Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Latent Dirichlet Allocation for Unsupervised Activity Analysis on an Autonomous Mobile Robot

Paul Duckworth, Muhannad Alomari,
James Charles, David C. Hogg, Anthony G. Cohn
School of Computing, University of Leeds, Leeds, LS2 9JT, UK.
{p.duckworth, scmara, j.charles, d.c.hogg, a.g.cohn}@leeds.ac.uk

Abstract

For autonomous robots to collaborate on joint tasks with humans they require a shared understanding of an observed scene. We present a method for unsupervised learning of common human movements and activities on an autonomous mobile robot, which generalises and improves on recent results. Our framework encodes multiple qualitative abstractions of RGBD video from human observations and does not require external temporal segmentation. Analogously to information retrieval in text corpora, each human detection is modelled as a random mixture of latent topics. A generative probabilistic technique is used to recover topic distributions over an auto-generated vocabulary of discrete, qualitative spatio-temporal code words. We show that the emergent categories align well with human activities as interpreted by a human. This is a particularly challenging task on a mobile robot due to the varying camera viewpoints which lead to incomplete, partial and occluded human detections.

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 (Marder-Eppstein et al. 2010; Hawes et al. 2016). 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, i.e. 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. Understanding what human activities occur in which regions and when, allows the robot to adjust its own behaviour, or assist in a task it believes is being undertaken.

The contribution of this work is in unsupervised activity analysis on continuous, unsegmented video sequences using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). We focus on simple human activities for daily living observable by a mobile robot in a human work place environment. Our framework outperforms recent work in the literature which uses manually segmented videos, where a single activity instance is present in each (Duckworth et al. 2016). We generalise that work with a method allowing for use of full length observations recorded from a mobile robot with no human filtering or segmentation. We propose a probabilistic generative approach that models each observation as a mixture of latent topics, where each topic recovered is a distribution over a vocabulary of discrete descriptors and is considered as a human activity class. Removing the requirement for temporal segmentation of the recorded observations, our robot is able to more quickly access larger quantities of data which otherwise would need human annotation to select interesting sequences of frames. Our work moves away from using a standard dataset, where each example contains a single pruned activity instance, to a more realistic setting where the robot is not told which sub-sequence of the observations to learn from. This loosely translates as removing the assumption that humans continuously perform interesting activities when being observed. We replace it with a more reasonable assumption that a human detection is modelled as a probabilistic mixture over an underlying number of latent topics, where some topics can be considered "interesting" human activities, and others more "mundane"; the definitions of these rely on the task-specifics of the mobile robot.

Further challenges when using human activity data captured from a 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 and noisy observations. ii) Each observed activity is likely to be carried out with particular variations, e.g. opening a door with different hands. Our framework helps alleviate these problems in two phases; first by utilising a state-of-the-art human pose estimator to improve the quality of observations, and secondly we use a *qualitative spatial representation* (QSR), which abstracts quantitative data to a discrete set of qualitative values, thus converting somewhat noisy observations of arbitrary spatial positions into semantic low level actions. 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.

Our methodology consists of first detecting and tracking humans from a mobile robot, then abstracting their pose estimates using multiple QSRs and encoding them as the occurrences of discrete qualitative descriptors. 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 distribution over the discrete vocabulary. To do this we perform LDA, a three-layer hierarchical Bayesian model where each observation is modelled as a finite mixture over an underlying set of topics, and each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities.

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 real-world human activity classes. In the following sections we provide formal details of the human pose estimates acquired by the robot, the qualitative abstractions used to generalise the observations and extract discrete features; a description of the generative learning process, experiments and their results; and our conclusions.

Related Work

Activity recognition from visual data is a mature sub-field of artificial intelligence. For a comparison on general activity recognition techniques, the reader is pointed to survey papers which cover the topic using RGB cameras (Turaga et al. 2008, Lavee et al. 2009, Weinland et al. 2011) and 3D RGBD cameras (Ye et al. 2013; Aggarwal and Xia 2014). Many common techniques perform supervised learning, where each data sample requires manual hand annotation with a ground truth label. This is not a feasable solution for a long term autonomous mobile robot.

More task-appropriate are unsupervised techniques which do not require time-consuming, offline manual annotations. Previous works have used probabilistic Latent Semantic Analysis (pLSA) and LDA for learning human activity categories in an unsupervised setting: authors have used low-level Space-Time Interest Point (STIP) features (Niebles et al. 2008); local shape context descriptors on silhouette images (Zhang and Gong 2010); and a combination of semantic and structural features (Wong et al. 2007, Liu et al. 2008). However, each has been performed without the variability of a mobile robot's frame of reference, and are restricted to a single person, performing a segmented action during the training phase, unlike our dataset.

