Grounding of Human Environments and Activities for Autonomous Robots

Muhannad Alomari[1*], Paul Duckworth[1*], Nils Bore[2], Majd Hawasly[1], David C. Hogg[1], Anthony G. Cohn[1]
[1]University of Leeds, United Kingdom.
[2]Royal Institute of Technology (KTH), Sweden
[1]{scmara, p.duckworth, m.hawasly, d.c.hogg, a.g.cohn}@leeds.ac.uk, [2]nbore@kth.se

Abstract

With the recent proliferation of human-oriented robotic applications in domestic and industrial scenarios, it is vital for robots to continually learn about their environments and about the humans they share their environments with. In this paper, we present a novel, online, incremental framework for *unsupervised* symbol grounding in real-world, human environments for autonomous robots. We demonstrate the flexibility of the framework by learning about colours, people names, usable objects and simple human activities, integrating state-of-the-art object segmentation, pose estimation, activity analysis along with a number of sensory input encodings into a continual learning framework. Natural language is grounded to the learned concepts, enabling the robot to communicate in a human-understandable way. We show, using a challenging real-world dataset of human activities as perceived by a mobile robot, that our framework is able to extract useful concepts, ground natural language descriptions to them, and, as a proof-of-concept, generate simple sentences from templates to describe people and the activities they are engaged in.

1 Introduction

To integrate in human environments, mobile robots with collaborative/assistive human-oriented tasks should be enabled to continuously learn about their environments, the people who inhabit these environments, and the activities that take place there. From an autonomous robot point of view, this requires incremental, unsupervised methods that operate on the outputs of various kinds of sensor modalities the robot might have, ranging from laser rangefinder and RGB-D cameras to voice recognition. The desired outcome of this process is learning a collection of grounded concepts of the robot's environment that are beneficial for the robot's specific task.

In this paper, we present a framework for symbol grounding for autonomously-extracted components of real-world, human environments for a mobile robot. The novelty of this framework is that it extends existing work in autonomous symbol grounding towards '*the wild*', i.e. from the typical lab settings towards more realistic, real-world scenarios, and from ideal sensing conditions to noisy, limited and changing perception of a mobile robot. Moreover, it does this in an unsupervised, incremental fashion. We presuppose that the robot can navigate and visually analyse the environment to extract a multitude of visual features in order to incrementally recover useful 'classes' of visual features, here named *concepts*. If natural language descriptions of the observations are also provided, they can be analysed along with the visual features to ground the words describing people, objects, activities, etc. to their most relevant perceptual concepts. One possible application of such a framework could be in the field of security or assistive robotics where robots need the ability to learn on-the-go how to describe new objects or situations in a human-understandable form in a lifelong setting.

As a proof of concept, this paper supports recognition of individuals, describing (some aspects of) their physical appearance using natural language, and commenting on the activities they are engaged in. We do not claim that these chosen concepts are exhaustive, more challenging or more relevant to human environments than any others, but they are provided here to give a flavour of the framework and its application. To this end, we integrate state-of-the-art object segmentation, human pose estimation and activity analysis into a flexible, incremental framework for learning to distinguish instances of faces, colours, objects, and activities in real-world complex scenarios. Moreover, we propose a simple language grounding framework to learn concept names for human-robot interaction purposes using natural language descriptions.

We concentrate on a small number of features and sensory data that are easily acquirable by capable mobile robots. To learn about humans we extract facial features using off-the-shelf face detectors/descriptors, and we acquire human pose estimates using a state-of-the-art pose machine. We also collect colour information from people's clothing in order to describe their appearance. For objects, we use automatic object segmentation and trajectory analysis to identify usable objects in the environment. For human activities, we use qualitative spatial-temporal representations to capture the interaction through the relations between body poses and object positions, which feed into a generative Bayesian model that learns activity classes in an unsupervised setting.

* The first two authors contributed equally to this paper.

Lastly, given textual descriptions, we propose a natural language grounding framework that deploys integer programming techniques to assign words to their physical representation in the visual domain, i.e. to the learned concepts in the numerous visual feature spaces.

2 Related Work

Enabling robots to share the human environment has been a goal of AI and robotics research, manifested in a vast array of active research areas e.g. continual learning, learning by demonstration, human-robot interaction, dialogue planning, compliant robotics, humanoid robots, etc.

In the robotics literature, grounding learned feature spaces focuses on fusing sensor modalities such as vision or haptics with natural language in order to teach robots useful concepts like object names, action labels, and spatial relations, e.g. [Beetz *et al.*, 2011; Spranger and Steels, 2015; Aksoy *et al.*, 2017], or the semantics of natural language navigation and manipulation commands, e.g. [Lauria *et al.*, 2002; Tellex *et al.*, 2011; Matuszek *et al.*, 2013; She *et al.*, 2014; Hemachandra *et al.*, 2015]. This is normally achieved in simple, controlled environments. In this work, we learn visual concepts and natural language groundings in a more challenging, real-world human environment.

