Proceedings Paper:
Alomari, M, Duckworth, P orcid.org/0000-0001-9052-6919, Bore, N et al. (3 more authors) (2017) Grounding of Human Environments and Activities for Autonomous Robots. In: IJCAI-17 Proceedings. 26th International Joint Conference on Artificial Intelligence, 19-25 Aug 2017, Melbourne, Austrailia. Lawrence Erlbaum Associates, Inc. , pp. 1395-1402. ISBN 9780999241103

Grounding of Human Environments and Activities for Autonomous Robots

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

Abstract

With the recent proliferation of 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 framework for autonomous, unsupervised learning from various sensory sources of useful human 'concepts'; including colours, people names, usable objects and simple activities. This is achieved by integrating state-of-the-art object segmentation, pose estimation, activity analysis and language grounding into a continual learning framework. Learned concepts are grounded to natural language if commentary is available, allowing the robot to communicate in a human-understandable way. We show, using a challenging, real-world dataset of human activities, that our framework is able to extract useful concepts, ground natural language descriptions to them, and, as a proof-of-concept, to generate simple sentences from templates to describe people and activities.

1 Introduction

For mobile robots to integrate in human environments, it is essential that they be equipped with the abilities to continuously learn about their environments, the people that inhabit these environments, and the activities that take place there. From an autonomous robot point of view, this requires incremental, unsupervised methods that can operate on the outputs of various kinds of sensor modalities the robot might have, which may include RGB and depth camera outputs, voice recognition, laser rangefinder measurements, etc. The desired outcome of this process is a collection of grounded concepts, such as colours, objects, people and activities that occur in the robot's environment.

In this paper, we present a framework for autonomous learning of human environments for a mobile robot. We presuppose that the robot can visually analyse the environment in order to extract a multitude of features and incrementally recover useful classes of features, or *concepts*. If natural language descriptions of the observations are provided, they can also be analysed, along with the other features, to ground the words describing people, objects and activities to their most relevant perceptual concepts. Thus, the framework supports recognition of individuals, describing their physical appearance using natural language, and classifying and commenting on the activities they are engaged in. To do this we integrate state-of-the-art object segmentation, pose estimation, activity analysis into a flexible, incremental framework for learning to distinguish instances of human-level 'concepts' (faces, colours, objects, and activities) in real-world complex scenarios. Moreover, we propose a simplified language grounding framework that works across multiple modalities to learn concept names for human-robot interaction purposes using natural language descriptions.

One possible application of such a framework can be in the field of assistive robotics where such a robot could help around the home, learning on-the-go how to describe new objects or situations in a human-understandable form.

We concentrate on a small number of features and sensory data that are easily acquirable by capable mobile robots. To learn about humans in the robot's environment, 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 motion analysis to identify potential objects in the environment. For human activities, we use qualitative spatial-temporal representations to capture the interaction through the relations between people's body poses and object positions, which feeds into a generative Bayesian model to learn activity classes in an unsupervised setting. Lastly, given textual descriptions, we use natural language grounding techniques to assign words and phrases to the learned concepts in the numerous feature spaces. Note that the chosen features are not intended to be exhaustive, but rather to demonstrate our approach.

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 including 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]. In this work, we learn concepts and natural language groundings in a real-world human environment, which is more challenging in nature than the simple controlled environments that are studied in the literature.

Researches have addressed incremental learning of simple elements in the robot's environment, e.g. object features [Sinapov *et al.*, 2014; Craye *et al.*, 2015]. On the other hand, other work 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.

3 Concepts

In this section we introduce our notion of *concepts*: abstractions of the feature spaces generated by the robot modalities which carry a human-level meaning. For example, concepts might include a colour represented as a cluster of values in the HSV colour space, or an object represented as a cluster of points in a 3D point cloud.

We present in the next section the sensors and feature spaces we use along with the unsupervised methods we employ to generate such concepts using a Scitos A5 mobile robot [MetraLabs, 2016] running ROS Indigo. 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.

For its basic operations, the mobile robot we use is equipped with a base-mounted laser scanner that is 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 Figure 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 Figure 1 (right). To estimate the human pose, we use a real-time depth-only tracker built on OpenNI [2016] along with a post-processing state-of-the-art 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, e.g.

