Unsupervised Learning of Human Activities by a Mobile Service Robot *

Paul Duckworth, Muhannad Alomari, Yiannis Gatsoulis, David C. Hogg, Anthony G. Cohn
School of Computing, University of Leeds, Leeds, UK
{p.duckworth, scmara, y.gatsoulis, d.c.hogg, a.g.cohn}@leeds.ac.uk

Abstract

We show that by using qualitative spatio-temporal abstraction methods, we can learn common human movements and activities from long term observation by a mobile service robot. Our framework encodes multiple qualitative abstractions of RGBD video from detected activities performed by a human as encoded by a skeleton pose estimator. Analogously to informational retrieval in text corpora, we use Latent Semantic Analysis (LSA) to uncover latent, semantically meaningful, concepts in an unsupervised manner, where the vocabulary is occurrences of qualitative spatio-temporal features extracted from video clips, and the discovered concepts are regarded as activity classes. The limited field of view of a mobile robot represents a particular challenge, owing to the obscured, partial and noisy human detections and skeleton pose-estimates. We show that the abstraction into a qualitative space helps the robot to generalise and compare multiple noisy and partial observations in a real world dataset and that a qualitative vocabulary of latent activity classes can be recovered.

1 Introduction

Autonomous mobile service robot platforms which are to operate in dynamic human populated environments have a need to update their own knowledge of the world based upon their observations and interactions with humans, using unsupervised learning frameworks. Such robots can be adaptable to their surroundings, the time of day, or to a specific task being observed. Understanding what activities occur in which regions and when, allows the robot to adjust its own behaviour and assist if it can.

The aim of our work is to understand human activities taking place from long term observation of real world scenarios. We present an unsupervised, qualitative framework for learning human activities in a real world environment, which is deployed on an autonomous mobile robot platform. The challenge is to learn semantically meaningful human activities from observing multiple people performing everyday activities, and learn a vocabulary to describe them.

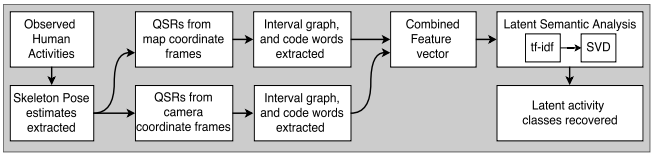

Figure 1: System architecture.

Two main challenges are: 1) The robot's on-board sensors only grant our system a partial and mobile view of the world, and obtains incomplete and noisy observations. 2) Each observed activity is likely to be carried out with particular variations, e.g. opening a door with opposite hands. Our framework helps alleviates these problems by using *qualitative spatial representations* (QSRs) as an effective abstraction method. This allows the system to compare observations based upon key qualitative features and learn common patterns in an abstracted space, instead of their metric details.

Briefly, the main steps of our learning methodology are shown in a flow diagram in Figure 1 which consist of:

a) Detection of humans using an RGBD sensor on a mobile robot; the system estimates and tracks the skeleton pose.
b) Transformation of the skeleton pose estimates into a qualitative space; qualitative calculi are used to abstract the metric coordinates of the person.
c) A code book of unique qualitative features is defined by extracting paths through an interval graph representation.
d) Finally, we use Latent Semantic Analysis (LSA) to recover semantically meaningful (latent) concepts which exist in the feature space.

Our learning methodology consists of first encoding a video clip (of a detected human) as the occurrences of its qualitative spatio-temporal features which is used as a feature vector to represent the activity. We analyse the collection of feature vectors analogously to a corpus of text documents, looking for semantically similar structures and features (words) that commonly co-occur; these define a vocabulary over which to describe human activities. Instead of documents containing multiple words, our video clips consist of multiple qualitative spatio-temporal features. The activity taking place in the video is akin to learning the document's context.

*This paper is a shortened version of a paper to appear in 22nd European Conference on Artificial Intelligence (ECAI) 2016.