An unsupervised approach, coupled with a qualitative representation, has been used in (Sridhar et al. 2010), where a qualitative spatial calculus is used to encode continuous videos containing aeroplane turnaround scenes. However, their videos consist of slow moving objects from a static camera frame of reference. Most similar is the work (Duckworth et al. 2016), where extracted qualitative features are used to encode human observations from a mobile robot. However, their approach requires a dataset of sample video clips, each containing a *single* "interesting" human activity sequence. They propose a discriminative learning approach where each segmented video is modelled as a single latent concept recovered from the dataset. In our work we model an observation as a probabilistic mixture of topics, removing the requirement for human temporal segmentation.

Knowledge Representation

Our aim is to understand human activities from long term observations of a human populated environment and for an autonomous robot to obtain a conceptual model of activities taking place. This level of understanding has the potential to be used by the robot to collaborate on joint tasks, and have a shared understanding of an observed scene (though this is beyond the scope of this paper). In this section, we first introduce the input data captured from our mobile robot, followed by details about the qualitative representation used to abstract the data, and finally we describe an auto-generated code book which is used as a discrete vocabulary resulting in a term-document matrix, similar to information retrieval settings.

Human Pose Estimates

Our 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. We represent the human pose estimates as ROS messages, where a single detected body joint location is represent as an (xyz) Cartesian coordinate in the camera coordinate frame along with the corresponding (xyz) position translated into the global map coordinate frame, i.e. $j = (id, x, y, z, x_m, y_m, z_m)$. The map frame coordinate relies on the robot being well localised within the map frame which is achieved by the robot being static during recordings. A *human pose* then comprises of a collection of body joint locations, i.e $p = [j_1, j_2, \ldots, j_{15}]$. For each human detected by the robot, we obtain a sequence of human poses over a time series of detections. 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.

Figure 1 (right) shows the Scitos A5 mobile robot used to observe the environment and (left) one section of its global map; semantically labelled with key regions and landmark objects in advance. Brightly coloured CAD (Blender) models can be seen where semantic objects are positioned in the environment (best viewed in colour), and are used to calculate qualitative spatial relations in the next section. Our framework would extend trivially to include dynamic objects detected in real-time by the robot.

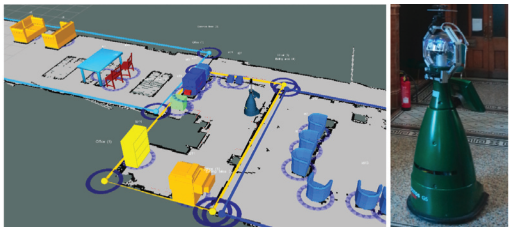


Figure 1: (left:) Semantic global map showing the kitchen region boundary in yellow. (right:) Scitos A5 mobile robot.

Qualitative Pose Sequences

Abstracting human pose sequences using 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 (xyz) coordinates of their hand or head are not important; it is the relative movement which captures the possible "waving" activity. Moreover, human activities often occur over differing durations. For example, the activity of standing still (where pose estimates are easy to estimate) can occupy a few thousand frames, whereas a more complex task such as opening the fridge can take less than a second (30 frames) and contain noisy pose estimates due to occlusions. The person also may only appear in the robot's field of view for a few seconds, during which limited time the pose estimates can be noisy and inaccurate. Conversely, a person might be performing a static activity and detected for thousands of frames (poses). This variation is a major difficulty in mobile robotics, which abstracting the data into a qualitative space helps to alleviate.

In this paper, we abstract human poses using three QSRs computed by a publicly available ROS library we co-authored (Gatsoulis et al. 2016b; 2016a): 1) Ternary Point Configuration Calculus (TPCC) qualitatively describes the spatial arrangement of a point, relative to two others, i.e. it describes the *referent*'s position relative to the plane created by connecting the *relatum* and *origin*, values are triples of $\langle$ { front, back }, { left, right, straight }, { distant, close } $\rangle$ (Moratz and Ragni 2008); 2) Qualitative Trajectory Calculus (QTC) represents the relative motion of two points with respect to the reference line connecting them, and is computed over consecutive timepoints (Delafontaine et al. 2011); it defines the following three qualitative spatial relations between two objects [1] 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). 3) Qualitative Distance Calculus (QDC) expresses the qualitative Euclidean distance between two points depending on defined distance thresholds (Clementini et al. 1997). 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 changes in the relations can help explain relative motion. An illustration of the three QSRs can be seen in Figure 2.

