Proceedings Paper:
Alomari, M, Duckworth, P orcid.org/0000-0001-9052-6919, Gatsoulis, Y et al. (2 more authors) (2017) A qualitative approach for online activity recognition. In: Advances in Cooperative Robotics. CLAWAR 2016: 19th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines, 12-14 Sep 2016, London, UK. World Scientific , pp. 747-754. ISBN 978-981-3149-12-0

https://doi.org/10.1142/9789813149137_0086

A Qualitative Approach for Online Activity Recognition

*Muhannad Alomari, *Paul Duckworth, Yiannis Gatsoulis,
David C. Hogg and Anthony G. Cohn

Intelligent Robotics Lab, School of Computing, University of Leeds, UK
{*scmara, P.Duckworth, Y.Gatsoulis, D.C.Hogg, A.G.Cohn*}*@leeds.ac.uk*
**The first two authors contributed equally to this work*

Keywords: Activity Recognition, Temporal Segmentation, HRI, QSR.

We present a novel qualitative, dynamic length sliding window method which enables a mobile robot to temporally segment activities taking place in live RGB-D video. We demonstrate how activities can be learned from observations by encoding qualitative spatio-temporal relationships between entities in the scene. We also show how a Nearest Neighbour model can recognise activities taking place even if they temporally co-occur. Our system is validated on a challenging dataset of daily living activities.

1. Introduction

Despite many years of research, Human activity recognition remains a challenging and ongoing area of research.[?] One particular challenge is the important problem of temporal segmentation which is especially difficult in video streams where multiple activities temporally co-occur or overlap. A variety of applications rely on learning activity models and temporal segmentation of video into human activities, such as: human robot interaction, smart surveillance systems, and semantic video database indexing. Advances in computer vision mean that a person's joints and many objects can be detected and tracked with a reasonable level of accuracy, enhancing the semantic information available to an activity recognition system.

In this work, we present an online activity recognition system that uses a novel, dynamic length temporal window in order to temporally segment an extended video which contains (possibly temporally overlapping) performed activities. Further, we learn activity models from generalising over multiple observations and encoding qualitative spatial relations between pairwise objects in the scene. This allows the system to detect a number of activities at once, which can temporally co-occur or overlap, or even be performed by multiple people in the scene at the same time.

Our objective in this work is to characterise movements of people and objects in a scene into semantically meaningful activities. For this purpose, we use RGB-D data and an object-centric abstraction method. Our input data is the positions of the human skeletal joints and of the objects in view. Figure 1 is an example of a daily living activity which displays minimum bounding rectangles (MBR) for the detected objects, and overlaid the skeletal joints (as red points and green lines) onto the RGB image. The images show one example of the "making cereal" activity and are taken from the Cornell Activities for Daily Living Dataset discussed in §6.[?]

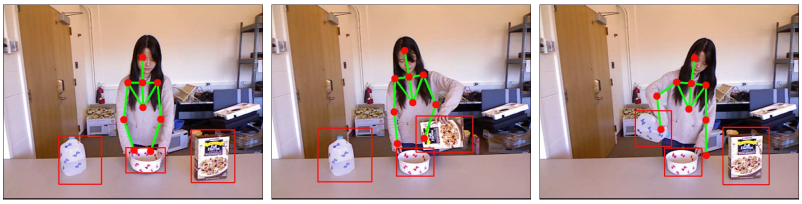

Fig. 1. Skeleton tracks and object detections on RGB images of "making cereal".

In this work we use a qualitative framework which allows us to abstract the exact metric positions of these joints and objects into a qualitative space. Abstracting into a qualitative space allows us to easily compare multiple observations, and draw conclusions even if they differ slightly in quantitative space. For example, the exact metric coordinates of a hand when it reaches for a mug is not important. A qualitative approach allows us to consider the spatial relationship between the mug and the hand, and represent this as a "moving towards" qualitative relation. This can then be compared to scene when a second person's hand also "moves towards" a mug, and we can consider them to be performing the action "reaching for a mug". We introduce the qualitative representations used in this work in §3.

