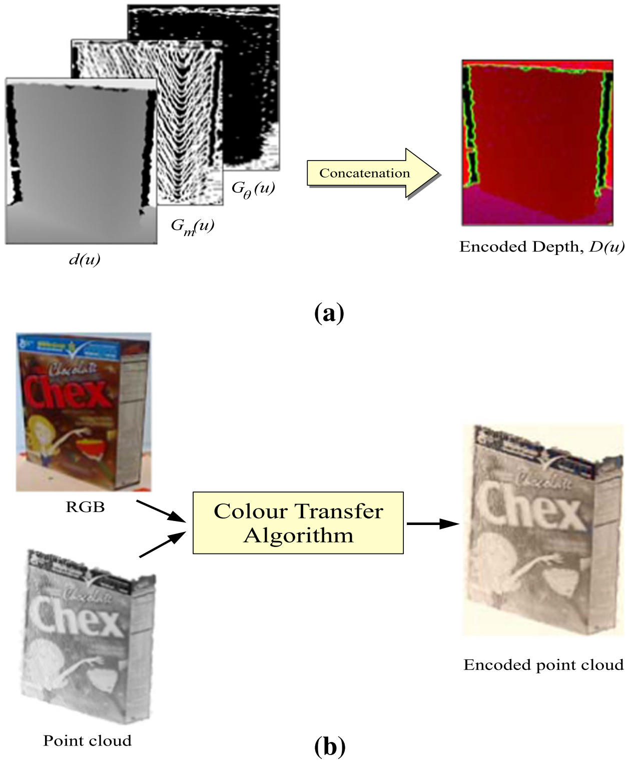

Fig. 2 Illustration of the proposed technique for CNN input encoding **a** depth image and **b** point cloud object-centric encoding before feature extraction

texture. The main advantage of this technique over existing encoding methods (Schwarz et al. 2015; Gupta et al. 2014) is that the colourization scheme is closely guided by the RGB images and is fully automatic. Figure 2b shows the effect of this technique. We can see that the resulting image closely resembles the corresponding RGB channel with the additional depth and shape information.

4 Feature fusion and inference with extreme learning machines

Existing methods that use CNN based features either directly use the feature vectors (Schwarz et al. 2015; Razavian et al. 2014) or use simple concatenation of feature vectors from convolution layers and fully connected neurons, F_{fc} (Hariharan et al. 2015; Liu et al. 2015a) as input to the classifiers (e.g. support vector machines). Such methods are straight-forward in implementation however, simple concatenation results in long feature vectors which increase the computational complexity of classification, especially when used with powerful classifiers with non-linear kernels (Huang et al. 2006). Moreover, since these features encode different modalities, the classifiers may need to make difficult decisions to weigh the relative importance of the features.

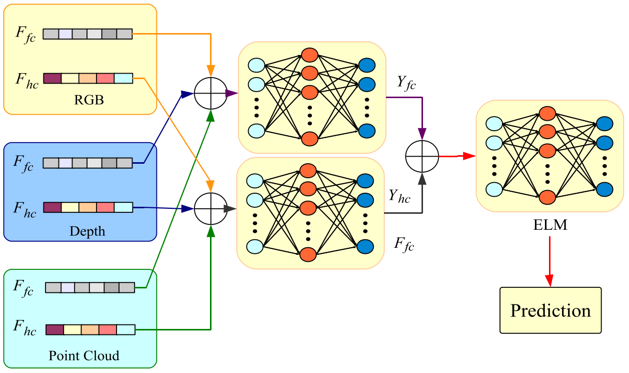

Fig. 3 Illustration of the proposed late fusion scheme to combine hypercube pyramid descriptor F_{hc} and the fully connected neurons F_{fc} for the combined RGB-D channels. ⊕ means concatenation

For robotic applications, the input to classifiers must be compact without sacrificing the discriminative properties of the features. We achieve this by employing the extreme learning machine classifier (Huang et al. 2006, 2012) and use ELM not only for multi-class object classification, but also as the feature fusion engine that combines the HP-CNN features F_{hc} with the fully connected layer F_{fc}. We use early and late fusion strategies to identify the most accurate classification scheme. Let $F_c = \{f_c^{(i)}, t^{(i)}\}, \in \mathbb{R}^D, i = 1, 2, \ldots, N$ represent the input feature vectors to the classifier, where N is the number of RGB-D images with labels t. In early fusion, F_c is the concatenation of F_{hc} and F_{fc} (i.e. $F_c = [F_{hc}, F_{fc}]$). ELM maps the feature vectors to the hidden layer to output $h = \sigma(\sum_{i=1}^{N} W_{in} f_c^{(i)} + b_{in}) \in \mathbb{R}^H$, where $\sigma(.)$, $W_{in} \in \mathbb{R}^{H \times D}$ and b_{in} are the piecewise sigmoid activation, randomized orthogonal input weight matrix and the bias vector, respectively. The hidden variables are then mapped to the target labels, parametrized by the output weight W_o and bias b_o giving the output variables $y = \sigma(\sum_{i=1}^{N} W_o t^{(i)} + b_o)\beta$.

The only parameter to be tuned is β which is achieved by optimizing a convex objective function of ELM that simultaneously minimizes the norm of the output weight and the loss between the actual output and the target labels as

$$\min_{\beta} \mathcal{J}_{ELM} = \frac{1}{2}\|\beta\|_F^2 + \frac{\lambda}{2}\|h\beta - T\|_2^2. \tag{5}$$

Equation 5 has a closed form solution using the linear least square, $\beta = h^{\dagger}T$, where $h^{\dagger}$ is the generalized Moore-Penrose pseudo-inverse of h. We compute $h^{\dagger}$ as $h^{\dagger} = (\mathbf{I}\lambda^{-1} + h^T h)^{-1} h^T$ or $h^{\dagger} = h^T(\mathbf{I}\lambda^{-1} + hh^T)^{-1}$, using an orthogonal projection method with the condition that $h^T h$ is non-singular (if $H > N$) or hh^T is non-singular (if $H < N$) (Huang et al. 2012). Here, $\mathbf{I}$ is an identity matrix. The coefficient λ is used for regularization and is chosen with cross validation such that it enhances the generalization of the ELM by acting as a solution stabilizer against overfitting (Huang et al. 2012).

In late fusion, we give F_{hc} and F_{fc} independently as inputs to the ELM classifiers and optimize Eq. 5. The ELM outputs are the vectors of class probabilities, y_{hc} and y_{fc}, for the HP-CNN and fully connected layer respectively. These vectores are concatenated (i.e. $F_c = [y_{hc}, y_{fc}]$) and used as input to another ELM that performs the final classification. This approach has the advantages that the feature dimension is reduced to only double the number of classes before the final classification and the learning/inference time of ELM is substantially reduced. Figure 3 illustrates the proposed late fusion method.

5 Multi-view 3D object pose model

To further highlight the importance of transfer learning and domain adaptation in RGB-D image recognition, we propose a multi-view 3D object pose (M3DOP) model to specifically fine-tune the CNN model of VGG-f (Chatfield et al. 2014) to the target datasets. However as mentioned earlier, the training data for depth images are deficient in terms of number of samples while manually acquiring and annotating the data are expensive. Additionally, RGB-D cameras only provide low resolution depth images, which are represented by 2.5D data and hence need to be carefully post-processed to remove the noise. To surmount this obstacle, we employ 3D CAD models which are freely available on the internet to render and generate synthetic RGB and depth images. These synthetic images are then used for fine-tuning the pre-trained RGB model for both the depth and colour channels. We intend to make the M3DOP model and rendered RGB-D dataset publicly available for the benefit of the research community.

