Artificial Intelligence 270 (2019) 67–92

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Unsupervised human activity analysis for intelligent mobile robots ☆

Paul Duckworth [a,*], David C. Hogg [b], Anthony G. Cohn [b,c]

[a] Machine Learning Research Group, University of Oxford, OX2 6ED, UK
[b] School of Computing, University of Leeds, LS2 9JT, UK
[c] Department of Computer Science and Technology in Tongji, College of Electronic and Information Engineering, University of Tongji, Shanghai, China

ARTICLE INFO

Article history:
Received 10 April 2018
Received in revised form 11 September 2018
Accepted 9 December 2018
Available online 7 January 2019

Keywords:
Human activity analysis
Mobile robotics
Qualitative spatio-temporal representation
Low-rank approximations
Probabilistic machine learning
Latent Dirichlet allocation

ABSTRACT

The success of intelligent mobile robots operating and collaborating with humans in daily living environments depends on their ability to generalise and learn human movements, and obtain a shared understanding of an observed scene. In this paper we aim to understand human activities being performed in real-world environments from long-term observation from an autonomous mobile robot. For our purposes, a human activity is defined to be a changing spatial configuration of a person's body interacting with key objects that provide some functionality within an environment. To alleviate the perceptual limitations of a mobile robot, restricted by its obscured and incomplete sensory modalities, potentially noisy visual observations are mapped into an abstract qualitative space in order to generalise patterns invariant to exact quantitative positions within the real world. A number of qualitative spatial-temporal representations are used to capture different aspects of the relations between the human subject and their environment. Analogously to information retrieval on text corpora, a generative probabilistic technique is used to recover latent, semantically-meaningful concepts in the encoded observations in an unsupervised manner. The small number of concepts discovered are considered as human activity classes, granting the robot a low-dimensional understanding of visually observed complex scenes. Finally, variational inference is used to facilitate incremental and continuous updating of such concepts that allows the mobile robot to efficiently learn and update its models of human activity over time resulting in efficient life-long learning.

1. Introduction

Advancements in the reliability of autonomous mobile robot platforms means they are well suited to continuously update their own knowledge of the world based upon their many observations and interactions [4,5]. Unsupervised learning frameworks over such long durations of time have the potential to allow mobile robots to become more helpful, especially when cohabiting human populated environments. By removing humans from the learning process, e.g. with no time-consuming data annotation, such robots can cheaply learn from greater quantities of available data (observations), allowing them to adapt to their surroundings and save time/effort hard-coding specific information. Maintaining an understanding of dy-

☆ Work carried out whilst at the University of Leeds. Parts of this work appear in conference proceedings [1], [2] and [3].

* Corresponding author.
E-mail addresses: paul.duckworth@eng.ox.ac.uk (P. Duckworth), d.c.hogg@leeds.ac.uk (D.C. Hogg), a.g.cohn@leeds.ac.uk (A.G. Cohn).

https://doi.org/10.1016/j.artint.2018.12.005

namic human environments, i.e. what human activities are occurring, in which regions at what times, allow a robot to adjust its own behaviour, or assist in a task being observed.

Our contributions are as follows: *i*) a qualitative spatial-temporal vector space framework for encoding observed human activities by an autonomous mobile robot; *ii*) methods for learning low dimensional representations of common and repeated patterns from multiple encoded visual observations using unsupervised probabilistic methods; *iii*) solutions to practical considerations when operating with long-term, autonomous mobile robots capturing continuous, unsegmented video sequences in a life-long learning setting.

Our methodology relies on first detecting and tracking human body movements from a single mobile robot's embedded sensors, along with learning the location of key objects in the environment using off-the-shelf techniques. Each human observation, originally recorded as a sequence of quantitative poses, is encoded using multiple qualitative calculi to abstract the exact spatial and temporal details of the observation, and finally represented as a vector of the occurrences of discrete qualitative descriptors (a vocabulary of which is learned from the data). We analyse the collection of encoded feature vectors analogously to a corpus of text documents containing multiple topics of interest. Multiple latent topics are recovered from the observations and considered as human activity classes, each defined as a multinomial distribution over an auto-generated vocabulary. Two techniques are presented to learn low-dimensional human activity representations. First, a non-probabilistic low-rank approximation approach is shown to work well with pre-segmented video sequences of observed human activity. Secondly, a more sophisticated probabilistic Latent Dirichlet Allocation (LDA) [6] technique is shown to somewhat remove the requirement for manual temporal segmentation of the recorded observations, allowing the robot to access large quantities of data which otherwise would need human annotation. LDA is a hierarchical Bayesian model where each observation is modelled as a mixture over an underlying set of topics, and each topic is, in turn, modelled as a mixture over the discrete vocabulary.

To the best of our knowledge, we are the first to combine a generative, probabilistic approach such as LDA with a qualitative spatial representation to recover low-dimensional representations of human activity observed from a real-world deployed mobile robot. This work moves away from using a standard dataset, where each data sample consists of a temporally segmented activity instance, to a more realistic setting where the instances are located in a longer observational sequence; this loosely translates as removing the assumption that humans continuously perform a sequence of interesting activities when being observed. A more reasonable assumption is that a human observation is modelled as a probabilistic mixture over an underlying number of latent topics, where some topics can be considered "interesting" human activities.

Specific challenges of using data captured from an autonomous mobile robot include: *i*) the robot's on-board sensors only grant a partial and changing viewpoint of the world, i.e. it obtains incomplete observations of activities being performed, which are often structurally noisy; *ii*) each observation is likely carried out in different ways, e.g. opening a door with opposite hands. The proposed framework helps alleviate these problems in two phases; first by utilising a state-of-the-art human pose estimator to improve the precision of observations, and secondly by using a *qualitative spatial representation* (QSR) with the ability to convert somewhat noisy observations of arbitrary spatial positions into semantic low level descriptors. For example, if a person reaches for a mug, the exact spatial position of the hand or mug are not as useful for learning the human activity "making coffee", as a qualitative representation of the hand "approaching" the mug.

Human activity analysis from mobile robots is a recent field of research, in part due to the advancements in navigation, localisation and planning using probabilistic robotics techniques [7]. This has allowed mobile robots to have more accurate and reliable estimates of their own location within their environment, and better able to perform actions based upon those estimates. This was highlighted by a successful indoor office marathon by a PR2 robot platform [8], in order to test the improved reliability of a navigation framework [4].[1] These capabilities allow mobile robots to co-exist with humans for long periods of time in dynamic environments, providing novel opportunity for human activity analysis on mobile robot platforms, and to learn from their own experiences.

2. Related work

There is a common distinction in the literature between vision-based human activity analysis, which extracts information from video (and depth) cameras using computer vision techniques, and sensor or wearable computing-based systems [10, 11]. Sensor-based systems often rely on the availability of small sensors, namely wearable sensors, smart phones, or radio frequency identification (RFID) tagged objects, that can be attached to a human under observation in order to obtain a representation of that person's movements. We focus on representing human activity from visual data, where the notion of *being observed* is restricted to a single camera's field of view. This is a mature sub-field of artificial intelligence and the reader is pointed to survey papers which cover the topic in detail using, largely static RGB cameras [12–14] or 3D depth cameras [15,16]. However, many of the common techniques in these surveys perform supervised learning, where

[1] Long-term robust reliability was also the focus of the EU funded STRANDS robotic project: strands-project.eu [5], where multiple MetraLabs mobile robots [9] were operational for a combined duration of 365 days autonomously travelling over 350 km and performing 23,000 defined tasks in long-term security and care installations over a four year period.

each training sample requires manual segmentation and annotating with a ground truth label. This is not a feasible solution for a long-term autonomous mobile robot which ideally, has minimal supervision whilst deployed in the real-world.

Unsupervised learning techniques are considered more appropriate for this task since they do not require time-consuming, offline manual annotations. Previous works have used Latent Semantic Analysis (LSA) [17], probabilistic LSA [18] and LDA [6] for learning low-dimensional human activity categories in an unsupervised setting: authors have combined these techniques with low-level Spatial Temporal Interest Point (STIP) features to learn action categories [19]; local shape context descriptors on silhouette images [20]; a combination of semantic and structural features to learn actions, faces and hand gestures [21]; and by fusing a vocabulary of local spatio-temporal volumes (cuboids) with a vocabulary of spin-images to capture the shape deformation of the actor [22]; However, a major problem cited in these works is the lack of spatial information about the human body captured by low-level image features, and the lack of more long-term temporal information encoded into the features restricts learning more complex actions. Descriptive spatial-temporal correlogram features have been used previously to address this issue [23], however, the approach still suffers from low-level image processing frailties, and the requirement for a single person to be modelled in the scene during a controlled training period. Another approach has been to learn the temporal relations between atomic actions in an unsupervised setting in order to accurately represent "composite" human activities [24]. However, the input videos for this technique require manual temporal segmentation into sequences of "overlapping fixed-length temporal clips", making it prohibitively expensive for life-long learning on an autonomous mobile robot. Further, each of these works have been performed without the variability of a mobile robot's frame of reference, and restricted to learning on temporally segmented video data during an offline training phase, unlike our work.

To address these issues, we abstract observed human and object estimated poses into a qualitative spatial representation. There is some evidence to suggest that there are dedicated areas of the brain to perform such abstractions [25]. It is therefore natural to attempt to embed this into a system to understand human behaviour in video data and ultimately, into autonomous robotic systems in order to represent behaviours performed in the environments they inhabit. Qualitative spatial and temporal calculi arise from a set of jointly exhaustive and pairwise disjoint relations. There are many types developed in the literature, some of the most popular include topological, directional and non-topological; a survey of popular calculi is given in [26]. Qualitative spatial representations are often used to represent visual, quantitative observational data in a low-dimensional and more semantically meaningful qualitative space, as in this paper. Often an object-based abstraction of a video sequence is performed, then common arrangements of the abstracted entities are learned using various relations, e.g. common table place settings for a meal [27]; simple activities for daily living from a static camera dataset [28]; predicted object categorisation [29]; to remove inconsistent visual observations from noisy video sequences [30]; and even performing reasoning about spatio-temporal events being observed [31]. Each of these methods was performed in a supervised learning setting unlike our proposed unsupervised methods. Further, they are extensively used for *qualitative reasoning* tasks and applied to many real-world domains [32,33]; however, qualitative reasoning is out of scope of this work.

Qualitative spatial relationships can either be manually specified in advance or they can be discovered from observational data. The benefit of learning relationships automatically is that they are instantly relevant to the behaviour of the domain under observation. However, a common limitation is that often all (or some representative sample) of the data must be observed before any representation or learning can take place, unlike specifying relations in advance. Composite spatial-temporal relations have been learned between tracked regions representing moving objects in real life domains, for example vehicles on a stretch of motorway [30], or similar moving-point objects represented as trajectories [34]. Non-topological relations have also been learned by creating a relative feature vector using distance and the rate of change of distance between pairs of moving point objects [35]. These feature vectors are then clustered to obtain component atomic events which are used to describe human manufacturing-like activities from an egocentric vision set-up. This approach relies on a known and fixed set of objects where interactions are recognised between wrist worn marker IDs and the set of tagged objects. Each of these approaches rely on analysing the observed data in an offline process where relations are learned by taking the entire dataset of interactions between entities and learning suitable discretisations to best represent the data. In this work, our goal is to learn incrementally in a life-long learning setting, so qualitative representations are manually defined in advance. However, we recently co-authored literature that learns a qualitative representation incrementally, using natural language to guide the segmentation of various continuous feature spaces extracted from observations, whilst simultaneously using that representation to describe the observations [36,37].

An object-centric and qualitative abstraction process of observed video data partially alleviates problems associated with low-level image features that have been used with probabilistic learning approaches in previous works [38]. Other work directly compared STIP features with qualitative features on three challenging ego-centric vision datasets and demonstrated that qualitative representations can outperform traditional image features when object tracks are available [39]. That is, the qualitative representation can maintain semantically meaningful relational sequences and information specifically relating to movements of interest. Qualitative features arguably encode more "longer term temporal information". We couple this with the common bag-of-words representation (where word ordering is often lost using image features), the temporal overlap within our features maintains important structure in the observation, while offering the full benefit of sparse, discrete representations. This allows our approach to learn latent patterns of commonly observed qualitative features.

An unsupervised approach, coupled with a single qualitative spatial calculus has previously been used to encode continuous video sequences of aeroplane turnaround scenes [40,41]. Here the granularity of activity classes learned using an

unsupervised technique is restricted by the perception challenges related to the abstracted input scenes, as is the case in our work. As in our work, a single camera location is used, but here slow moving objects are observed which lack the variability of dynamic human movements. An egocentric camera is used to learn a similar qualitative representation of human body pose movements in [42]. However, both of these approaches learn activities in an offline and batch process, where the goal of our work is to use approximate variational methods to address practical considerations relevant to mobile robotics performing life-long learning of human activities.

3. Quantitative representation

The goal is to understand human activities taking place from long-term observation of a human populated environment by a mobile robot. In this section we describe the quantitative input data captured by the robot. This section is organised as follows: first we define what we consider as a human activity and the specific activity domains the robot is required to operate in; then we present details of how the robot encodes each human observation as a quantitative *human body pose* sequence. Finally, we describe how the robot interprets its environment and learns key object locations which provide some human functionality.

3.1. Human activities

We introduce the term *activity* to relate to a temporally dynamic configuration of some *agents*, where the agents can be grounded in the real-world, or could be online agents, etc. In this research we aim to i) understand human activities as patterns performed by humans in real environments, and ii) for the system to scale to allow continual learning. We focus only on single human activities. To do this we explore the interaction between the human agent and environment, namely between a human and key objects which provide functionalities [43]. We therefore define a *human activity* to be a temporally dynamic configuration of a human agent relative to close-by *key objects* in the environment. We make the following assumptions and definitions related to human activities:

- A *key object* is a semantic entity with a fixed location in an environment which provides some functionality that may be required for the execution of certain activities of interest in that environment [44].
- A *human activity* is considered as a partially ordered sequence of sub-activities (or repeated patterns) between positions of a person's body joints relative to key objects. In turn, these patterns (or sub-sequences) can be thought of as one or more simple qualitative relations holding between a person's body joints and/or a number of objects in the environment. For example, a person "picking up a cup" might comprise of the sequence: "reaching", "grasping" and "lifting" performed by the person's hand with respect to a cup.

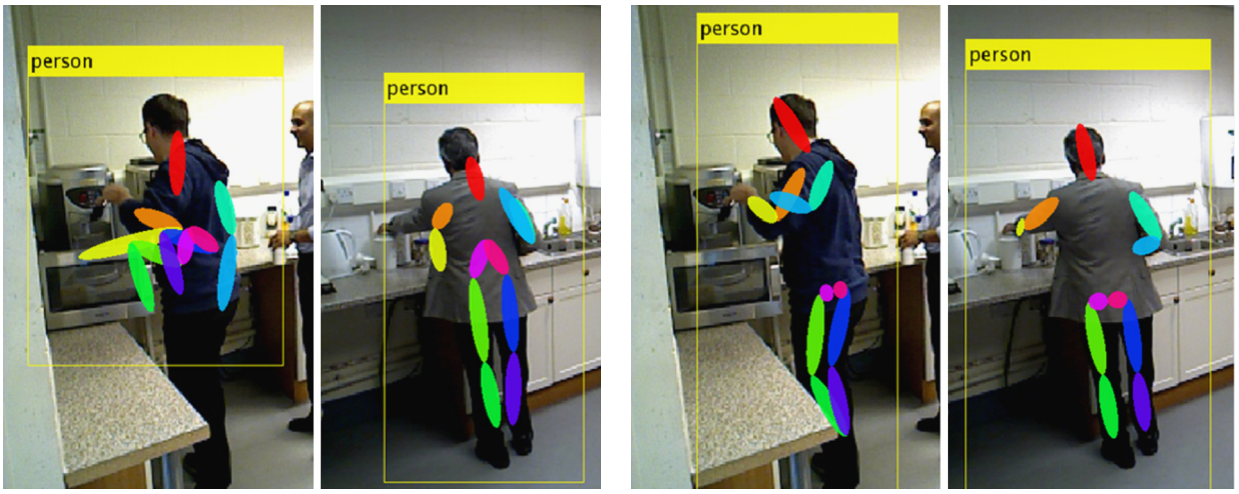
These assumptions are common in the literature [13,45,46]. This can also be considered under the framework of "Object-Action Complex" as introduced in [47].