2 Related Work

There is a considerable amount of literature which aims to understand, recognise or detect human motions and activities from video data; accompanied by multiple survey papers on the topic [Lavee *et al.*, 2009; Weinland *et al.*, 2011] and more recently, from RGBD data [Aggarwal and Xia, 2014]. Activity recognition using a mobile robot has previously been performed, albeit in a strictly supervised setting. Simple, whole body activities have been learned and recognised using the position of a person's detected face [Govindaraju and Veloso, 2005]. More recently, the locations of estimated skeletal joints have been abstracted using qualitative 3D cone bins to create histograms [Xia *et al.*, 2015], pose trajectory descriptors [Chrungoo *et al.*, 2014] and also joint location covariance descriptors [Hussein *et al.*, 2013]. Skeleton pose features have been used along with optical flow and STIP features to learn human activities from a spontaneous-actions dataset [Gori *et al.*, 2015], and also to perform early recognition of actions [Ryoo *et al.*, 2015].

To the best of our knowledge, we are the first to combine LSA with a qualitative spatial representation to learn human activities from a challenging and realistic mobile robot activity dataset. Our qualitative representation abstracts metric observations and takes inspiration from [Sridhar *et al.*, 2010; Duckworth *et al.*, 2016]. Previous works have used LSA and pLSA to learn activity categories in a similar unsupervised setting, although not from a mobile robot using qualitative features. Approaches have been developed using low-level 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]. These approaches are not performed with the variability of a mobile robot's frame of reference, and restricted to a single person in the scene during the training phase, unlike ours that can encode multiple people in the scene simultaneously. Further, a major problem cited in the literature is "The lack of spatial information provides little information about the human body, while the lack of longer term temporal information does not permit us to model more complex actions that are not constituted by simple repetitive patterns" [Niebles *et al.*, 2008]. Descriptive spatial-temporal correlogram features were used previously to attempt to address this issue [Savarese *et al.*, 2008], however, these still suffer from low-level image processing frailties, and a single person in the scene during training. We address and partially alleviate this problem by using semantically meaningful qualitative features extracted from multiple observations. Such qualitative features encode "longer term temporal information" than used in the previous works. Finally, our code book of features is auto-generated and therefore adaptable to the environment of the particular service robot; it can contain qualitative features with semantic landmarks which help understand more complex interactions with key regions, or objects.

3 Knowledge Representation

Our aim is to understand human activities taking place from long term observations over a large environment. In this section, we first introduce the input data into the system, followed by the qualitative representation used, and finally describe the auto-generated codebook of qualitative features which results in a term-document matrix representation.

3.1 Skeleton Pose Estimates

A mobile robot is used to detect humans as they pass within the field of view of its RGBD sensor. The system is hardware independent and modular, although we use an OpenNi skeleton tracker [OpenNI organization, 2016] to first detect the human, then estimate and track 15 joint positions (at approximately 25Hz). An example is given in Figure 2 (left). Also shown (right) is one part of the global map which is semantically labelled with key regions and landmark objects in advance. The detected person in camera frame coordinates is transformed into the map frame of reference; using the robot's location and camera orientation (fitted atop a pan-tilt unit).

Formally, we define one skeleton *joint pose* as an xyz Cartesian coordinate in the camera coordinate frame along with a corresponding xyz position in the map coordinate frame, i.e. $j = (id, x, y, z, x_{map}, y_{map}, z_{map})$. A *skeleton pose* then comprises of a collection of joint poses, one for each estimated skeletal joint of the detected human, i.e $p = [j_1, j_2, \dots j_n]$, where $n = 15$ using the OpenNi tracker. For a detected human, we obtain a sequence of skeleton poses over a time series of detections, and generate a *skeleton activity*. This is defined as $S = [p_1, p_2, \dots, p_t, \dots]$, where each p_t is the detected skeleton pose at timepoint t. Note that there are no restrictions placed on t, i.e. a skeleton activity comprises of an arbitrary number of frames and skeleton poses. This depends only upon the length of time the person is detected by the robot's sensors, and this variation is a major difficulty when using real world data on a mobile robot.