§6 presents a validation against a daily living activities dataset. The results (how much our predicted temporal segmentation overlaps with the ground truth, along with a classification analysis) demonstrate the effectiveness of our qualitative representation and sliding window methodology.

2. Related Work

Due to the availability of cheap depth sensors, activity recognition systems are increasingly using data from RGB-D cameras to analyse and predict what a person is doing in a scene. Human activity recognition from 3D data has been reviewed in detail,[1] with a focus on data abstraction methods.

Temporal segmentation has also been researched previously, e.g. in tracking ballistic movements to form "units of human movement planning"[?] where a Bayesian framework is used to temporally segment videos containing actions into atomic movements. Although it works well for rapid and efficient movements, it has difficulty with hesitations, or slow, laboured movements.

In the literature, qualitative representations have been used to abstract visual data to capture and reason with higher level semantics, e.g.[2] The current state of the art work[3] on our chosen dataset uses a mixture of qualitative and quantitative features. However, this is processed entirely offline with no attempt at temporal segmentation.

Similarly to this work, a maximum entropy Markov model[?] has been used to detect activities, where sub-activity models have been learned in advance. However, this is also performed offline, and no temporal segmentation of live data is performed. Moreover, sub-activities are constrained to specific locations, which we do not need to do in our approach. Finally, popular approaches are often based on STIP features;[4] however the low level pixel values can easily be distorted by motion-blur or lighting variations.

3. Qualitative Spatial Representations

Many different qualitative spatial representations (QSRs) have been developed covering different aspects of space.[5] We use an open source software library, QSRLib,[6] to calculate the QSR values from the MBRs and joint positions. We focus on the distance relationship between objects and detected skeleton joints, and therefore use the *Qualitative Distance Calculus (QDC)* as our qualitative spatial representation, with three relations: *Touch*, *Near* and *Infinite*. In QDC representations, the thresholds between relations usually need to be manually set. Here, we learn the threshold values from labelled observations from the specific dataset. This allows our system to employ relations automatically tuned to the semantics of the domain.[?] For example, if we hand selected the distance threshold for the spatial relation *Near* and *Far* suitable for describing the interaction between a hand and a mug, the same threshold is unlikely to be useful to abstract visual data containing an aeroplane and an airport terminal, as semantics for *Near* and

Far would need to be different.

3.1. *Combining Qualitative Temporal & Spatial Knowledge*

Our implemented system must consider a continuous stream of video input data, and just as spatial positions may not repeat exactly over repeated instances of the same action, so also there may be temporal variations. We thus employ a qualitative temporal representation, the *Interval algebra (IA)*[7] which consists of 13 relationships (e.g. *meets, before* between pairs of intervals. We encode all the qualitative spatial and temporal information into a graphical structure[2,8] called a *Qualitative Spatial Temporal Activity Graph (QSTAG)*. Layer 1 of a QSTAG consists of the spatial entities observed and tracked in the video. Layer 2 nodes are *episodes* which are the maximal intervals of time over which some QSR relation holds between a pair of layer 1 objects. Layer 3 nodes specify the qualitative temporal relation between episodes. Figure 2 (bottom left) illustrates a QSTAG for a pair of objects (hand, mug). Figure 2 (top) shows the timeline and the QSRs which hold, and the episode boundaries. The episodes thus compress a sequence of frames with the same QSR.

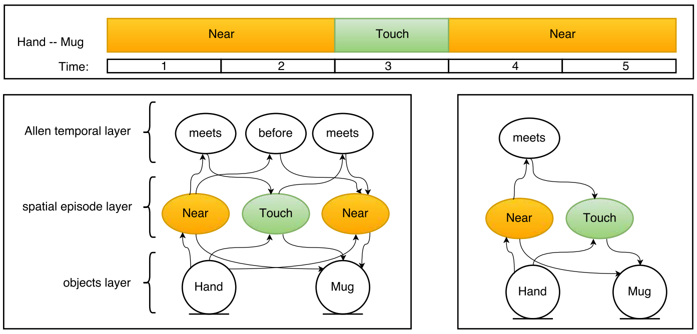

Fig. 2. Example of a time-line of QSRs, a QSTAG representation and a single graphlet