A major challenge is the resolution of human activities that can be learned is somewhat limited by the available perception or sensory inputs. This paper provides a framework for a mobile robot, and therefore the perception is limited by its sensors and field of view capabilities. This is a key limitation to our system; since the performance of state-of-the-art robot perception is still far from human level perception. This affects the robot's ability to detect objects (static or moving) within its environment and only learn activity patterns at a particular level of granularity. Recent work in activity plan understanding has used detected hand movements and their contact points with objects in the environment [48,49] to learn from video data, or unconstrained video from the web [50]. However, these works rely on a much closer view point than afforded to our autonomous mobile robot, and often use pre-trained hand or object neural networks for classification.

3.2. Human pose estimates

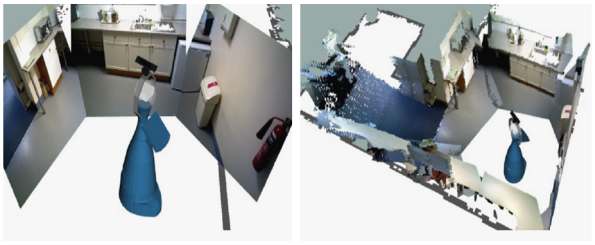
The mobile robot detects humans and infers their 3D pose (15 body joint locations) as they pass within the field of view of its RGBD sensor. A common approach is to use the OpenNI tracker [51] to detect multiple persons and infer their 3D pose in real-time from the sensors' depth stream. It is especially important to obtain reliable pose estimates in cases of human-object interaction from difficult viewpoints. Unfortunately, these interactions cause most pose estimation errors from OpenNI, where the object is inadvertently considered part of the person/foreground and/or the person is backward facing during an observation, see Fig. 1(a). To mitigate this problem, we leverage RGB colour data to help distinguish between object and person and resolve backward facing poses. Our pose estimation system operates in a two phase approach, firstly, the efficiency of OpenNI is utilized to produce person bounding boxes per frame. Secondly, person bounding boxes and the RGB frame are fed as input into a state-of-the-art convolutional network (ConvNet) 2D human pose estimator [52]. Subsequently, the (x, y) coordinates of OpenNI body joint positions are replaced with the superior 2D body joint coordinates provided by the ConvNet, see Fig. 1(b).

We represent the human pose estimates as ROS messages, where a single detected body joint location is represented by 3D Cartesian coordinates in a camera frame of reference along with the corresponding position transformed into the global

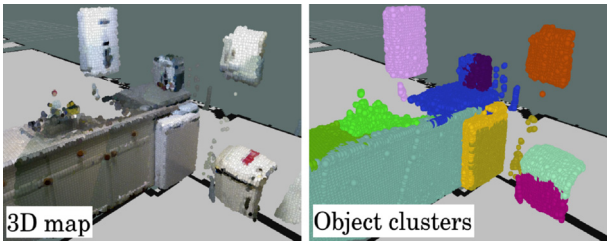


(a) Inaccurate OpenNI pose estimates (b) Improved pose estimates

Fig. 1. Improved human pose estimates. Image best viewed in colour.



(a) Generating a 3D representation of the environment. (left:) The robot moves its pan-tilt multiple times capturing an RGB image and a point cloud at each angle. (right:) The robot fuses together the registered point clouds to create a single 3D representation.



(b) Segmenting candidate object locations. (left:) Surfel representation of the robot's 3D environment. (right:) Candidate object locations automatically segmented as clusters.

Fig. 2. Quantitative representation of robot environment. Image best viewed in colour.

map frame of reference using the estimated location of the robot, i.e. $j = (id, x, y, z, x_{map}, y_{map}, z_{map})$. A *human pose* then comprises of a collection of body joint locations, i.e. $p = [j_1, j_2, \ldots, j_{15}]$, using the OpenNI/ConvNet implementation. For each human detected by the robot, we obtain a sequence of human poses over a time series of detections (camera frames). We define a *human pose sequence*, $S = [p_1, p_2, \ldots, p_i, \ldots]$, where each p_i is the detected human pose at timepoint i, and no restrictions are placed upon the length of the recorded sequences. This variation in length is a major difficulty when using real-world data to learn activities on a mobile robot.

3.3. Object representation

A second key component in the robot's environment are *objects*. In this work we focus on objects which people interact with in daily living and which provide some functionality for human activities. For example, a person might walk up to a printer-copier machine, stop in front of the machine in order to perform an action (swipe a key card to log in) and whilst doing so their body joints spatially interact with the object. For this reason our representation of human activity includes relative positions of people with respect to key objects within the robot's environment. However, detecting and tracking arbitrary objects in real time from a robotic platform is a difficult and unsolved problem. Therefore to learn the position of interesting objects within an environment, the robot first pre-builds a 3D model of its environment by fusing together multiple RGBD images. The process can be seen in Fig. 2a: 1) the robot moves its pan-tilt multiple times capturing an RGB image and a corresponding depth point cloud for each angular position and registers each pixel in the depth point cloud with an RGB value from the corresponding RGB image. This process is known as a *sweep*; 2) multiple sweeps are performed to create a large registered point cloud representation of the robot's entire environment (covered by the sweeps).

Once the robot has its 3D point cloud representation of its environment, it extracts locations of potential objects by rendering the surface using *surfels* (surface elements) [53] and extracting clusters. State-of-the-art performance in extracting semantically meaningful segments can be achieved using an unsupervised segmentation algorithm which is grounded in the convexity of common human objects. This is demonstrated in [54], and we use a method similar to that presented in [55]. Locating an object in 3D scenes is a challenging and well studied topic; as such we do not consider

Fig. 3. An example human body pose observation relative to the environment. (Left) RGB image corresponding to a single human body pose detection. (Right) The human body pose estimate is translated into the map coordinate frame using the localised position of the robot and overlaid as a person model where the two hand joint locations are shown as pink cubes. Also overlaid are the segmented key object clusters. Image best viewed in colour. (For interpretation of the colours, the reader is referred to the web version of this article.)

it a contribution of this work. An example of the surfel representation of the 3D environment can be seen in Fig. 2b (left) and the resulting segments (right). We consider these segments as candidate key objects in the robot's environment. An example human observation represented in the map frame of reference, plus segmented key objects, can be seen in Fig. 3.

4. Qualitative representation

Abstracting human pose sequences into a qualitative spatial representation (QSR) allows the robot to learn common and repeated patterns being performed over multiple observations, even if they vary quantitatively in their execution. For example, if a person raises their hand above their head and waves, the exact (x, y, z) coordinates of their hand or head are not important; it is the relative movement which captures the possible "waving" activity. A challenge when learning human activities is that they often occur over very different durations of time, e.g. opening a fridge vs standing still, and some activities will have accurate pose estimates whereas others may be noisier, e.g. due to occlusions or fast paced movements. These variations provide a major difficulty which abstracting the observed data into a qualitative space helps to alleviate. In this section we present the qualitative representations used, and the auto-generated codebook of qualitative features (descriptors) that results in a term-document representation. This is ideal to formulate the human activity analysis as an information retrieval problem.

In this manuscript, we use three qualitative calculi to abstract observation instances into a qualitative space. Two of these calculi require no manual tuning of parameters, they are: 1) Ternary Point Configuration Calculus (TPCC) [56] which qualitatively describes the spatial arrangement of an entity, relative to two others, i.e. it describes the *referent*'s position relative to the *relatum* and *origin* and possible values are triples of ⟨ {front/back}, {left/right/straight}, {distant/close} ⟩; 2) Qualitative Trajectory Calculus (QTC) [57] represents the relative motion of two points with respect to the reference line connecting them, and is computed over consecutive timepoints. It defines the following three qualitative spatial relations between two entities o_1, o_2: o_1 is moving towards o_2 (represented by the symbol $-$), o_1 is moving away from o_2 ($+$), and o_1 is neither moving towards or away from o_2 (0). The third calculus, the Qualitative Distance Calculus (QDC) [58] expresses the qualitative Euclidean distance between two points depending on defined distance thresholds, Δ and does rely on parameterised thresholds. The intuition is based on the assumption that human motion can be partially explained using distance relative to key landmarks. A set of QDC relations localises a person with respect to reference landmarks, and a change in the relations can help explain relative motion. Although QDC relies on pre-defined thresholds, we perform a detailed sensitivity analysis where various parameter values are explored.

A simplified diagram of each of the three calculi can be seen in Fig. 4. They are computed from observed (x, y, z) data over a series of timepoints (one per camera frame), i.e. a quantitative human pose sequence is abstracted into multiple sequences of qualitative relations (one per calculi used) using a publicly available ROS library we co-developed [59,60]. Each representation captures semantic information to describe human movements qualitatively, however, it is not exhaustive and other qualitative calculi could be explored.

4.1. Interval representation

One hypothesis in this work is that many human activities can be explained by a sequence of primitive actions over some duration of time. In order to learn these sequences, independent of the exact time or duration, the spatially abstracted data is also *temporally* abstracted using a temporal calculus. The QSR relations computed in the previous section represent observed qualitative relations holding between entities, one collection of relations for each pose or timepoint of an observed sequence (one per set of entities, per calculi). We consider these as a time series of observational data and compress

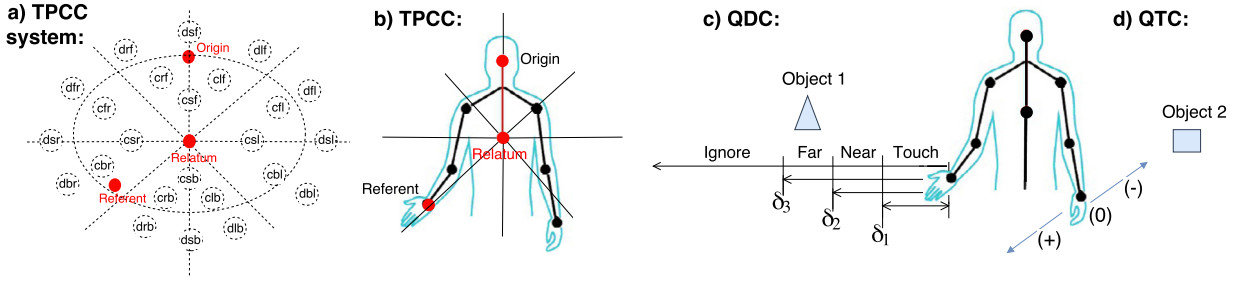


Fig. 4. (**a**) TPCC. Relations are triplets of the letters ⟨ {f/b}, {l/r/s}, {d/c} ⟩ which represent: front, back, left, right, straight, distant, close, respectively. The system represents the qualitative location of the *referent* relative to the *origin* and *relatum* (all shown as red points in 2D). Here, the referent is considered close, back and right or "cbr". (**b**) The TPCC system applied to a human body pose. The location of the *right hand* body joint is described with respect to the origin and relatum which are fixed to the *head* and *torso* body joints respectively. The hand's location is represented as "cbr". (**c**) An illustration of QDC (relative distance) applied to a human body pose between the *right hand* and *Object* 1 (triangle), computed using three boundaries: $\Delta = [\delta_1, \delta_2, \delta_3]$. (**d**) An illustration of QTC (relative motion) applied to a human body pose between *left hand* and *Object* 2 (square). Blue arrows denotes motion towards $(-)$, or away from $(+)$ the static object.

repeated QSR relations which hold between entities across consecutive timepoints, i.e. when a relation is stable for some period of time. The resulting encoding is an *interval representation*, which can be described as a temporally connected set of semantic intervals, each maintaining a duration of time over which a qualitative relation holds between entities. This representation is an abstraction of a QSTAG (qualitative spatial-temporal activity graph) first introduced in [1], and is closely related to an intermediate representation the authors developed in [59–61]. In parallel, researchers have considered "Semantic Scene Graphs", where discontinuous transitions of spatial relations are considered *moments* in a semantic event chain [62,63].

Consider the position of a body joint pose sequence encoded as QTC relations relative to a single object in the environment, e.g. a hand relative to a fridge for a duration of time. If the body joint appears to be moving towards the object, o_1, (QTC relation: '$-$'), for some consecutive number of frames τ, and then is static (0) with respect to that object for τ' further frames, we can compress the sequence into an interval representation consisting of two intervals: $I = \{\iota_1, \iota_2\}$ where each interval can be represented as a tuple consisting of a relation followed by a duration, i.e. $\iota_1 = (\text{`}-\text{'}, [0, \tau - 1])$, and $\iota_2 = (\text{`0'}, [\tau, \tau + \tau' - 1])$. Each interval $\iota \in I$ maintains a QSR value that holds between entities and the start and end timepoints of the interval during which the relation held. The interval representation of this example is shown as the top row of Fig. 5a; however, an interval representation of a complete human pose sequence contains a row for each body joint (or pairwise joints with each key object).

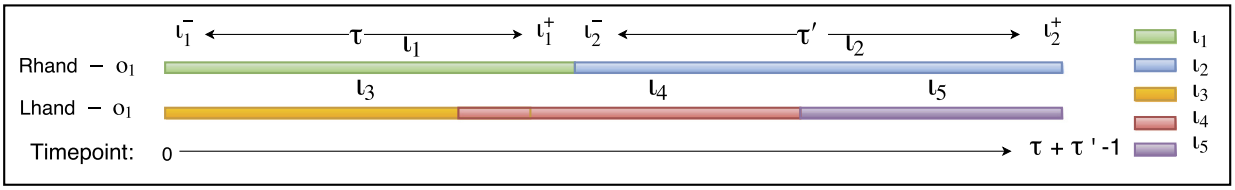
We introduce the terminology *lesser end point* as the start timepoint of an interval, ι, and represent it by ι^-, and similarly, the *greater end point* as the end point of the interval and represent it by ι^+, where $\iota^- < \iota^+$. To perform a temporal abstraction, the robot uses Allen's Interval Algebra (IA) [64] to abstract the exact temporal aspects of observed QSR sequences as encoded in an interval representation. This allows the robot to compare multiple observations irrespective of their exact temporal durations. IA defines 13 possible qualitative temporal base relations between any pair of intervals ι_1 and ι_2. Fig. 5b illustrates the 7 temporal interpretations of possible IA relations (based upon discrete time) using interval endpoints values (ι_1^-, ι_1^+) and (ι_2^-, ι_2^+) to define the temporal relations between intervals ι_1 and ι_2.

4.2. Interval graph

To extract descriptors from a human observation, each interval representation is *temporally* abstracted into an *interval graph* [65] using IA relations that hold between pairwise intervals, and then decomposed into *graph paths*. An example can be seen in Fig. 6 (left), which encodes both rows present in Fig. 5a. Formally, we say an interval graph $G = (V(G), E(G))$ comprising of nodes V and arcs E. Here, a node is used to represent an interval and contains only the QSR value (or set of values if using multiple qualitative calculi) that hold between entities, and the entities themselves. The exact timepoints are not explicitly depicted in the node, e.g. node ι_1' in Fig. 6 (left) contains $[Rhand, O_1, \text{`}-\text{'}]$ information temporally abstracted from interval ι_1. Nodes in the interval graph are linked by directed arcs if their intervals are temporally connected, i.e. there exists no temporal break between a pair of intervals. Directed arcs are labelled with the IA relation that holds.

4.2.1. Arc restrictions

By linking only *temporally connected* nodes with arcs means that there are no *before* or *after* relational arcs; these relations are superfluous within the graph as the IA composition table means these can be inferred. A second design characteristic to note, where two intervals occur at the beginning or end of the video clip (and therefore beginning or end of the interval representation), there is insufficient temporal information to infer the IA relation and therefore no arc is encoded between these nodes in the graph. For example in Fig. 6(left) there is no arc between nodes i_1' and i_3', since their corresponding intervals i_1 and i_3 in the interval representation occur at the start of the observation.