5.1 Rendering synthetic multi-view depth images

To this end, we use the newly introduced dataset of ModelNet40 (Wu et al. 2015) which consists of 3983 full 3D polygonal meshes organized into 40 categories. The training and validation split of datasets follows the procedure of Wu et al. (2015) with balanced distribution among mesh categories i.e. for each category, 80 meshes are used for training and 20 meshes for validation. In the case of insufficient number of meshes such as in the category *bowl* and *cup*, then 20 meshes are selected for validation while the rest are used for training. The object meshes are rendered under a perspective projection. The reflection model of Phong (1975) is used and a random colour palette is applied to generate the RGB images while the depth values are determined by taking the distance from the facets of polygon meshes to the virtual cameras. Shapes of the objects are fit into the viewing volume by uniform scaling. It is worth noting that the objects contained in this dataset have significantly different distribution to those in the evaluation datasets in Sect. 6. However, as we will demonstrate experimentally, the model generalizes well to different datasets and applications.

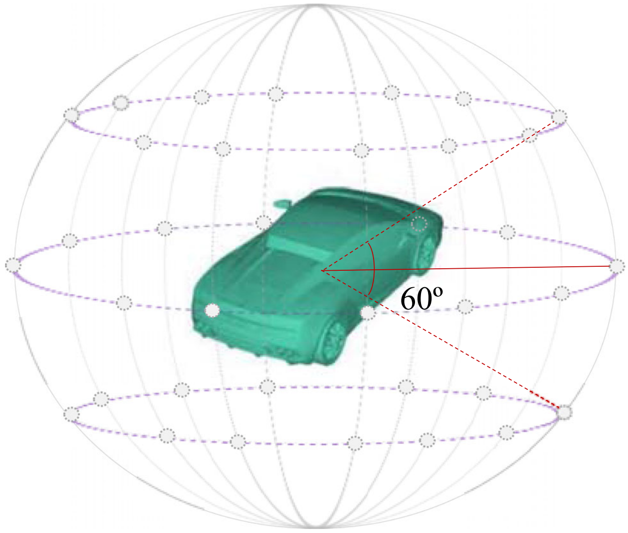

Fig. 4 Illustration of the synthetic data generation for the learning of multi-view 3D object pose model. Each point on the sphere corresponds to the virtual cameras placement looking towards the centre of the 3D CAD object mesh. The cameras that lie on the same horizontal ellipsoid are the multi-view cameras with the same elevation angle

To generate multi-view RGB-D images, we place several virtual cameras on a globe enclosing the object of interest (see Fig. 4). With the assumption that the object is always in an upright orientation along z-axis, we employ virtual cameras at three different elevation angles off the x-axis (30°, 0° and − 30°). The setting of the camera positions is chosen to emulate the vision of the robot where the robot needs to deal with different viewpoints of the objects in real world scenes. Moreover, multi-view representation of the polygon meshes enables us to learn a view-invariant model. Note that the same flexibility of rendering dynamically from negative elevation angle is limited in manual data acquisition [such as in the work of Lai et al. (2011), Browatzki et al. (2011)], whereas the viewpoints can be arbitrary using our method.

Then, we place the cameras along the ellipsoid at every 30° step size, making it a total of 36 views per object mesh (12 views per elevation angle). Besides providing full 3D representation of an object, this rendering technique has also important benefits over manually gathered 3D datasets using RGB-D cameras (Lai et al. 2011; Browatzki et al. 2011; Silberman and Fergus 2011; Song et al. 2015) in terms of arbitrary viewpoint capturing, efficiency and comes at absolutely zero operational cost. Moreover, it is becoming increasingly intricate to capture multi-view RGB-D images of huge and enormous items such as desks, bookshelves and wardrobe using low-cost Kinect camera. The final dataset consists of $3983 \times 36 = 143,388$ RGB-D images in total. Samples of the 3D meshes with corresponding rendered depth images are given in Fig. 5.

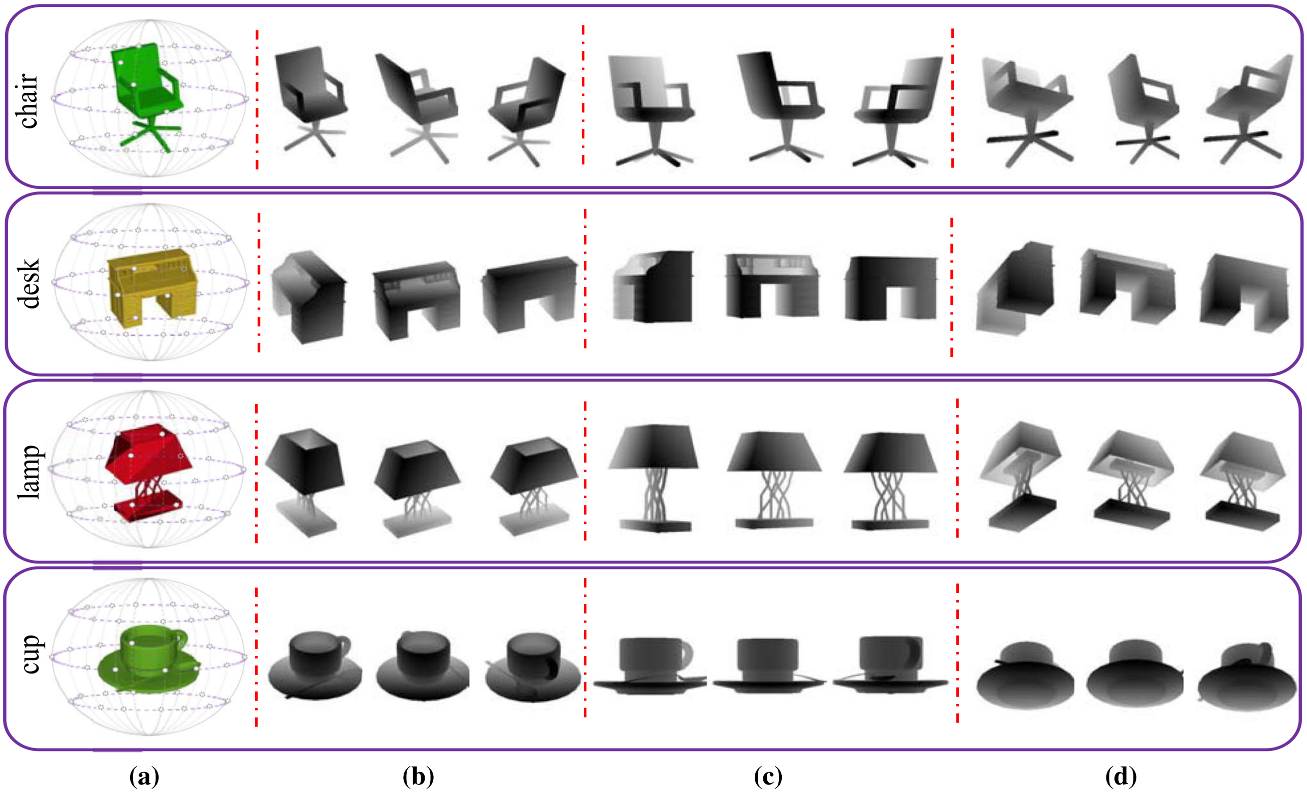

Fig. 5 Samples of the multi-view synthetic depth images rendered from **a** polygon meshes in ModelNet40 (Su et al. 2015). The synthetic data are generated from three different elevation angles. **b** 30°, **c** 0° and **d** − 30°. The depth images with corresponding RGB images are then used to train our view-invariant M3DOP model

5.2 Model architecture and learning

The learning process of the M3DOP model is performed by initializing the network using the VGG-f model and fine-tuning on the rendered RGB and depth images. Note that fine-tuning is done separately for RGB and depth images. We use the same encoding technique of depth images before feeding those images as input to our model as detailed in Sect. 3.2. Specifically, the network architecture follows the architecture of VGG-f with the exception of the last fully connected layer where we replace the layer with a new randomly initialized vector of dimension N, where N corresponds the number of categories to be classified. In order to prevent overfitting on the deep network, we also add a dropout layer (Srivastava et al. 2014) between fully connected layers (first fully connected, penultimate, and the final layer).