4. Learning Qualitative Models

Our aim is to learn and detect activities which may vary in complexity and length. For training purposes, we assume the video is segmented and labelled with the corresponding activity to be learned (potentially these segments could overlap). For each labelled activity we generate a QSTAG, and then learn an activity model consisting of a histogram of unique sub-graphs called *graphlets*.[2,8] The unique graphlets are mined from the QSTAG and can be thought of as small QSTAGs themselves; here we restrict these to *one* change of qualitative relation between a pair of objects, e.g. a hand initially *touching* an object and then distancing itself from the object, becoming *near*. An example graphlet is shown in Figure 2 (lower right), and these can be thought of as characterising primitive actions.

Given a set of detected objects and skeletal joints O in the scene, and a qualitative calculus with q relations, then there are

$$(|O| * \frac{(|O| - 1)}{2}) * (|q| * (|q| - 1)),$$

possible unique graphlets (given there is always two object nodes, two spatial nodes and a IA temporal node of *meets*). We ignore the *Infinite* spatial relation from QDC to reduce the total number of graphlets, so that $|q| = 2$ rather than 3. This gives us a fixed length attribute vector (in practice we only use the graphlets actually observed in the training data) and a single training instance is represented as a histogram over this feature vector. The set of all training instance histograms can then be used to classify new instances via a Nearest Neighbour Search (NNS).

5. Temporal Segmentation and Activity Detection

Temporal segmentation is a challenging task as every new frame yields a possible start or end for a candidate activity (and therefore all frames must be checked). We exploit the qualitative segmentation of time in a QSTAG to simplify the search. Once our activity models have been trained as above, our system is able to detect activities in real time using a dynamic length temporal sliding window over the k most recent episode boundaries (for a user defined k). Call this sliding window w. These episodes naturally vary in absolute frame-length, but this is abstracted away in the QSTAG representation. For each contiguous sub-sequence s of episodes in w (there are in general $(k^2 + k)/2$ possible such subsequences, including w itself), we classify s using NNS over the set of training instances (which may possibly yield no classification for some s if there is no activity "near enough" (see next section). Figure 2 is an example time-line representation of three episodes (six video frames). Every interval between episode start/end points becomes a candidate for the start of an activity. Hence, our system will classify six candidate temporal windows for activities, assuming the "current time" is set as just after the end of the third episode in Figure 2. Since the temporal windows can overlap, multiple activities can potentially be recognised even when co-occurring. Notice that the search space induced by our qualitative, episodic, representation of time is more coarse and hence more efficient to search than a typical frame-by-frame windowing segmentation; in the example above, there are six candidates since we consider every subsequence of length 1, 2 or 3 here (but this is still far fewer than considering all subsequences of actual frames).

6. Experiments

We test our system on a benchmark activities dataset which includes skeletal joint tracks and objects.[?] The dataset consists of four human participants, each performing 10 daily living activities thrice, totalling 120 different videos. We provide two separate analyses on this dataset.

Our first analysis is designed to highlight our system's ability to run online, over periods of time much longer than an individual activity and perform effective temporal segmentation of multiple activities. For this purpose we stitch all the videos from each participant together, into four long streams of video (each approximately 12,000 frames and 10 minutes in duration). We perform four-fold cross-validation (cv) where three participants are used to train our activity models, and the fourth (unseen) participant's video stream is analysed in an online setting. We perform a supervised Nearest Neighbour Search (NNS), with a variable radius applied to each training vector (based upon the Frobenius norm of two vectors). Applying this radius allows our system to recognise multiple activities co-occurring and similarly recognise that nothing is occurring. There are two ways we might recognise simultaneous activities: (1) if the exact same subsequence of episodes is within 2 activity radii, or (2) if two overlapping episode subsequences both have activities within radius.

Figure 3 shows the temporal segmentation of the activity videos for participant 3. The x-axis is a time-line over the combined video for the participant and each colour represents a different activity; vertical overlaps of the same colour represent the correctly classified frames. In this dataset as it happens there are no simultaneous activities but the method has the potential to detect them. Empirical recognition results, over all four participants, are presented in Table 1 and show an average 0.88 $F1$ score over all 48340 frames.