(a) Interval representation of five intervals $I = \{\iota_1, \iota_2, \ldots, \iota_5\}$, between two pairs of entities. The intervals maintain the QSR relation that holds between a human body joint and a single object, o_1. Image best viewed in colour.

IA Relation	Equivalent endpoints definition
ι_1 before ι_2	$\iota_1^+ < \iota_2^- - 1$
ι_1 equals ι_2	$(\iota_1^- = \iota_2^-)$ & $(\iota_1^+ = \iota_2^+)$
ι_1 meets ι_2	$\iota_1^+ = \iota_2^- - 1$
ι_1 overlaps ι_2	$(\iota_1^- < \iota_2^-)$ & $(\iota_1^+ > \iota_2^-)$ & $(\iota_1^+ < \iota_2^+)$
ι_1 during ι_2	$((\iota_1^- > \iota_2^-)$ & $(\iota_1^+ \leq \iota_2^+))$ or $((\iota_1^- \geq \iota_2^-)$ & $(\iota_1^+ < \iota_2^+))$
ι_1 starts ι_2	$(\iota_1^- = \iota_2^-)$ & $(\iota_1^+ \neq \iota_2^+)$
ι_1 finishes ι_2	$(\iota_1^+ = \iota_2^+)$ & $(\iota_1^- \neq \iota_2^-)$

(b) Allen's Interval Algebra [64] defined on endpoint intervals, (ι_1^-, ι_1^+) and (ι_2^-, ι_2^+), assuming discrete time.

Fig. 5. Temporal abstractions applied to sequences of qualitative spatial relations.

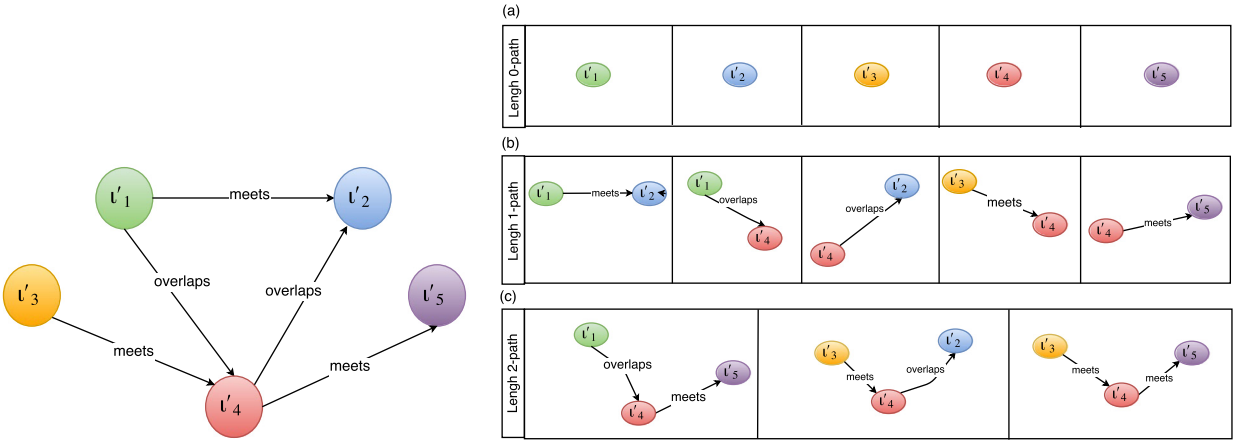


Fig. 6. (Left) Interval graph, G, encoded from interval representation I in Fig. 5a. A node ι_i' represents a temporally abstracted interval $\iota_i \in I$. Directed arcs represent the IA relation that holds between temporally connected, pairwise intervals. No arc is computed between two intervals if both are in the starting or both in the ending interval sets, I^- or I^+. (Right) All graph paths extracted from the interval graph using $\eta = 2$ and $\rho = 2$ parameters. Image best viewed in colour.

This restriction is formalised by defining a minimum and maximum timepoint of an interval representation I, i.e. $\iota^{--} = min(\{\iota^- \mid \forall \iota \in I\})$ and $\iota^{++} = max(\{\iota^+ \mid \forall \iota \in I\})$ respectively. Further, two sets of intervals known as the *starting set*, I^-, and *ending set*, I^+ are defined. An interval is a member of either set if it shares an interval endpoint with the observation's start or end timepoints ι^{--} or ι^{++} respectively, i.e. $I^- = \{j \mid \forall j \in I, j^- = \iota^{--}\}$, and $I^+ = \{j \mid \forall j \in I, j^+ = \iota^{++}\}$. No IA relation is computed between two intervals if both are in the starting or both in the ending interval sets, I^- or I^+. For example, intervals ι_1 and ι_3 in Fig. 5a both occur at the start of the interval representation, i.e. $\iota_1^- = \iota_3^- = \iota^{--}$, therefore both are in the starting intervals set, $I^- = \{\iota_1, \iota_3\}$, and hence no IA relation is computed, resulting in no arc.

4.2.2. Qualitative descriptors

A *path* of length n is defined as a set of n nodes that forms a path through an interval graph G by following directed arcs in its edge set, $E(G)$. The nodes and arcs in a path form a sub-graph of the original interval graph and is defined as a *graph path*, i.e. a graph path w is defined as: $w = \{\iota_1', \iota_2', \ldots, \iota_n'\}$ s.t. $\{\iota_i', \iota_{i+1}'\} \in E(G)$, $\forall\ 1 \leq i < n$. All graph paths up to some fixed length η (≥ 0) are evaluated. Recall each node ι' in an interval graph represents the set of entities (objects) and the qualitative spatial relations that holds between them over an interval of time. To help reduce computation, the number of graph paths are often also limited to a maximum number, ρ, of encoded entities or objects pairs. However, combinations of entities overlap between different rows in the interval representation and so implicitly the graph paths maintain information between all entities.

This process results in a collection (or bag) of potentially overlapping graph paths used to represent the interval graph, e.g. $D = \{w^1, w^2, \ldots, w^{N_D}\}$. We define D as a bag-of-words, where our novel overlapping graph-paths are considered as words. In the text analysis literature, word positional arrangement is ignored; however our graph paths represent partially

overlapping sequences of temporally connected QSR intervals and therefore maintain local temporal structure within the representation. An example bag-of-words is shown in Fig. 6 (right), where all paths through the interval graph are illustrated using $\eta = 2$ and $\rho = 2$.

4.2.3. Code book

We represent an interval graph as a bag-of-words of constituent graph paths. The terms *graph path* and *qualitative descriptor* are used interchangeably throughout. The set of unique graph paths in bag D is given by: $V_D = \bigcup_{w^i \in D} w^i$. We extend this to a collection of M human observations, where each observation is encoded as an interval graph G, i.e. $\mathcal{G} = [G_1, G_2 \ldots, G_M]$, and similarly a bag, i.e. $\mathcal{D} = \{D_1, D_2 \ldots, D_M\}$. Unique graph paths from all bags are automatically generated and form a *code book* $\mathcal{V}_\mathcal{D}$. Formally, we say: $\mathcal{V}_\mathcal{D} = \bigcup_{D \in \mathcal{D}} V_D$ where each graph path $w_i \in \mathcal{V}_\mathcal{D}$ is unique and defined as a *codeword*, i.e. $\mathcal{V}_\mathcal{D} = \{w_1, w_2, \ldots, w_N\}$. Note that we use subscripts to denote the index of a codeword, w_i, in the code book and superscripts, w^j, to indicate an observed graph path in a bag-of-words.

The code book represents the unique set of graph paths extracted from all the observed interval graphs where the total number, N, is not known or fixed in advance. In Section 5 we introduce techniques to dynamically update the code book given incremental observations, i.e. N can increase on observing new codewords. Note, the code book specifically depends upon the qualitative descriptors extracted from the encoded interval graphs, i.e. altering the encoded observations, or changing the graph path parameters may result in a different code book and a different representation. To efficiently determine whether two graph paths are identical, a distance based graph kernel is used to approximately represent each as a 32 bit hash value, described in [66]. This reduces the graph matching problem to efficient integer comparisons. This graph matching technique has been implemented in the open-source software library, QSRLib [59,60].

4.2.4. Codeword histograms

Given a code book $\mathcal{V}_\mathcal{D}$ of length N obtained from a collection of M human observations, each observation is encoded into a bag-of-words and represented as an N-dimensional feature vector that describes the frequency of each codeword in the bag. We refer to this feature vector as a *codeword histogram*, $\mathbf{h}$, which is defined over a specific code book $\mathcal{V}_\mathcal{D}$. In this case, the codeword histogram is considered a sparse feature vector representation since it may contain many zero codeword counts. The codeword histograms are each N-dimensional vectors and therefore once vertically stacked produce a *term–frequency matrix* C representing an entire corpus of M human observations as an $M \times N$ matrix, e.g. $C = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]$.

5. Unsupervised learning for human activities

Encoding a corpus of human observations into such a term-frequency matrix allows latent structure can be recovered in an unsupervised setting. The aim is to learn low-dimensional representations of repeated structure encoded as qualitative descriptors (graph paths) across multiple similar observations. To do this, information retrieval techniques are used. We focus on Latent Semantic Analysis (LSA) [17] and a more sophisticated, probabilistic method, Latent Dirichlet Allocation [6]. Both were developed for understanding large corpora of encoded text documents and used to recover distributions of latent topics or themes present in data. In this section we first introduce both methods and how each is applied to the encoded term-frequency matrix. Secondly, we introduce, and propose solutions to, some often-ignored practical considerations of autonomous mobile robots, namely, i) the unavailability of temporal segmentation applied to video sequences and ii) the challenges of life-long or incremental learning.

5.1. Low rank approximations for human activities

The aim is to learn a low-dimensional representation of an encoded term-frequency matrix by finding redundancy within the set of qualitative descriptors observed. The most discriminative descriptors are those that contain the most variation. The assumption is that by reducing the dimensionality of the matrix, but maintaining as much variance within the columns as possible, it is possible to represent the corpus of observations with a relatively small number of human activity classes. The process is performed using Latent Semantic Analysis (LSA) which computes linear combinations of columns to create new composite features containing as much variation as possible. Sorting the new features by their ability to discriminate the observations, the most redundant are removed to leave a low-dimensional representation and latent classes encoded in the data are recovered.

Given a term-frequency matrix C, LSA comprises two stages: First, compute and apply a *term frequency–inverse document frequency* (tf–idf) weighting to each column based upon its variation in the training samples, with the assumption that the most descriptive columns have the largest variation. The weighting increases proportionally to the number of times a codeword appears in a codeword histogram and is inversely proportional to the frequency of the codeword in the entire corpus. That is, it is a measure of how much information observing a codeword provides. Secondly, to find a lower dimensional representation of a matrix we compute a *low-rank approximation*. We do this by finding a second matrix C_r, of rank r, and requiring it to be as similar as possible to the original matrix based on the *Frobenius norm*. That is, to miminise the Frobenius norm of the matrix difference $X = C - C_r$, defined to be: $||X||_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (X_{ij})^2}$, Singular Value Decomposition

(SVD) is used. This process factorises C as $U_{(M\times M)}\Sigma_{(M\times N)}V^T_{(N\times N)}$, where U and V are orthogonal matrices comprising of the singular vectors of C, whilst Σ is a non-increasing diagonal matrix containing the squared eigenvalues of C.

The aim is to recover a small number of latent concepts from the encoded data. The assumption is that common human activities relate to repeated patterns of discriminative qualitative descriptors encoded within the observations. Examining the decomposition, the non-zero eigenvalues in the diagonal matrix Σ represent the r most discriminative new compositional features, known as *concepts*. These latent concepts can be thought of as the activity classes encoded in the original term-frequency matrix. The columns of the left singular $(M \times M)$ matrix U contain the eigenvectors of CC^T, since $CC^T = U\Sigma\Sigma^T U^T$, and provides information, as a linear combination, about the weighting of each concept to each observation, specifying its latent activity class (concept). The columns of the right singular $(N \times N)$ matrix V contain the eigenvectors of C^TC, since $C^TC = V\Sigma^T\Sigma V^T$, and specify a linear combination of weights for each qualitative descriptor (codeword) used to describe each latent concept.

5.1.1. Limitations

In Section 6.3 we show that LSA provides a relatively good method to recover discriminative latent concepts in an unsupervised setting that are embedded in a term-frequency matrix; along with a code book of descriptors used to describe them. However, there are limitations to this non-probabilistic technique. Given the matrix decomposition, i.e. the left/right singular matrices describe the linear combinations of observations to concepts U, and codewords to concepts V; one limitation is that both U and V are orthogonal matrices. The implication of the orthogonal matrices is that any concepts extracted cannot share columns, e.g. a specific codeword cannot be significant in two separate concepts.

A second limitation is that LSA is a batch learning algorithm, which requires the entire term-frequency matrix C to be encoded before the training process occurs. New observations can be represented by their similarity to already learned concepts, but they cannot contribute to the model and affect the concepts, unless the SVD decomposition is re-performed, which is inefficient for a life-long learning setting. Finally, selecting the most appropriate number of eigenvalues (i.e. rank) to best represent the low-rank approximate matrix C_r is often challenging. One technique for selecting a *good* value of r is to plot the variation of each eigenvalue, in a non-increasing *scree plot* that ideally shows a steep curve followed by a bend, often called the "elbow point", followed by a more flat line indicating any further features add little variance. This technique allows a good value of r to be ascertained, however, the exact number can often depend upon the task. Solutions to each of these limitations are proposed in the following section by using a generative probabilistic model.

5.2. Probabilistic topic distributions for human activities

One intuition is that an observation of a human should be modelled in such a way that allows for multiple, overlapping classes of activity to occur and for the activity classes themselves to overlap somewhat and share certain descriptors. For this reason we introduce Latent Dirichlet Allocation (LDA) which is commonly referred to as Topic Modelling. The key idea is two fold: a *topic* is defined as a multinomial distribution over a vocabulary of codewords (code book) and describes a particular thematic structure present in the corpus; a *document* (codeword histogram) is represented as a probabilistic mixture over topics, by inferring a *proportions* or *mixing* vector. The assumption is that similar documents use similar groups of co-occurring codewords, and therefore the co-occurrence can be used to identify the latent thematic topics. This framework allows for each observation of a human to be modelled as a mixture of activity classes occurring, and to simultaneously recover the latent activity classes as distributions over the code book. In Section 6.3 we show that LDA can learn low-dimensional representations for each human observation and for the activity classes themselves, with substantial benefits during the training phase when compared with the non-probabilistic LSA method described. First we describe how LDA is applied to observations.

5.2.1. Generative LDA model of human activity

Probabilistic generative models are based on probabilistic sampling and can be interpreted as a model of how the observed data was generated from a set of underlying latent variables. In this case, the collection of observations are assumed to be generated from latent topic distributions, topic assignments and mixing vectors, where the aim is to learn the best fit of these latent variables (assuming that the model generated the data). Fig. 7 shows the intuition behind the LDA generative process for a single human observation. For each observed bag-of-words D (or codeword histogram), the underlying process can be characterised as follows:

1. sample a per-document topic proportions vector, $\boldsymbol{\theta}_D$, from a prior Dirichlet distribution parameterised by α, i.e. $\boldsymbol{\theta}_D \sim Dir(\alpha)$. An example sample drawn over three topics is shown far right as a cartoon histogram in Fig. 7. This represents a multinomial distribution drawn from the Dirichlet simplex. These are the mixing proportions for the codeword histogram.
2. For each of the N_D codewords in the bag-of-words D:
   - draw a per-word topic assignment, $z_{D,n}$, from the proportions vector, i.e. sample an assignment coin $z_{D,n} \sim Multinomial(\boldsymbol{\theta}_D)$. For example, the pink topic in Fig. 7 is sampled first and shown with a pink coin, followed by yellow, then pink again etc. This allows each codeword in the bag to be drawn from different topic distributions, respecting the topic mixing proportions in $\boldsymbol{\theta}_D$ and facilitating the mixing of topics within an observation.