3.1 Extracting Concepts

Concepts are learned automatically by clustering the low-level sensory input of each of the sensor modalities of the robot after an appropriate encoding. This clustering operation results in a collection of classes that are candidate concepts within each feature space. Because we assume no

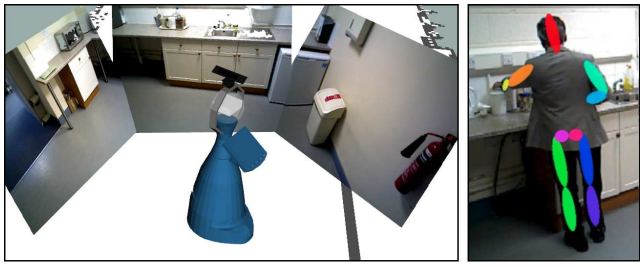

Figure 1: (**left**) Mobile A5 robot gathering data shown overlaid on its 2D map. (**right**) Example detected human pose.

pre-knowledge of the structure of the sensor feature spaces, we employ probabilistic modeling 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 can be detected in single observations. For example, simple visual concepts like colours can be represented as Gaussian components in a Gaussian Mixture Model over the HSV space [Alomari *et al.*, 2017]. Similarly, objects are simple concepts that can be segmented from fused 3D point clouds using geometrical and textural cues [Bore *et al.*, 2017].

On the other hand, *complex concepts* 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 more sophisticated clustering mechanism are needed [Duckworth *et al.*, 2017].

The robot first abstracts each observed human pose sequence using a qualitative representation and obtains clusters using a hierarchical Bayesian model, Latent Dirichlet Allocation [Blei *et al.*, 2003]. It translates the detected pose sequence into a relatively small number of logical spatial relations that can be used to qualitatively describe the interactions taking place. The topics recovered from this process are considered human activity concepts which the robot grounds to words in natural language.

In this paper, we demonstrate extracting four 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 around the location of the head joint using the pose estimate is automatically cropped from the visual feed for every person detection. We detect the presence of a face in the cropped images using a cascade of boosted classifiers with Haar features [Lienhart and Maydt, 2002] along with OpenCV generic face model. Then, we extract the Eigenvalues for the n most prominent Eigenfaces [Turk and Pentland, 1991]. This transforms a face into a much-smaller n-dimensional data point. Then, 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 Figure 2 (faces).

Colours: We cluster the 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 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 estimeated by the average of sampled pixel colours from the triangle of the two shoulders and the torso, and the colour of the lower garment is sampled from the triangle between the torso and the knees, as shown in Figure 2 (colours). The extracted colours are projected into Hue-Saturation-Value (HSV) space to increase the robustness under varying lighting. Examples of six clusters extracted can be seen in Figure 2 (colours).

Objects: The robot constructs a 3D model of its environment by fusing RGB-D images into *surfels* [Pfister *et al.*, 2000], from which it generates clusters of "objects of interest". 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 split the scene into a collection of *supervoxels* [Papon *et al.*, 2013] over which an adjacency graph 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. The end result is a collection of point cloud segments as illustrated in Figure 2 (objects).

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

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 objects of interest. Then, the observations are encoded into a number of qualitative spatio-temporal abstractions as in [Duckworth *et al.*, 2017]. This condenses noisy observations of arbitrary spatial positions into semantic low-level qualitative descriptors. For example, in a "making coffee" activity, the exact spatial position of a person reaching for a mug is not as useful for learning the activity as a qualitative representation of a hand approaching a mug.

We briefly introduce the Qualitative Spatio-temporal Representations (QSRs) used to encode detected pose-object sequences. A QSR is an abstraction from exact quantitative ob-

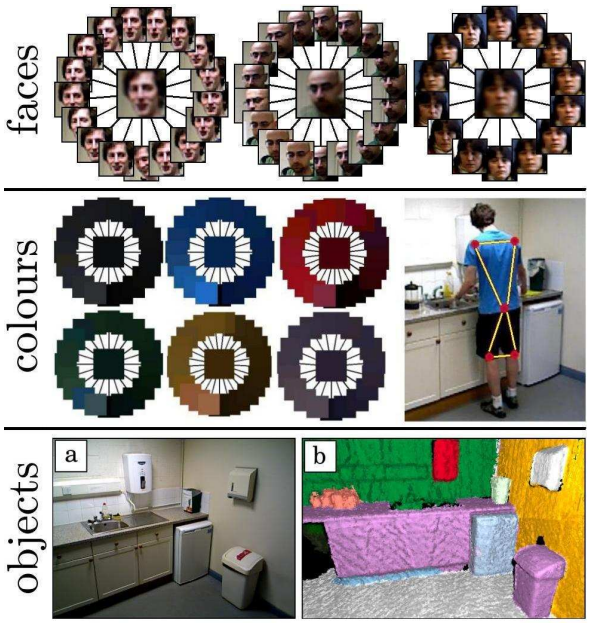

Figure 2: **(faces)** Examples of face clusters, with the averaged (mean) face shown in the center of each cluster. **(colours)** Left: Examples of different colour clusters, with the averaged (mean) colour shown in the center of each cluster. Right: defining upper and lower garments using human pose estimate. **(objects)** Processing of RGB-D feed (a) One image in a 3D sweep (b) Segmented surfel map.