The three QSRs are computed from (xyz) data of particular body joint positions over a series of timepoints. That is, a human pose sequence S is abstracted into multiple sequences of qualitative relations (one per calculi being used) and represented as a QSRLib response message (implementation details of which are given in the Experiments section). We believe these three QSR are appropriate to qualitatively describe the kind of human activities we are interested in; however, it is not an exhaustive list and other qualitative calculi could be explored (Chen et al. 2015).

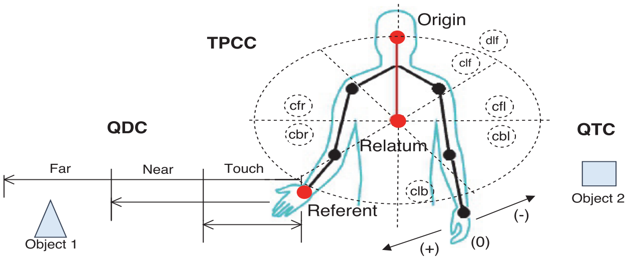


Figure 2: QSRs: (left:) QDC (relative distance) between right hand-object 1. (centre:) TPCC system between right hand-(relatum-origin) plane (for the full TPCC system see (Moratz and Ragni 2008). (right:) QTC (relative motion) between left hand-object 2 pair.

Extracting Qualitative Code Words

Many human activities observed by a robot can be explained by a sequence of primitive actions over a duration of time. Here, we describe how we temporally abstract over a time sequence of qualitative relations to generate a vector over a discrete code book (vocabulary) for each observed activity.

We first abstract the sequence of body joint positions S into a sequence of QSR values Q (one per calculi used), then compress repeated values to obtain an interval representation I of an activity [2]. For example, if the right hand appears to be moving towards the head (QTC relation: '$-$'), for τ consecutive frames and then is static (0) with respect to the head for τ' further frames, we compress this into an interval representation consisting of two intervals: $I_{(\text{head, Rhand})} = \{i_1, i_2\} : i_1 = \{`-', (0, \tau - 1)\}$ and $i_2 = \{`0', (\tau, \tau + \tau' - 1)\}$. Each interval i maintains the QSR value (or set of values; one per calculi used) in addition to the start and end timepoints, see first row of Figure 3 (top). An interval representation I of a complete human pose sequence contains a row for each pair of body joints or landmarks.

Given a set of human pose sequences, encoded from a number of observed activities, we compute an interval representation I for each and extract a set of unique qualitative features (*code words*) which are used to describe the observations. We compute an *interval graph* (de Ridder et al. 2016) for each interval representation I in our dataset by applying a subset of Interval Algebra (IA) (Allen 1983) to abstract the temporal relations between the observed intervals. IA is used to represent and reason with temporal intervals and defines 13 qualitative relations (for a complete list of the relations refer to (Allen 1983)). An example interval graph can be seen in Figure 3 (bottom), which encodes both rows present in Figure 3 (top). Here, a node i' is used to represent an interval i and contains only the QSR value (or set of values) that hold between the objects of that interval, and the objects themselves. The exact timepoints are not explicitly depicted in the node, e.g. node i'_1 in Figure 3 (bottom) contains (*head, Rhand*,'$-$') information temporally abstracted

[1] Note that objects are abstracted to their centroids when computing these QSRs.

[2] In the literature, this is closely related to a *Qualitative Spatial Temporal Activity Graph* (QSTAG) (Gatsoulis et al. 2016b), and similar to the representation used in (Duckworth et al. 2016).

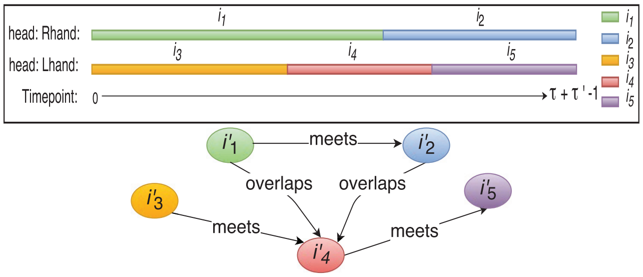


Figure 3: (top:) Interval representation of two pairs of body joints. (bottom:) Interval Graph. (Best viewed in colour.)

from i_1. Taking any two intervals, it is possible to calculate the temporal relation which holds between them using IA and is represented as a labelled, directed arc between the corresponding nodes, as seen in Figure 3. For example, the IA relation that holds between the i_1 and i_2 intervals is "meets"; and between i_1 and i_4 is "overlaps". Nodes are only linked if their intervals are temporally connected, i.e. there exists no temporal break between a pair of intervals, hence there are no arcs with IA relations *before* or *after*. Note, where two intervals occur at the beginning or end of the video clip (and therefore beginning or end of the interval representation), e.g. i_1 and i_3, there is insufficient temporal information to abstract over the intervals and there is no arc between the corresponding nodes in the interval graph, e.g. i'_1 and i'_3.