For the learning, a softmax regressor is added at the end of the network which is given by:

$$\arg\min_{\theta_D} \frac{1}{N} \sum_{j=1}^{N} \mathcal{L}\left(g(\mathbf{x}_D^{(j)}; \theta_D); \theta_D), t^{(j)} \right) , \quad (6)$$

where $\theta_D = \{\mathbf{W}_D, \mathbf{b}_D\}$ are the parameters of the CNN and $\mathcal{L}(.)$ denotes the conditional log-likelihood softmax loss function. During learning, standard stochastic gradient descent with backpropagation is used to optimize the objective function and update the model's parameters. We set relatively small learning rates of 0.00001 and 0.0001 for global and final fully connected layer respectively. The momentum is set to 0.9 and the weight decay is 0.0005 as suggested by Chatfield et al. (2014). We train the entire network for 67 and 31 epochs to learn the CNN models of depth and RGB images respectively. We let the algorithm run until the validation curve converges and stop the learning process when the curve stabilizes. The efficiency of the learning process is ensured by running the algorithm on a Tesla K40c graphics card. Each epoch takes approximately 30 min to complete. The training data are augmented using horizontal mirroring with 0.5 probability and normalized using the mean image of training data in ImageNet dataset (Deng et al. 2009).

6 Experimental setup and datasets

We extensively evaluated the proposed methods on benchmark RGB-D object recognition datasets including the Washington RGB-D (WRGB-D; Lai et al. 2011) and 2D3D (Browatzki et al. 2011) datasets. We also evaluated the proposed methods for cross domain adaptation by performing experiments on two challenging visual scene recognition datasets including SUN RGB-D Scene Dataset (Song et al. 2015) and NYU v1 Indoor Scene Dataset (Silberman and Fergus 2011). The implementation of the proposed algorithm based on the pre-trained CNN model (Sect. 3) was performed using the MatConvNet solver (Vedaldi and Lenc 2014; Vedaldi and Fulkerson 2010). As a baseline representation, we extract the channel-specific first fully connected neurons (fc_6) and concatenate the vectors as a final RGB-D representation prior categorization using ELM classifiers. Note that we only use the point cloud for object recognition as the canonical view of the point cloud generated from the depth images in scene recognition datasets is ill-defined. However, without any point cloud data, the proposed algorithm outperforms existing methods with a significant improvement as we will experimentally show in the next section.

The parameters of the ELM classifier including the number of hidden neurons H and regularization coefficient λ were determined by using grid search technique and cross-validated on the training set. Next, we will briefly describe the datasets with corresponding experimental protocols used for evaluation.

WRGB-D dataset contains RGB-D images of 300 household objects organized into 51 categories. Each image was captured using an ASUS Xtion Pro Live camera on a revolving turntable from three elevation angles (30°, 45° and 60°). We conducted two experiments following the experimental protocol of Lai et al. (2011), namely object category recognition and object instance recognition. In category recognition, a "leave-one-instance-out" procedure is adopted and the accuracy is averaged over ten trials. We use the same training-testing splits and the cropped images as suggested by Lai et al. (2011)[2] to ensure a fair comparison with other methods. For instance recognition, the images of objects captured at 45° were used as testing while remaining are used for training.

2D3D object dataset has relatively fewer images. It has 163 objects organized into 18 categories. The dataset consists of highly textured common objects such as drink cartons and computer monitors. We use the experimental protocol defined by Browatzki et al. (2011) for this dataset to ensure a fair comparison with existing state-of-the-art methods. The protocol requires the *fork*, *knife* and *spoon* classes to be combined into one class of *silverware* and the *phone* and *perforator* classes to be excluded due to their small sample numbers. This brings the final number of classes to 14 with 156 object instances for category recognition. For evaluation, six instances per class are randomly chosen for training and the remaining instances are used for validation. Only 18 RGB-D frames per instance are randomly selected for both sets. Except for categories that have less than six instances (e.g. *scissors*), in total in which case we use at least one instance is used for testing.

SUN RGB-D scene dataset (Song et al. 2015) is a benchmark suite for scene understanding and the most comprehensive and challenging RGB-D scene dataset to date. We adopted the same training/testing split for scene classification as suggested by the dataset authors. Specifically, the evaluation involves 19 scene/place categories which contain more than 80 images in each category. The final number of images for training and testing sets are 4845 and 4659 RGB-D images respectively. Complex indoor scenes with various degrees of object clutter and small inter-class variability make this dataset substantially challenging for recognition.

NYU v1 indoor scene dataset contains 2284 samples from seven scene classes. The experimental setup by Silberman and Fergus (2011) is used. In particular, we exclude the class *cafe* for its low number of samples and split the datasets into disjoint training/testing sets of equal size. Care has been taken to ensure that the frames captured from the same scene appear either in the training, or in the test set. In this paper, we do not use the ground-truth segmentation labels provided with the dataset and rely only on the categorical label of each scene frame for evaluation.

7 Model ablation analysis

We analyse the contribution of individual modules of the proposed method towards recognition accuracy to find the best-performing representation. We compare the accuracy with the baseline of our model which is the first fully connected neurons ($fc_6 \in \mathbb{R}^{4096}$) after the non-linear transformation by rectified linear units (ReLU). This representation has shown to encode highly discriminative features from a CNN model (Chatfield et al. 2014; Razavian et al. 2014). Next, we compare the accuracy of the proposed convolutional hypercube pyramid (HP-CNN) with fc_6 and the fusion techniques used to combine both features, denoted by simple concatenation (early fusion) and late fusion as discussed in Sect. 4. For multi-channel recognition task, we take the concatenation of the individual channel features as feature representation without performing any expensive dimensionality reduction method such as principal component analysis (PCA). In addition, the same set of experiments is conducted using the features extracted from the multi-view 3D object pose (M3DOP) model as discussed in Sect. 5.

[2] Available: http://rgbd-dataset.cs.washington.edu/.

Table 1 Effects of depth images augmentation with point cloud images towards categorization accuracy for object categorization in 2D3D dataset (Browatzki et al. 2011)

Features	Depth (D)	Point Cloud (P)	D + P
Washington RGB-D			
fc_6 + HP-CNN (late fusion)	79.5 ± 2.6	70.3 ± 2.3	85.0 ± 2.1
2D3D			
fc_6 + HP-CNN (late fusion)	90.3	84.8	92.9

7.1 Effects of augmenting depth images with point cloud images

Firstly, let us examine the effects and contributions of augmenting depth images and point cloud images as detailed in Sect. 3.2 towards categorization accuracy. We simply augment the final late fusion features extracted from both depth and point cloud images by concatenating them as a long vector prior classification. Table 1 tabulates the performance of object category recognition on WRGB-D and 2D3D dataset. Evidently, the accuracy consistently increases across these datasets when these features are combined together.

In WRGB-D, using HP-CNN representation with the late fusion scheme, the accuracy of using only the depth images or point clouds was 79.5 and 70.3% respectively. However, when both channels were fused together, the accuracy increases to 85% as depicted in Table 1. Similarly, this augmentation step accounts to approximately 2.6% accuracy increment in 2D3D dataset. This indicates that depth images and point clouds contain complementary information and augment each other to provide richer 3D information resulting in improved recognition performance.

7.2 Tuning optimal ELM parameters

In this section, we discuss the tuning of ELM classifier's parameters to produce optimal performance for categorization. Particularly, we would want to examine the effects of different settings of hyper-parameters including the number of hidden neurons, H and regularization coefficient, λ towards classification accuracy of fully connected layer activations, hypercube, and late fusion representation. Figure 6 shows the accuracy recorded for each point in a grid search scheme for RGB scene categorization task in SUN RGB-D dataset.