Existing research has addressed incremental learning of simple elements in the robot's environment, e.g. object features [Sinapov *et al.*, 2014; Craye *et al.*, 2015; Young *et al.*, 2016] or patterns of human occupancy over time [Jovan *et al.*, 2016]. Other work has focused on learning and grounding more complex elements non-incrementally, e.g. human actions from image motion features [Song *et al.*, 2016]. In this work, we incrementally and simultaneously learn and ground multiple elements of the robot's environment (objects, people, and human activities) in an unsupervised manner.

There is a wealth of research in computer vision that relates to learning the concepts we concentrate on here (e.g. activity analysis, deep learning object segmentation, face recognition, etc.), of which we present one variation. Note that our focus in this work is on the continual learning framework that integrates various learning elements and runs online with mobile robots' onboard computation, rather than on advancing the state-of-the-art in learning any individual element, and our approach can take advantage of advances in these individual areas.

3 Concepts

In this section we introduce our notion of *concepts*: abstractions of the perceptual feature spaces generated by the robot's sensory modalities which carry a human-level meaning. For example, concepts might include a colour represented as a cluster of values in the HSL colour space (Hue-Saturation-Lightness), or an object represented as a cluster of points in a 3D point cloud. We present next the sensors and feature spaces we use, for a Scitos A5 mobile robot [MetraLabs, 2016], along with the unsupervised methods we employ to generate such concepts. Note that our framework does not rely on any particular robot or any specific sensors; rather it is flexible to what the modalities of the robot can support.

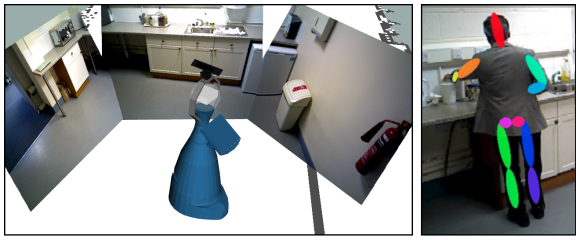

Figure 1: (**left**) Mobile Scitos A5 robot gathering RGB-D data. (**right**) An example of a human pose estimate.

For its basic operations, the mobile robot we use is equipped with a base-mounted laser scanner used to model the physical environment as a 2D occupancy grid where occupied cells indicate static objects, allowing localisation, mapping and navigation. Also, the robot is equipped with two RGB-D sensors, one over-head and one chest-mounted, that allow collecting 640×480 RGB video streams in addition to depth point clouds. These sensors are used to generate a 3D map of the robot's environment as shown in Fig. 1 (left). The robot detects and tracks humans as they pass within the field of view of its head-mounted RGB-D sensor. We define a human pose as the estimated 3D position of the person's 15 body joint locations at a single timepoint, see Fig. 1 (right). To estimate the human pose, we use a real-time depth-only tracker built on OpenNI [OpenNI, 2016] along with a post-processing state-of-the-art RGB pose estimation [Wei *et al.*, 2016]. For each human detected by the robot, a sequence of human pose estimates over a time series of frames is acquired.

3.1 Extracting Concepts

Concepts are learned automatically by clustering the low-level input of each of the robot's sensor modalities after an appropriate encoding. This clustering operation results in a collection of classes that are candidate concepts in each feature space. Because we assume no prior knowledge of the structure of the sensor feature spaces, we employ probabilistic modelling techniques to each feature space independently to elicit meaningful classes that are supported by the observed data.

We differentiate between two kinds of concepts. *Simple concepts* are ones that are time-independent and can be detected from a single or a small number of observations. For example, simple visual concepts like colours can be represented as Gaussian components in a Gaussian Mixture Model over the HSL space [Alomari *et al.*, 2017]. Similarly, static objects are simple concepts that can be segmented from fused 3D point clouds using geometrical/textural cues [Bore *et al.*, 2017].

On the other hand, *complex concepts* exhibit a temporal dimension and manifest over longer sequences of observations. For instance, temporally-extended human activities are one example of complex concepts. For these, a more elaborate encoding and a more sophisticated clustering mechanism are needed [Duckworth *et al.*, 2017]. Namely, the robot first abstracts each observed human pose sequence using a qualitative representation [Chen *et al.*, 2015], translating the detected quantitative pose sequence into a sequence of abstract,

spatio-temporal descriptors which qualitatively describe the interaction. For example, in a "drinking coffee" activity, the exact spatial position of a person's hand is not as useful for learning the activity as a qualitative relation between the hand and a coffee mug. Then, a probabilistic mixture over that qualitative space is obtained using a hierarchical Bayesian generative model, namely Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003]. We consider the components of that mixture (called topics) as human activity concepts.

In this paper, we demonstrate extracting four kinds of concepts; three simple ones: faces (to learn to distinguish people and later learn their names), colours (to describe people's attire) and objects (to learn their function), and one complex concept: human activities. We briefly introduce each of the feature spaces we use and show how the robot clusters observations in each of them to obtain candidate concepts.

