Unsubstantiated Anthropological Motion Exploration for Smart Mobile Robots

Reeja R Rajan[1] Ramya T V[2] Manoj M[3]
[1,2,3]Assistant Professor
[1,2,3]Jawaharlal College of Engineering & Technology, India

Abstract— The victory of smart mobile robots functioning and co-operating with persons in day-to-day living atmospheres depends on their capability to simplify and learn anthropological activities, and acquire mutual understanding of a detected scene. In this paper we aim to identify anthropological activities being executed in real-life atmospheres from enduring surveillance from an independent mobile robot. For our purposes, a anthropological doings is defined to be a varying spatial configuration of an individual's body cooperating with main items that deliver some functionality inside an atmosphere. To ease the perceptual restrictions of a movable robot, controlled by its concealed and missing sensory modalities, possibly loud graphic comments are charted into an abstract qualitative cosmos in order to specify outlines invariant to precise quantifiable positions within the physical biosphere. A number of qualitative spatial-temporal illustrations are used to seizure different aspects of the associations among the anthropological focus and their atmosphere. Analogously to statistics recovery on text volumes, a reproductive probabilistic method is used to improve latent, semantically-meaningful ideas in the encrypted opinions in an unsubstantiated method. The slight amounts of ideas exposed are considered as anthropological motion modules, yielding the robot a low-dimensional understanding of visually detected multifaceted acts. As a final point, dissimilarity inference is used to assist incremental and continuous updating of such ideas that permits the mobile robot to proficiently study and update its models of anthropological motion over period ensuing in effective life-long learning.

Keywords: Smart Mobile Robots, LDA model

I. INTRODUCTION

Innovations in the reliability of independent mobile robot stages worth they are well-matched to endlessly update their own information of the world based upon their many annotations and interactions [4,5]. Unsubstantiated learning structures over such long periods of time have the likely to allow mobile robots to become more supportive, mainly when cohabiting human populated situations. By eliminating humans from the learning procedure such robots can inexpensively swot from bigger extents of existing data (annotations), letting them to acclimatize to their environments and save time/work hard-coding specific facts. Sustaining an understanding of vibrant anthropological surroundings allow a robot to fine-tune its own conduct, or help in a job being perceived.

The aids we offer are as follows: i) a qualitative spatial-temporal vector space structure for converting perceived anthropological events by an self-directed mobile robot; ii) techniques for learning low dimensional demonstrations of public and frequent arrangements from various encoded graphic annotations using unsubstantiated probabilistic methods; iii) clarifications to real-world thoughts when working with enduring, self-directed mobile robots bagging nonstop, unsegmented video categorizations in a life-long learning situation.

Our methodology depend on first perceiving and tracking anthropological body actions from a solo mobile robot's embedded sensors, beside with learning the position of key objects in the environment using off-the-shelf methods. Each anthropological remark, originally documented as an order of quantifiable poses, is encoded using various qualitative calculi to abstract the exact spatial and temporal information of the observation, and at last denoted as a vector of the occurrences of discrete qualitative descriptors .The pool of encoded feature vectors is examined analogously to a corpus of text documents having several areas of interest. Multiple dormant matters are recovered from the annotations and considered as anthropological activity classes with a multinomial distribution over an self-generated vocabulary. Two methods are presented to study low-dimensional anthropological motion demonstrations. First, a non-probabilistic low-rank approximation method is shown to work well with pre-segmented video sequences of perceived anthropological activity. Secondly, a more sophisticated probabilistic Latent Dirichlet Allocation (LDA) [6] method is shown to somewhat eliminate the necessity for manual time-based segmentation of the recorded annotations, permitting the robot to access large amounts of data which otherwise would need anthropological annotation. LDA is a hierarchical Bayesian model where each observation is exhibited as a mixture over an underlying set of topics, and each topic is, in turn, modelled as a mixture over the discrete vocabulary.