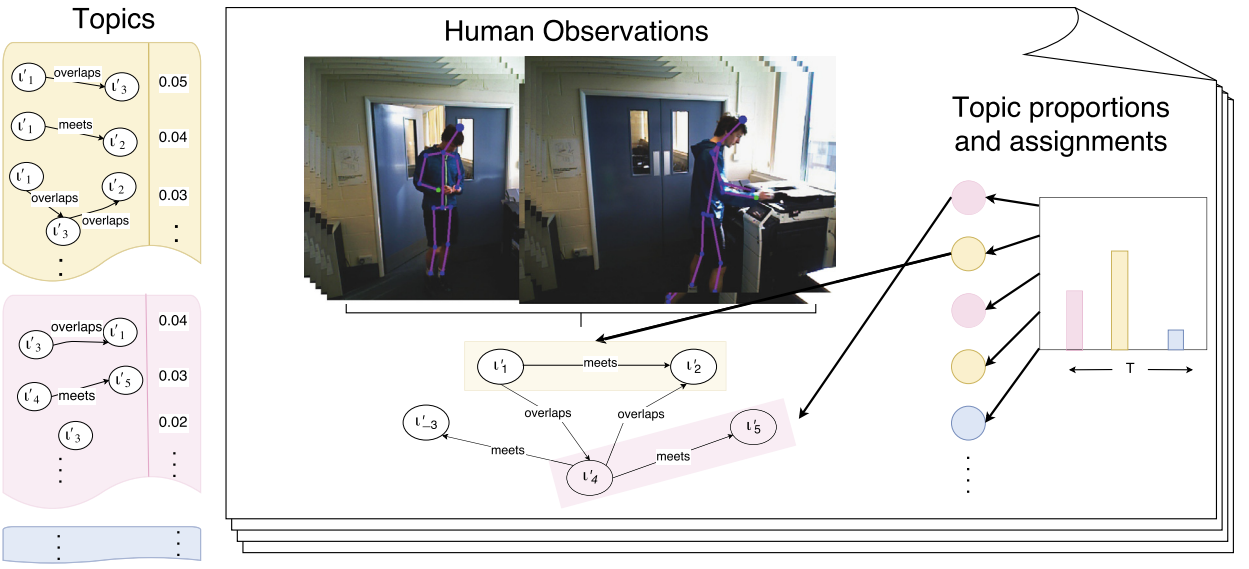


Fig. 7. Generative LDA model of human activity. (Left) Three topic distributions over the code book (yellow, pink and blue) along with three top-probable codewords in the yellow and pink topics. (Centre) Generated interval graph and bag-of-words obtained from encoding one human observation. (Right) Topic proportions vector (pink, yellow, blue histogram) and codeword assignments as a column of sample coins drawn (pink coin, yellow, pink, etc.). Image best viewed in colour. (For interpretation of the colours, the reader is referred to the web version of this article.)

- for each topic assignment, draw a word, $w_{D,n}$, from the multinomial topic distribution conditioned on the topic assignment $z_{D,n}$, i.e. sample a codeword $w_{D,n} \sim Multinomial(\boldsymbol{\phi}_{z_{D,n}})$, where each $\boldsymbol{\phi}_i$ represents a topic distribution over the code book, also drawn from a Dirichlet simplex parameterised by β, i.e. $\boldsymbol{\phi}_i \sim Dir(\beta)$. For example, from the pink topic assignment (coin) the codeword (ι'_4 *meets* ι'_5) is drawn, which can be seen far left in Fig. 7 as a highly probable codeword in the pink topic. Then, the codeword (ι'_1 *meets* ι'_2) is drawn from the yellow topic assignment, etc. This process generates the bags-of-words with mixing proportions $\boldsymbol{\theta}_D$.

3. This process repeats to generate M bags-of-words.

5.2.2. LDA as a graphical model

In reality, the robot only observes the bags-of-words and not the mixing proportions vectors or the assignment of each codeword into topics. This is the latent structure (variables) of the model that we aim to infer, i.e. p (topic distributions, proportions, assignments | codeword histograms). Given a collection of codeword histograms, our task corresponds to inferring the three sets of latent variables:

- $\Phi = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T]$ per-corpus topic distributions, where each $\boldsymbol{\phi}_i$ is a distribution over the code book $\mathcal{V}_{\mathcal{D}}$;
- $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M]$ per-document topic proportions vectors, where each $\boldsymbol{\theta}_D$ is a distribution over T topics;
- $\mathbf{z}$ is the assignment of all observed codeword tokens to topics, for all observations.

This is equivalent to inferring $p(\Phi, \Theta, \mathbf{z}|\mathcal{D})$. To do this, we present LDA as a Bayesian Network or Directed Acyclic Graph (DAG). This is an intuitive way of representing and visualising the relationships that exist between the variables that make up a topic model and corresponds to a specific factorization of the joint probability distribution (JPD).

In a DAG, nodes represent random variables and directed edges between nodes reflect conditional dependencies between variables. A common presentation technique is to depict observed random variables using shaded nodes, and non-shaded nodes for latent variables. Plate notation is used to highlight replicated random variables, with the number of random variables marked in the lower right corner of the plate. The DAG representing the LDA model for human activities is shown in Fig. 8. It is considered a three-layer Bayesian model since: i) the Dirichlet hyperparameters, α and β, are corpus-level parameters; ii) the topic proportions variables, $\boldsymbol{\theta}_D$ for each bag-of-words D in the corpus, $\mathcal{D}$, are codeword histogram-level parameters, sampled once per codeword histogram; iii) the variables $z_{D,n}$ and $w_{D,n}$ are codeword-level and sampled once for each codeword in a bag-of-words.

Formally, for M observed bags-of-words, each containing N_D codewords, and a set of T topic distributions, the joint probability distribution over the observed and latent variables within the DAG is defined as:

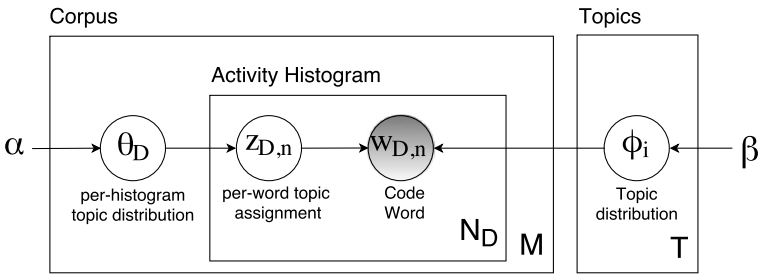


Fig. 8. DAG representation of LDA. Nodes represent random variables, links between nodes are conditional dependencies, plates are replicated components, and shaded nodes are observed random variables. For bag-of-words D: $\boldsymbol{\theta}_D$ (topic proportions vector for bag D); $w_{D,n}$ (nth observed codeword, shaded grey, in bag D); $z_{D,n}$ (nth codeword topic assignment in bag D); $\boldsymbol{\phi}_i$ (ith topic distribution over code book); α, β (Dirichlet hyperparameters).

$$p(\boldsymbol{\phi}_{1:T}, \boldsymbol{\theta}_{1:M}, \mathbf{z}, D_{1:M}|\alpha, \beta) = \prod_{i=1}^{T} p(\boldsymbol{\phi}_i|\beta) \prod_{D=1}^{M} p(\boldsymbol{\theta}_D|\alpha) \left(\prod_{n=1}^{N_D} p(z_{D,n}|\boldsymbol{\theta}_D) p(w_{D,n}|\boldsymbol{\phi}_{1:T}, z_{D,n}) \right). \tag{1}$$

The three representations (i) the intuitive generative process, ii) the Bayesian DAG and iii) the JPD) are equivalent ways of describing the probabilistic assumptions behind the LDA model. The robot uses posterior expectations to perform inference and learn the latent structure of the LDA model, described next.

5.3. Approximate inference

Given the joint probability distribution in Equation (1), we can answer possible inference queries by marginalization, i.e. summing out over irrelevant variables. Given M observed bags-of-words $D_{1:M}$, as encoded as codeword histograms in a term-frequency matrix, inference allows us to estimate the latent variables, i.e. the topic distributions and the mixing proportions vectors that best fit the observations. This can be considered as finding latent patterns in the data which best separate it into meaningful topics or concepts; its thematic structure. This translates as computing the posterior distribution of the latent variables given a collection of bags-of-words:

$$p(\boldsymbol{\phi}_{1:T}, \boldsymbol{\theta}_{1:M}, \mathbf{z}_{1:M}|D_{1:M}, \alpha, \beta) = \frac{p(\boldsymbol{\phi}_{1:T}, \boldsymbol{\theta}_{1:M}, \mathbf{z}_{1:M}, D_{1:M}|\alpha, \beta)}{p(D_{1:M}|\alpha, \beta)}. \tag{2}$$

This posterior distribution is intractable to compute in general. More details about how to marginalise over variables can be found in [6] and [67]. We use Collapsed Gibbs Sampling [68,69] as an approximate inference technique that is based upon Markov chain Monte Carlo (MCMC) sampling. The key idea is to generate posterior samples from the conditional distribution by iterative sampling using a Markov Chain [70], and is described in detail in [68].

5.3.1. Variational inference for incremental learning

In order to operate in real-world environments however, it is not efficient to repeatedly perform batch learning on an ever increasing corpus of recorded video sequences, e.g. using techniques such as LSA or Collapsed Gibbs Sampling (standard LDA inference). Ideally, an incremental learning method could update its learned distributions based upon only new observations and does not require re-computing for previously analysed data. Therefore, we propose to use Variational Bayes (VB) approximate inference which aims to optimise a simplified, parametric distribution in order to fit the LDA model posterior using mini batches of new observations. This allows the robot to continually learn about human activities based upon incrementally updating the LDA posterior. This method was first used with LDA to analyse massive corpora containing millions of natural language text documents where batch algorithms were prohibitively computationally expensive [71].

The basic idea of variational inference is to formulate the computation of a marginal or conditional probability in terms of an optimization problem. This, generally intractable problem, is then “relaxed”, yielding a simplified optimization problem that depends on a number of free parameters, known as variational parameters. Solving for the variational parameters gives an approximation to the conditional probabilities of interest, in our case, the conditional distribution that defines the LDA posterior. The method we propose to use here is Variational Bayes inference which optimises a simplified parametric distribution based upon the Kullback–Leibler divergence to the posterior [71,72]. It has been shown to converge faster and be as accurate as MCMC sampling methods [73], and therefore it is ideal as a practical solution for a life-long learning setting where the number of observations is unknown and could become computationally intractable for batch methods.

In practice, we incrementally update the topic distribution estimates that represent activity classes of human behaviour. For a new observation the process of updating the topic model is threefold:

1. any new codewords in the observations are first appended to the current code book $\mathcal{V}_\mathcal{D}$ and to the topic distributions Φ with zero probability;

2. a multinomial distribution over the current set of topics/activities for the new observation is computed, $\boldsymbol{\theta}$, that represents the mixture of topics observed;
3. finally, the topic distributions over the vocabulary, Φ, are updated using this new observation, or mini-batch of observations.

This allows the robot to efficiently update its model of human activities using a single pass over new observations, optimising both storage and computation complexity. Each observation can therefore be maintained as a low-dimensional distribution over the set of topics considered human activities. We present an experiment to demonstrate the efficacy of this approach compared to Gibbs Sampling in Section 7.2.

5.4. Probabilistic mixture of activities

One further, often overlooked, practical consideration when using visual data recorded from a real-world mobile robot, is that any video sequence is likely to contain multiple overlapping human activity instances or incomplete activities performed within the robot's narrow field of view. That is, each human observation will not be a single, temporally segmented instance of an activity class; unless manual segmentation of the sequences into clips is performed. Further, without such temporal segmentation, many observations may contain humans performing no interesting (or repeated) activities, e.g. a common occurrence is a detected human simply walking straight past the robot.

Given the many sequences of images recorded by the robot, it is not ideal or always possible for a human to manually segment out "interesting" sub-sequences in order to learn a representation of only these activities. Therefore, we propose LDA applied to our encoded qualitative representations in order to handle these challenging longer sequences. We demonstrate with the use of two experiments the hypothesis that LDA models each observation as a probabilistic mixture of emergent topics, and therefore assumes that multiple activities are occurring in each observation. This allows the robot to learn coherent activity classes (topic distributions) even when video sequences are not temporally segmented into clips focused on a single activity instance. In Section 7.1 we provide experiments that use different manual temporal segmentation methods applied to the recorded observations to demonstrate that this approach works well. We conclude with an experiment that uses no manual temporal segmentation, and (many-to-many) map the learned Topic distributions to annotated multi-labelled video sequences.

6. Evaluation

In this section we empirically evaluate the proposed unsupervised learning methods on two human pose video datasets that differ in complexity. We demonstrate that consistent activity classes can be learned from both datasets when compared to human-annotated ground truth labels after temporal segmentation and that the performance of LSA and LDA is superior to simple clustering approaches in this high dimensional and complex setting. The first dataset, which is publicly available and popular in recent literature, consists of 124 video sequences where each is scripted in advance and contains a single temporally segmented activity class instance (along with a label) recorded from a static camera set-up. The second dataset is more challenging; it is recorded from an autonomous mobile robot observing an unstructured, real-world university common area over a one week duration with no restrictions on the types of interactions observed; and is also available publicly. Further experiments are performed in Section 7, which address the two identified practical robot learning problems.

6.1. Datasets

6.1.1. Cornell activities for daily living dataset

As a benchmark, we evaluate our unsupervised learning framework on a popular and freely available human activities dataset consisting of activities for daily living from Cornell University [74], known as CAD120. It consists of 124 RGBD videos, acted out by four different actors. Each video clip comprises of a single actor performing one high level activity class instance out of 10 classes, resulting in a video clip with a corresponding ground truth class label. The activity classes are predefined as: *arranging objects*, *cleaning objects*, *having a meal*, *making cereal*, *microwaving food*, *picking objects*, *stacking objects*, *taking food*, *taking medicine* and *unstacking objects*. Many videos contain real-time object detections at each timepoint and correspond to dynamic objects in the scene. This allows us to evaluate our qualitative framework using dynamic object locations as well as static key objects (present in the mobile robot dataset). The dataset has a hierarchical structure and each of the high level activity classes is comprised of multiple lower level activities, such as: *pouring*, *eating*, *opening*, *placing*, *reaching*, *moving*, *cleaning*, *drinking* and *closing*. The occurrence of these overlap within the 10 high level activities, which provides some difficult inter-class similarities. The activities are performed facing a fixed camera, with minor background clutter, and where the subject is positioned in the centre of the camera frame to allow for good human pose estimates (15 joint positions) and auto and ground truth object detections.

This dataset is a large and challenging datasets of human daily living activities in recent literature. It contains real-world activities that occur in one's daily life and are considered useful for a robot to learn about. Further, the human body poses are estimated using OpenNi which struggles with body joint occlusions, even with simple reaching or placing activities.