servations in a particular feature space into qualitative states that hold between the human's pose and objects in the environment. The three representations used in this paper are: 1) Ternary Point Configuration Calculus (TPCC) [Moratz and Ragni, 2008] qualitatively describes the spatial arrangement of a point relative to two others. That is, it describes the *referent*'s position relative to the plane created by connecting the *relatum* and *origin*. Relations are triples of $\langle$ { front, back }, { left, right, straight }, { distant, close } $\rangle$. 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 $(0)\}$. 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 person with respect to reference landmarks, while changes in the relations can help explain relative motion. An illustration of the three QSRs relative to two objects can be seen in Figure 3 (top).

Once each human pose-objects sequence is converted into a set of qualitative relations (one per frame), we perform a temporal abstraction using Allen Interval Algebra (IA) [Allen, 1983]. This compresses repeated qualitative relations at adjacent frames into an *interval* representation, maintaining the relation and duration information. Secondly, IA

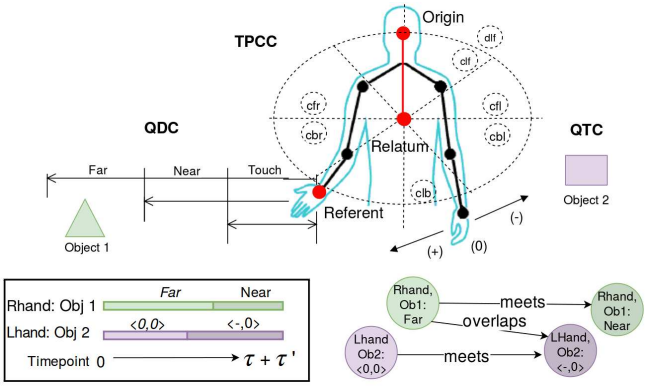

Figure 3: QSRs and Interval representations; (**top left**) QDC (relative distance) between right hand and *object* 1. (**top centre**) subset of the TPCC system between right hand and (relatum-origin) plane. (**top right**) QTC (relative motion) between left hand and *object* 2. (**bottom left**) Interval representation. (**bottom right**) Interval Graph.

relations are computed between temporally connected intervals to create an *Interval Graph* where nodes represent intervals (relations holding between a set of objects) and directed arcs link nodes with the IA relation. An example interval representation and Interval Graph can be seen in Figure 3 (bottom left and bottom right, respectively).

Given a corpus of Interval Graphs, one per human detection, a set of unique k-length paths are extracted from the graphs as *code words* for some small k (usually ≤ 4), where a code word represents a small set of temporally-connected spatial relations between some objects (likewise ≤ 4). This unique set of code words is considered as a discrete *vocabulary*, and thus bag-of-words descriptors of activities (called *activity feature vectors*) can be computed for each detection. This bag-of-words representation is different from the traditional bag-of-words used normally in document analysis in that it maintains some temporal information through the structure of the code words.

We use Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], a three layer hierarchical Bayesian generative model of a collection of discrete data, to discover *topics* in the activity feature vectors. This model has proved successful in problems with large corpora not exclusive to document analysis, e.g. [Duckworth *et al.*, 2017]. The graphical model representation can be seen in Figure 4, where α, β are the model-level Dirichlet hyperparameters, T is the number of topic distributions, M the number of videos in the corpus, and N_d is the number of code words in video d. A topic, a probability distribution over the vocabulary of code words, is a conceptual model of a human activity, and thus it is considered as a candidate concept.

4 Grounding Natural Language to Concepts

In this section we describe how the robot performs grounding of natural language sentences to the automatically-learned concepts, in order to enable the robot to communicate effectively with the humans in its environment.