From the set of all interval graphs, we define a *code book* V as the set of unique code words $V = [\gamma_1, \gamma_2, \dots]$, which are extracted by enumerating all paths through all interval graphs, up to and including some fixed path-length k. For example, for $k = 2$ the unique code words extracted from the interval graph shown in Figure 3 (bottom) are generated by taking all paths of length 1 and 2, i.e. { i'_1, i'_2, i'_3, i'_4, i'_5, (i'_1 meets i'_2), (i'_1 overlaps i'_4), (i'_2 overlaps i'_4), (i'_3 meets i'_4), (i'_4 meets i'_5)}. The length of the code book V depends upon the number of unique paths and is affected by the number of objects encoded, the QSR calculi used along with their values, plus the path-length k. The experimental details of these choices are given later. The code words generated using this technique represent combinations of qualitative relation intervals specifically observed within the data and is akin to observing a particular set of words (or n-grams) in a document. This makes it an efficient and intuitive method for representing observed human activities.

For each observed activity, we encode a sparse vector (with length $|V|$) describing the frequency of each code word in that observation and call this an observed activity "histogram". This is similar to a *Bag of Words*, where the code words extracted from the video ignores positional arrangement.

Latent Dirichlet Allocation

In this section we draw comparisons with document analysis and use Latent Dirichlet Allocation (LDA); a generative probabilistic model of a collection of discrete data and use Collapsed Gibbs Sampling (Lynch 2007) which extracts a set of interesting topics from the corpus. LDA extends Latent Semantic Indexing (LSI) (Deerwester et al. 1990) and probablistic LSI (Hofmann 2001).

Similarly, we use LDA to uncover human activities (topics) from videos (documents) using our encoding as activity histograms over a code book (discrete vocabulary). In this setting each human observation is analogous to a document and modelled as a random mixture over latent topics. Each topic is a latent multinomial variable and characterised by a distribution over a vocabulary of code words. This framework allows us to model each observation as a mixture of topics, where each code word is sampled from a multinomial distribution over the vocabulary. This translates as allowing a mixture of activity classes to be encoded within the same observation, removing the requirement for temporal segmentation of human observations into "interesting" sequences and hence generalising the work in (Duckworth et al. 2016).

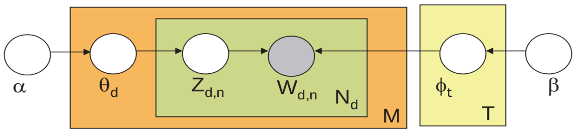


Figure 4: Graphical model representation of LDA using plate notation. Nodes represent random variables, links between nodes are conditional dependencies, plates are replicated components, and shaded nodes are observations.

LDA is a probabilistic topic model that generates a set of M documents D, from a set of underlying T topics, where each topic is a distribution $\phi = \{\phi_1, \dots, \phi_T\}$ over the vocabulary of unique code words V (extracted in the previous section). Each document $d \in D$ consists of a set of N_d words W_d, and is composed of a distribution $\theta = \{\theta_i : i \in D\}$ over the topics. Figure 4 shows a graphical model representation of the three-layer hierarchical Bayesian model using plate notation. The variables ϕ and θ, as well as $\mathbf{z}$ (the assignment of word tokens to topics) are the three sets of latent variables that we would like to infer. We briefly introduce the main random variables here, (for $d \in D$ and $n \in W_d$):

i. θ_d: topic proportions of document d,
ii. $Z_{d,n}$: per-word topic assignment,
iii. $W_{d,n}$: observed words (shaded grey),
iv. $\phi_t \in \phi$: word proportions of topic t,
v. α, β: Dirichlet hyperparameters.

The generative process can be characterised by first sampling a Dirichlet(α) distribution over topics, then sample a topic, finally sampling a word from that topic. The three-layers of the Bayesian model are described by the parameters (α, β) which are corpus level parameters. The variables θ_d (for $d \in D$) are document level parameters, sampled once per document, and finally, the variables $z_{d,n}$ and $w_{d,n}$ are word-level and sampled once for each word in a document. Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta).$$

Given this generative model, our aim is to determine the latent topics based on the observed words that appear in our observations. This translates as computing the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)},$$

which is intractable, and so we use Collapsed Gibbs Sampling; an approximate inference algorithm based on the Monte Carlo Markov Chain (MCMC) technique where the idea is to generate posterior samples from its conditional distribution.