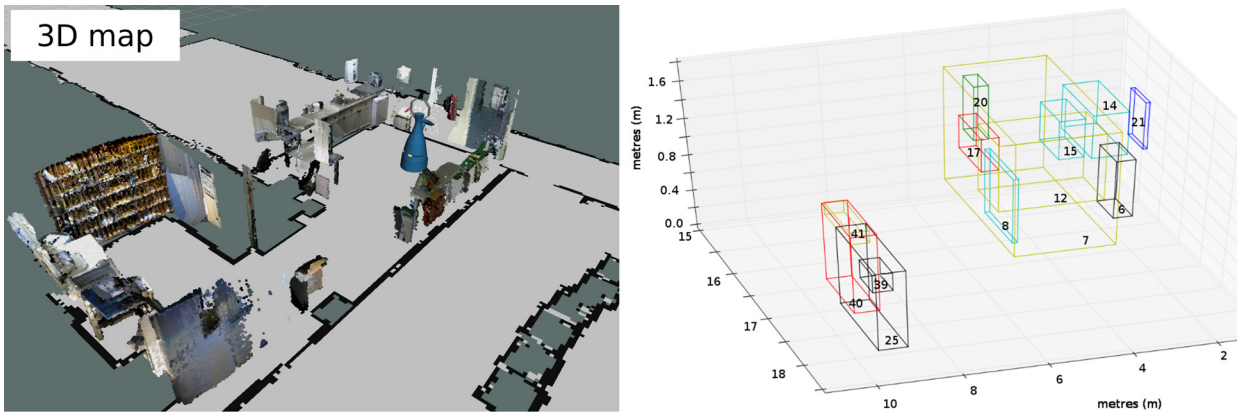


Fig. 9. Human Activity Dataset environmental set-up and autonomously learned object locations. (Left) 3D object clusters extracted from fused point clouds, overlaid onto the metric map. (Right) Sub-set of object clusters selected using analysis of human trajectories. The axes refer to the metric map frame relative to the map origin. Image best viewed in colour.

However, it is not considered complex compared to a robot's 'in-the-wild' human observations. The activity classes are clearly defined, scripted and acted with slow body joint movements. It has very balanced activity classes in terms of the number of instances within each class; ≈ 12 repeats of each. Also, the videos are temporally localised, i.e. each clip focusing only on a single activity instance.

The rationale for including analysis on this fixed camera, highly scripted dataset, is that it provides a known set of activity classes at a particular activity-granularity in order to test the presented qualitative framework. Further, the availability of dynamic object tracks also provides an interesting extension to using static objects learned from a mobile robot.

6.1.2. Mobile robot human body pose dataset

This real-world human observation dataset was captured by an autonomous Metralabs Scitos A5 mobile robot deployed in a human-populated university environment.[2] Images and human pose estimates are captured using a head-mounted ASUS Pro-Live Xtion RGBD camera observing members of staff and students performing every day activities in the kitchen and student common areas. No restrictions were placed upon the observations or activities that occur in the environment, meaning that the robot observed many partial, incomplete, and fast-paced interactions between people and objects, where body poses are challenging to infer. The dataset also provides many difficult variations due to lighting, various viewpoints and many occlusions, usually from multiple people in the environment at the same time.

Objects: The robot first performs a 3D metric sweep of the environment to generate a registered 3D point-cloud, then extracts a set of key object clusters from this representation and overlays them onto the metric maps as described in Section 3.3. The object clusters learned using this process can be seen as point cloud clusters in Fig. 9 (left). A sub-set of objects are selected using the analysis of human trajectories, i.e. where people stop and what locations their body joints interact with. This resulted in a set of 41 key object locations shown in Fig. 9 (right), which somewhat correspond to locations of real-world objects, e.g. object ID 6 corresponds to the trash bin, 21 to the paper towel dispenser, and 25 to part of the printer-copier machine. The locations of the autonomously learned key objects are evaluated with respect to manually specified objects that people interact with in [3]. This evaluation provides an alternate set of 12 manually specified key object locations encoded and which we also use to compare our learning framework to determine how dependent the system is on accurate key object locations. The manually selected objects include: *shelves, microwave, water cooler, tea/coffee pot, sink, kettle, fridge, paper tray, printer screen, paper towel dispenser* and two *waste bins*. We evaluate using both objects sets.

Human body poses: The robot was tasked with observing from pre-defined topological nodes (map frame x, y locations) and observing the environment from various different viewpoints. Given a detected person in the robot's field of view, the camera records RGB images along with the estimated human body pose as described in Section 3.2. The human body pose is estimated from the camera image, first using the depth image (OpenNi2), then the RGB image post-process (CPM). Each body joint position is then translated into the robot's map coordinate frame of reference using the localised position of the robot and the pan-tilt angle, i.e. where the camera is pointed. Obtaining an accurate position of the body joints in the map frame relies on the robot being well localised within the map. The visual SLAM algorithm is not always accurate when the robot is moving, so we restrict the human observations to when the robot is static. The dataset was collected over the period of one week. The robot observed 287 individuals during the process and estimated a human body pose sequence for each. These sequences contain arbitrary number of poses with high variance, with a mean number of 513 poses and

[2] The dataset and software repository is available at: http://doi.org/10.5518/86.

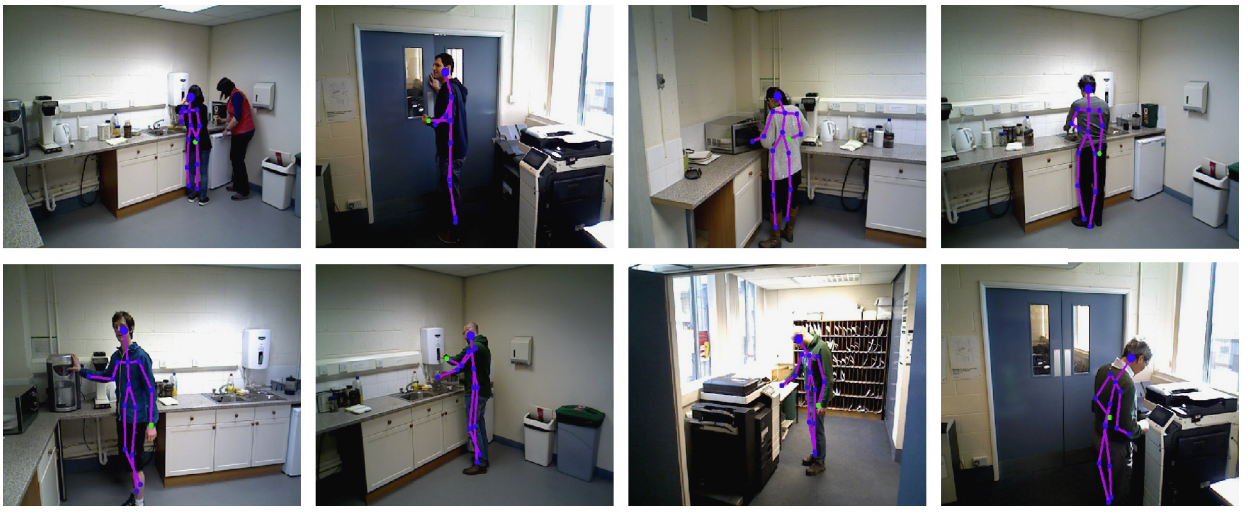


Fig. 10. Example RGB images with human pose estimate overlaid from the Mobile Robot Human Body Pose Dataset. Image best viewed in colour.

standard deviation of 588 poses, indicating a very large spread (and long tail), representative of the nature of varied human observations.

Activity ground truth: To obtain a ground truth (GT) activity class label for the human observations, each recorded sequence was manually inspected by volunteers. A set of common and repeated activities present in the recordings was agreed upon and this defined the set of activity classes annotated. The activity granularity of the defined classes was somewhat limited by the data available from the robotic vision component; in particular no object tracker was available, and hand tracks are not always reliable; meaning that activities involving small objects or complex hand movements were indistinguishable, and thus only human activities involving static objects were used as ground truth labels.

The occurrences of each instance of each activity class within the observed videos was temporally segmented by the volunteers to delineate the activities, creating multiple shorter video sequences containing only a single activity instance in each. The full list of the activity classes annotated, along with the number instances in each class are as follows: Wash cup (82); Take object from fridge (81); Use the kettle (70); Throw trash in bin (65); Take paper towel (45); Take tea/coffee (35); Use printer interface (35); Use the water cooler (26); Take printout from tray (24); Microwave food (19); Opening double doors (11); No Activity Label (N/A) (77); Total 493 (570). A total of 493 individual activity class instances were extracted, (with 77 observations containing no activity class). These segmented clips are shorter than the full recorded sequences, with mean number of 137 poses and standard deviation of 190 poses, and temporally focused on a single activity instance taking place. We consider these sequences as *interesting* human activities, as defined and segmented by the volunteer annotators. The experiment presented in this section uses these pre-segmented clips (we move to a more realistic setting in Section 7).

6.2. Experimental procedure

Here we present the implementation details of our framework that are common to both datasets, with the aim of learning coherent human activity classes in an unsupervised setting and using the ground truth labels for evaluation purposes only. Details include: i) the qualitative representations used in the following experiments and ii) the parameters used to compute codewords, resulting in a codeword histogram representation for each video clip and a term-frequency matrix per dataset.

A segmented video clip m (from either dataset) is first represented as a human pose sequence $S_m = [p_1, \ldots, p_i, \ldots p_t]$ of length t, where each p_i is the human body pose at timepoint i (corresponding to the ith frame) and contains both the camera frame and map coordinate frame 3D position of 15 body joint locations, i.e. as 15 joint poses. Recall that t is arbitrary and varies for each observation. Abstracting this sequence of body poses into a qualitative representation is performed in a two stage process, first the *camera* frame relationships are computed, followed by translating and computing the *map* frame relations.

TPCC calculi is then used to abstract the person's body joint positions relative to the head-torso (origin-relatum) 2D line in the *camera* coordinate frame. The sequence of TPCC relations Q_{cam} of length t contains TPCC relations between the centre-line and the left/right hands and shoulders joint positions. Other joints were omitted for efficiency and also because we assume their movements do not contribute to the kinds of human activities in our two datasets. However, they could obviously be added. Next, to abstract the body joint locations in the *map* coordinate frame we use a combined QDC and QTC calculus. QTC captures relative motion and QDC captures relative distances of the poses to key objects. These QSRs are used

to represent the relative movements of the person's left/right hand body joints and torso location, relative to the key objects. A sequence of combined QDC and QTC pairs is produced, Q_{map}, of length $t-1$ (since QTC relies on pairs of consecutive poses so the QDC value at $t=1$ is unused). QDC values greater than the largest boundary in Δ are not considered, this produces a sparse interval representation and an efficient process with fewer codewords per observation.

For each sequence Q_{cam} and Q_{map}, we apply a median filter (window size of 10), which smooths rapid flipping between relations. This is not uncommon when abstracting into a qualitative representation, for example, pair-wise distances that exist on or close to a QDC boundary can constantly flip between relations. The aim is to retain only those relational changes which form intentional human body movements. We create an interval representation and interval graph for each sequence separately and extract all graph paths from both using graph path parameters η and ρ given per dataset below. Since the qualitative relations differ between the two sequences, Q_{cam} and Q_{map}, we merge the two collections of extracted graph paths (words) together in order to create a single bag-of-words to represent each observation, i.e. $D_m = \{w^1, w^2, \ldots, w^{N_D}\}$, where each w^i is an observed graph path. Finally, after observing and encoding M sequences, a codebook of N unique codewords is generated.

Given an $M \times N$ term-frequency matrix C representing an entire training dataset, we learn activity classes and compare them to the ground truth annotated labels. To learn the best number of latent topics or concepts, (or a suitable low-rank r in the LSA method), a small number of "large" singular values are extracted using the "elbow point" after plotting the decreasing eigenvalues as described in Section 5.1. The optimal values found are 10 latent concepts for the CAD120 dataset, and 11 for the Mobile Robot Dataset. However, we propose a method to alter this number dynamically during an incremental learning process in the next section, which more accurately reflects the requirement of real-world robotics. In subsequent experiments, different temporal segmentation methods are applied to the activity video clips in order to evaluate the efficacy of our proposed methods. However, we set the LDA hyperparameters α, β to 0.5 and 0.03 respectively for all experiments. These hyperparameter settings reflect our prior belief on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ that each observation is likely to consist of only a small number of topics, and that topic distributions consist of a relatively small number of codewords with probability mass. A more detailed analysis into the LDA hyperparameters is performed in [75].

CAD120 dataset: In each video sequence in the CAD120 dataset a single person is situated near the centre of the camera frame performing the activity, and the objects are situated close to the person. For this reason, the QDC relation thresholds (introduced in Section 4) are set to $\Delta = [0.15\text{ m}, 0.4\text{ m}, 0.8, 1.0\text{ m}]$, creating five semantic regions which can be labelled as *touch* [0–0.15 m], *near* (0.15–0.4 m], *medium* (0.4–8.0 m], *far* (0.8–1.0 m] and *ignore* (>1 m). These were experimentally chosen to distinguish the body pose movements in the more simple set-up. Also note, for the dynamic objects, the abstract object class is used as the object ID in the interval representation and interval graph. A code book $\mathcal{V}_\mathcal{D}$ of codewords (unique graph paths) is computed, where $|\mathcal{V}_\mathcal{D}| = 5{,}520$ codewords (29,016 in total), and a low-pass filter is applied reducing this to $|\mathcal{V}_\mathcal{D}| = 958$, using QSTAG parameter choices: $\eta = 3$ and $\rho = 1$, i.e. all paths of up-to length three are computed that include only one pair of objects. Finally, a codeword histogram is computed for each video clip and an $M \times N$ term-frequency matrix is computed where $M = 124$ and $N = 958$.

Mobile robot human body pose dataset: The video sequences recorded from the mobile robot are much more varied and challenging. We evaluate the learning methods when using two sets of keys objects. First, the set of 14 most interacted with key objects were obtained from the 3D sweeps and trajectory analysis described in Section 3.3; these align reasonably well to real objects in the environment, and they are not labelled with any prior semantic knowledge. Secondly, we evaluate using the set of 12 manually specified key object locations in [3]. This allows us to determine how important obtaining the exact location of key objects is to our framework. QDC thresholds used in this experiment are initially defined as $\Delta = [0.25\text{ m}, 0.5\text{ m}, 1.0\text{ m}]$ creating four semantic regions which can be labelled as *touch* [0–0.25 m], *near* (0.25–0.5 m], *medium* (0.5–1.0 m] and *ignore* (>1 m), which were experimentally chosen to best distinguish activities in this more complex environment. However it is possible to learn the threshold values from observations in an unsupervised setting [35], and in Section 6.3.2 we provide a sensitivity analysis of altering these values, and the effect it has upon the activity topics.

A code book $\mathcal{V}_\mathcal{D}$ is generated, initially using QSTAG parameters $\eta = 4$ and $\rho = 2$, however a detailed sensitivity analysis is provided in Section 6.3.2. Here, using the 14 autonomously learned key objects, $|\mathcal{V}_\mathcal{D}| = 20{,}637$, which is reduced to 2,876 when a low-pass filter (frequency = 5) is applied. Similarly, $|\mathcal{V}_\mathcal{D}| = 22{,}829$, reduced to 3,594, for the case when using the set of manually defined key objects. The extra codewords highlight that there are more unique qualitative relationships between the human pose estimates and the manually defined object locations and therefore the possibility of more discriminative codewords which can help the unsupervised learning performance. A codeword histogram is computed for each video resulting in an $M \times N$ term-frequency matrix. We show in Section 7.2 that a more efficient incremental update of the code book with new codewords as they are observed in a life-long learning setting is most practical.

6.3. Experiment 1: learning activities from activity clip

Clustering metrics: Using an unsupervised learning framework, there is not likely to be a one-to-one mapping from the learned topics (or clusters) to each ground truth activity class. This is more likely to be a many-to-many mapping, especially when dealing with highly unbalanced classes. Therefore, we provide results using popular clustering metrics where the aim

Table 1
Experiment 1. (Top) CAD120 Dataset. (Bottom) Mobile Robot Dataset using autonomously learned key objects. (Manually defined object results shown in brackets). Cluster metrics obtained comparing the ground truth labels segmented video clips encoded using a qualitative framework, against the learned, emergent human activity classes. The table shows methods of increasing sophistication: unsupervised k-means clustering; low-rank approximate LSA; Generative LDA; compared against random chance and a supervised SVM as an intuitive lower and upper bound respectively.