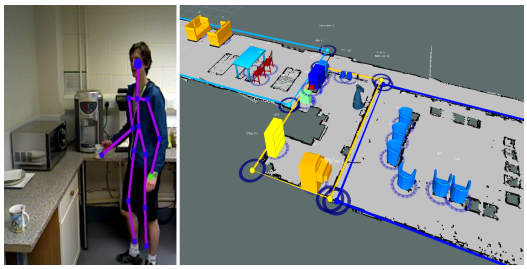

Figure 2: (left:) Skeleton pose estimate at a single timepoint. (right:) Semantic global map frame. (Best viewed in colour).

3.2 Qualitative Representation

Abstracting the metric skeleton data using a qualitative representation allows the robot to learn common and repeated activities being performed over multiple observations, even if they vary metrically in their execution. In this paper we use qualitative spatial representations (QSRs) that allow for the comparison of multiple observations, and the potential to draw similarities which can be used to understand the activities being observed. 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. In this paper, we use three QSRs to abstract our skeleton data: **Ternary Point Configuration Calculus** (TPCC) qualitatively describe the spatial arrangement of an object, 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*} [Moratz and Ragni, 2008]. **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 $o1, o2$: $o1$ is moving towards $o2$ (represented by the symbol $-$), $o1$ is moving away from $o2$ ($+$), and $o1$ is neither moving towards or away from $o2$ (0). **Qualitative Distance Calculus** (QDC) expresses the qualitative Euclidean distance between two points depending on defined region boundaries [Clementini *et al.*, 1997]. The intuition is based on the assumption that human motion can be partially explained using distance relative to a key landmark; a set of QDC relations localises a person with respect to a reference landmark, and a change in the QDC relations can help explain relative motion.

The three QSRs are computed from xyz data over a series of timepoints, i.e. a skeleton activity is abstracted into multiple sequences of qualitative relations (one per calculi being used), using a publicly available ROS library we developed [Gatsoulis *et al.*, 2016b; 2016a]. Each representation chosen is appropriate to describe human actions qualitatively; experimental details are presented in § 6.

3.3 Qualitative Features (code book)

Many human activities observed by a mobile robot can be explained by a sequence of primitive actions over a duration of time. In this section, we describe how we represent the time series of qualitative detections as a feature vector in a qualitative feature space (over the code book). To do this, we use Allen's Interval Algebra (IA) [Allen, 1983] to abstract the temporal relations of the observed QSRs. We first abstract the metric skeleton joint coordinates in each timepoint as above, then compress repeated relations into an interval representation. For example, if the right hand appears to be moving towards the head (QTC relation: $-$), for n' consecutive frames, and then is static (0) with respect to the head for n'' further frames, we compress this into an interval representation consisting of two intervals: $i_1 = \{\text{`}-\text{'}, (0, n'-1)\}$ and $i_2 = \{\text{`}0\text{'}, (n', n'+n''-1)\}$, each maintaining the QSR value (or set of values; one per calculi used) and the start and end timepoints. The interval representation of this example is shown in the first row of Figure 3 (top); however an interval representation of a complete skeleton activity contains a single row for each joint (or pairwise joints with landmarks) that QSRs are encoded to represent.

Taking any two intervals in this representation, it is possible to calculate the temporal relation which holds between them using IA temporal abstractions. IA is used to represent and reason with temporal intervals and defines 13 qualitative relations (for a list of the relations refer to [Allen, 1983]). 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".

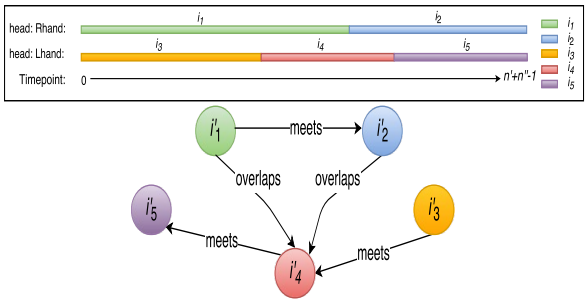

Figure 3: (top:) Interval representation of the relations between two skeletal pairwise joints. (bottom:) Interval Graph representation. (Best viewed in colour).

Given the interval representation of a video clip, we extract a set of unique qualitative features which are used to describe the observation. We define a *code book* as the set of unique features extracted from all observed skeleton activities, defined as $[\gamma_1, \gamma_2, \dots]$, where each γ_i represents a code word observed in at least one skeleton activity. The occurrence of each code word within a skeleton activity allows us to encode a sparse feature vector describing it (with equal length to the code book). This is similar to the *Bag of Words* technique, where words are represented by qualitative features extracted from the videos, and ignores their positional arrangement.