Experimental Procedure

We begin this section by describing the observed data captured and recorded from an autonomous mobile robot. Then, we describe three experimental settings intended to highlight the improvement of our methodology over (Duckworth et al. 2016), and present results in the next section.

Dataset

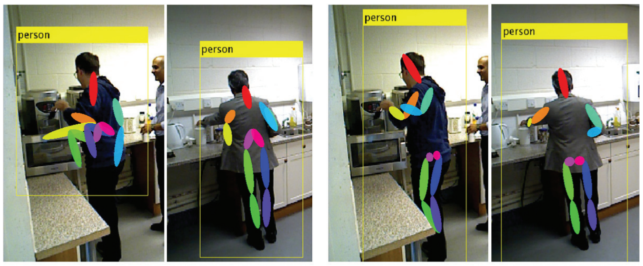
To validate our methodology, we use a publicly available dataset recorded from an autonomous mobile robot observing human environments over a one week duration [3]. The robot, fitted with a headmounted ASUS Xtion Pro-Live RGBD camera and OpenNI2 human pose estimator, was tasked with patrolling pre-defined waypoints and observing a kitchen and student common area with varying view points. The robot observed 287 individuals during the one week process and created a human pose sequence for each. These sequences contain arbitrary number of poses with high variance; mean (μ) and std $(\sigma) = (513, 588)$.

For the purpose of obtaining a ground truth, each recorded sequence was temporally segmented by volunteers into multiple shorter sequences containing only a single activity class. The following is a list of the activity classes annotated, along with the number of occurrences: Microwave food (19); Take object from fridge (81); Use the water cooler (26); Use the kettle (70); Take paper towel (45); Throw trash in bin (65); Wash cup (82); Use printer interface (35); Take printout from tray (24); Take tea/coffee (35); Opening double doors (11). The granularity of the activity schema was determined by the data available from the robotic vision component; in particular no object tracker was available, and hand tracks were unreliable. A total of 493 individual activity instances were segmented. These sequences are much shorter, $(\mu, \sigma) = (137, 191)$, and temporally focused on the activity instance taking place. We consider these sequences as each containing a single "interesting" activity, as defined and segmented by the volunteer annotators.

Note, 77 (of 287) observations were deemed to contain none of the above activity classes by the annotators, and given the label "NA". This is a considerable percentage of noise and provides a major difficulty when no manual segmentation or filtering of the data is provided.

Pose Sequence Improvements

A common approach to capture human pose estimates is to use the OpenNI tracker (OpenNI organization 2016) to detect multiple persons and infer their 3D pose in real-time from the sensors' depth stream. For our work however, 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 interaction, see Figure 5(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 (in real-time on the CPU). Secondly, person bounding boxes and the RGB frame are fed as input into a state-of-the-art convolutional neural network pose machine (CPM) (Wei et al. 2016) to better estimate the 2D human pose (on a midrange GPU). Subsequently, we take the improved (xy) coordinates of body joint positions from the CPM, and the depth coordinate (z) from the original OpenNI detection, see Figure 5(b).



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

Figure 5: Improved human pose estimates. (View in colour.)

Experimental Setup

We conducted three experiments using the above dataset; each differs in the level of temporal segmentation of the observations provided to the system.

Experiment 1: We first present a direct comparison between our methodology and that of the most similar literature (Duckworth et al. 2016), where Latent Semantic Analysis (LSA) is used to recover semantic concepts from a term-document matrix. The temporally segmented human activities dataset of 493 activity instances is used with the CPM improved pose estimates and we present results using only the OpenNI pose estimates in brackets. We validate our learned topics by comparing the cluster metrics obtained when using a supervised algorithm, as an upper bound; a linear-SVM is trained using 5-fold cross validation and has access to the ground truth labels during the learning process. A comparison to a standard unsupervised clustering algorithm, k-means, and random chance is also presented; each as an average over 10 repeats.

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

Experiment 2: By concatenating multiple segmented clips together, we highlight the probabilistic mixing properties of our LDA model. This set of video clips is formed by taking the temporally segmented sequences used in experiment 1, and concatenating them back together, excluding the temporally surrounding poses. This results in a set of, possibly discontinuous, video sequences where only the annotated "interesting" activity classes take place and the other poses are removed. There are 210 such sequences containing 2.3 segmented clips on average (max=10), where each sequence now has multiple associated ground truth labels. In this setting, the same sequences of data are provided as experiment 1, however multiple sequences are now present in each observation and requires modelling the observation as a mixture of topics. We also present a comparison to (Duckworth et al. 2016) using these video sequences.