	Random clustering	Unsupervised k-means	Unsupervised LSA	Unsupervised LDA	*Supervised* SVM
CAD120 Dataset					
V-measure	0.18	0.63	0.76	0.74	0.84
Mutual Information	0.41	1.33	1.72	1.66	1.93
Normalised MI	0.18	0.63	0.76	0.75	0.84
Accuracy	0.10	–	–	–	0.81
Mobile Robot Dataset					
# classified	493	493	493	487	493
V-measure	0.05	0.30 (0.41)	0.68 (0.65)	0.63 (0.71)	0.71 (0.73)
Mutual Information	0.12	0.31 (0.80)	1.46 (1.42)	1.40 (1.51)	1.59 (1.63)
Normalised MI	0.05	0.54 (0.41)	0.68 (0.64)	0.63 (0.71)	0.71 (0.73)
Accuracy	0.12	–	–	–	0.78 (0.80)

is to generate "coherent" clusters that are composed of the same activity class labels, and that all instances of a class are assigned the same cluster (concept or topic). For this purpose we use the two metrics, V-measure [76] and (Normalised) Mutual Information (NMI) [77]. V-measure is a combination of the homogeneity and completeness clustering metrics, given two sets of labels. Homogeneity evaluates whether all the predicted clusters contain only data points which are members of the same class; whereas completeness evaluates whether the member data points of a given class are all elements of the same predicted cluster. Both values range from 0 to 1, with higher values desirable. NMI is an normalization of the Mutual Information (MI) score between two sets of clusters, ranging from 0 (no mutual information) and 1 (perfect correlation).

Upper and lower bounds: To add perspective to the human activity classes learned using our proposed unsupervised framework, we present multiple comparison techniques. First, we propose a *supervised* learning technique as a hypothetical upper bound on performance. We learn human activity classes using a supervised Support Vector Machine (SVM) algorithm on the rows of the encoded training term-frequency matrix with corresponding ground truth labels. The SVM is trained using 5-fold cross validation, with a linear kernel, and where the code book is trained once across the whole dataset (as opposed to recomputing the code book at each training fold). This supervised technique has access to the ground truth labels during the learning process and so we naturally expect its to outperform the unsupervised approaches. Secondly, we present the results of random clustering as an average over 10 repeated runs as a lower bound. We expect each of our proposed learning methods to outperform this. Finally, we compare against the simple unsupervised clustering k-means algorithm, where the number of clusters is set to the optimum number found using the eigenvalue decomposition. Due to variations in the initialisation of the algorithm, the presented results are an average over 10 repeated runs, each initialised randomly.

6.3.1. Experiment 1 results

Results for the unsupervised learning of activity classes from the Cornell Activities for Daily Living Dataset and the Mobile Robot Human Pose Dataset are presented in Table 1. For the case where a video is assigned a multinomial distribution of concepts/topics, the highest value is taken as its classified value to compare against the ground truth and cluster metrics are computed.[3] The results show that the unsupervised learning methods are sophisticated enough to separate the observations into coherent classes that are consistent with human annotated ground truth labels; and that the qualitative framework used to abstract the observations can be considered somewhat viewpoint invariant and can handle noise and some variation during the observational phase. The proposed unsupervised learning methods significantly outperform the simple unsupervised k-means clustering algorithm and the random assignment. We interpret this as LSA and LDA better generalising observations than the more simple methods, since they can consider codewords features with similar meaning, i.e. identifying synonymy between encoded dimensions unlike the simple methods. Further, the performance of these unsupervised methods is comparable with a supervised SVM, which performs only slightly better, however it uses the ground truth labels to compute its decision boundaries. Confusion matrices of the emergent LDA topic assignments vs the ground truth class labels are shown in Fig. 11.

6.3.2. Experiment 1 sensitivity analysis

In Table 2 we present how the cluster metric (V-measure) performs on the Mobile Robot Dataset after altering hyperparameters in our experimental procedure. We take the approach of varying each of our qualitative representation hyperparameters, over a viable grid of values. The results highlight where specific choices of parameters may influence

[3] For the Mobile Robot Dataset only the highest topic proportion > 0.3 probability threshold is selected, where the "# classified" row specifies the number of observations classified above this threshold.

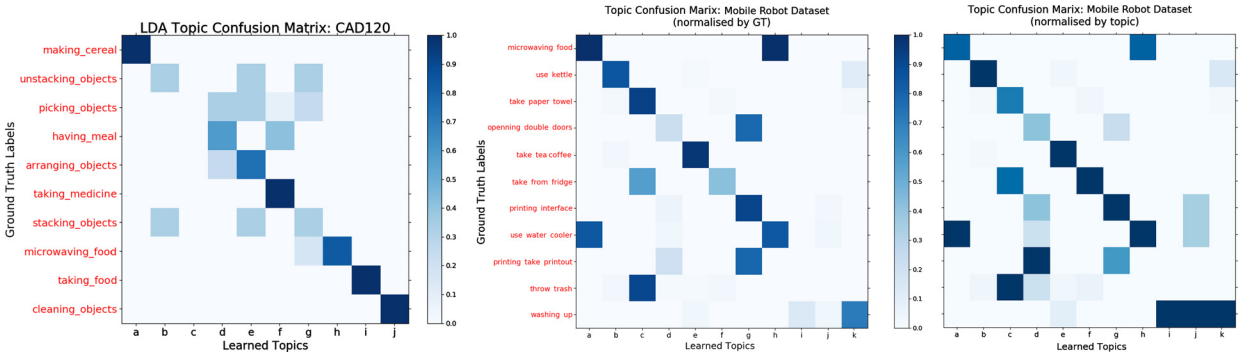


Fig. 11. Experiment 1 (Left) Confusion Matrix for CAD120 Dataset. Normalised by ground truth labels. (Centre and right) Confusion Matrix for Mobile Robot Dataset (using manually defined objects). Normalised by ground truth labels (left) and topic assignment (right).

Table 2
Experiment 1. Sensitivity Analysis of parameters in Experiment 1. (Top) We vary the QSR calculi used, using just one calculus. (* When using only QTC_{BC} no low-pass filter was used on the codebook); (Middle top) We then exclude one calculus, leaving two remaining; (Middle bottom) We vary the QDC parameter Δ; (Bottom) We vary the QSTAG path parameters, η and ρ. In all cases, we present the V-score metric for the unsupervised LSA and LDA; and a supervised SVM.

Sensitivity parameter	Parameter value	Codebook size	Unsupervised LSA	Unsupervised LDA	*Supervised* SVM
QSR Calculi	QDC	1917	0.66	0.66	0.70
	TPCC	1083	0.20	0.23	0.34
	QTC_{BS}	75*	0.65	0.12	0.04
QSR Excluded Calculi	¬ QDC	1094	0.21	0.17	0.34
	¬ TPCC	1915	0.66	0.66	0.70
	¬ QTC_{BS}	3597	0.62	0.69	0.72
QCD thresholds (Δ)	[.15,.5,1.0]	3010	0.53	0.60	0.69
	[.25,.5,1.0]	3594	0.65	0.71	0.73
	[.4,.5,1.0]	3693	0.67	0.65	0.69
	[.25,.35,1.0]	2916	0.60	0.59	0.74
	[.25,.75,1.0]	3582	0.65	0.68	0.72
	[.25,.5,.75]	3148	0.62	0.61	0.73
	[.25,.5,1.25]	3491	0.66	0.64	0.73
QSTAG (ρ, η)	$\eta = 3$, $\rho = 1$	316	0.61	0.68	0.76
	$\eta = 5$, $\rho = 1$	488	0.61	0.63	0.76
	$\eta = 3$, $\rho = 2$	2701	0.60	0.65	0.73
	$\eta = 4$, $\rho = 2$	3594	0.65	0.71	0.73
	$\eta = 4$, $\rho = 3$	6566	0.69	0.63	0.69

the framework more than others. We conclude that, given this representative set of QSR calculi, hyperparameter changes do not greatly affect the clustering methods.

The top three rows of Table 2 where just a single QSR calculi is used to encode the video clips, we see that the resulting codebooks are much smaller and that using TPCC alone yields a poor representation. Next, we compare to Table 1 and remove one of the three calculi (leaving two calculi remaining). Here we observe that QDC provides a large amount of the discriminatory information to the framework. This is seen when we remove these relations from the qualitative representation, the performance of both the supervised and unsupervised methods is greatly reduced. Given this importance of the QDC relations, we present a sensitivity analysis varying each QDC threshold value, $\delta_i \in \Delta$, over a small grid. Finally, in the bottom rows of Table 2 we consider different values of the QSTAG graph path parameters. We see that this also greatly affects the codebook size, but performance is somewhat resilient given the representation of all three calculi.

It should be noted that the chosen three QSRs cover the space of possible QSRs well. As pointed out in [26] there are various aspects of space that QSRs abstract: 1) mereotopology, 2) direction, 3) distance, 5) motion, and 5) shape. Using TPCC covers (2), QDC (3) and QTC (4). In the domain of interest, mereotopology is not so relevant since it largely is concerned with parthood relations, and our physical objects do not interpenetrate, whilst we cannot discern object shape well enough to want to consider shape calculi (and moreover, except for human poses, our objects do not change shape).

6.4. Discussion

CAD120 dataset: The confusion matrix shows that when using the unsupervised LDA, some activities such as *making cereal*, *taking medicine*, *taking food*, *cleaning objects* etc. are separated very well and all instances of these classes are coherently classified within their own topic distribution. However, some activity classes are sometimes confused, such as, *stacking objects* and *unstacking objects*, which is somewhat expected given the high visual similarities between these activities. Classifying a video sequence by selecting only the highest probable concept/topic does not adequately demonstrate the mixture over the topics and can negatively effect the attempted mapping between ground truth labels and emergent topics.

Mobile robot dataset: We can see that manually specifying the key object locations across the environment helps to obtain more coherent clusters of human activity with respect to ground truth labels. One hypothesis for this improvement is that given more accurate key object locations in an environment, a more rich qualitative representation between the human pose and the objects can be encoded, leading to more discriminative codewords for the same video sequence (and a larger code book). It is also interesting to note that LDA outperforms LSA when using the manually defined key objects, i.e. when the intra-class similarities are much higher (compared to the CAD120 Dataset). We believe this is because the LDA model is able to model a mixture in the training samples better, so even in training samples that are difficult to classify, the LDA topic distributions are more representative.

Due to the unbalanced activity classes, i.e. the number of instances of each class, we normalise the confusion matrices (in Fig. 11) by both the frequency of ground truth labels (centre) and of topic assignments (right). From this presentation, we can interpret that the activity class *washing up*, is almost entirely classified into Topic ID 10 (centre), but also that Topic ID 8 and 9 consist of mainly these ground truth videos (right). Topic k is a highly frequent topic that represents the human activity, however, topics i and j may represent this activity being performed in a slightly different visual way. This variety is less frequent, as only a small number of videos are classified into i or j.

The Mobile Robot Dataset is recorded from an unstructured human environment, however, in this section we have included a pre-processing step to segment the video sequence clips. There remain several practical considerations that limit the effectiveness of the framework in a life-long robotic deployment setting, such as, how to autonomously obtain human observations, how to deal with continuous streams of video data and employ learning methods that are incremental and more efficient. In the next section we discuss some of these issues and attempt to remove some of the assumptions in order to allow for a better learning framework to be deployed in a continuous, life-long learning setting.

7. Practical considerations for life-long learning

Practical considerations that arise from deploying a mobile robot in a real-world setting are not often focused on in much of human activity analysis research, i.e. it is common to use popular heavily, pre-processed video datasets. In this section we propose methods to address two limitations of many activity analysis frameworks. These limitations impede popular techniques from being applicable in a continuous, life-long learning settings; such considerations are required for real-world deployed mobile robots [5]. The assumptions we address in this section are:

1. The recorded human pose sequences are temporally focused around a single human activity instance occurring, as opposed to multiple overlapping activities, or similarly, no interesting activity occurring in the video sequence at all.
2. The robot has unlimited time and computing resource to learn human activity classes by repeatedly performing a batch learning process over an ever increasing corpus of observations, instead of using more efficient, incremental learning methods to build upon its knowledge over long periods of time.

We introduce each of these two challenges and propose a solution and experiments to validate our approach.

7.1. Learning from continuous video

This is the assumption that human observations recorded by a robot consist of a consecutive sequence of images where a person is performing a single “interesting” activity, and that the body pose can be inferred correctly, i.e. a person enters the robot’s frame of view, performs a typical human activity in such a way to not occlude anybody joints, and then exits the camera frame. In practice, for robots deployed in real-world environments, this is not what they observe. Often the robot’s observations consist of people performing multiple overlapping activity classes, or performing only part of an activity within the robot’s limited field of view. Alternately, observations may contain no repeated behaviours. In this section we propose experiments to evaluate our claims in Section 5.4, namely the hypothesis that probabilistic LDA models assume each observation as a mixture of emergent topics apriori, and therefore model multiple topics (activities) occurring in each sample (observation). This allows the robot to learn coherent topics even when the video sequences are not temporally focussed on a single activity or where the video clips do not precisely delineate the start and end of an activity (albeit that this itself is a subjective notion). We provide two experiments to demonstrate that this approach works well.

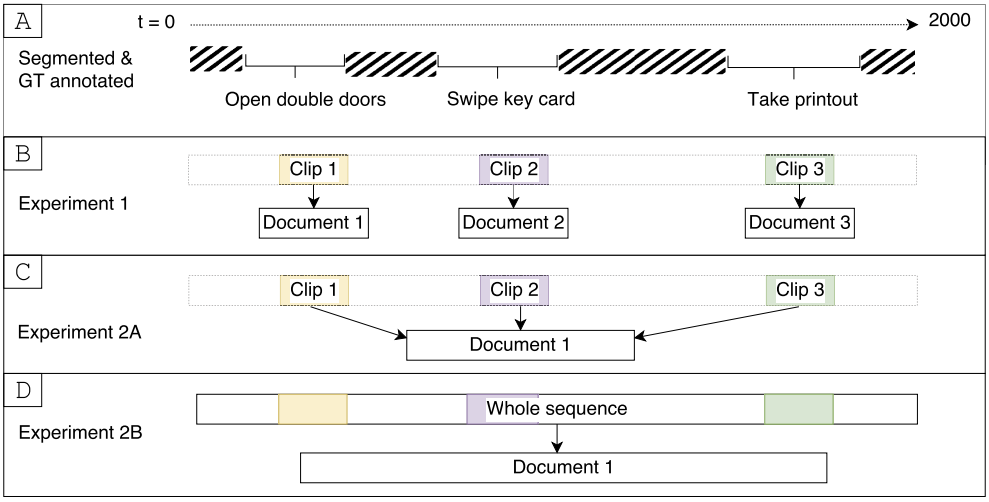


Fig. 12. Demonstration of three different temporal segmentation methods applied to a recorded video sequence containing three activities as annotated by volunteers.

7.1.1. A word on temporal segmentation

The Mobile Robot Dataset described in Section 6.1.2, is recorded using OpenNi2 [51] in real-time to detect humans in the robot's field of view and estimate their body pose per frame. Volunteers temporally segment (manually) the video sequences into shorter video clips focused around a single human performing an activity. This is a common practice with video datasets. As seen in Experiment 1, this segmentation steers the learning methods towards representing the repeated qualitative behaviours present in these clips as emergent activity classes. However, when no temporal segmentation is applied there are many different interactions within the video sequences that are repeated or common that were not considered "interesting" activities by the annotators and therefore excluded from the segmented clips. For example, most of the observations will contain codewords that may relate to the human body pose when stood idle, or walking. From the 293 human observations recorded in the dataset, 77 contained none of the ground truth activities (as annotated by volunteers), however they all contain a human detection and a sequence of estimated body poses. This is a major challenge when using real-world, non-segmented, video sequences recorded "in-the-wild". Similarly, the repetition of activities in the dataset is one of the reasons the volunteers annotated the sequences with a ground truth label, i.e. the set of labels derives from interesting activities that were commonly repeated. However, other activities occur in the recorded sequences but with low frequency, for example one recorded observation contains a person *cutting a birthday cake*, however this was a unique instance of this activity and therefore was not included in the list of ground truth classes.