The procedure of our parameter tuning proceeds as follows. Firstly, we run a grid search to find an optimal combination of H and λ for fully connected layer activation. Next, the same procedure is conducted for hypercube representation. Using these hyper-parameters, we then run a grid search to find an optimal H and λ for the late fusion representation. As shown in Fig. 6, for all feature representation, the accuracy peaks at *overcomplete* representation i.e. number of hidden neurons is larger than the number of input neurons. This larger number of features gives the classifier many non-linear projections of the input data. Hence, in contrast to simple linear projections which have limited representational power, non-linear projections can make data closer to linearly separable and therefore easier to classify (Coates et al. 2011). As for the regularization coefficient λ, we find that the accuracy peaks at values of 1e2 and 1e3 for fully connected layer activations and hypercube, respectively. Again, λ is an important element in ELM for high generalization capability and avoiding overfitting (Huang et al. 2012).

7.3 RGB-D object recognition

As shown in Table 2 (2D3D column, rows without the highlighter) and Table 3, the combination of our HP-CNN and the fully connected layer activations (fc_6) with the late fusion technique consistently outperforms the other three alternative modules by a significant margin for all evaluation datasets. This is mainly credited to the fusion technique which uses the class probability distributions obtained from different feature representations as the new feature vectors for classification.

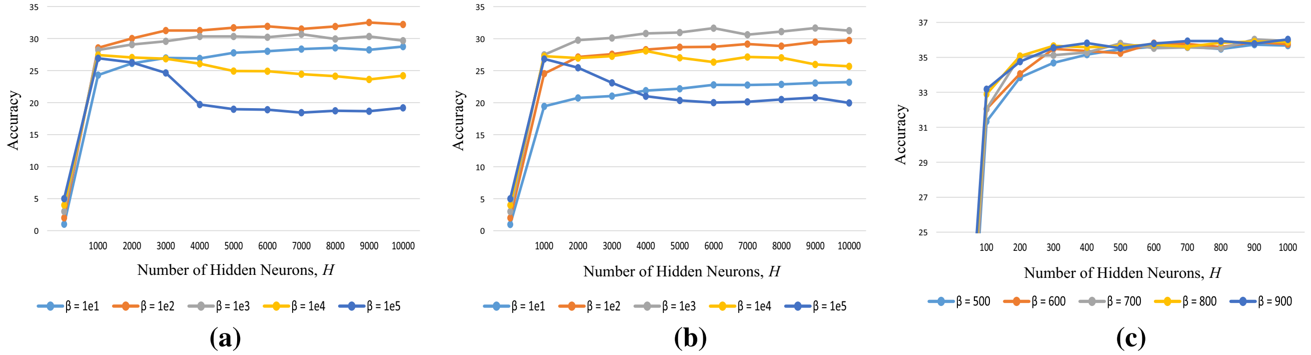

Fig. 6 Categorization accuracy using fully connected layer activations, hypercube and late fusion representation with different ELM parameters. **a** Fully connected layer, **b** hypercube and **c** late fusion

Table 2 Comparison of recognition accuracy (in %) in 2D3D object dataset (Browatzki et al. 2011), SUN RGB-D scene dataset (Song et al. 2015) and NYU v1 indoor scene dataset (Silberman and Fergus 2011) of the baseline method (fc_6), the proposed hypercube pyramid (HP-CNN), and the HP-CNN extracted from the multi-view 3D object pose (M3DOP) model (HP-CNN-T) with various fusion methods

	2D3D			SUN RGB-D			NYU v1		
Features	RGB	D	RGB-D	RGB	D	RGB-D	RGB	D	RGB-D
Baseline									
fc_6	87.4	88.4	90.6	32.0	21.1	33.3	73.1	52.1	71.6
fc_6 (extracted from M3DOP)	92.5	94.2	95.8	33.1	26.4	34.9	72.3	62.7	77.6
Hypercube Pyramid									
HP-CNN	91.6	88.9	92.3	30.3	23.9	33.7	67.8	51.3	68.3
HP-CNN-T	92.4	94.9	94.5	31.7	24.3	33.8	69.9	56.4	73.8
Early Fusion (Concatenation)									
fc_6 + HP-CNN	90.8	88.9	92.2	31.6	22.8	32.0	72.4	52.8	72.1
fc_6 + HP-CNN-T	91.8	95.2	94.8	31.6	25.0	32.3	74.1	64.7	77.2
Late Fusion With ELM									
fc_6 + HP-CNN	92.0	92.9	94.7	36.0	25.5	40.5	75.1	56.6	74.3
fc_6 + HP-CNN-T	**94.4**	**96.2**	**97.3**	**38.8**	**28.5**	**42.2**	**77.5**	**66.3**	**79.4**

Maximum values are given in bold

Table 3 Model ablation results in terms of classification accuracy (in %) of the proposed method for object and instance recognition in Washington RGB-D dataset (Lai et al. 2011) (the reported accuracy is the average accuracy over ten splits)

Task	Category Recognition		
Features	RGB	D	RGB-D
fc_6	85.5±2.1	79.6±1.8	87.6±1.7
HP-CNN	85.1±2.0	77.1±2.2	85.0±1.9
fc_6 + HP-CNN (Early Fusion)	85.9±1.9	81.2±2.1	87.9±1.8
fc_6 + HP-CNN (Late Fusion)	**87.6±2.2**	**85.0±2.1**	**91.1±1.4**
fc_6 + HP-CNN-T (Late Fusion)	86.9±1.7	84.8±2.2	90.2±1.5
Task	Instance Recognition		
Features	RGB	D	RGB-D
fc_6	95.1	48.0	94.6
HP-CNN	94.0	39.1	91.5
fc_6 + HP-CNN (Early Fusion)	94.8	28.1	86.7
fc_6 + HP-CNN (Late Fusion)	**95.5**	50.2	**97.2**
fc_6 + HP-CNN-T (Late Fusion)	95.2	**50.3**	96.6

Maximum values are given in bold

The testing time for late fusion features is only 6.3×10^{-5} s for one image using MATLAB on a 64-bit, 2.5 GHz machine. The accuracies of fc_6 and HP-CNN are comparable for all tasks, which depicts that the earlier convolutional layers activations also contain strong semantic cues, which can be used as a powerful representation for recognition tasks given an appropriate encoding scheme. The results also show that conventional fusion schemes using simple concatenation (early fusion) are less effective for combining the features originating from different sources. This is probably because of the difficulty faced by the classifier in suitably weighing the inputs that carry different sets of information.

For category recognition task in 2D3D dataset, note that the HP-CNN consistently outperforms fc_6 for RGB, depth and combined channels. In a deep network like CNN, it has been shown that earlier layers capture the low-level visual features such as oriented edges, lines and textures while more abstraction is modelled going deeper into the network (Zeiler and Fergus 2014; Bengio et al. 2013, 2007; Coates et al. 2011). As this dataset contains various highly textured objects, earlier convolutional layer activation encoded using HP-CNN is more representative than the fc_6 features. Therefore, we conjecture that any visual recognition task involving subtle inter-class discrimination should consider encoding earlier layers' activations of the deep network in the representation.

For instance recognition in WRGB-D dataset, we observe an interesting pattern in the classification accuracy where the performance severely drops when the HP-CNN representation is combined with the fc_6 using early fusion for depth-only and RGB-D recognition tasks. This trend shows that while the early fusion representation is powerful for categorical classification, it is less effective for more fine-grained tasks such as instance recognition. Nevertheless, the accuracy increases when the late fusion scheme is used to combine the features, showing that the two representations contain complementary information.

7.4 RGB-D scene recognition

The results of model ablation on the RGB-D scene recognition task are tabulated in Table 2 (SUN RGB-D and NYU v1 columns, rows without the highlighter). Remarkably, the same classification accuracy pattern is observed in this task as in the object recognition, although the datasets have significantly different distributions from the ImageNet (Deng et al. 2009) where the CNN model was trained on. Therefore, the transfer learning of a well-trained CNN model across applications, which is now commonplace in RGB based image recognition (Razavian et al. 2014), is also feasible in the context of 3D image recognition.