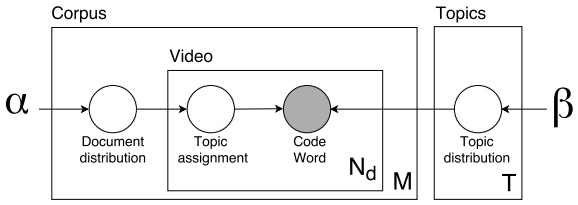

Figure 4: Graphical model representation of LDA using plate notation. Nodes represent random variables, links between nodes are conditional dependencies, plates are replicated components, and shaded nodes are observations.

First, it is essential that the robot gets a natural language description of what it is learning about to perform grounding. Ideally we would like our robot to have a speech recognition modality and the capacity to ask people for particular objects, qualities or actions, but this remains an ambition for the future. At present, we collect multiple natural language descriptions of video snippets recorded by the robot using Amazon Mechanical Turk. The descriptions are parsed into grammar trees using NLTK [Bird *et al.*, 2009] 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" we obtain a grammar tree, and from these trees we extract all verbs, nouns, and adjectives, e.g. "Andy" is a noun, "blue" is an adjective, etc. After applying a low-pass filter (remove words with low occurrences), this becomes the set of words used to ground to concepts.

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: $K(c,w) = \max\left(\frac{\#(c,w)}{\#(c)}, \frac{\#(c,w)}{\#(w)}\right)$, where $\#(.)$ refers to the count function. This translates to computing 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 the chosen associations as a function $\mathcal{A}$ where $\mathcal{A}(c,w) = 1$ if the association (c,w) is chosen and 0 otherwise, we can formulate the problem as solving an integer program with the objective function:

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

We maximise that objective function while: i) keeping sparsity of the associations by forcing the number of selected associations to be below some small $\lambda\%$ (set between 5 and 10%) of the total number of possible associations: $\sum_{\mathcal{C}\times\mathcal{W}} \mathcal{A}(c,w)/mn < \lambda\%$, and ii) forcing the selection to assign at least a single word to each of the concepts, $\sum_{\mathcal{W}} \mathcal{A}(c,w) \geq 1, \forall c \in \mathcal{C}$. Solving this integer program re-

sults in assigning a number of highly-correlated words to each concept. The error in this process gets rectified through continual learning.

5 Continual Learning

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

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 (W-statistic and Hotelling's T^2 test) to decide whether a new measurement is part of the currently learned components or not. If they are, the component is updated. Otherwise, a new component in the feature space is created, i.e. a new concept is created.

For human activity concepts, we incrementally update our generative LDA model using Variational Bayes Inference (VB) [Hof, 2010]. For new observations the process is twofold: i) the multinomial distribution representing the observed activity over the current set of topics is computed, then ii) the topic distributions over the vocabulary are updated using this new observation. New code words can be added to the vocabulary if they do not already exist, and the topic distributions are uniformly initiated. This allows the robot to efficiently update its model of activity concepts using a single pass over the data, optimising both storage and computation complexity.

For 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) updating the frequency measure of every observed word and concept in the correlation matrix K, ii) adding new rows and columns to K, corresponding to newly learned concepts and newly observed words, then iii) re-solving the integer program to generate the new associations. Thus, the previous data needs not be stored.

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 a one week period by a mobile robot from multiple view points (Dataset: http://doi.org/10.5518/86). The dataset contains 493 video clips each containing a single human performing a simple activity in a kitchen area of an office environment, the activities include, for example, heating food, preparing hot drinks, using a multi-function printer, throwing trash and washing up. 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. A total of almost 3000 descriptions were collected (6 per clip in average)[1]. Example images from a video clip are shown in Figure 5 along with a subset of the descriptions obtained.

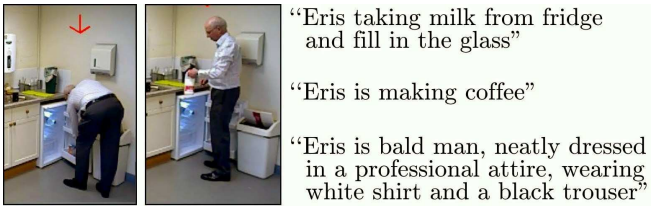

Figure 5: Example images from a video clip, with natural language descriptions.

Concept Extraction Evaluation We incrementally extract concepts in each of the feature spaces; namely faces, colours, objects, and activities over the 5 days in the dataset.

Since the learning is performed in an unsupervised setting, we use two popular clustering metrics to evaluate the performance: *V-measure* [Rosenberg and Hirschberg, 2007] and normalised *Mutual Information* [Vinh *et al.*, 2009]. 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. 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. 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 9 colours, 12 objects, 17 names, and 11 activities extracted manually from the dataset.