Note if the test video (represented as a histogram of graphlets) has no neighbour within the variable radius threshold, our system classifies nothing is happening. However, the ground truth for this dataset suggests that the activities start immediately after one another, which is not accurate as there is a period of no movement at the beginning of each video; hence our results maybe "more accurate" than the ground truth. Finally, we only tested on the activity classes that had all the objects tracked.

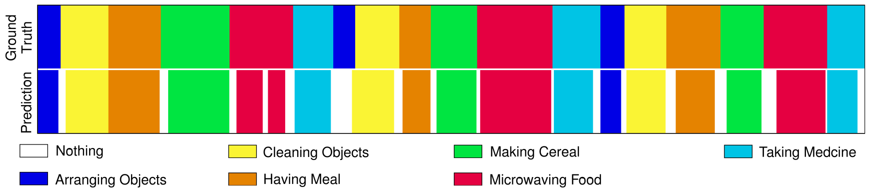

Fig. 3. Online temporal segmentation of participant 1's videos combined back to back.

test participant	frames	precision	recall	$F1$
P 1	11185	0.9699	0.7715	0.8594
P 2	12730	1.0	0.8552	0.9219
P 3	10283	1.0	0.8048	0.8918
P 4	14142	0.9191	0.8198	0.8667
Avg:		0.9723	0.8128	0.8849

In our second analysis, we perform a classification task over each video individually. We used four-fold cv (where each participant's videos are a fold). Each video from the test participant is then classified (using the same supervised NNS with a variable radius) into a predicted activity class.

Out of 72 videos with object detections, 63 were correctly classified leading to an average of 87.5% accuracy. It can be seen that our system performs very well, especially given the challenges in the dataset, such as; the variations in the activities being performed between participants (intra-class variation); the similarity of different classes (between-class similarity); the testing methodology (on an unseen participant); and finally, the input skeleton joint and objects tracks, which are less than perfect.

7. Conclusion

In this paper, we presented a novel qualitative, dynamic length sliding window (suitable for deployment on a mobile robot) to learn and temporally segment observed activities. Qualitative spatial relations between observed entities are abstracted into a relational graph. An activity model is then built by representing training instances of activites as histograms over graphlets which occur in the training instances. A dynamic qualitative temporal window is used to reduce the search space during recognition which allows the system to understand complex scenes. The system therefore has the potential to recognise multiple activities temporally co-occurring with different scales of complexity; or multiple people performing those activities.

Planned future work includes obtaining data to learn a hierarchy of activities, and then representing these higher level activities as QSTAGs over lower level activities rather than QSR episodes. This would enable the representation and detection of complex levels of the daily living activity-hierarchy such as multiple people performing activities, or activities such as preparing a meal consisting of a temporally overlapping instances of our current activity models.

References

1. J. Aggarwal and L. Xia, *Pattern Recognition Letters* **48**, 70 (2014).
2. M. Sridhar, A. G. Cohn and D. C. Hogg, Unsupervised learning of event classes from video, in *AAAI*, 2010.
3. J. Tayyub, A. Tavanai, Y. Gatsoulis, A. G. Cohn and D. C. Hogg, Qualitative and quantitative spatio-temporal relations in daily living activity recognition, in *ACCV 2014*, 2015.
4. M. S. Ryoo and J. K. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in *ICCV*, 2009.
5. J. Chen, A. G. Cohn, D. Liu, S. Wang, J. Ouyang and Q. Yu, *The Knowledge Engineering Review* **30**, 106 (2015).
6. Y. Gatsoulis, P. Duckworth, C. Dondrup, P. Lightbody and C. Burbridge, QSRlib: A library for qualitative spatial-temporal relations and reasoning(Jan 2016), qsrlib.readthedocs.org.
7. J. F. Allen, *Communications of the ACM* **26**, 832 (1983).
8. P. Duckworth, Y. Gatsoulis, N. Jovan, F.and Hawes, D. Hogg and A. Cohn, Unsupervised learning of qualitative motion behaviours by a mobile robot, in *AAMAS*, 2016.