The simple concatenation of the HP-CNN features and fc_6 consistently gives slight degradation of performance in most channel-specific tasks. Besides the problem of *curse of dimensionality* which is a major source of overfitting (Yang et al. 2009; Yang and Ramanan 2015), both features represent different sets of information in the context of scene recognition. For example, the globally designed fc_6 feature is more representative of scenes with a high degree of *spatial envelope* (Torralba et al. 2003), while less discriminative for scenes with object clutter and distributed scene elements (Yang and Ramanan 2015) which can be more appropriately captured by the lower layers of CNN. Nonetheless, our proposal of projecting the features onto a supervised space in the late fusion scheme mitigates this problem which is reflected by the improved classification accuracy on all subtasks.

7.5 Effect of domain adaptation using the proposed multi-view 3D object pose model (M3DOP)

In this section, we compare the performance of features extracted from the CNN model that is fine-tuned using the technique discussed in Sect. 5 and the features extracted from the pre-trained CNN using the same experimental setup. The model for depth images is trained on the rendered depth images from ModelNet40 while we include the training images from 2D3D, SUN RGB-D and NYU v1 datasets to aid the learning of the model for RGB images. This results in the dimension of final fully connected layer of $N = 40$ and $N = 40+14+6+19 = 79$ for depth and RGB model, respectively.

As depicted in Table 2 (rows with the highlighter), the performance of fc_6 significantly boosts for the majority of tasks as a result of fine-tuning. However, we observe that there is little performance gain for HP-CNN after fine-tuning in the scene recognition datasets. In some cases, the early fusion representation degrades after the fine-tuning process. This result can be explained by referring to the learning rates used during learning the model. As the learning rate is relatively small for convolutional layers compared to the fully connected layers, the evolution of the model occurs very slowly and requires a lot of training epochs to converge. On the other hand, using a small global learning rate is crucial in our recognition task to avoid overshooting the gradient-based optimization process (Bengio 2009; Hinton et al. 2006; Hinton 2012) as the distribution of the target dataset is significantly contrary to the dataset the model was initialized with.

For all cases and datasets, projecting the features onto a supervised space before classification using late fusion scheme consistently gives highest performance compared to other modules. Comparing the two late fusion representations (the last two rows), it is clear that fine-tuning incredibly helps the recognition tasks. Interestingly, the significant performance gains are recorded for the depth-only recognition in scene datasets (21.1–26.4 % for SUN RGB-D and 52.1–62.7 % for NYU v1). Since there is no training data used from these datasets for learning the M3DOP model for depth images, the results open up possibilities of domain adaptation and transfer learning not only across modalities, but also across different applications.

It is worthy to note that using M3DOP model to extract the features and perform classification task improves the recognition accuracy for all testing datasets, except for WRGB-D dataset. The main reason is that the number of categories in ModelNet is relatively lower than WRGB-D dataset (40 vs 51). While a model learned from a dataset with larger number of categories can generalize to other datasets with fewer categories, the inverse is not true (Azizpour et al. 2016). Increasing the number of categories for model learning might improve the accuracy. However, we leave this for further investigation in the future.

8 Comparative analysis against state-of-the-art methods

8.1 Results on Washington RGB-D object dataset

To compare the accuracy of our algorithm with the state-of-the-art methods, we benchmark HP-CNN against ten related algorithms including EMK-SIFT (Lai et al. 2011), depth Kernel (Bo et al. 2011), CNN-RNN (Socher et al. 2012), CKM (Blum et al. 2012), HMP (Bo et al. 2012), semi-supervised learning (SSL; Cheng et al. 2014), subset-based deep learning (subset-RNN; Bai et al. 2015), CNN-colourized (Schwarz et al. 2015), CaRFs (Asif et al. 2015b) and LDELM (Zaki et al. 2015). The results are included in Table 4. All results in this section are taken from the original publications.

The results depict the superiority of our proposed method which constitutes state-of-the-art for several subtasks for

Table 4 Performance comparison in terms of recognition accuracy (in %) of the proposed hypercube pyramids with state-of-the-art methods on Washington RGB-D object dataset (Lai et al. 2011). The accuracy is reported is an average over 10 trials

Recognition type		Category recognition			Instance recognition		
Method		RGB	D	RGB-D	RGB	D	RGB-D
EMK-SIFT [a]	ICRA ’11	74.5 ± 3.1	64.7 ± 2.2	83.8 ± 3.5	60.7	46.2	74.8
Depth Kernel [a]	IROS ’11	77.7 ± 1.9	78.8 ± 2.7	86.2 ± 2.1	78.6	54.3	84.5
CNN-RNN	NIPS ’12	80.8 ± 4.2	78.9 ± 3.8	86.8 ± 3.3	–	–	–
CKM	ICRA ’12	–	–	86.4 ± 2.3	–	–	90.4
HMP [a]	ISER ’13	82.4 ± 2.1	81.2 ± 2.3	87.5 ± 2.9	92.1	51.7	92.8
SSL	ICPR ’14	81.8 ± 1.9	77.7 ± 1.4	87.2 ± 1.1	–	–	–
subset-RNN	Neurocomp.’15	82.8 ± 3.4	81.8 ± 2.6	88.5 ± 3.1	–	–	–
CNN-colourized	ICRA ’15	83.1 ± 2.0	–	89.4 ± 1.3	92.0	45.5	94.1
CaRFs [a]	ICRA ’15	–	–	88.1 ± 2.4	–	–	–
LDELM [a]	DICTA ’15	78.6 ± 1.8	81.6 ± 0.7	88.3 ± 1.6	92.8	**55.2**	93.5
HP-CNN [a]	this work	**87.6 ± 2.2**	**85.0 ± 2.1**	**91.1 ± 1.4**	**95.5**	50.2	**97.2**

Maximum values are given in bold
The methods [a]the experiments were conducted using the same training/testing splits

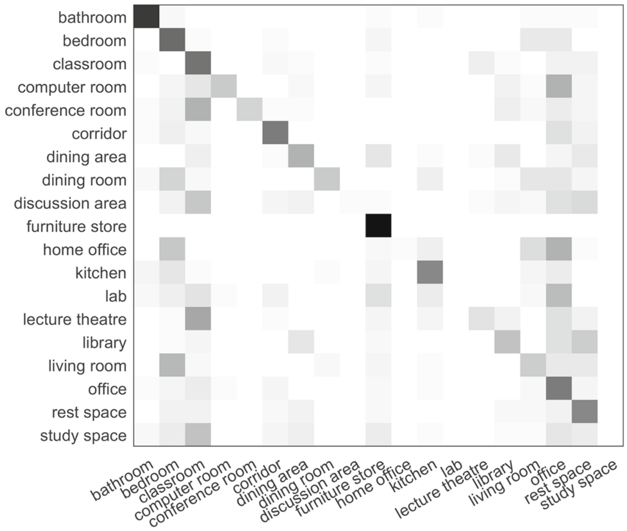

Fig. 7 Confusion matrix for one of the object categorization trials using the proposed hypercube pyramid representation and late fusion scheme on the Washington RGB-D object dataset (Lai et al. 2011). This figure is best viewed with magnification

WRGB-D. For object category recognition, our method outperforms other methods by a significant margin. Other methods which extract features from additional derivative channels such as gradient channels (Zaki et al. 2015), surface normals (Bo et al. 2012) and point cloud surfels (Asif et al. 2015b) do not perform as good as our three-channel feature extraction. Additionally, our choice of features substantially reduces the processing time needed to extract them from depth and point cloud channels.