Metric	Faces	Colours	Objects	Activities
V-measure (SVM)	*0.95*	*0.74*	–	*0.71*
V-measure	0.69	0.66	0.69	0.62
Homogeneity Score	0.90	0.91	0.71	0.60
Completeness Score	0.55	0.54	0.68	0.64
Mutual Information	1.87	1.27	1.21	1.34
Normalised MI	0.71	0.70	0.69	0.62

Table 1: Experimental results of unsupervised concept extraction. In addition to a supervised SVM as an upper limit.

Table 1 presents results of our incremental, unsupervised concept extraction when compared against ground truth classes (assigned by volunteers). We use the most likely component in a mixture as a label if the prediction is multinomial, like in the case of activity topics. The robot managed to recover 35 face concepts, 13 colour concepts, 14 objects of interest, and 13 activity classes from this real-world dataset with changing view points, lighting conditions and occlusions. The results show the majority of the instances observed are successfully clustered into consistent concepts. As an upper bound, we also show the V-measure results obtained when

[1]This collection of descriptions will be made public.

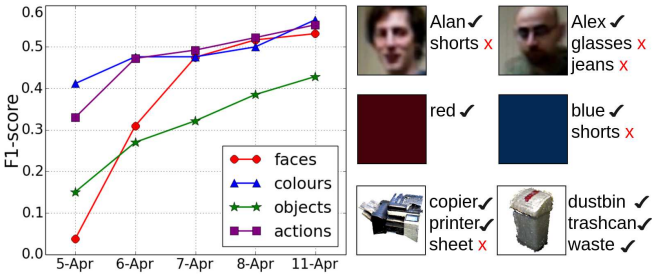

Figure 6: (**left**) F1-score for incremental grounding over 5 days. (**right**) Examples of learned concepts from the three simple feature spaces along with their grounded words. Note that in each case one or more groundings are correct.

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.

using a supervised SVM (with 4-fold cv) which has access to the ground truth labels during training; this marginally outperforms our unsupervised techniques. Note that the objects do not have SVM results as they are segmented from surfels and do not have a feature vector representation.

Given the limited size of the dataset, we seed the activity model by learning topics using Collapsed Gibbs Sampling [Gelman *et al.*, 2014] on a batch of observations first (day 1), and then incrementally adding new data using Variational Bayes with a regular mini-batch size of 5 videos to allow frequent updating. To pick the number of topic distributions T, we start with the number of discovered objects and increase this number by one each day to allow new activities to appear over time, and remove any unused topics.

Grounding Language Evaluation We evaluate the system's ability to acquire correct groundings for words from pairs of short video clips and their corresponding descriptions. We aim to learn all the possible groundings of words to their corresponding concepts. For ground truth, we manually annotated all correct word-concept groundings in the dataset. As a metric, we compute the F1-score [Van Rijsbergen, 1977] of the grounding results in each feature space separately. Batches of videos recorded by day are fed to the system, effectively updating the robot's groundings in the evening of each day. Figure 6 (left) shows the results of the incremental grounding over the 5 days in the dataset. The graph shows an improving trend in the F1-score of the word groundings in each feature space as more data is observed. 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. Examples of learned concepts and their grounded words are shown in Figure 6 (right).

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 leave 10 video clips out of the training data (one at a time), and pass each of them to the robot after training on the remaining videos. The robot is provided with natural language sentence templates that have placeholders for concept names to describe an activity or a person. The two templates we use are "*name/face* has a *colour* top and a *colour* lower garment" and "The person is *activity* using a(n) *object*". The robot extracts known concepts from the test videos 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 their video clips can be seen in Figure 7.

7 Conclusion

We present a framework for autonomous learning of human concepts for a mobile robot. The framework continually acquires and updates learned concepts in an unsupervised manner and grounds natural language words to them. The main challenge of autonomous learning using mobile robots include the partial, noisy and changing viewpoints of the world using on-board sensors, in addition to limited computing power.

We learn both simple and complex concepts, where the difference relates directly to the richness of the feature spaces in which the concepts are embedded. For language grounding we depend on human descriptions of the visual scenes, however, perception limitations e.g. varying lighting conditions, cause errors in the grounding. On the other hand, as more data is observed, continual learning rectifies the associations.

One improvement to the framework could be to remove words with a strong association to a concept from consideration, or words that are not consistent to concepts. This would boost scalability of the continual grounding over time. Finally, during evaluation we used a collection of segmented activity videos. Extending this to use an unsegmented visual feed is possible using the same methods described in [Duckworth *et al.*, 2017]. However, correlating natural language annotations to the unsegmented video would be more challenging.