It combines a propagative, probabilistic approach such as LDA with a qualitative spatial representation to recover low-dimensional representations of anthropological activity perceived 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 motion instance, to an additional 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 realistic supposition is that an anthropological observation is modelled as a probabilistic mixture over an underlying number of latent topics, where some topics can be measured "interesting" anthropological activities.

Specific tasks of using data taken from an self-directed mobile robot include: i) the robot's on-board sensors only award a partial and changing viewpoint of the world, i.e. it obtains incomplete annotations of activities being executed, which are often structurally noisy; ii) each

observation is likely carried out in different ways, e.g. opening a door with contrasting hands. The proposed structure helps ease these problems in two parts; first by utilising a state-of-the-art human pose estimator to progress the correctness of annotations, and secondly by using a qualitative spatial representation (QSR) with the ability to transform somewhat noisy annotations of arbitrary spatial positions into semantic low level descriptors.

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

II. LITERATURE REVIEW

There is a public difference in the literature among vision-based anthropological motion examination, which excerpts statistics from cinematographic cameras by means of PC visualization methods, and sensor or wearable computing-based structures [10, 11]. Sensor-based structures often depend on on the obtainability of small sensors, specifically wearable sensors, smart mobiles, or RFID appended things, that can be attached to anthropoid under surveillance in order to get illustration of that person's activities. We emphasis on representing anthropoid motion from filmic records, where the belief of being detected is limited to a solo camera's field of view. This is a matured sub-field of artificial intellect and the reader is pointed to review documents which cover the subject in detail using, RGB cameras [12–14] or 3D cameras [15, 16]. Though, a lot of common methods in these surveys do supervised learning, where every training model needs manual division and annotating with a ground truth tag. This is not a practicable resolution for a long-term independent mobile robot which ideally, has nominal supervision at the same time as deployed in the practical world.

Unsubstantiated learning methods are considered more suitable for this mission since they do not need time-consuming, offline labour-intensive observations. Preceding works have used LSA (Latent Semantic Analysis) [17], probabilistic LSA [18] and LDA [6] for learning low-dimensional anthropological motion classes in an unsubstantiated background. They joint these methods with low-level STIP (Spatial Temporal Interest Point) features to study action classes [19]; local shape context descriptors on silhouette images [20]; a mixture of semantic and mechanical features to study actions, faces and hand gestures [21]; and by combining a vocabulary of local spatio-temporal capacities with a language of spin-images to capture the outline deformation of the artist [22]; Though, a main difficulty cited in these papers is the lack of spatial facts about the anthropological body taken by low-level image features, and the nonexistence of more long-term progressive statistics encrypted into the features limits learning additional multifaceted movements. Descriptive spatial-temporal correlogram features have been used before for this issue [23], though, the method still suffers from low-level image processing infirmities, and the prerequisite for a person to be modelled in the act during a well-ordered training. A different method has been to study the temporal relations among atomic actions in an unsubstantiated setting in order to exactly represent "composite" anthropological actions [24]. On the other hand, the input videos for this method require labour-intensive temporal segmentation into orders of "overlapping fixed-length temporal clips", make it expensive for life-long education on an independent mobile robot. More, the works have been achieved without the changeability of a mobile robot's frame of locus, and limited to learning on temporally segmented video records during an offline training phase, unlike our work. To address these problems, we abstract observed anthropoid and object assessed postures into a qualitative spatial illustration. There is some indication to propose that there are devoted areas of the brain to do such ideas [25]. It is thus usual to try to implant this into a system to know anthropological behaviour in video and finally, into independent robotic systems in order to represent behaviours done in the surroundings they live. Qualitative spatial and temporal calculi grow from a set of reciprocally extensive and pair-wise disjoint associations. Many types were established in the literature, amongst the most popular contains topological, directional and non-topological; a review of popular calculi is specified in [26]. Qualitative spatial representations are frequently used to signify visual, quantitative observational figures in a low-dimensional and additionally semantically expressive qualitative space, as per in this paper. Regularly an object-based idea of a video categorization is done, then common provisions of the abstracted bodies are erudite by several relations, e.g. dining table [27]; simple events for daily activity from a stationary camera dataset [28]; forecast object categorisation [29]; to remove inconsistent visual observations from noisy videos [30]; and smooth performance about spatio-temporal actions being practical [31].

