Unsupervised human activity analysis for intelligent mobile robots

Ms. HINA HASHMI
Assistant Professor, TMU,
hina.computers@tmu.ac.in

Mr. SIDHARTH JAIN
BCA, TMU, Moradabad
siddharthj@tmu.ac.in

Abstract— **Research & Development in Artificial intelligence is growing day by day on a very large scale. Artificial Neural Network is a process of representation human mind that try to simulate its learning process. This paper shows the surveys on artificial intelligence that 1. What is Artificial Intelligence? 2. Comparison between human mind and Artificial intelligence. 3. How beneficial it can in long run or can also bring destruction in Human workforce.**

Keywords— **Artificial intelligence, Artificial, Neural, Network, Machine**

I. INTRODUCTION

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

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

Our methodology relies on first detecting and tracking human body movements from a single mobile robot's embedded sensors, along with learning the location of key objects in the environment using off-the-shelf techniques. Each human observation, originally recorded as a sequence of quantitative poses, is encoded using multiple qualitative calculi to abstract the exact spatial and temporal details of the observation, and finally represented as a vector of the occurrences of discrete qualitative descriptors (a vocabulary of which is learned from the data). We analyse the collection of encoded feature

vectors analogously to a corpus of text documents containing multiple topics of interest. Multiple latent topics are recovered from the observations and considered as human activity classes, each defined as a multinomial distribution over an auto-generated vocabulary. Two techniques are presented to learn low-dimensional human activity representations. First, a non-probabilistic low-rank approximation approach is shown to work well with pre-segmented video sequences of observed human activity. Secondly, a more sophisticated probabilistic Latent Dirichlet Allocation (LDA) [6] technique is shown to somewhat remove the requirement for manual temporal segmentation of the recorded observations, allowing the robot to access large quantities of data which otherwise would need human annotation. LDA is a hierarchical Bayesian model where each observation is modelled as a mixture over an underlying set of topics, and each topic is, in turn, modelled as a mixture over the discrete vocabulary

II. Related work

There is a common distinction in the literature between vision-based human activity analysis, which extracts information from video (and depth) cameras using computer vision techniques, and sensor or wearable computing-based systems [10, 11]. Sensor-based systems often rely on the availability of small sensors, namely wearable sensors, smart phones, or radio frequency identification (RFID) tagged objects, that can be attached to a human under observation in order to obtain a representation of that person's movements. We focus on representing human activity from visual data, where the notion of *being observed* is restricted to a single camera's field of view. This is a mature sub-field of artificial intelligence and the reader is pointed to survey papers which cover the topic in detail using, largely static RGB cameras [12–14] or 3D depth cameras [15,16]. However, many of the common techniques in these surveys perform supervised learning, where each training sample requires manual segmentation and annotating with a ground truth label. This is not a feasible solution for a long-term autonomous mobile robot which ideally, has minimal supervision whilst deployed in the real-world.

Unsupervised learning techniques are considered more appropriate for this task since they do not require time- consuming, offline manual annotations. Previous works have used Latent Semantic Analysis (LSA) [17], probabilistic LSA [18] and LDA [6] for learning low-dimensional human activity categories in an unsupervised setting: authors have combined these techniques with low-level Spatial Temporal Interest Point (STIP) features to learn action categories [19]; local shape context descriptors on silhouette images [20]; a combination of semantic and structural features to learn actions, faces and hand gestures [21]; and by fusing a vocabulary of local spatio-temporal volumes (cuboids) with a vocabulary of spin-images to capture the shape deformation of the actor [22];

However, a major problem cited in these works is the lack of spatial information about the human body captured by low-level image features, and the lack of more long-term temporal infor- mation encoded into the features restricts learning more complex actions. Descriptive spatial-temporal correlogram features have been used previously to address this issue [23], however, the approach still suffers from low-level image processing frailties, and the requirement for a single person to be modelled in the scene during a controlled training period. Another approach has been to learn the temporal relations between atomic actions in an unsupervised setting in order to accurately represent "composite" human activities [24]. However, the input videos for this technique require manual temporal segmen- tation into sequences of "overlapping fixed-length temporal clips", making it prohibitively expensive for life-long learning on an autonomous mobile robot. Further, each of these works have been performed without the variability of a mobile robot's frame of reference, and restricted to learning on temporally segmented video data during an offline training phase, unlike our work

III. QUANTITATIVE REPRESENTATION

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

Human activities

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

A *key object* is a semantic entity with a fixed location in an environment which provides some functionality that may be required for the execution of certain activities of interest in

that environment [44].

A *human activity* is considered as a partially ordered sequence of sub-activities (or repeated patterns) between positions of a person's body joints relative to key objects. In turn, these patterns (or sub-sequences) can be thought of as one or more simple qualitative relations holding between a person's body joints and/or a number of objects in the environment. For example, a person "picking up a cup" might comprise of the sequence: "reaching", "grasping" and "lifting" performed by the person's hand with respect to a cup.