Faces: To learn and recognise people's faces, a small patch from the RGB feed is automatically cropped around the location of the head joint of the human pose estimate for every person detected. We check for the presence of a face in the cropped image using a cascade of boosted classifiers with Haar features [Lienhart and Maydt, 2002] along with the OpenCV generic face model. Then, we compute the Eigenvalues for the n most prominent Eigenfaces [Turk and Pentland, 1991] (extracted from a collection of face images.) This transforms a face into a much-smaller n-dimensional data point in the space of Eigenfaces. Finally, we fit a Gaussian mixture model in that space with an optimal number of components selected using the Bayesian Information Criterion (BIC) [Posada and Buckley, 2004]. The resulting Gaussian components are used as candidate concepts to represent people. Examples of such clusters are shown in Fig. 2 (faces).

Colours: We cluster the HSL colour values of the upper and lower garments of each person detection using a Gaussian mixture model. The number of Gaussian components is selected automatically using the BIC. The colours of the upper and lower garments are extracted from the visual feed using the human pose estimate, where the colour of the upper garment is estimated by taking the average pixel colour from the triangle of the two shoulders and the torso, and the colour of the lower garment is estimated by taking the average pixel colour of the triangle between the torso and the knees, as shown in Fig. 2 (colours). The extracted colours are projected into HSL space to increase the robustness under varying lighting conditions. Examples of six clusters extracted can be seen in Fig. 2 (colours).

Objects: The robot constructs a 3D model of its environment by fusing RGB-D images into *surfels* [Pfister *et al.*, 2000]. As demonstrated in [Schoeler *et al.*, 2015], an unsupervised segmentation algorithm grounded in the convexity of common human objects can achieve state-of-the-art performance in extracting semantically meaningful segments. We use a similar method to that presented in [Bore *et al.*, 2017], which first splits the scene into a collection of *supervoxels* [Papon *et al.*, 2013] over which an adjacency graph

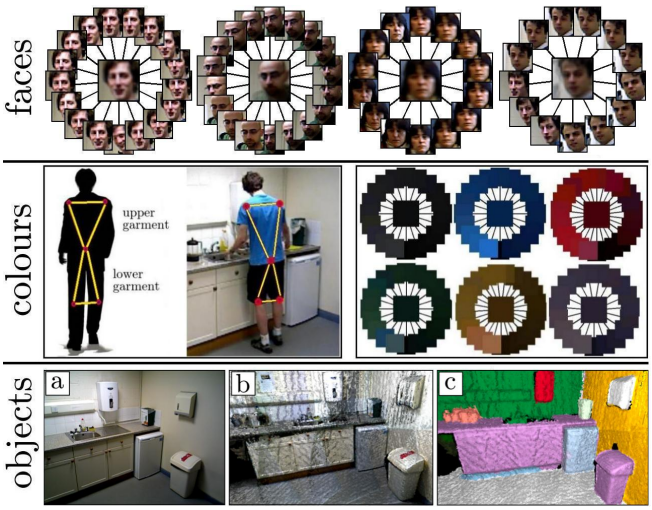

Figure 2: **(faces)** Examples of face clusters, with the averaged (mean) face shown in the centre of each cluster. **(colours)** left: Upper and lower garments extracted from a human pose estimate. right: Examples of different colour clusters, averaged (mean) colour shown in the centre. **(objects)** Processing of RGB-D feed (a) RGB image from a 3D sweep (b) registered 3D point cloud of the kitchen (c) segmented surfel map.

is formed. Then, weights are assigned to the edges based on local convexity of the point cloud and colour differences between segments. Finally, to segment the point cloud, iterative graph cuts are performed to separate parts with concave boundaries and/or large colour differences. This results in a collection of point cloud segments which relate to the objects in the environment, as illustrated in Fig. 2 (objects).

To concentrate attention on only the objects that are part of the observed activities, the 3D human pose estimate sequences are analysed to extract the locations where people stop more frequently in the environment. Then, the candidate objects are scored according to their proximity to people's hands in these locations. The highest scoring candidate objects are considered as object concepts.

Human Activities: To learn temporally-extended human activities, the pose of humans within the environment is detected and tracked along with the positions of the learned object concepts. Then, the observations are encoded into a number of Qualitative Spatio-temporal Representations (QSRs). A QSR is an abstraction of exact quantitative observations in a particular feature space into qualitative states, condensing noisy observations of arbitrary spatial positions into higher-level relational descriptors. We briefly introduce the three QSRs used to encode object-human pose sequences in this paper, which we compute using the publicly available library QSRlib [Gatsoulis *et al.*, 2016]:

1) *Ternary Point Configuration Calculus (TPCC)* [Moratz and Ragni, 2008] qualitatively describes the spatial arrangement of a point (the *referent*) relative to two other points (a line connecting the *relatum* and *origin*). Relations are triples of $\langle$ { front, back }, { left, right, straight }, { distant, close } $\rangle$. An illustration of this QSR can be seen in Fig. 3 (top centre).