Experiment 3: Finally, in this experiment no temporal segmentation of the observations is provided to the system, and the topics generated are described with respect to the multi-labelled ground truth activities taking place. The 287 full length human pose sequences contain the above "interesting" sequences, but also contain many more interactions, e.g. people walking, standing, chatting, and are much longer, (μ, σ) = (513, 588) and more varied. For this reason, we do not expect the learned topics to match one-to-one to the ground truth labels annotated by the volunteers. However, we present a mapping between the learned topics and the annotated labels, which help us understand what the topic distributions refer to. This leads to the interesting question of "what constitutes an activity class"?

Implementation Details

Here we describe the implementation details used in the above experiments. Computing the QSRs from the human pose sequences is performed in a two stage process. First abstracting the person's relative body joint positions in the camera frame of reference, and secondly, abstracting the person relative to pre-defined semantic landmarks in the map coordinate frame of reference. Each observation (regardless of the segmentation provided), is represented as a human pose sequence $S_m = [p_1, p_2, \dots p_t]$, of body joint positions p_i.

To calculate TPCC relations for a pose, we fix the *origin* and *relatum* to the *head* and *torso* joint positions respectively to generate a person's "centre line", see Figure 2 (right). A sequence Q_{cam}, of length t, is produced containing TPCC relations between this plane and the left/right hands, knees and shoulder joint positions. Similarly, to encode a person's position in the global map frame we use QDC and QTC calculi, see Figure 2 (left). These QSRs are used to describe the relative position of the person's torso body joint and left/right hand positions, relative to a set of 12 landmark objects in the "kitchen" semantic region, see Figure 1. A sequence Q_{map}, length $t-1$, of QDC and QTC pairs is produced [4]. Since the landmarks are static, we use the QTC$_{B11}$ variant of QTC (Delafontaine, Cohn, and Van de Weghe 2011). The threshold values used for the QDC relations are: *touch* [0-0.25m], *near* (0.25-0.5m], *medium* (0.5-1.0m] and *ignore* (>1m][5]. For example, in an observed activity where a person opens a fridge, a possible sequence for the *hand-fridge* pair in Q_{map} is: [*('+', 'Near'), ('+', 'Near'), ('+', 'Medium'),* $\dots$], where $+$ is from the QTC calculi and Near, Medium from QDC.

For each sequence Q_{cam} and Q_{map}, we apply a median filter (window size = 10 frames), which smooths rapid flipping between relations, owing to noise in the pose sequences. We then create an interval representation and interval graph for each, by compressing repeated relations and abstracting temporally using IA, as described above. Since the number of paths increases exponentially with the number of interval nodes (each encoding two objects), we use path-length $k = 4$ and restrict the nodes on a path to encode *at most* 4 different objects in total. Enumerating all paths, we generate a single code book V which contains relations from TPPC, QTC and QDC. We apply a binary low-pass filter to remove any code words that occur in fewer than 5 observed activities, which we regard as noise. Note $|V|$ varies between the three experiments due to the increase in observed unique interval paths when no segmentation is performed (plus the effect of the binary low pass filter), i.e. in the three experiments $|V|$ = 4565, 4337 and 18,001, respectively. Finally, we create an activity histogram for each recorded activity, consider them documents and perform LDA with hyperparameters $\alpha, \beta = (0.5, 0.03)$.

Results

In this section we provide results for the three experimental settings described above. Using an unsupervised learning framework, it is not possible to map the learned topics (or clusters) directly to each activity class in the ground truth labels. This can be a many-to-many mapping, especially when dealing with highly unbalanced classes. Therefore, we provide results using popular clustering metrics where the aim is to generate clusters composed of the same activity class label, and that all instances of a class are present in the same cluster. For this purpose we use the two metrics, V-measure (Rosenberg and Hirschberg 2007) and (Normalised) Mutual Information (NMI) (Vinh et al. 2009). The 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).

Experiment 1 results are presented in Table 1, along with a comparison to the current state-of-the-art technique, Latent Semantic Analysis (LSA) (Duckworth et al. 2016). Our method outperforms LSA when both methods are provided with temporally segmented activity clips, containing a single activity instance in each. Results are provided using only

[4] QTC relies on pairs of consecutive poses, so we remove the QDC value at $t = 1$ to obtain $t - 1$ pairs.