III. QUANTITATIVE REPRESENTATION

We can understand the human activities taking place from long-term observation with the help of a mobile robot. Here the quantitative input data is captured by the robot. Initially, consider a human activity and the specific activity domains is required to operate with robot.In the next step robot can encode the human observation into a quantitative human body sequences.during the last step, robot can interpret the environment with key object locations which can analyse human functionality.

A. Activities performed by humans

The term activity is related with the temporally dynamic configuration of some agents, whereas the agents are stranded in the real-world. In this paper we can analyse i)

knowing about human ii) continual learning. Inorder to achieve this, there should be an interaction is needed between the human agent and environment, specifically between a human and key object. As a result of this, we can define the activity of humans to be a temporally dynamic configuration of a human agent. The following assumptions are related with human activities

A key object is the semantic entity with a fixed position in an environment which is used to provide certain functionality that is required for the execution of certain activities of interest.

The activities performed by object is considered as a moderately ordered sequence of sub-events (or repeated patterns) between positions of a person's body joints relative to the key objects. The repeated patterns (or sub-events) can be considerd as simple qualitative relations holding the human's body joints with a number of objects in the particular environment.

Human pose estimates

The main purpose of a mobile robot is to detect the human and gathers their three-dimension view with fifteen joint locations of the body. The three dimensional view is passed to an RGBD sensor.we can use OpenNItracker inorder to get multiple person's three dimensional view. When we are representing the three dimensional view,the real challenge is to get the human object interaction in difficult viewpoints,these are the major causes of errors occurred in our system. The pose estimation system has two phases. (1)OpenNI can produce bounding boxes per frame for a person (2)the state-of-the-art convolutional network (ConvNet) 2D human pose estimator can take RGB frame as input. Afterward, the X and Y coordinates of OpenNI body joints are changed with the superior 2D body joint coordinates. The human pose can be evaluated with ROS message, whereas the body joint location can be represented as a 3D Cartesian coordinates in a camera frame

Fig. 3.1: Incorrect OpenNI pose estimates

B. Representation of objects

Objects provide some functionality for human activities. The human activity includes relative positions of people with respect to key objects within the robot's environment. noticing and tracing arbitrary objects in real time from a robotic platform is difficult and unexplained problem. In order 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. This process is known as a sweep phase; 2) multiple sweeps are used to create large cloud representation of the robot's whole environment. unsupervised segmentation algorithm extracts locations of potential objects by rendering the surface using surfels (surface elements)

Fig. 3.2: Developed pose estimates

C. Representation of objects

Objects provide some functionality for human activities. The human activity includes relative positions of people with respect to key objects within the robot's environment. noticing and tracing arbitrary objects in real time from a robotic platform is difficult and unexplained problem. Inorder to 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. This process is known as a sweep phase; 2) multiple sweeps are used to create large cloud representation of the robot's whole environment. unsupervised segmentation algorithm extracts locations of potential objects by rendering the surface using surfels (surface elements)

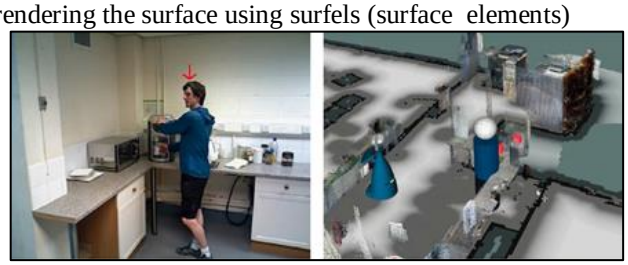

Fig. 3.3: Human body pose observation relative to the environment. (Left) RGB image matching to a single human body pose detection. (Right)

IV. 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 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). The third calculus, the Qualitative Distance Calculus (QDC) [58] expresses the qualitative Euclidean distance between two points depending on defined distance thresholds, O 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.