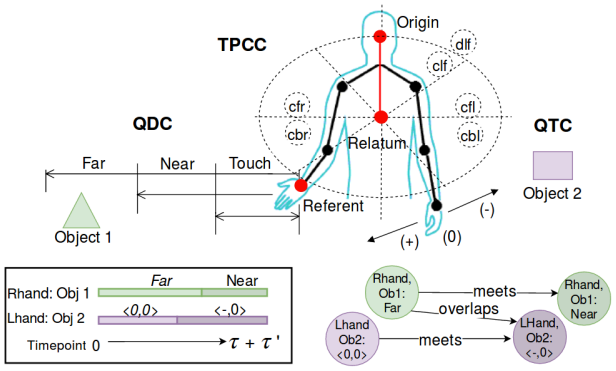

Figure 3: QSRs and Interval representations; (**top left**) QDC (relative distance) between right hand and *object* 1, with three possible relations: *touch*, *near* and *far*. (**top centre**) Subset of the TPCC system between right hand and the torso-head line. 6 of the 12 possible relations are shown, where the symbols *f*, *b*, *l*, *r*, *d*, *c* stand for front, back, left, right, distant and close, respectively (**top right**) QTC (relative motion) between left hand and *object* 2, where relations could be one of $(+)$ for moving away, $(-)$ for moving closer, or (0) for being static (**bottom left**) Interval representation for an example scenario, using QDC (green) and QTC (purple). (**bottom right**) Interval Graph of the interval representation.

2) *Qualitative Trajectory Calculus (QTC)* [Delafontaine *et al.*, 2011] represents the relative motion of two points with respect to the reference line connecting them, and is computed over consecutive timepoints. For two objects o_1, o_2, it defines the following three relations: $\{o_1$ is moving towards o_2 (symbol $-$), o_1 is moving away from o_2 $(+)$, o_1 is neither moving towards or away from o_2 $(0)\}$. An illustration of this QSR for a joint relative to an object can be seen in Fig. 3 (top right).

3) *Qualitative Distance Calculus (QDC)* [Clementini *et al.*, 1997] expresses qualitative Euclidean distance between two points based on defined distance thresholds. A set of QDC relations localises a joint with respect to reference landmarks (e.g. object locations in an environment). Thus, changes in the relations can help explain relative motion. An illustration of this QSR for a joint relative to an object is in Fig. 3 (top left).

By encoding a human pose sequence as a QSR abstraction, we obtain a set of qualitative relations (one per calculi used) that hold between each body joint pose and each automatically segmented object concept, per time point. We then perform a temporal abstraction which compresses repeated qualitative relations at adjacent frames into an *interval* representation, maintaining only the relation and duration information. An example interval representation between joints and objects can be seen in Fig. 3 (bottom left). Then, we employ Allen's Interval Algebra (IA) [Allen, 1983] to create an *Interval Graph*, where nodes represent intervals (relations holding between joints/objects) and directed arcs connect temporally-adjacent intervals with IA relations. An example Interval Graph can be seen in Fig. 3 (bottom right).

Given a corpus of Interval Graphs, one per human detection, k-length paths are extracted from the graphs as *code words* for some small k (usually $k \leq 4$). Thus, a code word encodes a small number (≤ 4) of temporally-connected nodes, i.e. spatial relations of joint-object/joint-joint pairs, capturing a snapshot of an activity. The set of all unique code words is considered as a discrete *vocabulary*, from which a descriptor analogous to bag-of-words representation (a *histogram*) can be computed for each detection. This histogram maintains counts of occurrence of code words in a detection. Note that, this representation is different from the traditional bag-of-words normally used in document analysis, in that it maintains some temporal information inside the code words.

We use Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], a three-layer hierarchical generative model for discrete data, to model the histograms. This Bayesian model has proved successful in problems with large corpora not exclusive to document analysis, e.g. [Duckworth *et al.*, 2017]. LDA simultaneously discovers *topics* in the 'corpus' of histograms and infers the topic proportions for each detection. A topic is a probability distribution over the vocabulary of code words, thus it is a conceptual model of a human activity and a candidate activity concept. The graphical model representation of LDA can be seen in Fig. 4.

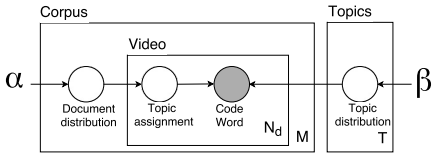

Figure 4: Graphical model representation of LDA using plate notation. Nodes represent random variables, links between nodes are conditional dependencies, plates are replicated components, and shaded nodes are observables.

Here, α, β are the model-level Dirichlet hyperparameters, T is the number of topic distributions (activity concepts), M the number of detections in the corpus, and N_d is the number of code words in a detection d. As the model shows, only the code words/histograms are observable, while the rest of the latent variables are inferred, namely the distribution $p(\text{topic distr., doc. distr., code word assignments} \,|\, \text{histogram})$, which captures a mixture of activity topics in a detection.

4 Grounding Natural Language to Concepts

In this section we describe the process of grounding natural language sentences to the learned concepts (e.g. faces, colours, human activities, etc.), in order to enable the robot to communicate effectively with the humans in its environment.

