Qualitative and Quantitative Spatio-Temporal Relations in Daily Living Activity Recognition

Jawad Tayyub, Aryana Tavanai, Yiannis Gatsoulis, Anthony G. Cohn and David C. Hogg

School of Computing, University of Leeds, Leeds, LS2 9JT, UK

Abstract. For the effective operation of intelligent assistive systems working in real-world human environments, it is important to be able to recognise human activities and their intentions. In this paper we propose a novel approach to activity recognition from visual data. Our approach is based on qualitative and quantitative spatio-temporal features which encode the interactions between human subjects and objects in an efficient manner. Unlike the state of the art, our approach uses significantly fewer assumptions and does not require knowledge about object types, their affordances, or the sub-level activities that high-level activities consist of. We perform an automatic feature selection process which provides the most representative descriptions of the learnt activities. We validated the method using these descriptions on the CAD-120 benchmark dataset, consisting of video sequences showing humans performing daily real-world activities. The method is shown to outperform state of the art benchmarks.

1 Introduction

One of the most challenging areas of research in the fields of computer vision and pattern recognition is learning and understanding human activities from observed visual data. The research question is, given a sequence of images with one or more people performing various activities, is an intelligent system capable of recognising the activities that are being performed? Despite its long research history [1–4] finding a universal semantic representation for activity analysis is still a difficult challenge due to the complexity of human activities and the variability of how these activities can be performed, even by the same person. Activity analysis is often investigated from a security domain perspective, as automatic recognition of human behaviour in sensitive areas is a critical issue for video surveillance [5–7]. Recently however, understanding daily human activities has also become popular in moving towards smart environments and robotic assistive living, where activity analysis is vital for effective operation.

In prior work, two distinct approaches have been adopted, those that first detect objects and then examine the spatial and temporal relationships between these objects [8], and those that examine patterns of image features directly without first detecting the objects [9]. In the object-level based approaches,

some use qualitative relations between pairs of objects (e.g. disjoint, partially-overlapping), whilst others work directly with quantitative relations such as distances.

In this paper, we propose a novel method for activity recognition that combines quantitative and qualitative representations, feature selection and a standard multi-class classifier. The method significantly improves on the state of the art performance on the publicly available activity CAD-120 dataset from Cornell.[1]. The prior state of the art on this dataset [10, 11] learns and recognises human activities by modelling the sub-activities from which they are composed and the affordances of the objects involved, as well as how these change over time and relate to one another. Although the recognition of sub-activities and object types/affordances may be important for some applications, we show that it is possible to achieve a very high level of recognition performance of high level activities without either of these.

The rest of the paper is organised as follows. Section 2 presents related work. Section 3 describes our proposed framework in detail. In Section 4 the experimental results are presented and discussed. The conclusions and future work are presented in Section 5.

2 Related Work

Qualitative spatio-temporal relations are primarily successful as they capture key spatial and temporal changes in visual data, and have become quite common in representing activities in various approaches. These approaches are analysed in this section.

In previous work [12–14] spatial relations based on the well established RCC spatial calculus [15–17] were combined with temporal relations based on Allen's Interval Algebra [18] to produce a qualitative spatio-temporal graph that represents an activity. Although previous results have demonstrated the effectiveness of this method, its lack of quantitative features that are not encapsulated by the qualitative ones makes the method unable to distinguish between events and activities where these quantitative features are important. In our experiments we demonstrate the importance of using quantitative features together with qualitative features.

The RCC spatial calculus together with Allen's Interval Algebra has also been used in [19], but in that work pre-defined knowledge of the object categories was also exploited and together with the spatio-temporal features and using Inductive Logic Programming (ILP) the developed system was able to learn and recognise observed human activities. Although the system demonstrated successful results and its ability to avoid over-fitting of the training dataset was a key strength, its reliance of prior knowledge about the categories of the objects in conjunction with its strict classification approach due to ILP, causes performance issues when these are missing.

[1]CAD-120: http://pr.cs.cornell.edu/humanactivities/data.php

A hierarchical approach using variable length Markov models and taking as observations the contour of a human body in terms of control points has been developed to learn and recognise human activities, and has been evaluated in exercise activities that require no object interactions [20, 21]. Given the nature of the feature vector which takes into account the contour of the object rather than actual spatial relations between spatial interest points, it is questionable whether it will be able to deal with activities that involve object interactions.

Other approaches [22, 23] have used interest point detectors, extracted a 3D cuboid at each interest point, computed descriptors for each of the cuboids and then clustered similar descriptors together hence forming a feature descriptor codebook similar to the traditional bag-of-words approach. An extension to this method is using a probabilistic approach that combines prior domain knowledge to model each activity as a distribution over the codewords and each video as a distribution over the activities [24]. Although the advantage of these approaches that use image descriptors is that they do not require skeleton or object tracks to describe the activity observed, they are unable to take into account spatio-temporal relations between the different relevant entities in the scene, which are important elements when learning and recognising human activities [25, 17]. To address this issue, the concept of a "spatio-temporal phrase" that is defined as a combination of local words in a certain spatial and temporal structure, including their order and relative positions is introduced [26]. This is a very similar approach to the graphical representation described before [12–14], however, the spatio-temporal phrase still does not include qualitative spatial relations and also the temporal relations are much fewer than the Allen's Interval Algebra used in the graphs method.