Our method also outperforms other methods for channel-specific category recognition. The accuracy of our RGB-only recognition improves state-of-the-art by 4.5%, which can be attributed to our proposed HP-CNN representation. The significant performance improvement for depth-only recognition is an interesting result. It shows that the features extracted from a pre-trained CNN on RGB-only images were powerful enough to achieve high accuracy even when the underlying data was coming from a different modality. Hence, using appropriate encoding and rendering techniques, such as our proposed depth and point cloud encoding (Sect. 3.2), seamless transfer of knowledge between modalities is possible.

Our technique also outperforms other methods for instance recognition by a large margin, except for depth-only recognition in which LDELM (Zaki et al. 2015) descriptor wins with a reported accuracy of 55.2%. While this can be attributed to the heavily tuned deep networks from different derivative depth channels, the accuracy of LDELM is largely inferior for RGB and RGB-D recognition compared to our proposed algorithm. We observe that for instance recognition, colour information provides better discrimination across intra-class instances while they generally share very similar shapes (e.g. balls are spherical, soda cans are cylindrical). Nonetheless, this problem can be effectively mitigated by considering colour and depth features in unison.

Figure 7 visualizes the confusion matrix for one of the category recognition trials on the WRGB-D dataset. The strongest off-diagonal element shows the misclassification of *mushroom* which is labelled as *garlic*. As depicted in Fig. 8a, this is due to both instances having similar appearance and shape which makes the recognition task difficult even for human experts. In addition, the category *mushroom* has a very low number of examples in the dataset, highlighting class imbalance problem which makes it hard for the classifier to construct a good model for inference. We conjecture that the performance can be further improved by performing data augmentation techniques such as jittering (Chatfield et al. 2014; Razavian et al. 2014) to increase the number

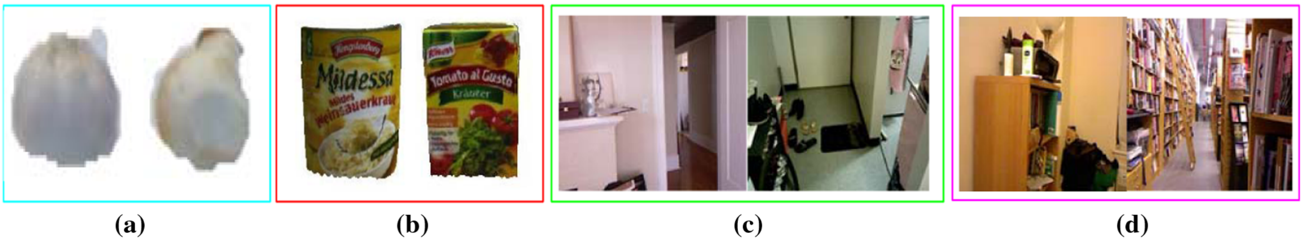

Fig. 8 Selected outliers for **a** WRGB-D (*mushroom* misclassified as *garlic*) **b** 2D3D (*drink carton* misclassified as *can*) **c** SUN RGB-D (*bedroom* misclassified as *living room*) and **d** NYU v1 (*bedroom* misclassified as *bookstore*)

Table 5 Performance comparison in terms of recognition accuracy (%) of the proposed hypercube pyramid (HP-CNN) with state-of-the-art methods on 2D3D object dataset (Browatzki et al. 2011)

Methods		RGB	D	RGB-D
2D + 3D	ICCVW ’11	66.6	74.6	82.8
HMP	ISER ’13	86.3	87.6	91.0
RICA	NIPS ’11	85.1	87.3	91.5
R^{2}ICA	ACCV ’14	87.9	89.2	92.7
Subset-RNN	Neurocomp. ’15	88.0	90.2	92.8
LDELM	DICTA ’15	90.3	91.6	94.0
HP-CNN	ICRA ’16	92.0	92.9	94.7
HP-CNN-T	This work	**94.4**	**96.2**	**97.3**

Maximum values are given in bold

of training samples and the accuracy is taken as an average prediction from all augmented images.

8.2 Results on 2D3D object dataset

For this dataset, we benchmark our HP-CNN against several state-of-the-art methods; combination of hand-crafted features (2D + 3D), HMP, reconstruction independent component analysis (RICA) (Le et al. 2011), R^{2}ICA, subset-RNN (Bai et al. 2015) and LDELM (Zaki et al. 2015). The depiction of comparison to other existing methods are reported in Table 5. The proposed HP-CNN outperforms all state-of-the-art methods for all subtasks with a considerable margin. The closest competitors, LDELM and subset-RNN, which are based on expensive channel-wise learning and subset generation procedure, lag 3.3 and 4.5 % in performance from our proposed HP-CNN. We credit this result mainly to the effectiveness of the depth and point cloud encoding which is also reflected in the higher accuracy achieved by the depth-only recognition compared to the RGB-only recognition.

Although the depth images captured using low-cost kinect-like sensors in this dataset are extremely noisy, with a lot of missing points and holes for reflective objects (e.g. silverware), the HP-CNN features extracted using the M3DOP model, which is learned from rendered 3D CAD models, successfully recognizes the object categories. Thus, we can adapt this technique of learning a powerful deep network from clean 3D CAD data for the purpose of designing robust real-time visual recognition capabilities which are essential for a robot’s online learning and recognition. A sample off-diagonal entry of confusion matrix (Fig. 9a) is depicted in Fig. 8b for qualitative analysis of a misclassification case for this dataset.

8.3 Results on sun RGB-D scene dataset

As this dataset was just recently introduced, we compare the proposed HP-CNN against the methods introduced by dataset creators including the scene-specific hand-crafted Gist descriptors (Torralba et al. 2003) and places (Zhou et al. 2014) which uses the features extracted from a CNN learned from a large-scale scene dataset. We also include the method of semantic regularized scene classifier (SS-CNN) (Liao et al. 2016) which is based on a CNN fine-tuned using scene datasets. As shown in Table 6, our HP-CNN outperforms other methods with a considerable margin, although the model does not implicitly use scene-specific data for learning, except a small portion in learning the M3DOP model for RGB images. Figure 9b shows the confusion matrix for this dataset using the fine-tuned HP-CNN. Our method still produces strong diagonal entries despite the challenging nature of the dataset, where some scene images are not entirely representative of their original scene classe and appear extremely similar to some other class, as depicted in Fig. 8c.

8.4 Results on NYU v1 indoor scene dataset

For NYU v1 Dataset, we compare our HP-CNN with recent approaches including bag-of-word using SIFT descriptors (BoW-SIFT; Silberman and Fergus 2011), spatial pyramid matching (SPM; Lazebnik et al. 2006), sparse-coding based SPM (ScSPM; Yang et al. 2009), RICA (Le et al. 2011) and R^{2}ICA (Jhuo et al. 2014). Our HP-CNN significantly outperforms all methods for all tasks as illustrated in Table 7. The proposed HP-CNN outperforms the deep learning based descriptor R^{2}ICA–which currently records the highest accuracy for this dataset to the best of our knowledge–by a significant margin, i.e. an accuracy improvement of up to 3.2%. Similar case holds for the channel-specific recognition tasks, especially for depth-only recognition although R^2ICA explicitly learns the deep model from depth patches. In contrast, our method transfers the knowledge of a model explicitly learned from the RGB domain to the domain of depth images, albeit including the fine-tuning process in M3DOP for improved accuracy. The confusion matrix for the dataset using the fine-tuned HP-CNN is shown in Figs. 9c and 8d depicts a sample misclassification.