First, it is essential to acquire a natural language description of what the robot is perceiving to perform the grounding. Ideally we would like our robot to have a speech recognition modality and the capacity to ask people about particular objects, qualities or actions using natural language, but this remains a goal for future work. At present, we use Amazon Mechanical Turk to collect multiple natural language descriptions of video clips recorded by the robot. The descriptions are parsed into grammar trees using NLTK and an off-the-shelf English grammar model [Schuster and Manning, 2016]. For example, parsing the sentence "Andy is tall and is wearing a blue shirt with black shorts" gives a grammar tree from

which we extract all verbs (e.g. "wear"), nouns (e.g. "Andy"), and adjectives (e.g. "blue"). After applying a low-pass filter on the frequency of extracted words (removing words with low occurrence – fewer than three times in the entire dataset), this becomes the set of words that will be used for language grounding to concepts. We attempt to ground verbs to activity concepts, adjectives to colour concepts and nouns to people names and object labels.

For grounding, we search for the highest correlations between words in a video clip description and the various concepts that feature in that clip, allowing multi-to-multi associations to preserve the richness of natural language. Given the set of m learned concepts $\mathcal{C}$ and the set of n unique words $\mathcal{W}$, the *concept-word correlation* matrix K is an $m \times n$ matrix computed using the maximum of two frequentist measures:

$$K(c,w) = \max\left(\frac{\#(c,w)}{\#(c)}, \frac{\#(c,w)}{\#(w)}\right), \quad c \in \mathcal{C}, w \in \mathcal{W}$$

where $\#(.)$ is the count function. This computes the number of times a concept and a word are observed together, normalised by either the number of times the word is observed or the concept is observed, i.e. the strength of associating the word to the concept or the concept to the word. The maximum of these two terms concentrates on the less-observed of the word and the concept, improving the quality of the multi-to-multi associations.

Defining a target function $\mathcal{A}$ where $\mathcal{A}(c,w) = 1$ if the association (c,w) is correct and 0 otherwise, we can formulate the problem of multi-to-multi concept-to-word association as solving a constrained integer program with the objective function:

$$\max_{\mathcal{A}} \sum_{\mathcal{C}\times\mathcal{W}} \mathcal{A}(c,w)\, K(c,w).$$

We optimise the objective function subject to the constraints:

- $\sum_{\mathcal{C}\times\mathcal{W}} \mathcal{A}(c,w)/mn < \lambda\%$, keeping sparsity of the associations by forcing the number of selected associations to be below some small percentage $\lambda\%$ (set between 5 and 10%) of the total number of possible associations.
- $\sum_{\mathcal{W}} \mathcal{A}(c,w) \geq 1,\ \forall c \in \mathcal{C}$, forcing the assignment of at least a single word to each of the concepts.

Solving this integer program results in assigning a number of highly-correlated words to each concept. The error in this process gets rectified through continual learning as more video is processed.

5 Continual Learning

In this section we describe the incremental techniques we use to update the visual concepts and activity topics with new observations, and the incremental language grounding process.

For the concepts extracted from 2D visual features (i.e. faces and colours), we use an Incremental Gaussian Mixture Model (IGMM) [Song and Wang, 2005] which uses statistical tests (namely, W-statistic and Hotelling's T^2 test) to decide whether a new measurement is part of a known component (representing a learned concept), and thus the component is updated with the measurement. Otherwise, a new component in the feature space is created, i.e. a new concept is learned.

For human activity concepts, we incrementally update the generative LDA model using Variational Bayes Inference (VB) [Hoffman *et al.*, 2010]. For new observations the process is threefold: i) any new code words in the observations are first appended to the vocabulary and to the topic distributions with zero probability, ii) the multinomial distribution over the current set of topics/activity concepts for the new detection is computed, then iii) the topic distributions over the vocabulary are updated with the new observations.

The use of an Incremental Gaussian Mixture Model and Variational Bayes inference allows the robot to incrementally learn new concepts in the environment, whilst efficiently updating its model of the previously-learned concepts with new observations. By processing a small number of observations at a time, both IGMM and VB optimise storage and computational complexity of the framework, avoiding the need to store or re-analyse previous observations.

For grounding of natural language, the integer programming association is performed again whenever new observations and text descriptions are available. This is vital as the richness of natural language and the possible noise in the data require continuous re-evaluation of the associations. This is achieved by i) adding new rows and columns to the correlation matrix K for new words and newly-learned concepts, ii) updating the frequency measure of every observed word and concept pair, then iii) re-solving the integer program to generate new associations. Again, there is no need to store anything more than the frequencies in K.