To extract the qualitative features, we compute an *interval graph* for each skeleton activity in our dataset [de Ridder and others, 2016]; an example can be seen in Figure 3 (bottom), which encodes both rows present in Figure 3 (top). Here, a node is used to represent an interval and contains only the QSR value (or set of values) that hold between a number of objects, and the objects themselves. The exact timepoints are not encoded in the node, e.g. node i'_1 in Figure 3 (bottom) contains (*head, Rhand*,'$-$') information abstracted from i_1. Nodes are linked by directed edges if their intervals are temporally connected, i.e. there exists no temporal break between a pair of intervals. The directed edges are labelled with their IA relation which holds between the two intervals. Thus there is no edge if the IA relation is *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), there is insufficient temporal information to abstract over the intervals and there is no edge between these nodes, e.g. there is no edge between i'_1 and i'_3, as both i_1 and i_3 occur at the start of the observation.

The code book is generated by enumerating all paths through the interval graph up to and including some fixed k. For example, the unique code words extracted (where $k = 2$) from the interval graph shown in Figure 3 (bottom) are generated by taking all paths up to length 2 and are: $\gamma_1 = i'_1$, $\gamma_2 = i'_2$, $\gamma_3 = i'_3$, $\gamma_4 = i'_4$, $\gamma_5 = i'_5$, $\gamma_6 = (i'_1$ meets $i'_2)$, $\gamma_7 = (i'_1$ overlaps $i'_4)$, $\gamma_8 = (i'_2$ overlaps $i'_4)$, $\gamma_9 = (i'_3$ meets $i'_4)$, $\gamma_{10} = (i'_4$ meets $i'_5)$.

The code words generated using this technique represent meaningful durations of qualitative relations specifically observed within the data. This makes it an efficient and intuitive method for representing observed human activities.

4 Latent Semantic Analysis

Each skeleton activity is represented as a feature vector over the auto-generated code book (by counting the occurrences each code word). We then draw comparisons with Information Retrieval systems and use Latent Semantic Analysis (LSA), which is often used to semantically analyse a "term-document matrix" defined as a matrix of word counts over a corpus of documents. In our case, the "terms" are the qualitative spatio-temporal features extracted in our code book, and each "document" in a corpus is a human activity video clip in our dataset. Therefore, creating a feature vector for each skeleton activity (as described in the previous sections) generates a term-document matrix.

Given a term-document matrix D of size $(m \times n)$, (where $m = |dataset|$ and $n = |codebook|$), we apply *term frequency-inverse document frequency* (tf-idf) [Rajaraman and Ullman, 2011] to weight the observations based upon the importance of the qualitative features observed. For example, if a qualitative feature is present in every video clip, it is given a very low score, as opposed to features which occur less frequently which have a higher score. It is calculated by the product of two statistics, *term frequency* and *inverse document frequency*. The tf-idf value increases proportionally to the number of times a word appears in a document, and is offset by the frequency of the word in the entire corpus, which is a measure of how much information that word provides. We use this weighting to adjust for the fact that some qualitative features appear much more frequently in general than others.

Once we have the dataset represented as tf-idf weighted feature vectors, the aim is to recover the latent concepts in the data. To do this, we use Singular Value Decomposition (SVD) to extract the singular values, and the left/right singular vectors. This step is key, since the decomposition provides information about the main concepts in the data, along with which features are prominent in each concept.

The technique is akin to finding the eigenvalues of the matrix. For an $(m \times n)$ tf-idf weighted matrix, the number of non-zero eigenvalues (and therefore singular values), is bounded by the rank of the matrix, i.e. at most $min(m, n)$. However, our aim is to recover a small number of latent concepts from our matrix and select only the largest singular values (assuming the activities are repeated a number of times).