A. 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

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.

B. Unsubstantiated learning for anthropological activities

Encoding a corpus of anthropological annotations into such a term-frequency matrix allows latent structure can be recovered in an unsubstantiated setting. The aim is to learn low-dimensional representations of frequent structure encoded as qualitative descriptors across numerous similar annotations. This is achieved by using information retrieval techniques. Latent Semantic Analysis (LSA) [17] and probabilistic method, Latent Dirichlet Allocation [6] are taken into consideration. These techniques helps to understand large corpora of encoded text documents and to recover distributions of latent areas or themes existing in data.

C. Low rank approximations for anthropological activities

The aim is to learn a low-dimensional representation of an encoded term-frequency matrix by finding severance within the set of qualitative descriptors perceived. The descriptors with most discrimination will contain the most variation. The supposition is that by dropping the dimensionality of the matrix, but maintaining as much change within the columns as possible, it is conceivable to symbolize the corpus of annotations with a comparatively small number of anthropological activity classes. The process is accomplished using Latent Semantic Analysis (LSA) which figures linear combinations of columns to generate new composite features having as much dissimilarity as possible. Sorting the new features by their ability to discriminate the annotations, the most redundant are detached to leave a low-dimensional representation and latent classes encoded in the data are improved.

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 grounded upon its dissimilarity in the training trials, with the supposition that the most descriptive columns have the major variation. The weighting upsurges proportionally

to the number of times a code- word 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 demonstration of a matrix a low-rank approximation is computed. 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. A small number of latent concepts can be recovered from the encoded data. The supposition is that common anthropological activities relate to frequent patterns of discriminative qualitative descriptors encoded within the annotations. By observing decomposition, the non-zero eigenvalues in the diagonal matrix Σ represent the r most discriminative new compositional features, known as concepts. These latent notions can be thought of as the activity classes encoded in the original term-frequency matrix. The columns of the left singular (M X M) matrix U contain the eigenvectors of CC T , since CCT $U\Sigma\Sigma^T$ U T , and provides information, as a linear combination, about the weighting of each concept to each surveillance, specifying its latent activity class (concept). The columns of the right singular (N N) matrix V contain the eigenvectors of CT C , since CT C V Σ^T ΣV T , and specify a linear combination of weights for each qualitative descriptor (codeword) used to describe each latent notion.

D. Limitations

There are limits to this non-probabilistic method,LSA. Given the matrix decomposition, i.e. the left/right singular matrices refer to the linear combinations of annotations to concepts U , and codewords to concepts V ; one drawback is that U and V are both orthogonal matrices. The repercussion of the orthogonal matrices is that any concepts extracted cannot share columns, e.g. a specific codeword cannot be significant in two different 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.

Finally, selecting the rank to best represent the low-rank approximate matrix C_r is frequently challenging. A good value of r can be selected by plotting the variation of each eigenvalue, in a decreasing scree plot that idyllically shows a abrupt curve trailed by a bend, known as "elbow point", trailed by a more flat line representing any further features add little variance. Generative probabilistic model provide solutions to each of these limitations .

E. Probabilistic distributions for anthropological activities

One intuition is that an observation of a anthropoid should be exhibited , that allows for numerous, overlying classes of activity to follow and share certain descriptors. Latent Dirichlet Allocation (LDA) which is commonly referred to as Topic Modelling,is introduced for this reason.. The key idea is twofold: a topic can be referred as a multinomial distribution over a code book and pronounces a particular thematic arrangement present in the corpus; a document (codeword histogram) is signified as a probabilistic mixture over topics, by deducing a mixing vector or proportions . The supposition made is that similar groups of co-occurring codewords are used for similar documents , and therefore the co-occurrence can identify the latent thematic topics. This framework permits for each remark of a anthropoid to be modelled as a mixture of activity classes stirring, and to concurrently improve the latent activity classes as distributions over the code book.