6 Empirical Evaluation

We present three experiments to evaluate the system's performance in: 1) unsupervised concept extraction, 2) unsupervised language grounding, and 3) simple sentence generation to describe previously unseen video clips. We use a publicly-available long-term human activity dataset collected over the period of five days by a mobile robot from multiple view points [1]. The dataset contains 493 video clips each containing a single human performing a simple activity in a kitchen area of an office environment (e.g. heating food, preparing hot drinks, using a multi-function printer, throwing trash, washing up, etc.) On top of the dataset, we collected natural language descriptions of each video clip using Amazon Mechanical Turk, where we requested 'Turkers' to describe the activity in the clip and the person's appearance (given a fabricated name). A total of almost 3000 descriptions were collected (6 per clip on average). Example images from a video clip are shown in Fig. 5 along with a subset of the descriptions obtained.

Concept Extraction Evaluation We incrementally extract concepts in each of the feature spaces; namely faces, colours, objects, and activities, over the 5-day dataset. Since the learning is performed in an unsupervised setting, we use two popular clustering metrics to evaluate the performance:

[1] Dataset: http://doi.org/10.5518/86

"Eris taking milk from fridge and fill in the glass"

"Eris is making coffee"

"Eris is bald man, neatly dressed in a professional attire, wearing white shirt and a black trouser"

Figure 5: Example images from the dataset with natural language descriptions. Fabricated names were provided to the annotators.

normalised *Mutual Information* [Vinh *et al.*, 2009] and *V-measure* [Rosenberg and Hirschberg, 2007]. Normalised Mutual Information is a measure of how many bits are needed in order to store predicted outcomes given that the true value is known. V-measure is a combination of *homogeneity* – whether each predicted cluster contains same-class data points, and *completeness* – whether the member data points of a given class are all elements of the same predicted cluster. Both metrics provide a measure of similarity of any two sets of class labels, where 0 indicates no mutual information and 1 indicates perfect correlation. For ground truth we use the sets of 17 names, 9 colours, 12 objects, and 11 activities, extracted manually from the dataset by paid volunteers. For objects, the closest from the 12 objects to the location where the person stopped in the video is chosen to be the ground truth object.

Table 1 presents results of our incremental, unsupervised concept extraction when compared against ground truth classes. We use the most likely component in a mixture as a label if the prediction is multinomial, as in the case of activity topics. The robot managed to recover 34 face concepts, 13 colour concepts, 14 object concepts, and 13 activity concepts from this challenging real-world dataset with multiple view points, changing lighting conditions and occlusions. The results show the majority of the instances observed are successfully clustered into consistent concepts.

As an upper bound and to provide a reference result, we also show the V-measure results obtained using a supervised (linear) support vector machine classifier (SVM) with 4-fold cross-validation. The SVM clearly has access to the ground truth labels during training. Still, it only marginally outperforms our unsupervised techniques.

Given the limited size of the dataset, we compute the most prominent 20 Eigenfaces from the observations of day 1, and use them after that to compute Eigenvalues in all later detections. Also, we first seed the activity model by learning topics using Collapsed Gibbs Sampling [Gelman *et al.*, 2014] on day 1 observations in batch mode. After that, we incrementally process new data using Variational Bayes with a regular mini-batch size of 5 videos to allow frequent updating. For the number of topics/activity concepts T, we first start with the number of discovered objects, then increase this number by one each day to allow new activities to appear over time. Also, we remove any unused topics.

It is worth noting that all data collection, processing and analysis were performed using midrange CPU and GPU units. Our robot has two PCs with i7 processors running ROS indigo, and a single GTX 1050 Ti GPU with 2 GB of memory on which the convolutional pose machine for human pose estimation runs. The use of incremental techniques (IGMM and VB) for concept learning allowed relatively less complex and more memory-efficient processing, making it possible for the full framework to run onboard.

Metric	Faces	Colours	Objects	Activities
Mutual Information	1.85	1.27	1.21	1.34
Normalised MI	0.70	0.70	0.69	0.62
Homogeneity Score	0.90	0.91	0.71	0.60
Completeness Score	0.55	0.54	0.68	0.64
V-measure	0.68	0.66	0.69	0.62
V-measure (SVM)	*0.75*	*0.74*	*0.77*	*0.69*

Table 1: Experimental results of unsupervised concept extraction showing four clustering metrics for face, colour, object and activity extraction. Also, we show the V-measure using a supervised SVM, that has access to the ground truth labels in the four datasets, as an upper limit.

Language Grounding Evaluation We evaluate the system's ability to acquire correct word groundings using pairs of short video clips accompanied with their corresponding descriptions. We aim to learn all the possible groundings of words to their corresponding visual concepts. For ground truth, we manually annotated all correct word-concept groundings in the dataset. As a metric, we compute the F1-score of the grounding results in each feature space separately. Daily batches of recorded videos are fed to the system, effectively updating the robot's groundings each evening.

Figure 6 (left) shows the improving trend in the F1-score of the word groundings in each feature space as more data is observed in the 5-day dataset. We hypothesise that extended observation of the environment will allow all the concepts in these pre-defined feature spaces to be correctly grounded in an unsupervised manner. Similarly, the concepts themselves will become better defined with more observations. Examples of learned concepts and the natural language words which are grounded to them are shown in Fig. 6 (right).