Given our tf-idf weighted, term-document matrix C, SVD performs the matrix decomposition: $C = U\Sigma V^T$, where U and V are the singular vectors (rotations around the axis), whilst Σ is a non-increasing diagonal matrix containing the squared eigenvalues of C (i.e. the scaling values in each dimension). Examining the decomposition, the eigenvalues in the diagonal matrix Σ are the latent concepts of the matrix, and can be thought of as the latent activity classes encoded in the data. Further, the columns of the left singular vector (U) contain the eigenvectors of C^TC, and hold information about which video clips are assigned to which latent activity class (concept). Finally, the columns of the right singular vector V contain the eigenvectors of CC^T and tell us which qualitative features are used to describe each activity class. This is akin to finding a vocabulary to best describe human activities performed in real world scenarios, in an unsupervised setting.

5 Activities Dataset

In this section, we present a real-world human activities dataset captured using an autonomously patrolling Metralabs Scitos A5 mobile robot. The dataset is captured by observing university members of staff and students performing a set of common every day activities in a real university environment. These are defined as activity classes. The dataset provides difficult intra-class variation due to different viewpoints and partial occlusions.

The robot is equipped with a laser range finder for mapping and localization and is equipped with two RGBD cameras; one chest mounted for the purpose of obstacle avoidance, the other head mounted and used to detect people in the environment using an OpenNi skeleton tracker (introduced in § 3.1). Given a detected person in the robot's field of view, the camera records RGB images along with the estimated skeleton pose, plus meta data about the detection i.e. date, time of day, odometry data, region of map. [1]

The dataset was collected over the period of one week. The robot patrolled a pre-mapped space which can be seen in Figure 2 (right). During the week, we detected 300 human instances performing 398 daily living activities. A selection of example detections can be seen in Figure 4, where the second row, shows three different views of the same activity class (as judged by the ground truth labels).

To reduce any bias, the dataset was annotated by a group of independent volunteers, who segmented each video clip into the human activities occurring. It is these segments which form the basis for the experiments in § 6. As anticipated, the dataset is unbalanced with respect to the number of each activity class observed (i.e. some activities were observed more frequently than others), and the durations of each instance vary greatly. The following is a complete list of the common activities annotated, along with the number of occurrences: a: Microwave food (17); b: Take object from fridge (52); c: Use the water cooler (26); d: Use the kettle (58); e: Take paper towel (35); f: Throw trash in bin (50); g: Wash cup (66); h: Use printer interface (28); i: Take printout from tray (22); j: Take tea/coffee (35); k: Opening double doors (9). These instances are used during the experiments sections below to evaluate the unsupervised learning methodology.

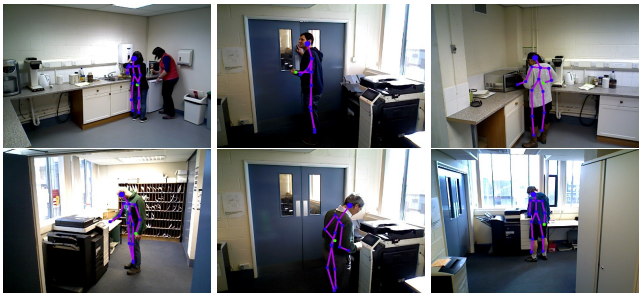

Figure 4: Human Activities Dataset examples. Second row are all examples from the same class. (Best viewed in colour).

[1] The dataset collected, along with meta-data and software repository, is available at: http://doi.org/10.5518/86.

6 Experimental Procedure

Our experiments comprise one main task; to learn a representation of human activities in an unsupervised setting. For this task, we use the activity instances from the human activities dataset captured from our mobile robot. The main steps of the flow diagram in Figure 1 have been explained above, and the implementation details are given here.

Our experiments and results section is presented using the "kitchen" region (defined by the yellow quadrilateral in Figure 2 (right)), which is considered the most interesting in the dataset. In this region there are a total of 10 semantic landmark objects including: *printer, shelves, microwave, water cooler, tea/coffee pot, sink, kettle, fridge, waste bin* and *paper towel dispenser*. For the purpose of evaluation and computational efficiency, we restrict the entire set of 15 skeletal joints to a discriminative subset of 8 including; the *head*, *torso*, left and right, *shoulders*, *hands* and *knees*. For generality when encoding our code book, we do not distinguish between "left" or "right" *shoulders*, *hands* and *knees*.