[5] We intentionally do not encode "ignore" intervals, creating a sparse interval representation and leading to a more efficient process.

the OpenNI pose estimates (in brackets), and performance improves when combined with the 2D CPM human pose estimates. To calculate the above metrics, only the highest topic proportion (> 0.5) is selected, and 383 observations are classified. As in the previous work, the number of topics is set to the number of activity classes in the dataset $T = 11$. Both techniques significantly out perform k-means and uniform random assignment (using 11 clusters), and the results of a supervised method (linear-SVM) which has access to the ground truth labels during the training process, and could be regarded as an upper bound on the possible performance.

Metric	LDA	LSA	k-means	random	*SVM*
V-measure	**0.69** (0.64)	0.54	0.27	0.05	*0.71*
NMI	**0.69** (0.64)	0.53	0.29	0.05	*0.71*
Accuracy	N/A	N/A	N/A	0.11	*0.77*

Table 1: Results using temporally segmented video clips: LDA (OpenNi/CPM vs OpenNi only in brackets), compared against LSA; unsupervised k-means clustering; random chance; and a supervised SVM as an upper bound.

Experiment 2 uses the concatenated sequences of clips, described above, and translates as multiple activities occurring in each observation (and nothing else). Figure 6 (left) presents a cosine similarity matrix of the 11 learned topic distributions when using these concatenated clips compared against the topics learned in experiment 1 (using temporally segmented activity clips). The strong diagonal indicates a one-to-one mapping between the two recovered sets of topic distributions and validates the mixture assumption of LDA, i.e. longer observations do not require temporal segmentation and that similar topic distributions are recovered from observations containing a mixture of activities. The average cosine similarity between the two sets of topic proportions is 0.82. Further, the average cosine similarity of the topic distributions using LSA is 0.60, which shows our method is able to handle a mixture of classes in each observation better.

In **Experiment 3** we provide no temporal segmentation and as a result the observations are much longer and considerably more varied. For this reason, there is not a one-to-one mapping present between the topics recovered and the activity class labels. In this setting, we sum the topic proportions of all documents which contain each annotated ground truth label (with low-pass filter > 0.5). This is represented as a matrix and shown in Figure 6 (right). For brevity, we merge the two "printing" rows together (these are not distinguishable), and remove the "opening double door" row since it has the least number of instances (11). We can see that the majority of the learned topic distributions correlate to a single or pair of activity class labels, e.g. topic ID 6 correlates highly with "take tea/coffee" and likewise ID 3 to "washing up". Topics such as ID 1 relate to a mixture of human labels such as "using kettle" and "using paper towel". This is intuitive, and based upon the activities that are often observed together, e.g. washing and drying a mug. However, some classes are being confused based upon their spatial arrangement in the environment, e.g. the microwave is 30cm away from the water dispenser. These objects are not commonly used together, however topic ID 10 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/water cooler; this is what this topic distribution represents. By using a low-pass threshold it is unclear what topic ID 0 or 7 relate to. However, from manual inspection it is clear that both topics relate to behaviours which occur across many labels below the required thresold; thus the set of human activities in this dataset can be distinguished using only 9 of the 11 topic distributions.

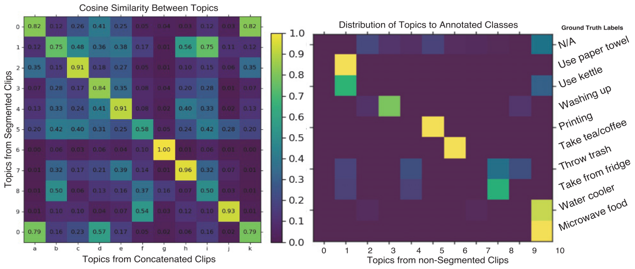


Figure 6: (left:) Similarity of 11 learned topics, segmented (Ex.1) vs concatenated (Ex.2) video sequences. (right:) Topic-proportions per activity class, using no temporal segmentation (Ex.3).

Conclusion

This paper focuses on learning human activities from long term mobile robot observations in an unsupervised setting. Our methodology is capable of improving upon and generalising our previous work. It first abstracts pose estimates of detected people using qualitative representations which help alleviate challenges arising from activity recognition on a mobile robot, in particular varying view points and noisy/partial detections. It then auto-generates a vocabulary of discrete qualitative spatio-temporal code words which is used to encode observations analogous to information retrieval settings, where observations are considered documents. A probabilistic topic model (LDA) is used to recover latent topics which are considered to represent human activities. Our methodology improves upon recent literature on learning human activities from a standard human activity dataset with large intra-class variations present. Further, we generalise the method for use on a mobile robot by performing learning using no temporal segmentation or manual filtering of observations. This is achieved by using a probabilistic generative approach where each observation is modelled as a mixture of latent topics.