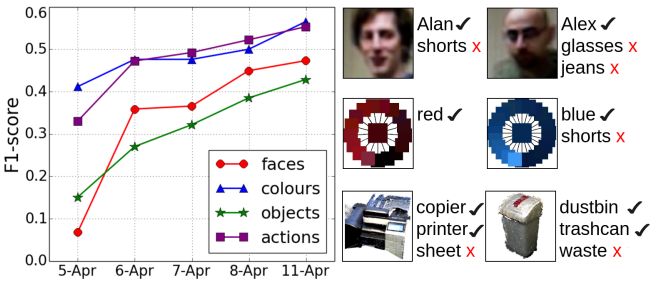

Figure 6: (**left**) F1-score for incremental grounding over five days. (**right**) Examples of learned concepts from the three simple feature spaces along with their grounded words. Note that in each case one or more groundings are correct. The rest are semantically related but are not the actual concept labels (e.g., Alex wears glasses and jeans in all the videos he appears in.)

Sentence Generation Evaluation Finally, we evaluate the soundness of both the learned concepts and word groundings by generating natural language sentences of previously unseen video clips. For this task, we removed a video clip from the training data and passed it to the robot after training on the remaining videos in order to generate a sentence describing the unseen video. We repeated this 10 times. The robot is provided with natural language sentence templates, to describe an activity or a person, with placeholders for concept labels. The two templates we use are "⟨*person*⟩ has a ⟨*colour*⟩ top and a ⟨*colour*⟩ lower garment" and "The person is ⟨*activity*⟩ using a(n) ⟨*object*⟩". The robot detects learned concepts from the test video and picks their most-highly associated words to fill in the sentence templates. In 10 videos our system was able to correctly generate/fill 46 out of the 50 available blank spaces. The correctness of the generated sentences were evaluated by an external volunteer. Examples of the generated sentences along with images from the test video clips can be seen in Fig. 7.

7 Conclusion

We have presented a novel framework for autonomous learning of human concepts in real-world scenarios, with partial, noisy and changing viewpoints of the world using on-board sensors and limited computing power of a mobile robot. The framework continually acquires and updates simple and complex concepts, differing in the richness of the feature spaces in which they are embedded, in an unsupervised manner and grounds natural language words to them. The perceptual side of the framework is presented in this paper, while exploiting the learned concepts in useful behaviour is an ambition for future work.

For language grounding, the framework depends on human descriptions of the visual scenes. We noticed that limitations in human perception, e.g. under varying lighting conditions, might cause errors in scene annotation leading to mistakes in grounding. Even though the process of continual learning is able to rectify erroneous associations in grounding as more data is observed, a direction of future work is to enable the robot to *enquire* about unknown entities/activities directly, removing the need of external annotations. An additional improvement to the grounding could be to remove from consideration words already having a strong association to a concept, or words that are not consistent to any particular concept. This would boost scalability of the continual grounding over time. Also, the interplay between concept learning and language grounding could be exploited to improve both processes.

Finally, the fixed experimental setup made it sufficient for the object segmentation to be performed with a single scan of the environment. A direction for future work could be to continually track the object locations in the environment by modelling the change in the object segmentation representation in multiple scans over time. This lifelong learning approach also requires handling a continuous, unsegmented visual feed, compared to the segmented videos used in this paper. This is possible using a method similar to [Duckworth *et al.*, 2017], but correlating natural language annotations to the unsegmented video stream would be more challenging.

a - *Andy* has a *purple* top and a *black* lower garment.
b - The person is *cooking* using a *microwave*.
c - *Alan* has a *blue* top and a *black* lower garment.
d - The person is *rinsing* using a *kettle*.

Figure 7: Examples of generated sentences from previously-unseen videos. (a-b) describing video 1, (c,d) describing 2.

Acknowledgments

We acknowledge the financial support provided by EU FP7 project 600623 (STRANDS).

References

[Aksoy *et al.*, 2017] Eren E. Aksoy, Ekaterina Ovchinnikova, Adil Orhan, Yezhou Yang, and Tamim Asfour. Unsupervised linking of visual features to textual descriptions in long manipulation activities. *IEEE RA-L*, 2017.

[Allen, 1983] James F. Allen. Maintaining knowledge about temporal intervals. *CACM*, 26(11):832–843, 1983.

[Alomari *et al.*, 2017] Muhannad Alomari, Paul Duckworth, David C. Hogg, and Anthony G. Cohn. Natural language acquisition and grounding for embodied robotic systems. In *AAAI*, 2017.

[Beetz *et al.*, 2011] Michael Beetz, Ulrich Klank, Ingo Kresse, Alexis Maldonado, Lorenz Mosenlechner, Dejan Pangercic, Thomas Ruhr, and M. Tenorth. Robotic roommates making pancakes. In *Humanoid Robots*, 2011.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3, 2003.

[Bore *et al.*, 2017] Nils Bore, Rares Ambrus, Patric Jensfelt, and John Folkesson. Efficient retrieval of arbitrary objects from long-term robot observations. *Robot Auton Syst*, 2017.