6.1 Qualitative Features

For each activity instance, we generate a skeleton activity and extract QSR features from the metric observations. We do this in a two stage process, first abstracting the person's relative joint positions in the camera frame, and secondly, abstracting the person relative to pre-defined semantic landmarks. This process generates two sequences of QSR values that are used to create our term-document matrix. For each skeleton activity S_m in our dataset of M observations, we have a sequence of t skeleton poses. This is referred to as $S_m = [p_1, p_2, \dots p_t, \dots]$, where each p_i contains both camera coordinate and map coordinate frame xyz positions of each joint being used.

For each skeleton joint, at each timepoint, we calculate a TPCC relation relative to a person centre line (defined by connecting the detected *head* and *torso* joint positions). This produces a sequence Q_{cam} of relations for each of the skeleton joints used (excluding the head and torso), of length t.

Similarly, to abstract the person's position in the global map frame we use QDC and QTC calculi to describe the relative position of key joints relative to semantic landmark objects. We create a sequence Q_{map} of QSR pairs (QDC and QTC) of length $|S_m| - 1$ (since QTC relies on pairs of consecutive timepoints, we remove the QDC value at $t = 1$ to obtain $|S_m| - 1$ pairs). Since the landmarks are static, we use the QTC_{B11} variant of QTC [Delafontaine *et al.*, 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]. We found the above intuitive considering the semantic landmarks in some regions are not particularly well spaced out. An example sequence for *hand-fridge* in Q_{map} is [*('+', 'Near'), ('+', 'Near'), ('+', 'Medium'),* ...], of length $|S_m| - 1$.

Generating the Code Book

For each sequence of QSRs, described above, we create an interval representation and interval graph by compressing repeated relations. During this process, we apply a median filter which smoothes any rapid flipping between relations, owing to visual noise. By using semantically meaningful QDC relations in the sequence Q_{map}, we do not encode interval graph nodes for any timepoints where the QDC value is "ignore". This has the effect of creating a sparse interval graph.

To produce our term-document matrix we extract the set of unique qualitative features (code words) by enumerating paths up to some length k, over all interval graphs in our dataset. Since the number of paths increases exponentially with the number of interval nodes, we use $k = 4$ and restrict the nodes on a path to encode at most 4 different objects. These paths allow us to capture overlapping qualitative features between multiple object pairs which occur in the observations. Finally, since the calculi used in each sequence are distinct, we merge the unique features into a single code book of length 6482 (using the parameters presented above).

Given the auto-generated code book, we encode each skeleton activity S_m in our dataset into a feature vector representation, as per § 3.3. We do this by counting the occurrence of each code word in the skeleton activity's interval graph (by comparing to the each extracted path). This process generates a vector of the occurrences of each path in each skeleton activity, which we stack to create the (m x n) term-document matrix D, (where $m = |dataset|$ and $n = |codebook|$).

7 Results

In this section, we present empirical results that the methodology presented in this paper learns common human activities from unsupervised observations. We demonstrate this by applying our learning methodology, and experimental procedure, to the challenging human activities dataset captured from a mobile robot, introduced in § 5. The structure of this section is as follows; firstly, we present the results of our unsupervised LSA learning framework. This is supplemented by a comparison to a commonly used supervised learning technique and an unsupervised technique used previously in [Duckworth *et al.*, 2016]. Secondly, we discuss the learned vocabulary over which each activity class is defined as the occurrences of qualitative features over the code book.

The results here are generated by performing LSA onto the term-document matrix D generated in the previous section. This involves applying the tf-idf weights to the term-document matrix D to obtain the weighted matrix C, and performing SVD as per § 4. Figure 5 shows the resulting singular values extracted. It can be seen that there is a limited number of "large" singular values where each represents a latent concept in the matrix; this is intuitive given our dataset contains 11 different activity classes. We threshold and use only the largest singular values from Σ (threshold shown in green in Figure 5) and the results for 10 latent concepts are presented in Table 1. The dataset contains 11 activity classes, and the decomposition recovers 10 latent concepts with large singular values. The unsupervised system does not know which ground truth labels match to each emergent concept; for this reason we use clustering metrics: V-measure [Rosenberg and Hirschberg, 2007] (comprising of Homogeneity and Completeness scores) and Mutual Information [Vinh *et al.*, 2009]. We use these to compare the ground truth labels (assigned by volunteers), to the emergent concepts.