F. Generative LDA model of anthropological activity

Probabilistic sampling are the main concept in probabilistic generative models and can be inferred as a model of how the detected data was generated from a set of fundamental latent variables. In this case, the pool of annotations are expected to be generated from latent topic distributions, mixing vectors and topic assignments. Fig. 7 shows the intuition behind the LDA generative process for a single human statement. For perceived codeword histogram, the underlying process can be categorized as follows:

1) a per-document topic proportions vector, θ_D ,is sampled from a prior Dirichlet distribution parameterized by α, i.e. θ_D Dir(α).
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 $z_{D,n}$
   -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 $w_{D,n}$ ~ Multinomial(φzD,n), where each φi represents a topic distribution over the code book, also drawn from a Dirichlet simplex parameterised by β, i.e. φi ~ Dir(β).
3) This process repeats to generate M bags-of-words.

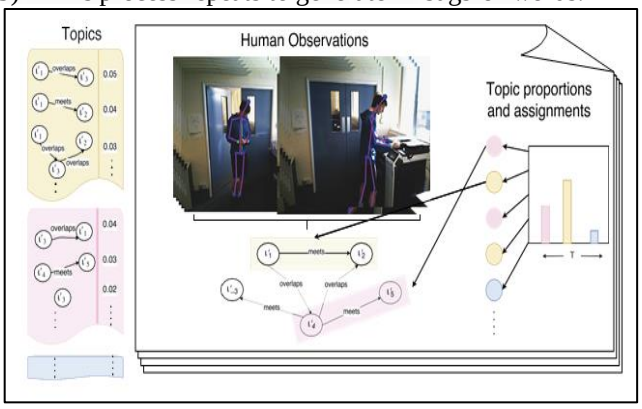

Fig. 4.1: Generative LDA model of anthropological activity. (Left) Three topic scatterings over the code book (yellow, pink and blue) along with three top-probable codewords in the yellow and pink topics. (Centre) Produced interval graph and bag-of-words gained from encoding one anthropoid observation. (Right) Topic proportions vector (pink, yellow, blue histogram) and codeword assignments as a column of sample coins drawn (pink coin, yellow, pink, etc.).

G. LDA as a graphical model

In reality, the robot does not observes the mixing proportions vectors or the assignment of each codeword into topics .It observes the bags-of-words only. Given a

collection of codeword histograms, inferring the three sets of latent variables is our task:

- 8 = [φ1, ..., φT] per-corpus topic distributions, where each φi is a distribution over the code book VD ;
- © = [θ 1, ..., θ M] per-document topic proportions vectors, where each θ D is a distribution over T topics;
- z is the assignment of all all observed codeword tokens to topics, for all observations.

Therefore it inferes p(8, ©, z). LDA is considered as a Directed Acyclic Graph (DAG) or Bayesian Network . In a DAG, nodes signify random variables and directed edges between nodes represents conditional dependencies between variables. A common presentation technique is to show observed random variables using shaded nodes, and non-shaded nodes for latent variables. Plate notation is used to highlight imitation random variables, with the number of random variables marked in the lower right corner of the plate. The DAG representing the LDA model for anthropological activities is shown in Fig. 4.2. It is a three-layer Bayesian model since: i) the Dirichlet hyperparameters, α and β, are corpus-level parameters; ii) the topic proportions variables, θ D for each bag-of-words D in the corpus, , are codeword histogram-level parameters, sampled once per codeword histogram; iii) the variables zD,n and w D,n are codeword-level and sampled once for each codeword in a bag-of-words.