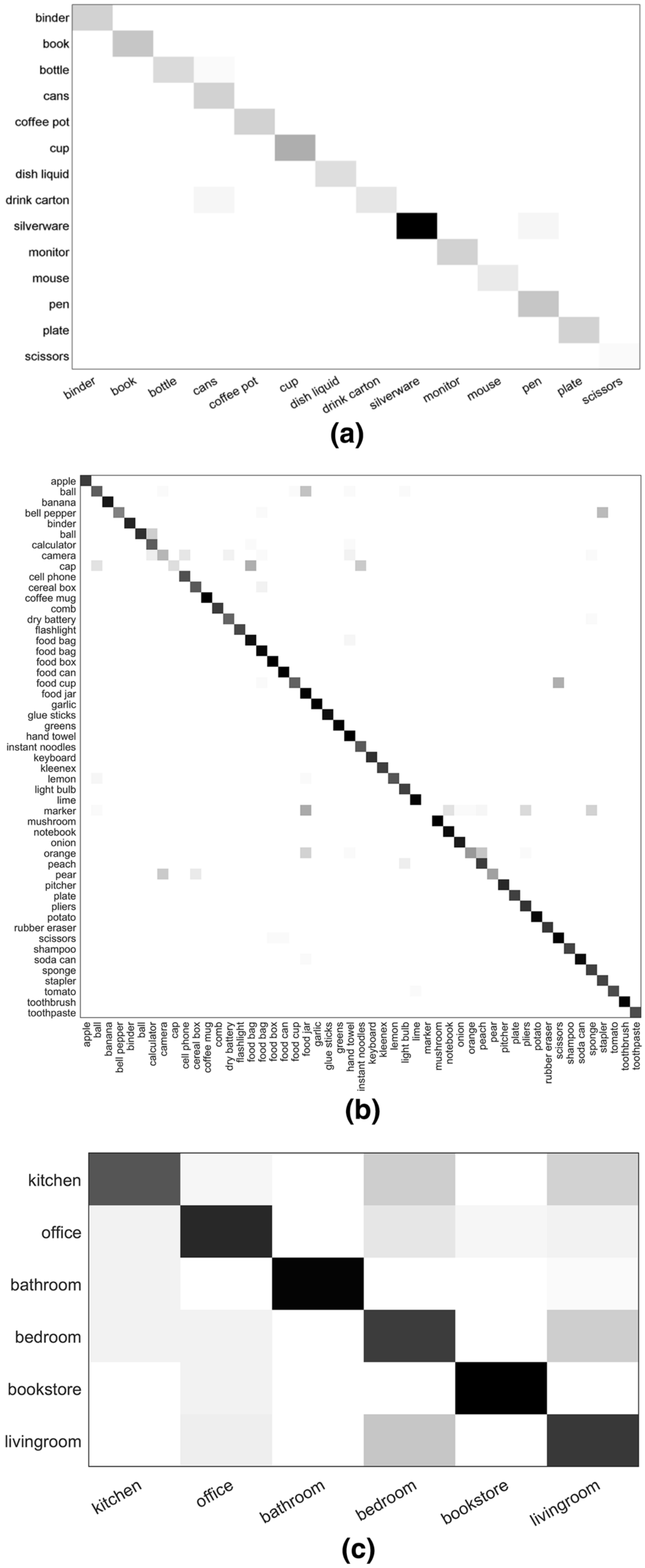

Fig. 9 Confusion matrices for image recognition using our proposed HP-CNN on the 2D3D (Browatzki et al. 2011), SUN RGB-D (Song et al. 2015) and NYU v1 (Silberman and Fergus 2011). Superior recognition accuracy is recorded as indicated by the strong diagonal entries. **a** 2D3D, **b** SUN RGB-D and **c** NYU v1

Table 6 Performance comparison in terms of recognition accuracy (%) of the proposed hypercube pyramid (HP-CNN) with state-of-the-art methods on SUN RGB-D Dataset (Song et al. 2015)

Methods		RGB	D	RGB-D
Gist + RSVM	CVPR '15	19.7	20.1	23.0
Places + LSVM	CVPR '15	35.6	22.2	37.2
Places + RSVM	CVPR '15	38.1	27.7	39.0
SS-CNN	ICRA '16	36.1	–	41.3
HP-CNN	ICRA '16	36.0	25.5	40.5
HP-CNN-T	This work	**38.8**	**28.5**	**42.2**

Maximum values are given in bold

Table 7 Performance comparison in terms of recognition accuracy (%) of the proposed hypercube pyramid (HP-CNN) with state-of-the-art methods on NYU v1 indoor scene dataset (Silberman and Fergus 2011)

Methods		RGB	D	RGB-D
BoW-SIFT	ICCVW '11	55.2	48.0	60.1
SPM	CVPR '06	52.8	53.2	63.4
RICA	NIPS '11	74.5	64.7	74.5
ScSPM	CVPR '09	71.6	64.5	73.1
R^2ICA	ACCV '14	75.9	65.8	76.2
HP-CNN	ICRA '16	75.1	56.6	74.3
HP-CNN-T	This work	**77.5**	**66.3**	**79.4**

Maximum values are given in bold

Table 8 Average computation time (in s) of several modules of our proposed framework for instance recognition task in WRGB-D (Lai et al. 2011). Note that only the time taken to resample one pyramid level is reported as this extraction step is highly parallel

Module	Time (s)
Feed forward	0.0327
Feature map resampling	0.1793
Maxpooling in four quadrants	0.008
Classification	0.0011
Total	0.2211

8.5 Computational cost

Our technique can be employed in real-time applications as it does not involve complex feature computation or computationally expensive testing phase. Our technique outperforms the current RGB-D object and scene categorization methods on the WRGB-D (Lai et al. 2011), 2D3D (Browatzki et al. 2011), SUN RGB-D (Song et al. 2015) and NYU v1 (Silberman and Fergus 2011) datasets by learning a view-invariant model using an independent training dataset without supervision from the target datasets. Therefore, in comparison to existing methods, the proposed HP-CNN-T is more general and can be used in online object and scene recognition systems. More precisely, the cost of adding a new object or scene class using our approach in an online system equals to the cost to train an ELM classifier.

As shown in Table 8, our method takes only 0.2211 s to classify one testing image or approximately 13 frames per second on a 3.4 GHz machine with 16 GB RAM. Convolutional feature map resampling consumes the majority of computation burden as each map needs to be resampled into the same dimension. This can be improved by employing tensor based resizing methods such as the technique based on 3D Discrete Cosine Transform (DCT; Uzair et al. 2015). Overall computational complexity can be further reduced by running the algorithm on a multi-threading machine and GPU. Moreover, as the algorithm was implemented in Matlab, we conjecture that using more efficient platforms such as C++ and OpenCV could further speed up the execution of each module in the pipeline.

9 Conclusions

We proposed a viewpoint invariant method for multi-modal object and scene recognition based on deep learning framework. We presented a powerful feature representation coined Hypercube Pyramid (HP-CNN) that encodes multi-scale features from all convolutional layers of a CNN. We also proposed a feature fusion technique to incorporate our HP-CNN and the activations of the fully connected layer leading to a compact representation and efficient prediction performance. Addressing the issue of limited training data in the RGB-D domain, we proposed a deep CNN model that represents RGB-D objects rendered from multiple viewpoints in a view-invariant high-level feature space. The end-to-end training of this model was performed using a large corpus of synthetically generated RGB-D training data from a repository of 3D models and the HP-CNN representation was extracted using this model. Experiments on benchmark RGB-D object recognition datasets demonstrate that the proposed method consistently outperforms state-of-the-art with a significant margin. We also evaluated the method on cross-application recognition where we conducted experiments on scene categorization. Although the CNN model was trained only on object poses, the proposed method outperformed several state-of-the-art methods that were specifically tuned for scene categorization.

Acknowledgements Funding was provided by Australian Research Council (Grant No. Australian Research Council (ARC) Discovery Project DP160101458).

References

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J. A. (2008). Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, *24*(5), 1027–1037.

Asif, U., Bennamoun, M., & Sohel, F. (2015). Discriminative feature learning for efficient rgb-d object recognition. In *IEEE/RSJ international conference on intelligent robots and systems (IROS), 2015* (pp. 272–279). IEEE.

Asif, U., Bennamoun, M., & Sohel, F. (2015). Efficient RGB-D object categorization using cascaded ensembles of randomized decision trees. In *Proceedings of ICRA*, (pp. 1295–1302).

Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., & Carlsson, S. (2016). Factors of transferability for a generic convnet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1790–1802. https://doi.org/10.1109/TPAMI.2015.2500224.

Bai, J., Wu, Y., Zhang, J., & Chen, F. (2015). Subset based deep learning for RGB-D object recognition. *Neurocomputing*, *165*, 280–292.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, *19*, 153.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE PAMI*, *35*(8), 1798–1828.

Blum, M., Springenberg, J.T., Wulfing, J., & Riedmiller, M. (2012). A learned feature descriptor for object recognition in RGB-D data. In *Proceedings of ICRA* (pp. 1298–1303).

Bo, L., Ren, X., & Fox, D. (2011). Depth kernel descriptors for object recognition. In *Proceedings of IROS* (pp. 821–826).

Bo, L., Ren, X., & Fox, D. (2012). Unsupervised feature learning for rgb-d based object recognition. In *Proceedings of ISER*.

Browatzki, B., Fischer, J., Graf, B., Bulthoff, H., & Wallraven, C. (2011). Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset. In *IEEE international conference on computer vision workshops (ICCVW)* (pp. 1189–1195).

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of BMVC*. arXiv preprint arXiv:1405.3531.

Cheng, Y., Zhao, X., Huang, K., & Tan, T. (2014). Semi-supervised learning for RGB-D object recognition. In *Proceedings of ICPR* (pp. 2377–2382).

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of AISTATS* (pp. 215–223).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition, 2009, CVPR 2009* (pp. 248–255). IEEE.

Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. In *Proceedings of ECCV* (pp. 345–360).

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of CVPR*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

Hinton, G., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G.E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599–619). Springer.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1), 489–501.

Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(2), 513–529.

Jhuo, I.H., Gao, S., Zhuang, L., Lee, D., & Ma, Y. (2014). Unsupervised feature learning for RGB-D image classification. In *Proceedings of ACCV* (pp. 276–289).

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS* (pp. 1097–1105).

Lai, K., Bo, L., Ren, X., & Fox, D. (2011). A large-scale hierarchical multi-view RGB-D object dataset. In *Proceedings of ICRA* (pp. 1817–1824).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *proceedings of CVPR* (Vol. 2, pp. 2169–2178).

Le, Q.V., Karpenko, A., Ngiam, J., & Ng, A.Y. (2011). Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in neural information processing systems* (pp. 1017–1025).

Liao, Y., Kodagoda, S., Wang, Y., Shi, L., & Liu, Y. (2016). Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In *2016 IEEE international conference on robotics and automation (ICRA)* (pp. 2318–2325). IEEE.

Liu, L., Shen, C., & van den Hengel, A. (2015). The treasure beneath convolutional layers: Cross convolutional layer pooling for image classification. In *Proceedings of CVPR*.

Liu, W., Ji, R., & Li, S. (2015). Towards 3D object detection with bimodal deep boltzmann machines over RGBD imagery. In *Proceedings of CVPR*.

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

Lowry, S., Snderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., et al. (2016). Visual place recognition: A survey. *IEEE Transactions on Robotics*, *32*(1), 1–19. https://doi.org/10.1109/TRO.2015.2496823.

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317.

Razavian, A.S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of computer vision and pattern recognition workshops (CVPRW)* (pp. 512–519).

Schwarz, M., Schulz, H., & Behnke, S. (2015). RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In *Proceedings of ICRA*.

Silberman, N., & Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 601–608). IEEE

Socher, R., Huval, B., Bath, B., Manning, C.D., & Ng, A. (2012). Convolutional-recursive deep learning for 3D object classification. In *Proceedings of NIPS* (pp. 665–673).

Song, S., Lichtenberg, S.P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 567–576).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 945–953).

Torralba, A., Murphy, K.P., Freeman, W.T., & Rubin, M.A. (2003). Context-based vision system for place and object recognition. In *Proceedings of ninth IEEE international conference on Computer vision, 2003* (pp. 273–280). IEEE.

Uzair, M., Mahmood, A., & Mian, A. (2015). Hyperspectral face recognition with spatiospectral information fusion and pls regression. *IEEE Transactions on Image Processing*, *24*, 1127–1137. https://doi.org/10.1109/TIP.2015.2393057.

Vedaldi, A., & Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on multimedia* (pp. 1469–1472). ACM.

Vedaldi, A., & Lenc, K. (2014). Matconvnet-convolutional neural networks for matlab. arXiv preprint arXiv:1412.4564.

Welsh, T., Ashikhmin, M., & Mueller, K. (2002). Transferring color to greyscale images. *ACM Transactions on Graphics*, *21*(3), 277–280. https://doi.org/10.1145/566654.566576.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1912–1920).

Yang, J., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *IEEE conference on computer vision and pattern recognition, 2009, CVPR 2009* (pp. 1794–1801). IEEE

Yang, S., & Ramanan, D. (2015). Multi-scale recognition with DAG-CNNS. In *Proceedings of the IEEE international conference on computer vision* (pp. 1215–1223).

Zaki, H.F., Shafait, F., & Mian, A. (2015). Localized deep extreme learning machines for efficient RGB-D object recognition. In *Proceedings of digital image computing: Techniques and applications (DICTA)* (pp. 1–8). https://doi.org/10.1109/DICTA.2015.7371280.

Zaki, H.F.M., Shafait, F., & Mian, A. (2016). Convolutional hypercube pyramid for accurate RGB-D object category and instance recognition. In *Proceedings of ICRA* (to appear).

Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision—ECCV 2014* (pp. 818–833). Springer.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Proceedings of NIPS* (pp. 487–495).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Hasan F. M. Zaki received his B.Eng. degree in Mechatronics in 2010 from International Islamic University of Malaysia (IIUM), Malaysia and the M.Eng. degree in Mechatronics in 2013 from University of Malaya, Malaysia. He completed his Ph.D. from The University of Western Australia in 2017. He is currently an Assistant Professor at Department of Mechatronics, Kuliyyah of Engineering, IIUM. His research interests include robotic vision, RGB-Depth object and scene recognition, 3D shape analysis and first person action recognition.

Faisal Shafait is working as the Director of TUKL-NUST Research & Development Center and as an Associate Professor in the School of Electrical Engineering and Computer Science at the National University of Sciences and Technology, Pakistan. Besides, he holds an Adjunct Senior Lecturer position at the University of Western Australia, Perth, Australia. He has worked for a number of years as a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), Germany and a visiting researcher at Google, California. He received his Ph.D. in computer engineering with the highest distinction from TU Kaiserslautern, Germany in 2008. His research interests include machine learning and computer vision with a special emphasis on applications in document image analysis and recognition. He has co-authored over 100 publications in international peer-reviewed conferences and journals in this area. He is an Editorial Board member of the International Journal on Document Analysis and Recognition (IJDAR), and a Program Committee member of leading document analysis conferences including ICDAR, DAS, and ICFHR. He is serving on the Leadership Board of IAPRs Technical Committee on Computational Forensics (TC-6) as well as the President of Pakistani Pattern Recognition Society.

Ajmal Mian completed his Ph.D. from The University of Western Australia in 2006 with distinction and received the Australasian Distinguished Doctoral Dissertation Award from Computing Research and Education Association of Australasia. He received the prestigious Australian Postdoctoral and Australian Research Fellowships in 2008 and 2011 respectively. He received the UWA Outstanding Young Investigator Award in 2011, the West Australian Early Career Scientist of the Year award in 2012 and the Vice-Chancellors Mid-Career Research Award in 2014. He has secured seven Australian Research Council grants and one National Health and Medical Research Council grant with a total funding of over $3 Million. He is a guest editor of Pattern Recognition, Computer Vision and Image Understanding and Image and Vision Computing journals. He is currently in the School of Computer Science and Software Engineering at The University of Western Australia. His research interests include computer vision, machine learning, 3D shape analysis, hyperspectral image analysis and pattern recognition.