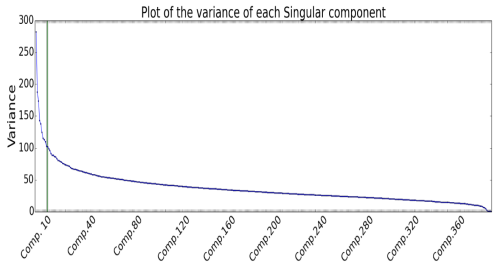

Figure 5: Singular values of LSA decomposition of the weighted, term-document matrix (C). The x-axis represents the singular values (maximum of rank(C)). Threshold limit for further analysis shown in green.

Table 1 presents a comparison between our LSA approach and two commonly used alternative methods, a supervised approach (an SVM) which uses the ground truth labels, and an unsupervised k-means used in previous work. The methods are compared using the same qualitative features. The SVM was trained using 5-fold cross validation, with a linear kernel, and where the code book was trained only once across the whole dataset. It can be seen that the supervised approach obtains 66.1% accuracy on the challenging dataset, and performs only slightly better than the LSA when evaluating using clustering metrics, even though it has access to labelled training instances to create decision boundaries. It can also be seen that LSA outperforms a standard k-means implementation using 10 cluster centres (average result presented over 10 runs, as with random chance classifier). We interpret this as LSA generalising observations better than k-means, since it considers qualitative features with similar meaning, i.e. identifying synonymy between dimensions, unlike k-means.

Metric	LSA	SVM	k-means	chance
V-measure	0.542	0.614	0.368	0.057
Homogeneity Score	0.520	0.617	0.280	0.057
Completeness Score	0.566	0.611	0.542	0.057
Mutual Information	1.180	1.407	0.637	0.130
Normalised MI	0.543	0.614	0.388	0.057
Accuracy	N/A	0.661	N/A	0.113

Table 1: Experimental results comparing LSA, with a supervised linear SVM, unsupervised k-means clustering and random chance clustering.

The results presented demonstrate 10 activity classes are recovered from a challenging, real world, mobile robot dataset. The dataset contains high intra-class variation, shown by multiple view points in Figure 4, and activities that are often occluded and partially observed. The results show the majority of these instances are successfully considered part of the same latent concepts (activity class). This shows the qualitative descriptors used in the abstraction are viewpoint invariant and can handle large amounts of noise and variation during the unsupervised learning phase.

7.1 Learned Vocabulary

As per § 4, the LSA decomposition recovers latent concepts in the non-increasing diagonal matrix Σ. In this section, we are interested in the right singular vector V^T. The vectors specify the rotations around the axes whereby the importance of each feature can be determined and a vocabulary defined for each activity class over the auto-generated code book. For our recovered 10 concepts from the above decomposition, V^T (with shape $|codebook|$x10), contains an assignment weight for each qualitative feature (code word), for each latent concept. Two of the right singular vectors from the above decomposition are plotted in Figure 6. We consider these singular vectors as the recovered vocabulary over the latent activity classes present in our dataset.

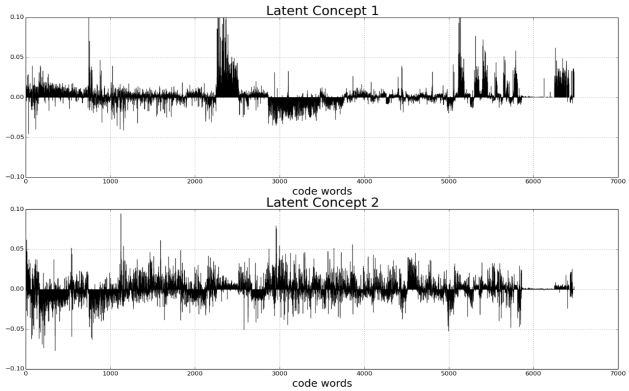

Figure 6: Two latent concepts from our LSA decomposition.