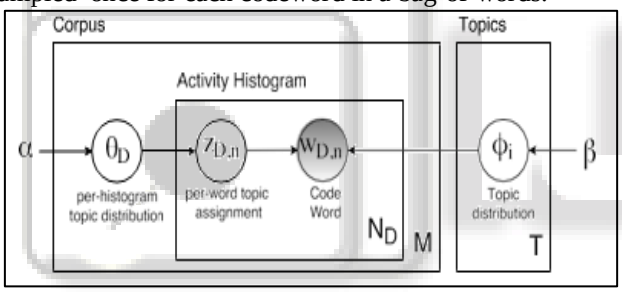

Fig. 4.2: DAG representation of LDA.

H. Approximate Inference

The joint probability distribution can be used to find possible inference queries by marginalization, i.e. summing out over inappropriate variables. Given M observed bags-of-words as D_1:M , as encoded as codeword histograms in a term-frequency matrix, inference allows us to evaluate the latent variables, i.e the mixing and the topic distributionsproportions vectors that best fit the observations. This can be considered as finding latent patterns in the data which best separate it into a meaningful topics; its thematic structure. This translates as computing the posterior distribution of the latent variables given a collection of bags-of-words:

I. Variational inference for incremental learning

It is not efficient to recurrently perform batch learning on increasing corpus of recorded video sequences. Ideally, an incremental learning technique could update its learned distributions based upon only new annotations which does not need recomputing for earlier examined data. Variational Bayes (VB) approximate inference was introduced for this purpose which aims to enhance a simplified, parametric distribution in order to fit the LDA model posterior using mini batches of new observations. This allows the robot to frequently learn about anthropological activities based upon incrementally updating the LDA posterior. This method was first used with LDA to analyze huge corpora containing millions of natural language text documents where batch algorithms were exorbitantly computationally expensive [71].

The progression of updating the topic model is threefold for a new observation:

1) any new codewords in the annotations are first affixed to the current code book V_D and topic distributions 8 with zero probability;
2) a multinomial distribution over the existing set of topics for the new statement is figured, θ , that symbolizes the combination of topics detected;
3) finally, the topic distributions, 8, are updated using new observation,

This allows the robot to efficiently apprise its model of anthropological activities using a single pass over new observations, enhancing both computation and storage complexity. Each observation can therefore be maintained as a low-dimensional distribution over the set of topics considered human activities

J. Probabilistic mixture of activities

Each anthropological observation will not be a solo, temporally segmented occurrence of an activity class; unless manual segmentation of the sequences into clips is achieved. Further, without such temporal segmentation, many annotations may cover individuals performing no interesting activities,.

It is not ideal for a anthropoid to manually segment out “interesting” sub-sequences from sequences of images recorded by the robot, LDA representations is used to handle these exciting longer sequences. We assume that numerous activities are occurring in each observation. Thus coherent activity classes can be learned by robots even when video sequences are not temporally segmented into clips focused on a single activity instance.

K. Evaluation

Here we are evaluating the unsupervised learning methods on two human pose video datasets which differ in complexity. The activity classes can be learned from the individual datasets with LSA and LDA is superior to simple clustering approaches The first dataset, which is publicly available and consists of 124 video sequences where each is scripted in advance and contains a temporally segmented activity class instance which is recorded from a static camera. The second dataset an independent mobile robot observing an unstructured, real-world university common area over a one week duration with no limitations on the forms of interactions observed; and is also available publicly.

L. Cornell Activities for Daily Living Dataset

It consists of 124 RGBD videos, with four different actors. The video clip comprises of a single actor executing one high level activity class instance out of 10 classes, resulting in a video clip with a equivalent ground truth class label. The activities are performed by facing a

fixed camera, with minor contextual clutter, and where the subject is situated in the centre of the camera frame to allow for good human pose estimates (fifteen joint positions) and auto and ground truth object detections.

This dataset is a large and challenging datasets of human daily living activities. 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 scraps with body joint obstructions, even with simple reaching or placing activities.