Acknowledgments

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

References

[Aksoy *et al.*, 2017] E.E. Aksoy, E. Ovchinnikova, A. Orhan, Y. Yang, and T. Asfour. Unsupervised linking of visual features to textual descriptions in long manipulation activities. *RA-L*, 2017.

[Allen, 1983] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[Alomari *et al.*, 2017] M. Alomari, P. Duckworth, D. C. Hogg, and A. G. Cohn. Natural language acquisition and grounding for embodied robotic systems. In *AAAI*, 2017.

[Beetz *et al.*, 2011] M. Beetz, U. Klank, I. Kresse, A. Maldonado, L. Mosenlechner, D. Pangercic, T. Ruhr, and M. Tenorth. Robotic roommates making pancakes. In *Humanoid Robots*, 2011.

[Bird *et al.*, 2009] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of ML Research*, 3, 2003.

[Bore *et al.*, 2017] N. Bore, R. Ambrus, P. Jensfelt, and J. Folkesson. Efficient retrieval of arbitrary objects from long-term robot observations. *Robotics and Autonomous Systems*, 2017.

[Clementini *et al.*, 1997] E. Clementini, P. Di Felice, and D. Hernández. Qualitative representation of positional information. *Artificial Intelligence*, 95(2):317 – 356, 1997.

[Craye *et al.*, 2015] C. Craye, D. Filliat, and J. Goudou. Exploration strategies for incremental learning of object-based visual saliency. In *ICDL-EpiRob*, 2015.

[Delafontaine *et al.*, 2011] M. Delafontaine, A. G. Cohn, and N. Van de Weghe. Implementing a qualitative calculus to analyse moving point objects. *Expert Systems with Applications*, 38(5):5187 – 5196, 2011.

[Duckworth *et al.*, 2017] P. Duckworth, M. Alomari, J. Charles, D. C. Hogg, and A. G. Cohn. Latent dirichlet allocation for unsupervised activity analysis on an autonomous mobile robot. In *AAAI*, 2017.

[Gelman *et al.*, 2014] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[Hemachandra *et al.*, 2015] S. Hemachandra, F. Duvallet, T.M. Howard, N. Roy, A. Stentz, and M.R. Walter. Learning models for following natural language directions in unknown environments. In *ICRA*, 2015.

[Hof, 2010] Online learning for latent dirichlet allocation. In *Advances in neural information processing systems*, 2010.

[Lauria *et al.*, 2002] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein. Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38(3):171–181, 2002.

[Lienhart and Maydt, 2002] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing*, 2002.

[Matuszek *et al.*, 2013] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, 2013.

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

[Moratz and Ragni, 2008] R. Moratz and M. Ragni. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1):75–98, 2008.

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

[Papon *et al.*, 2013] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *CVPR*, 2013.

[Pfister *et al.*, 2000] H. Pfister, M. Zwicker, J. Van Baar, and M. Gross. Surfels: Surface elements as rendering primitives. In *Computer Graphics and Interactive Techniques*, 2000.

[Posada and Buckley, 2004] D. Posada and T. 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] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, 2007.

[Schoeler *et al.*, 2015] M. Schoeler, J. Papon, and F. Worgotter. Constrained planar cuts-object partitioning for point clouds. In *CVPR*, 2015.

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

[She *et al.*, 2014] L. She, S. Yang, Y. Cheng, Y. Jia, J.Y. Chai, and N. Xi. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Meeting of the Special Interest Group on Discourse and Dialogue*, 2014.

[Sinapov *et al.*, 2014] J. Sinapov, C. Schenck, and A. Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *ICRA*, 2014.

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

[Song *et al.*, 2016] Y.C. Song, I. Naim, A. Al Mamun, K. Kulkarni, P. Singla, J. Luo, D. Gildea, and H. Kautz. Unsupervised alignment of actions in video with text descriptions. In *IJCAI*, 2016.

[Spranger and Steels, 2015] M. Spranger and L. Steels. Co-Acquisition of Syntax and Semantics - An Investigation in Spatial Language. In *IJCAI*, pages 1909–1905. Palo Alto, US, 2015.

[Tellex *et al.*, 2011] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

[Turk and Pentland, 1991] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 1991.

[Van Rijsbergen, 1977] C. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119, 1977.

[Vinh *et al.*, 2009] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *ICML*, 2009.

[Wei *et al.*, 2016] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.