8 Conclusion

This paper has presented a novel, unsupervised framework for learning a qualitative vocabulary for many common daily living activities, from a mobile robot. We demonstrated its effectiveness at learning activities in a real-world human activity dataset, with large intra-class variation and noise.

Our methodology abstracts the exact metric coordinates of a detected person and landmarks, using a qualitative representation. It generates a code book, from observations, comprised of qualitative descriptors which work well with the occluded, and changing field of view afforded by a mobile robot's sensors. Latent Semantic Analysis (LSA) is used to decompose the tf-idf weighted, term-document matrix and recover latent concepts which are regarded as the activity classes observed. The results presented validate our methodology, showing we learn $10/11$ activity classes from a challenging dataset, and define a vocabulary for each activity class over the qualitative code words.

Further work includes making an online framework, and also increasing the complexity of the dataset by adding more data pertaining to more varied activities with hierarchical dependencies. A hierarchical structure of activity classes could be learned by successively relaxing the number of concepts.

Acknowledgments

We thank colleagues at the University of Leeds and in the STRANDS project consortium (http://strands-project.eu) for their input. We also acknowledge the financial support provided by EU FP7 project 600623 (STRANDS).

References

[Aggarwal and Xia, 2014] J.K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70 – 80, 2014.

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

[Chrungoo *et al.*, 2014] Addwiteey Chrungoo, SS Manimaran, and Balaraman Ravindran. Activity recognition for natural human robot interaction. In *Social Robotics*. Springer, 2014.

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

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

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

[Duckworth *et al.*, 2016] P. Duckworth, Y. Gatsoulis, F. Jovan, N. Hawes, D. C. Hogg, and A. G. Cohn. Unsupervised learning of qualitative motion behaviours by a mobile robot. In *AAMAS*, 2016.

[Gatsoulis *et al.*, 2016a] Y. Gatsoulis, M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, and A. G. Cohn. QSRlib: a software library for online acquisition of qualitative spatial relations from video. In *29th International Workshop on Qualitative Reasoning (QR16), at IJCAI-16*, New York, USA, 2016.

[Gatsoulis *et al.*, 2016b] Y Gatsoulis, P Duckworth, C Dondrup, P Lightbody, and C Burbridge. QSRlib: A library for qualitative spatial-temporal relations and reasoning. qsrlib.readthedocs.org, Jan 2016.

[Gori *et al.*, 2015] I. Gori, J. Sinapov, P. Khante, P. Stone, and JK Aggarwal. Robot-centric activity recognition in the wild. In *Social Robotics*. Springer, 2015.

[Govindaraju and Veloso, 2005] Dinesh Govindaraju and Manuela Veloso. Learning and recognizing activities in streams of video. In *AAAI workshop on learning in computer vision*, 2005.

[Hussein *et al.*, 2013] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *IJCAI*, 2013.

[Lavee *et al.*, 2009] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(5):489–504, Sept 2009.

[Liu *et al.*, 2008] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. In *CVPR*, 2008.

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

[Niebles *et al.*, 2008] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.

[OpenNI organization, 2016] OpenNI organization. OpenNI user guide. www.openni.org/documentation, 2016.

[Rajaraman and Ullman, 2011] A. Rajaraman and J. D. Ullman. Data mining. In *Mining of Massive Datasets*, pages 1–17. Cambridge University Press, 2011.

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

[Ryoo *et al.*, 2015] M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me'. In *HRI*, 2015.

[Savarese *et al.*, 2008] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, and Li Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and video Computing WMVC*, 2008.

[Sridhar *et al.*, 2010] M. Sridhar, A. G. Cohn, and D. C. Hogg. Unsupervised learning of event classes from video. In *AAAI*, 2010.

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

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

[Wong *et al.*, 2007] Shu-Fai Wong, Tae-Kyun Kim, and Roberto Cipolla. Learning motion categories using both semantic and structural information. In *CVPR*, 2007.

[Xia *et al.*, 2015] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo. Robot-centric activity recognition from first-person RGB-D videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

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