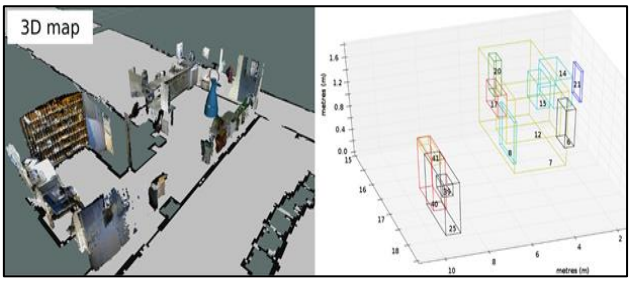

Fig. 4.3: Human Activity Dataset. (Left) 3D object clusters extracted from fused point clouds, (Right)

M. Mobile Robot with Human Body Pose Dataset

Objects: The robot implements a 3D metric sweep of the environment inorder to generate a registered 3D point-cloud, then extracting a set of key object clusters from this and overlays them onto the metric maps .The object clusters can be seen as point cloud clusters in Fig. 4.3 (left). A sub-set of objects are selected using the analysis of human paths, i.e. where people stop and what locations their body joints interacted there. This resulted in a set of fourty one key object locations shown in Fig. 4.3 (right),

Human body poses: The human body pose is estimated from the camera image, using the depth image (OpenNi2), then the RGB image post-process (CPM). Each body joint position of human is translated into the robot's map coordinate frame of reference with the localised position of the robot and the pan-tilt angle, . The visual algorithm is not always correct when the robot is moving.so we can analyse more accurate values when the robot is static.the simple analysis can be collected as a dataset over a period of one week or more.

Fig. 4.4: Example RGB images with human pose estimate

Activity ground truth: Inorder to obtain a ground truth (GT) activity class label for the human observations, each recorded sequence are manually inspected by the volunteers. A group of common and repeated actions in the recordings agreed and defined a set of activity classes.

The incidences of each instances of each activity class in the observed videos was temporally segmented by the volunteers to outline the activities, creating multiple smaller video arrangements containing only a single activity instance.

N. Experiment

The implementation details the framework are common to both datasets, with the aim of learning coherent human activity classes in an unsupervised learning technique and using the ground truth labels for assessment purposes. Details are i) the qualitative representations ii) the constraints used for computing codewords, resulting in a codeword histogram illustration for video clip and a term-frequency matrix for each dataset.

A segmented video clip m is first represented as a human pose sequence as Sm p1, ..., pi , ... pt of length t, whereas each pi is the human body pose at time point i and contains both the camera frame and map coordinate frame 3D position of 15 physique joint locations, i.e. as 15 joint poses. Recall that t is random and varies for each observation. Conceptualizing this arrangement of body poses into a qualitative representation, first the camera frame relationships are calculated, followed by translating and calculating the map frame relatives.

TPCC calculi is used for abstracting a person's full body joint positions comparative to the head-torso 2D line in the camera organized frame. The TPCC relations Qcam of length t contains the TPCC relations among the centre-line and the left/right hands and shoulders joint positions. Other joints omitted for efficiency, we assume that their movements do not contributing to the kinds of human activities in two datasets. However, they could clearly be added. Next is to abstract the body joint locations in the map coordinate frame. Here we are using a combined QDC and QTC calculus. QTC is used to capture relative motion and QDC is used for capturing relative distances of the poses. QSRs are used to characterize the relative movements of the person's left/right hand body joints and torso position, relative to the key objects. The combined QDC and QTC pairs is produced, Q map , of length t 1 QDC values larger than the biggest boundary in O are not measured, this produces a sparse interval illustration and an efficient process with less codewords .

For the sequence Qcam and Q map, apply a median filter ,which can smoothen rapid flipping among relations. This is common when we are abstracting into qualitative representation. We produce an interval representation and an interval graph for each sequence distinctly and extract all graph paths using graph path parameters η and ρ given as per the given dataset. The qualitative relations differ between the two sequences, Qcam and Q map , we can merge the two collections of extracted graph paths composed to create a single bag-of-words to characterize each observation, i.e. Dm w1, w2, ..., w ND , whereas each wi can be observed as a graph path. After observing and encoding M sequences, the codebook of N unique codewords is produced.