[Chen *et al.*, 2015] Juan Chen, Anthony G. Cohn, Dayou Liu, Shengsheng Wang, Jihong Ouyang, and Qiangyuan Yu. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(01):106–136, 2015.

[Clementini *et al.*, 1997] Eliseo Clementini, Paolino Di Felice, and Daniel Hernández. Qualitative representation of positional information. *Artif Intell*, 95(2):317 – 356, 1997.

[Craye *et al.*, 2015] Céline Craye, David Filliat, and Jean-François Goudou. Exploration strategies for incremental learning of object-based visual saliency. In *ICDL-EpiRob*, 2015.

[Delafontaine *et al.*, 2011] Matthias Delafontaine, Anthony G. Cohn, and Nico Van de Weghe. Implementing a qualitative calculus to analyse moving point objects. *Expert Syst Appl*, 38(5):5187 – 5196, 2011.

[Duckworth *et al.*, 2017] Paul Duckworth, Muhannad Alomari, James Charles, David Hogg, and Anthony Cohn. Latent Dirichlet Allocation for unsupervised activity analysis on an autonomous mobile robot. In *AAAI*, 2017.

[Gatsoulis *et al.*, 2016] Yiannis Gatsoulis, Muhannad Alomari, Chris Burbridge, Christian Dondrup, Paul Duckworth, Peter Lightbody, Marc Hanheide, Nick Hawes, and Anthony G. Cohn. QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video. In *Workshop on Qualitative Reasoning, at IJCAI*, 2016.

[Gelman *et al.*, 2014] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, 2014.

[Hemachandra *et al.*, 2015] Sachithra Hemachandra, Felix Duvallet, Thomas M. Howard, Nicholas Roy, Anthony Stentz, and Matthew R. Walter. Learning models for following natural language directions in unknown environments. In *ICRA*, 2015.

[Hoffman *et al.*, 2010] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for Latent Dirichlet Allocation. In *NIPS*, 2010.

[Jovan *et al.*, 2016] Ferdian Jovan, Jeremy Wyatt, Nick Hawes, and Tomáš Krajník. A Poisson-Spectral Model for Modelling the Spatio-Temporal Patterns in Human Data Observed by a Robot. In *IROS*, 2016.

[Lauria *et al.*, 2002] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, and Ewan Klein. Mobile robot programming using natural language. *Robot Auton Syst*, 2002.

[Lienhart and Maydt, 2002] Rainer Lienhart and Jochen Maydt. An extended set of Haar-like features for rapid object detection. In *Image Processing*, 2002.

[Matuszek *et al.*, 2013] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, 2013.

[MetraLabs, 2016] MetraLabs. www.metralabs.com, 2016.

[Moratz and Ragni, 2008] Reinhard Moratz and Marco Ragni. Qualitative spatial reasoning about relative point position. *J Visual Lang Comput*, 19(1):75–98, 2008.

[OpenNI, 2016] OpenNI. www.openni.org, 2016.

[Papon *et al.*, 2013] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *CVPR*, 2013.

[Pfister *et al.*, 2000] Hanspeter Pfister, Matthias Zwicker, Jeroen Van Baar, and Markus Gross. Surfels: Surface elements as rendering primitives. In *SIGGRAPH*, 2000.

[Posada and Buckley, 2004] David Posada and Thomas R. Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004.

[Rosenberg and Hirschberg, 2007] Andrew Rosenberg and Julia Bell Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, 2007.

[Schoeler *et al.*, 2015] Markus Schoeler, Jeremie Papon, and Florentin Worgotter. Constrained planar cuts-object partitioning for point clouds. In *CVPR*, 2015.

[Schuster and Manning, 2016] Sebastian Schuster and Christopher D. Manning. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *LREC*, 2016.

[She *et al.*, 2014] Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Y. Chai, and Ning Xi. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Meeting of the SIG on Discourse and Dialogue*, 2014.

[Sinapov *et al.*, 2014] Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *ICRA*, 2014.

[Song and Wang, 2005] Mingzhou Song and Hongbin Wang. Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering. In *Defense and Security*, 2005.

[Song *et al.*, 2016] Young C. Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry Kautz. Unsupervised alignment of actions in video with text descriptions. In *IJCAI*, 2016.

[Spranger and Steels, 2015] Michael Spranger and Luc Steels. Co-Acquisition of Syntax and Semantics - An Investigation in Spatial Language. In *IJCAI*, pages 1909–1905. 2015.

[Tellex *et al.*, 2011] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis G. Banerjee, Seth Teller, and Nicholas Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

[Turk and Pentland, 1991] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J cognitive neurosci*, 3(1), 1991.

[Vinh *et al.*, 2009] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *ICML*, 2009.

[Wei *et al.*, 2016] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[Young *et al.*, 2016] Jay Young, Valerio Basile, Lars Kunze, Elena Cabrio, and Nick Hawes. Towards lifelong object learning by integrating situated robot perception and semantic web mining. In *ECAI*, 2016.