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

Human pose estimates

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

convolutional network (ConvNet) 2D human pose estimator [52]. Subsequently, the *(x, y)* coordinates of OpenNI body joint positions are replaced with the superior 2D body joint coordinates provided by the ConvNet.

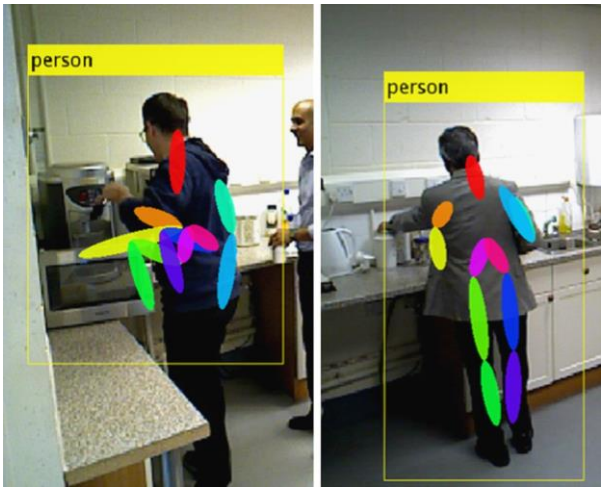
We represent the human pose estimates as ROS messages, where a single detected body joint location is represented by 3D Cartesian coordinates in a camera frame of reference along with the corresponding position transformed into the global

IV. Unsupervised learning for human activities

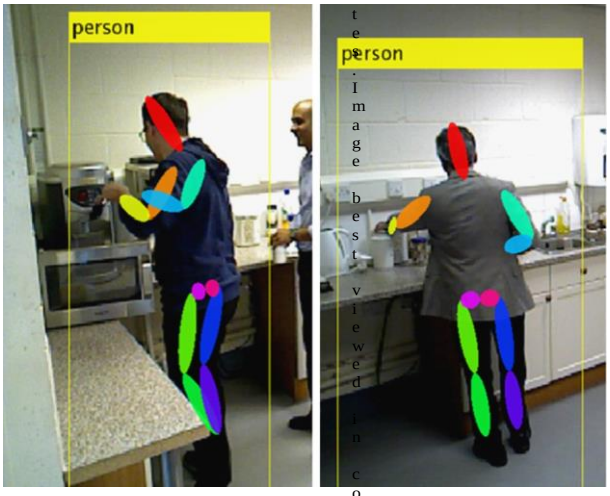
Encoding a corpus of human observations into such a term-frequency matrix allows latent structure can be recovered in an unsupervised setting. The aim is to learn low-dimensional representations of repeated structure encoded as qualitative descriptors (graph paths) across multiple similar observations.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

Low rank approximations for human activities

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

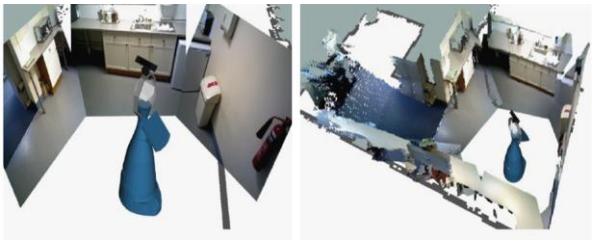


(a) Inaccurate OpenNI pose estimates



(b) Improved pose estimates

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



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

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

map frame of reference using the estimated location of the robot, i.e. j (id, x, y, z, xmap , ymap , zmap). A human pose then robot, we obtain a sequence of human poses over a time series of detections (camera frames). We define a human pose sequence, S p1, p2, ..., pi , .. . , where each pi is the detected human pose at timepoint i, and no restrictions are

The aim is to recover a small number of latent concepts from the

comprises of a collection of body joint locations, i.e. p j1, j2, , using the OpenNI/ConvNet implementation. For each human detected by the

placed upon the length of the recorded sequences. This variation in length is a major difficulty when using real-world data to learn activities on a mobile robot.

encoded data. The assumption is that common human activities relate

to repeated patterns of discriminative qualitative descriptors encoded within the observations. Ex- amining the decomposition, the non-zero eigenvalues in the diagonal matrix Σ represent the r most discriminative new compositional features, known as concepts. These latent concepts can be thought of as the activity classes encoded in the original term-frequency matrix. The columns of the left singular (M M) matrix U contain the eigenvectors of CC T , since CCT UΣΣT U T , and provides information, as a linear combination, about the weighting of each concept to each ob-servation, 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 ΣT ΣV T , and specify a linear combination of weights for each qualitative descriptor (codeword) used to describe each latent concept.

V. Limitations

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

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

VI Conclusion

In summary, we have introduced a

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

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

A possible future direction of research could be to extend this to many months of observational data. This would allow for totally new topics to be discovered, possibly from the robot entering entirely new environments. Also, a “learning-rate” parameter could be updated online given new environments explored by the robot in order to more quickly converge on new human activities

being observed. Any topics removed, or not updated, could be considered as the robot "forgetting" a particular human activity.

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

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.