Given M N term-frequency matrix C which represents an entire training dataset, learning activity classes

and associate them to the ground truth annotated labels. The optimal values found are 10 latent ideas for the CAD120 dataset, and 11 for the Mobile Robot Dataset. However, we suggest a method to alter this number dynamically in an incremental learning process .Different temporal segmentation methods are functional to the activity video clips in order to estimate the efficancy of the proposed methods. However, the set the LDA hyperparameters α, β to 0.5 and 0.03 respectively for all the experiments. These hyperparameter settings reflect the prior belief on θ and φ that each observation is likely to consist of only a minor number of subjects, and that subject distributions consist of a relatively minor number of codewords with probability mass.

CAD120 dataset: Each video sequence in the CAD120 dataset is a single person situated near the centre of the camera frame carrying out his daily activity, and the objects are placed close to the person. The QDC relation thresholds are set to O 0.15 m, 0.4 m, 0.8, 1.0 m , creating 5 semantic areas which is labelled as touch [0–0.15 m], close (0.15–0.4 m], middle (0.4–8.0 m], faraway (0.8–1.0 m] and neglect (>1 m).In the dynamic objects, the abstract object class is used as an object ID in the interval illustration and interval graph. A code book VD of codewords is calculated, where |VD | = 5,520 codewords (29,016 entire), and a low-pass filter is applied for reducing the |VD | = 958, by QSTAG parameter choices: η = 3 and ρ = 1, therefore. all paths of up-to length three are computed which include only one pair of objects. Finally, a codeword histogram is calculated for each video clip and an M × N term-frequency matrix is calculated where M = 124 and N = 958.

Mobile robot human body pose dataset: The video sequences recorded with mobile robot are very much varied and stimulating. We can evaluate the learning methods when using two sets of keys of objects. initially, the set of 14 most interrelated key objects were attained from the 3D sweeps and trajectory analysis is performed. Next , we can evaluate the set of 12 manually specified key object locations,which determines how important to obtaine the exact location of key objects. m (0.5–1.0 m] and ignore (>1 m), which are experimentally chosen to best differentiate activities in the more complex environment.It is possible to study the threshold values from observations in an unsupervised setting.

A code book is produced, originally using QSTAG parameters η 4 and ρ 2.

The 14 autonomously learned key objects,20,637, which is reduced to 2,876 when the low-pass filter (frequency five) is used. Also 22,829, reduced to 3,594, for the case when we are using a set of physically defined key objects. The additional codewords highlight are more unique qualitative relationships between the human pose estimates and the physically defined object places and therefore the opportunity of more discriminative codewords which can help the unsupervised learning performance. The codeword histogram is calculated for each video resulting in an M N term-frequency matrix.

V. CONCLUSION

In swift, we have make known to a novel context whereby low-dimensional illustrations of anthropological annotations from a mobile robot are learned. We determine that by first theorizing explanations using qualitative spatial relations among traced entities in a visual section and secondly carrying out probabilistic unverified learning techniques, efficient topic circulations can be cultured representing human actions. As a key influence, we have providing a formal illustration of human annotations as developed by a mobile robot, qualitative abstractions to generalise these, and methods to extract discrete features as arrangements of observed qualitative relationships. Multiple unconfirmed methods to learn low- dimensional images of anthropological happenings 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.

A possible future way of research could be to cover this to many months of observational data. This would allow for totally new topics to be exposed, probably from the machine ingoing entirely new surroundings. Also, a "learning-rate" parameter could be reorganized online given new surroundings discovered by the machine in order to more quickly join on new anthropological actions being observed. Any topics removed, or not updated, could be considered as the robot "forgetting" a particular human activity.

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)
[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.