Fig. 12 demonstrates how a single video sequence (A) consisting of 2000 poses (camera frames) and three ground truth activities can be represented using three different temporal segmentation approaches (B, C and D). Here, the annotators believed three activity instances occurred, and annotated the sequences accordingly. It is not uncommon for large temporal gaps between the annotated activities, where no activity is believed to be occurring. These excluded sequences of frames, marked with hashes in Fig. 12(A), are extremely difficult to handle for an activity learning framework because any human behaviour, not restricted to the set of ground truth classes, could be observed.

Fig. 12(B) is an example of full manual temporal segmentation, where each activity class instance is segmented into a video clip and encoded as a codeword histogram (document); (D) is an example of where no manual temporal segmentation is applied. Here, the entire human detection sequence is encoded into a single codeword histogram (document). This is the most realistic and challenging setting; (C) represents the in-between case, where manual segmentation of the clips in (A) are concatenated back together into a possibly discontinuous sequence that excludes the surrounding (hashed) frames. Experiment 1 in Section 6.3 used segmentation method (B) for the Mobile Robot Human Pose Dataset. We present two further experiments on this dataset using methods (C) and (D) next.

7.1.2. Experiment 2A: concatenated activity clips (method C)

This experiment is designed to evaluate the capability of our framework to learn coherent human activity classes based upon video sequences that contain multiple activity instances, i.e. a step towards using recorded video with no manual segmentation. The idea is to compare the emergent LDA topic distributions when manually segmented video clips are concatenated back together to achieve a sequence containing multiple activity instances, excluding the surrounding video frames. The Mobile Robot Dataset consists of 287 recorded video sequences manually segmented into 493 annotated clips. We concatenate the clips back together into 210 concatenated sequences (recall that 77 recorded sequences contained none of the ground truth activities). This results in a set of 210 sequences that contain a (mean) average of 2.3 segmented clips

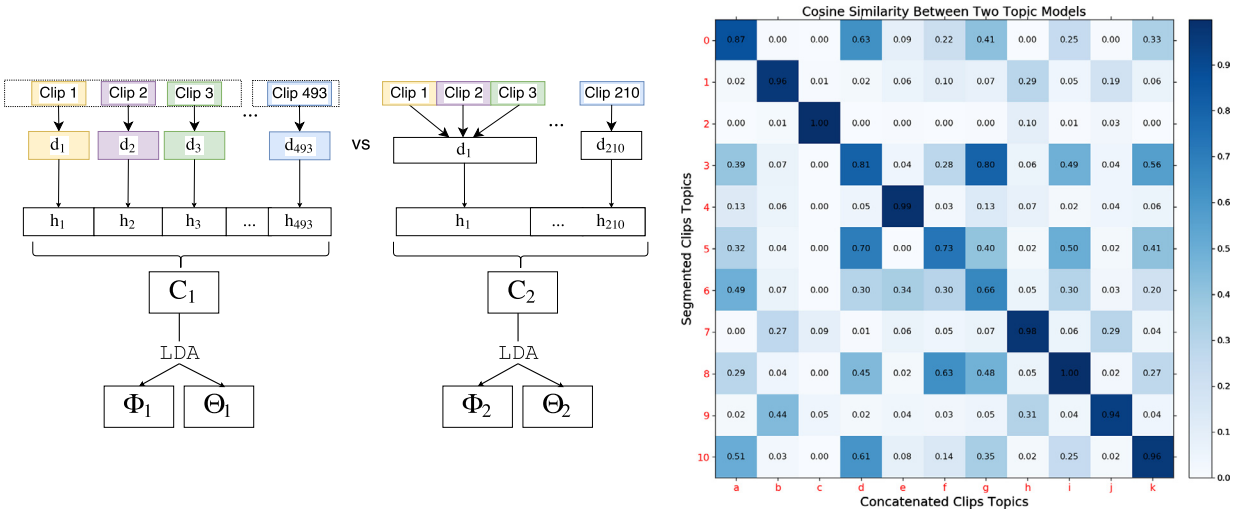


Fig. 13. (Left) Experiment 2A set-up. Comparing LDA topic distributions using segmented video clips (Experiment 1) versus using concatenated sequences containing multiple activity instances. (Right) Similarity matrix comparing the Cosine Similarity of two topic distribution matrices Φ_1 and Φ_2.

per sequence (max $= 10$), where each clip is associated to a ground truth annotated label, so each concatenated sequence corresponds to multiple ground truth activity labels.

In this experiment it is the same estimated human body pose sequences that are used as in Experiment 1, however multiple activities are now present in each observation (codeword histogram) and so we require the learning framework to be flexible and model each as a mixture of classes. This translates as assuming a multinomial distribution over the topics within each observation, removing the requirement for manual temporal segmentation of observations into shorter clips where only a single activity occurs. We follow the same experimental procedure as in Experiment 1 and the goal here is to compare the LDA topic distributions obtained to those obtained in Experiment 1. That is, using the temporal segmented clips an $(493 \times 3{,}543)$ term-frequency matrix C_1 is computed (Experiment 1) then, using the concatenated sequences a $(210 \times 3{,}436)$ term-frequency matrix C_2 is computed. The slight variation in the number of codewords is due to the low-pass filter, i.e. a codeword must appear in a minimum of 5 codeword histograms. The experimental set-up is shown in Fig. 13 (left). The idea is to compare (Φ_1, Θ_1) with (Φ_2, Θ_2) and evaluate the effects of the extra manual temporal segmentation applied to the video sequences encoded in C_1, i.e. compare segmentation method (B) with (C). We do this by examining the Cosine Similarity between the two emergent topic distribution matrices Φ_1 and Φ_2, that represents the difference of the angle between the (unit-normalised) topic distributions.

Experiment 2A results: In Fig. 13 (right) we present the Cosine Similarity matrix between the topic distributions learned using the temporally segmented clips against those learned using the concatenated sequences. We use the Munkres Hungarian algorithm [78] in order to match the highest corresponding topic distributions together, i.e. an assignment problem. We can see clearly that the learned topic distributions are very similar, even though the input video sequences are very different, i.e. C_2 is encoded using multiple activity classes in each observation. The strong diagonal indicates a good one-to-one mapping between the two recovered topic distributions and demonstrates that the framework presented is able to recover coherent topic distributions representing activity classes from the observations containing a mixture of activities.

The average Cosine similarity between the two assigned sets of topic distributions is 0.90. This value drops to 0.56 when using the low-rank approximation method LSA to learn the emergent concepts from the two term-frequency matrices. This demonstrates that the probabilistic, generative LDA method is able to better handle human observations that contain a mixture of activities occurring in each observation. This is a very desirable property as we move towards using no manual segmentation applied next.

7.1.3. Experiment 2B: no manual segmentation (method D)

This experiment is designed to evaluate the consistency of learned topic distributions with respect to annotated ground truth activity labels when no manual temporal segmentation of the recorded sequences is applied. That is, the robot encodes a bag-of-words and a corresponding codeword histogram for each sequence of recorded human pose estimates it observes whilst observing "in-the-wild". This is a very challenging task, given the dynamic and real-world environment the robot is required to observe.

All the recorded video sequences in the Mobile Robot Dataset are used in this experiment, i.e. $M = 287$ (where 77 contain no labelled activity class instance). These sequences are much longer with mean: 513 poses and std of 588. They are more varied and contain many human behaviours that were not repeated consistently throughout the dataset, or are

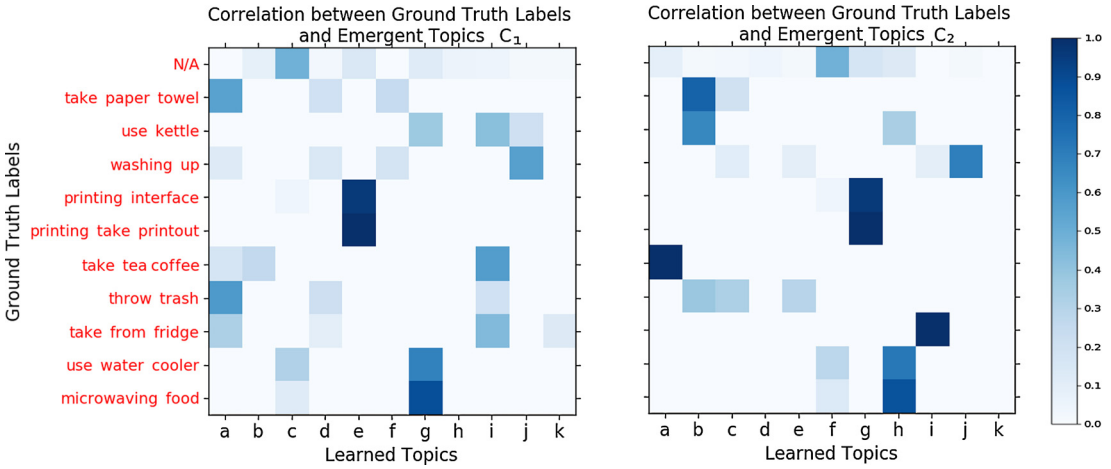


Fig. 14. Topic distributions learned using no temporal segmentation of video sequences correlating to each annotated ground truth activity class. (Left) C_1 computed using low-pass filter size of 10 and (right) C_2 with filter of 15.

not considered "interesting" by the annotators, i.e. those poses hashed out in Fig. 12 and include *cutting a birthday cake* for example. Each sequence can correspond to multiple ground truth activity classes taking place and therefore multiple ground truth activity labels. In this experiment, there is no manual input into the segmentation process.

The same experimental set-up as Experiment 1 is used, however there are many more unique codewords (graph paths) extracted from the much larger interval graphs, i.e. 48,172 unique codewords extracted from the observations. With no segmentation the code book is much larger and many of the codewords represent qualitative relations that are not consistently observed but which are present in at least one observation. For this reason, we experiment using two alternate low-pass filters to reduce the vocabulary size; one with the filter size of 10 and a second larger filter of 15, resulting in two term-frequency matrices: (287×6521) matrix C_1 and (287×4531) matrix C_2 respectively. The term-frequency matrices represent each observation as a histogram over the two different code books computed.

Experiment 2B results: Here no manual temporal segmentation was applied to the recorded video sequences and as a result the observations are much longer and considerably more varied. For this reason, we do not expect a one-to-one mapping present between the learned topic distributions and the annotated activity class labels. However, we present a "consistency mapping" between the learned topics and the annotated labels, which help us understand the consistent topics with respect to the human activities taking place. To present the consistency, we sum the topic distributions corresponding to each codeword histogram which contain each annotated ground truth label (using a low-pass filter > 0.5). This is represented as a matrix and shown in Fig. 14, using C_1 (left) and C_2 (right).

We can clearly see (from both consistency matrices) that some activities are not distinguishable from each other, such as the two activity classes that both involve interacting with the printer-copier machine. For brevity, we have removed the "opening double doors" activity class since it has the fewest instances (11) and is not learned. It can be seen that using a smaller code book, the mapping between learned topics and ground truth activities is more clearly defined. Therefore, we shall discuss the results specific to C_2 (right) from here on. The correlation matrix shows that the majority of the learned topic distributions (columns) correlate to a single or pair of activity class labels (rows). For example topic *a* correlates highly with *take tea/coffee* and likewise topic *j* to the activity class of *washing up*. Topics such as *b* relate to a mixture of human activities such as *using kettle* and *taking paper towel*. This is intuitive, and based upon the activities that are often observed together (both temporally and spatially), e.g. washing and then drying a mug is a common pattern observed in the dataset when the videos are not segmented.

However, some classes are being confused based upon their spatial arrangement in the environment, for example, the microwave is roughly 30 cm away from the water dispenser. These objects are not commonly used together, however topic *h* contains a mixture of these classes (plus *N/A*, and *use kettle*). From manual inspection, we see that people usually stand waiting for the microwave and the water cooler in a similar manner, this is what this topic distribution represents and it is unable to distinguish between the two ground truth activity classes.

Experiment 2 discussion: We have demonstrated that a strong assumption of video input sequences being manually temporal segmented is not required using the probabilistic LDA method, i.e. it does not require perfectly temporally segmented activity datasets, with a single activity instance present in each. The model assumes each observation is modelled as a mixture of latent topics, allowing the robot to learn activity classes from data that is not pre-processed by humans in advance. This translates as removing the requirement for manual labour-intensive temporal segmentation tasks, and allows the robot to learn from a larger quantity of recorded video sequences.

Experiment 2A specifically demonstrated that similar topic distributions are learned when clips are concatenated back together and multiple activities are forced into each observation. Experiment 2B showed that somewhat coherent activity classes could be learned from video streams where no manual temporal segmentation is applied and that these topic distributions can be interpreted with respect to ground truth activity classes. Clearly, this setting results in more varied observations being encoded and a direct one-to-one mapping to the annotated labels (of only segmented clips) is unattainable.

7.2. Experiment 3: continual human activity learning

Our final experiment aims to address the practical consideration that it is not efficient to repeatedly perform batch learning on-board an autonomous mobile robot for weeks or months at a time using techniques such as low-rank approximations (LSA) or Gibbs Sampling (standard LDA). We propose to use an incremental learning method that can update its learned activity classes based upon only new observations and does not require re-computing for previously analysed data. Introduced in Section 5.3.1, we use Variational Bayes (VB) approximation method that aims to optimise a simplified, parametric distribution in order to fit the LDA model posterior using mini batches of observations allowing for efficient updating. VB efficiently iterates between analysing a mini-batch of observations and updating dataset-wide parameters. This is particularly relevant to human activity analysis on an autonomous robot "in-the-wild", since it may obtain an intractable quantity of observations to perform standard MCMC sampling approaches (Gibbs Sampling techniques). This method saves memory by not storing the exact quantitative observations and instead maintain a lower dimension distribution over the learned topics (and the topic distributions themselves).

Experimental procedure: Here, we repeat Experiment 1 and learn human activity classes from the challenging Mobile Robot Dataset using the incremental VB method. The idea is to compare how coherent the activity classes obtained are when using the more efficient incremental approach, against the standard approach for fitting the LDA posterior (Collapsed Gibbs Sampling) which requires multiple passes over all the training samples in order to converge on topic distributions. The same segmented video sequences and experimental set-up as Experiment 1 is used, along with the 12 manually defined key object locations in order to produce the same qualitative representation to encode codeword histograms. The term-frequency matrix C is generated of size $(493 \times 3{,}594)$, where each row has an associated ground truth label from the set of 11 ground truth activity classes assigned by volunteers in Experiment 1.

Given the limited size of the dataset, we seed the activity model by learning topics using Collapsed Gibbs Sampling [79] on a batch of observations representing the first of the observations (day 1) (using LDA hyperparameters $\alpha = 0.005$ and $\beta = 0.01$). This equates to the first 146 observations. Then, we incrementally add new codeword histograms using Variational Bayes with a regular mini-batch size of 5 observations to allow for frequent updating of the topic distributions. To pick the number of topic distributions, we employ a simple method that starts with the number of key objects in the environment, i.e. 12 in this case, and increase by one each day to allow new activities to be learned over time. We also remove any topic distribution that is not used sufficiently in order to maintain a reasonable number of topics. When performing batch LDA, the least observed codewords are easily removed using a low-pass filter (window $= 5$). This allows the learning to focus on the most descriptive and repeated qualitative codewords across the dataset. However, a significant challenge with an incremental approach is that there is no clearly efficient way of applying a low-pass filter to an ever increasing vocabulary. As a result the code book $\mathcal{V}_{\mathcal{D}}$ in Experiment 3 grows to a less manageable 22,829 unique codewords, which may negatively effect performance.

Experiment 3 results: Table 3 presents results comparing the incremental, unsupervised topic extraction when compared against ground truth classes. We use the most likely component in a mixture as a label since the proportions vector is multinomial. The method converges to 13 emergent activity classes from the real-world dataset with challenging unstructured behaviour, varying view points, lighting conditions and occlusions. The results show the majority of instances observed are successfully clustered into consistent activity classes using both the VB algorithm and Collapsed Gibbs Sampling. There is a small performance drop when using the efficient incremental method, however slightly more video sequences are classified into classes (above the 0.3 threshold as shown in the "# classified row"), which is an advantage. As an upper bound, we also present results obtained when using a supervised (linear) SVM (with 5-fold cross validation) which has access to the ground truth labels during training and marginally outperforms the unsupervised techniques as in Experiment 1.