Given the complex nature of human environments, one limitation of our method is that the learned topic distributions cannot evolve over time (if activities evolve), and that the number of topics should be set in advance. A possible direction to look into is the use of Dynamic Topic Models (Blei and Lafferty 2006). It is also the case that object-specific terms in the QSRs would benefit from further abstraction, allowing us to learn more general, object-independent topics, such as "picking" and 'carrying".

Acknowledgments

We thank colleagues in the School of Computing Robotics Lab, in particular Majd Hawasly, and in the STRANDS project consortium (http://strands-project.eu) for their valuable input. We also acknowledge the financial support provided by EU FP7 project 600623 (STRANDS).

References

Aggarwal, J., and Xia, L. 2014. Human activity recognition from 3D data: A review. *Pattern Rec. Letters* 48:70 – 80.

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

Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *Proc. 23rd Int. Conf. on ML*, 113–120. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of ML research* 3:993–1022.

Chen, J.; Cohn, A. G.; Liu, D.; Wang, S.; Ouyang, J.; and Yu, Q. 2015. A survey of qualitative spatial representations. *The Knowledge Engineering Review* 30:106–136.

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

de Ridder, H. N., et al. 2016. Information System on Graph Classes and their Inclusions (ISGCI). www.graphclasses.org (Interval Graphs).

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391.

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

Duckworth, P.; Alomari, M.; Gatsoulis, Y.; Hogg, D. C.; and Cohn, A. G. 2016. Unsupervised activity recognition using latent semantic analysis on a mobile robot. In *22nd European Conf. on. Artificial Inteligence (ECAI)*.

Gatsoulis, Y.; Alomari, M.; Burbridge, C.; Dondrup, C.; Duckworth, P.; Lightbody, P.; Hanheide, M.; Hawes, N.; and Cohn, A. G. 2016a. QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video. In *Workshop on Qualitative Reasoning (QR16), at IJCAI-16*.

Gatsoulis, Y.; Duckworth, P.; Dondrup, C.; Lightbody, P.; and Burbridge, C. 2016b. QSRlib: A library for qualitative spatial-temporal relations and reasoning. qsrlib.readthedocs.org.

Hawes, N.; Burbridge, C.; Jovan, F.; Kunze, L.; Lacerda, B.; Mudrová, L.; Young, J.; Wyatt, J. L.; Hebesberger, D.; Körtner, T.; Bore, R. A. N.; Folkesson, J.; Jensfelt, P.; Beyer, L.; Hermans, A.; Leibe, B.; Aldoma, A.; Faulhammer, T.; Vincze, M. Z. M.; Al-Omari, M.; Chinellato, E.; Duckworth, P.; Gatsoulis, Y.; Hogg, D. C.; Cohn, A. G.; Dondrup, C.; Fentanes, J. P.; Krajník, T.; Santos, J. M.; Duckett, T.; and Hanheide, M. 2016. The STRANDS project: Long-term autonomy in everyday environments. *EEE Robotics and Automation Magazine* In Press.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 42(1-2):177–196.

Lavee, G.; Rivlin, E.; and Rudzsky, M. 2009. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39(5):489–504.

Liu, J.; Ali, S.; and Shah, M. 2008. Recognizing human actions using multiple features. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Lynch, S. M. 2007. *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media.

Marder-Eppstein, E.; Berger, E.; Foote, T.; Gerkey, B.; and Konolige, K. 2010. The office marathon. In *IEEE Conf. on Robotics and Automation (ICRA)*.

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

Niebles, J. C.; Wang, H.; and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. *Int. Journal of Computer Vision* 79(3):299–318.

OpenNI organization. 2016. www.openni.org/.

Rosenberg, A., and Hirschberg, J. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2010. Unsupervised learning of event classes from video. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*.

Turaga, P.; Chellappa, R.; Subrahmanian, V. S.; and Udrea, O. 2008. Machine recognition of human activities: A survey. *IEEE Trans. on Circuits and Systems for Video Technology* 18(11):1473–1488.

Vinh, N. X.; Epps, J.; and Bailey, J. 2009. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proc. of the 26th Annual Int. Conf. on Machine Learning*.

Wei, S.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Weinland, D.; Ronfard, R.; and Boyer, E. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding* 115(2):224–241.

Wong, S.; Kim, T. K.; and Cipolla, R. 2007. Learning motion categories using both semantic and structural information. In *IEEE Conf. on Computer Vision and Pattern Rec. (CVPR)*.

Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; and Gall, J. 2013. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer. 149–187.

Zhang, J., and Gong, S. 2010. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding* 114(8):857–864.