Experiment 3 discussion: We have demonstrated that incremental Variational Byes approximate learning methods can be used to update the posterior distribution of a Topic model in order to perform life-long learning, where the number of observations may grow beyond the computing resources available using batch methods. Experiment 3 showed comparable performance between VB and Collapsed Gibbs Sampling for our task. However, we believe that when using VB the posterior distribution did not fully converge due to lack of observations and the larger dimensionality of the code book. A further challenge of incremental methods is the specific order in which the observations are observed. Ideally, the robot would observe a random assignment of different activity classes to clearly separate them within the topic distributions early in the incremental learning process. However in practice, i.e. when the robot is observing the real-world it is often static during the recording process, and many consecutive observations are recorded of a similar activity class, before new locations are investigated. This is a major challenge for incremental learning approaches.

Table 3
Experiment 3: Results comparing fitting an LDA posterior using Collapsed Gibbs Sampling versus incremental Variational Bayes inference.

Cluster Metric	Standard LDA	Incremental VB LDA	Supervised SVM
# classified	486	493	493
V-measure	0.71	0.66	0.73
Mutual Information	1.51	1.42	1.63
Normalised MI	0.71	0.66	0.73
Accuracy	–	–	0.80

8. Conclusion

In summary, we have introduced a novel framework whereby low-dimensional representations of human observations from a mobile robot are learned. We demonstrate that by first abstracting observations using qualitative spatial relations between tracked entities in a visual scene and secondly performing probabilistic unsupervised learning techniques, efficient topic distributions can be learned representing human activities. As a key contribution, we have provided a formal representation of human observations as acquired by a mobile robot, qualitative abstractions to generalise these, and methods to extract discrete features as sequences of observed qualitative relationships. Multiple unsupervised methods to learn low-dimensional representations of human activities have been compared, along with experiments and results to validate our approach. Lastly, the framework has been shown to work well given real-world practical challenges of mobile robotics less often reported on.

We have shown that from multiple human observations in real-world environments, it is possible to learn consistent and meaningful patterns of detailed 3D human body pose sequences using unsupervised learning methods applied to our novel qualitative representation of human observations. Models of human activities are learned with the presence of dynamic objects in a staged static camera set-up dataset (CAD120), as well as a more challenging, real-world, environment with object locations automatically learned. We presented a comparison between our proposed unsupervised methods to a standard supervised method in order to add a perspective to the learning performance. It was shown that the performance of LSA and LDA in these settings is comparable to the supervised technique, without requiring ground truth training labels. Finally, we proposed solutions to interesting and as yet unsolved practical problems in the field of human activity analysis from a mobile robot deployed in real-world environments. We have shown that by using more sophisticated learning methods, it is possible to address some of the practical limitations surrounding life-long human activity learning from a mobile robot. Namely, that manual temporal segmentation is not required and that Variational Bayes inference can be applied for incremental and life-long learning settings.

A possible future direction of research could be to extend this to many months of observational data. This would allow for totally new topics to be discovered, possibly from the robot entering entirely new environments. Also, a “learning-rate” parameter could be updated online given new environments explored by the robot in order to more quickly converge on new human activities being observed. Any topics removed, or not updated, could be considered as the robot “forgetting” a particular human activity.

Open source software has been developed (DOI: qsrlib.readthedocs.org), and a mobile robot dataset has been made openly accessible, (DOI: http://doi.org/10.5518/86). It is our hope that the work presented in this paper will help human activity analysis researchers move away from standard offline approaches applied to static, pre-processed visual datasets. In favour of solutions, such as ours, developed to generalise to real-world environments that mobile robots actually inhabit. These solutions are more practical for the evolution of mobile robotics research in the long-term.

Acknowledgements

The authors greatly thank current and ex-colleagues at the University of Leeds School of Computing Robotics lab, specifically Muhannad Alomari, Majd Hawasly, Yiannis Gatsuolis and James Charles; also, the STRANDS project consortium (http://strands-project.eu), and in particular Nils Bore for his work on object-segmentation. We also acknowledge the financial support provided by EU FP7 project 600623 (STRANDS).

References

[1] P. Duckworth, M. Alomari, Y. Gatsoulis, D.C. Hogg, A.G. Cohn, Unsupervised activity recognition using latent semantic analysis on a mobile robot, in: 22nd European Conference on Artificial Intelligence, ECAI, 2016.
[2] P. Duckworth, M. Alomari, J. Charles, D.C. Hogg, A.G. Cohn, Latent Dirichlet allocation for unsupervised activity analysis on an autonomous mobile robot, in: Proc. of Association for the Advancement of Artificial Intelligence, AAAI, 2017.
[3] P. Duckworth, M. Alomari, N. Bore, M. Hawasly, D.C. Hogg, A.G. Cohn, Grounding of human environments and activities for autonomous robots, in: 26th International Joint Conference on Artificial Intelligence, IJCAI, 2017.
[4] E. Marder-Eppstein, E. Berger, T. Foote, B. Gerkey, K. Konolige, The office marathon, in: IEEE Conference on Robotics and Automation, ICRA, 2010.
[5] N. Hawes, P. Duckworth, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, et al., The strands project: long-term autonomy in everyday environments, IEEE Robot. Autom. Mag. 24 (3) (2017) 146–156.

[6] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[7] S. Thrun, W. Burgard, D. Fox, Probabilistic Robotics, MIT Press, 2005.
[8] PR2 Robot Platform, http://wiki.ros.org/Robots/PR2.
[9] MetraLabs, www.metralabs.com/en.
[10] L. Chen, J. Hoey, C.D. Nugent, D.J. Cook, Z. Yu, Sensor-based activity recognition, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 42 (6) (2012) 790–808.
[11] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2013) 1192–1209.
[12] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.
[13] G. Lavee, E. Rivlin, M. Rudzsky, Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 39 (5) (2009) 489–504.
[14] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vis. Image Underst. 115 (2) (2011) 224–241.
[15] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications, Springer, 2013, pp. 149–187.
[16] J. Aggarwal, L. Xia, Human activity recognition from 3D data: a review, Pattern Recognit. Lett. 48 (2014) 70–80.
[17] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391.
[18] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (1–2) (2001) 177–196.
[19] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.
[20] J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Underst. 114 (8) (2010) 857–864.
[21] S. Wong, T.K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2007.
[22] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[23] S. Savarese, A. DelPozo, J.C. Niebles, L. Fei-Fei, Spatial-temporal correlatons for unsupervised action classification, in: IEEE Workshop on Motion and Video Computing, WMVC, 2008.
[24] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: unsupervised understanding of actions and relations, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[25] P.X. Amorapanth, P. Widick, A. Chatterjee, The neural basis for spatial relations, J. Cogn. Neurosci. 22 (8) (2010) 1739–1753.
[26] J. Chen, A. Cohn, D. Liu, S. Wang, J. Ouyang, Q. Yu, A survey of qualitative spatial representations, Knowl. Eng. Rev. 30 (2015) 106–136.
[27] K.S. Dubba, M.R.d. Oliveira, G.H. Lim, H. Kasaei, L.S. Lopes, A. Tome, Grounding language in perception for scene conceptualization in autonomous robots, in: AAAI Spring Symposium Series, 2014.
[28] J. Tayyub, A. Tavanai, Y. Gatsoulis, A.G. Cohn, D.C. Hogg, Qualitative and quantitative spatio-temporal relations in daily living activity recognition, in: 12th Asian Conference on Computer Vision, ACCV, 2015.
[29] L. Kunze, C. Burbridge, M. Alberti, A. Thippur, J. Folkesson, P. Jensfelt, N. Hawes, Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding, in: IEEE International Conference on Intelligent Robots and Systems, IROS, 2014.
[30] J. Fernyhough, A.G. Cohn, D.C. Hogg, Building qualitative event models automatically from visual input, in: IEEE International Conference on Computer Vision, ICCV, 1998, pp. 350–355.
[31] K. Dubba, M. Bhatt, F. Dylla, D.C. Hogg, A.G. Cohn, Interleaved inductive–abductive reasoning for learning complex event models, in: International Conference on Inductive Logic Programming, Springer, 2011, pp. 113–129.
[32] A.G. Cohn, S. Li, W. Liu, J. Renz, Reasoning about topological and cardinal direction relations between 2-dimensional spatial objects, J. Artif. Intell. Res. 51 (2014) 493–532.
[33] M. Crouse, K.D. Forbus, Elementary school science as a cognitive system domain: how much qualitative reasoning is required? in: Proceedings of Fourth Annual Conference on Advances in Cognitive Systems, 2016.
[34] M. Michael, N. Bernd, Understanding object motion: recognition, learning and spatiotemporal reasoning, in: Special Issue: Toward Learning Robots, Robot. Auton. Syst. 8 (1) (1991) 65–91.
[35] A. Behera, A. Cohn, D. Hogg, Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations, in: Advances in Multimedia Modeling, 2012, pp. 196–209.
[36] M. Alomari, P. Duckworth, D.C. Hogg, A.G. Cohn, Semi-supervised natural language acquisition and grounding for robotic systems, in: Proc. Association for the Advancement of Artificial Intelligence, AAAI, 2017.
[37] M. Alomari, P. Duckworth, D.C. Hogg, A.G. Cohn, Semi-supervised natural language acquisition and grounding for robotic systems, in: AAAI Spring Symposium, 2017.
[38] J.C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2007, pp. 1–8.
[39] G. Bleser, D. Damen, A. Behera, G. Hendeby, K. Mura, M. Miezal, A. Gee, N. Petersen, G. Maçães, H. Domingues, D.C. Hogg, A.G. Cohn, et al., Cognitive learning, monitoring and assistance of industrial workflows using egocentric sensor networks, PLoS ONE 10 (6) (2015) e0127769.
[40] M. Sridhar, A.G. Cohn, D.C. Hogg, Unsupervised learning of event classes from video, in: Association for the Advancement of Artificial Intelligence, AAAI, 2010.
[41] M. Sridhar, Unsupervised Learning of Event and Object Classes From Video, Ph.D. Thesis, The University of Leeds, 2010.
[42] A. Behera, D.C. Hogg, A.G. Cohn, Egocentric activity monitoring and recovery, in: Asian Conference on Computer Vision, ACCV, 2012.
[43] P.E. Agre, D. Chapman, Pengi: an implementation of a theory of activity, in: Proc. Association for the Advancement of Artificial Intelligence, AAAI, 1987.
[44] D. Kirsh, The intelligent use of space, Artif. Intell. 73 (1–2) (1995) 31–68.
[45] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, C. Isbell, A novel sequence representation for unsupervised analysis of human activities, Artif. Intell. 173 (14) (2009) 1221–1244.
[46] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. 43 (3) (2011) 16.
[47] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Worgotter, A. Ude, T. Asfour, D. Kraft, D. Omrcen, A. Agostini, R. Dillmann, Object–action complexes: grounded abstractions of sensory–motor processes, Robot. Auton. Syst. 59 (10) (2011) 740–757.
[48] K. Zampogiannis, K. Ganguly, C. Fermüller, Y. Aloimonos, Extracting contact and motion from manipulation videos, preprint, arXiv:1807.04870.
[49] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, M. Pfeiffer, Prediction of manipulation actions, Int. J. Comput. Vis. 126 (2018) 358–374.
[50] Y. Yang, Y. Li, C. Fermüller, Y. Aloimonos, Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web, in: Association for the Advancement of Artificial Intelligence, AAAI, 2015, pp. 3686–3693.
[51] OpenNI organization, www.openni.org/, 2016.
[52] S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[53] H. Pfister, M. Zwicker, J. Van Baar, M. Gross, Surfels: surface elements as rendering primitives, in: Computer Graphics and Interactive Techniques, 2000.
[54] M. Schoeler, J. Papon, F. Worgotter, Constrained planar cuts-object partitioning for point clouds, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[55] N. Bore, R. Ambrus, P. Jensfelt, J. Folkesson, Efficient retrieval of arbitrary objects from long-term robot observations, Robot. Auton. Syst. 91 (2017) 139–150.
[56] R. Moratz, M. Ragni, Qualitative spatial reasoning about relative point position, J. Vis. Lang. Comput. 19 (1) (2008) 75–98.
[57] M. Delafontaine, A.G. Cohn, N. Van de Weghe, Implementing a qualitative calculus to analyse moving point objects, Expert Syst. Appl. 38 (5) (2011) 5187–5196.
[58] E. Clementini, P. Di Felice, D. Hernández, Qualitative representation of positional information, Artif. Intell. 95 (2) (1997) 317–356.
[59] Y. Gatsoulis, P. Duckworth, C. Dondrup, P. Lightbody, C. Burbridge, QSRlib: a library for qualitative spatial-temporal relations and reasoning, qsrlib.readthedocs.org, Jan. 2016.
[60] Y. Gatsoulis, M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, A.G. Cohn, QSRlib: a software library for online acquisition of qualitative spatial relations from video, in: Workshop on Qualitative Reasoning, at IJCAI, 2016.
[61] P. Duckworth, Y. Gatsoulis, F. Jovan, D.C. Hogg, A.G. Cohn, Unsupervised learning of qualitative motion behaviours by a mobile robot, in: Proc. of the 15th International Conference on Autonomous Agents & Multiagent Systems, AAMAS, 2016.
[62] E.E. Aksoy, A. Abramov, F. Wörgötter, B. Dellen, Categorizing object-action relations from semantic scene graphs, in: 2010 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2010, pp. 398–405.
[63] E.E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, F. Wörgötter, Learning the semantics of object–action relations by observation, Int. J. Robot. Res. 30 (10) (2011) 1229–1249.
[64] J.F. Allen, Maintaining knowledge about temporal intervals, Commun. ACM 26 (11) (1983) 832–843.
[65] H.N. de Ridder, et al., Information System on Graph Classes and their Inclusions (ISGCI), (Interval Graphs) www.graphclasses.org, 2016.
[66] F. Costa, K. De Grave, Fast neighborhood subgraph pairwise distance kernel, in: Proc. of the 26th International Conference on Machine Learning, ICML, Omnipress, 2010, pp. 255–262.
[67] J.M. Dickey, J.M. Jiang, J.B. Kadane, Bayesian methods for censored categorical data, J. Am. Stat. Assoc. 82 (399) (1987) 773–781.
[68] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. USA 101 (suppl 1) (2004) 5228–5235.
[69] S.M. Lynch, Introduction to Applied Bayesian Statistics and Estimation for Social Scientists, Springer Science & Business Media, 2007.
[70] M.I. Jordan, Learning in Graphical Models, vol. 89, Springer Science & Business Media, 1998.
[71] M. Hoffman, F. Bach, D. Blei, Online learning for Latent Dirichlet Allocation, in: Advances in Neural Information Processing Systems, NIPS, 2010.
[72] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, J. Am. Stat. Assoc. 112 (518) (2017) 859–877.
[73] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh, On smoothing and inference for topic models, in: Proc. of Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 27–34.
[74] J. Sung, H. Koppula, B. Selman, A. Saxena, Cornell activity datasets: CAD-60 & CAD-120, http://pr.cs.cornell.edu/humanactivities/, Jun 2014.
[75] P. Duckworth, Unsupervised Human Activity Analysis for Intelligent Mobile Robots, Ph.D. Thesis, University of Leeds, 2017.
[76] A. Rosenberg, J. Hirschberg, V-measure: a conditional entropy-based external cluster evaluation measure, in: EMNLP–CoNLL, 2007.
[77] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary? in: Proc. of the 26th Annual International Conference on Machine Learning, ICML, 2009.
[78] J. Munkres, Algorithms for the assignment and transportation problems, J. Korea Soc. Ind. Appl. Math. 5 (1) (1957) 32–38.
[79] A. Gelman, J. Carlin, H. Stern, D. Rubin, Bayesian Data Analysis, vol. 2, Chapman & Hall/CRC, Boca Raton, FL, USA, 2014.