An alternative approach is using convolutional deep learning methods to learn templates of the patterns of the activities and then be able to recall them [27–29]. Although deep learning methods have demonstrated impressive results in visual pattern matching, they require large training datasets and training is very computationally expensive. Furthermore, like the bag-of-words approaches for activity recognition, deep learning methods have to-date operated at the image level and do not consider rich spatio-temporal relations among the relevant entities in the scene. Other approaches [30, 31] make use of low-level optical flow input and build high-level spatio-temporal representations of the activity. Ryoo [30] extracts feature points from a video and describes the scene by modelling spatio-temporal relations between these feature points. Similarly, Brendel [31] builds on representing activities as spatio-temporal graphs generated from pixel intensities and motion properties in the video. These approaches show promising results but suffer from image-level distortions, such as motion-blur, lighting changed etc., and do not capture high-level scene reasoning.

In our approach, we make use of the CAD-120 dataset. Much work has been done using this dataset. Benchmark setting approaches developed by Koppula [10, 11], Rybok [32] have focused on modelling activities using generalized description of objects. Koppula [10, 11] made use of object affordances, i.e. the purpose of an object, in order to build stronger models of activities, supporting

the hypothesis that the use of object affordances instead of specific object descriptions is more beneficial since it is more important to know what an object does rather than what an object is. Rybok [32], similarly, generalizes object modelling by representing regions in a scene where objects are interacting through detection of salient object features rather than complete objects themselves. These approaches, however, still heavily model objects in order to recognize activities. In our work, we give equal weight to modelling all interactions amongst elements (all skeleton joints and objects) in a scene thus removing heavily weighted bias towards object modelling alone.

3 Framework

We propose that in order for an intelligent system to effectively recognise observed human activities, we encode both the qualitative and quantitative spatio-temporal relations of the relevant entities in the scene. For this we research and develop a method that allows an intelligent system to learn and recognise high-level activities by selecting the most important and discriminative features from a set of feature templates that were designed based on qualitative and quantitative spatio-temporal feature representations (QQSTR) of the activities. The resulting selected features are then used to train a multi-class support-vector machine (SVM) for future prediction. These steps are shown in Figure 1.

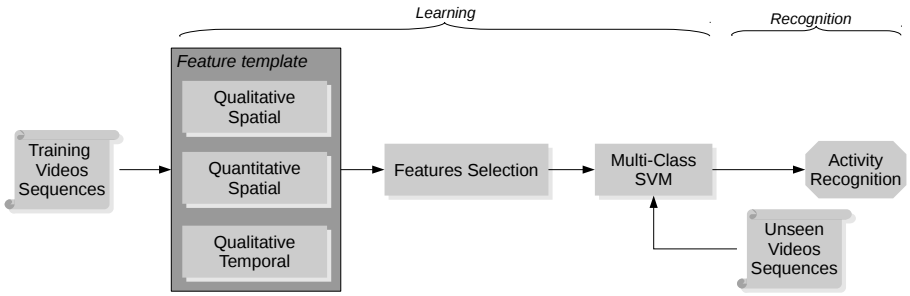

Fig. 1. Flowchart showing the high-level steps of the QQSTR

Before the detailed explanation of the model is given, it is useful to briefly describe the key terms in the QQSTR: spatial, temporal, qualitative and quantitative.

Spatial: These features describe properties and relations between objects that exist in space. Examples of spatial features are poses of objects, relative poses of objects with respect to other objects, absolute and relative direction of motion, etc.

Temporal: These features describe properties and relations of the objects or activities themselves in the time domain. Allen's interval algebra is an example of temporal relation features between two events. Other examples of temporal features are (a) the time an activity starts and its duration, (b) the time before a car runs out of fuel, etc.

Qualitative: The term qualitative is defined as "relating to, measuring, or measured by the quality of something rather than its quantity"[2]. For example a qualitative spatial feature is that two objects are partially overlapping each other without specifying the proportion of overlap. A qualitative temporal feature example is that an activity starts and finishes before another activity starts. Both, RCC and Allen's interval algebra are qualitative relational frameworks.

Quantitative: In contrast to qualitative, the term quantitative is defined as "relating to, measuring, or measured by the quantity of something rather than its quality"[2]. An example of a spatial quantitative feature is saying that two objects overlap each other by 30%; an example of a temporal quantitative feature is saying that an activity finishes 5 minutes before another activity starts.

Our feature set F consists of three components, namely qualitative spatial, quantitative spatial and qualitative temporal components. All three components comprise of histograms and statistical measures both of which are noise resilient. We have chosen not to include a quantitative temporal component as we found that the qualitative temporal components encodes sufficient temporal information in the domain under consideration and including quantitative component would result is unnecessary additional complexity to the feature space in F.

$$F = \langle F_1, F_2, F_3 \rangle \tag{1}$$

In equation 1, F_1 is the set of qualitative spatial features, F_2 is the set of qualitative temporal features and F_3 is the set of quantitative spatial features. The complete feature set F then undergoes minimum-redundancy maximum-relevancy (MRMR) [33] feature selection in order to identify features from each set F_1, F_2 and F_3 that have a significant contribution. This selection step provides the minimal and most discriminative representation of an activity.

3.1 Qualitative Spatial Representations (F_1)

The qualitative spatial representation (QSR) used is based on the well-established Region Connection Calculus-5 (RCC-5) [16], which is a binary mereological calculus containing 5 relations. We use a still coarser representation which we refer to as RCC-3. It contains the relations DR (discrete), PO (partial overlap) and PiP {the union of the RCC-5 relations PPi, PP, EQ} (Part, Part inverse and equality). These form a Jointly Exhaustive and Pairwise Disjoint (JEPD) set of relations and, as with RCC-5, RCC-3 holds between pairs of entities (tracked

[2] http://www.oxforddictionaries.com

objects and human body parts) in n-dimensional Cartesian space. All three relations are symmetric, though in our work we arrange the use of PiP such that the first argument is always a part of the second (or equal to it, however in practice equals rarely occurs). This is effectively the representation used in [13] and for representational convenience we use D and P rather than DR and PiP in the rest of the paper. These RCC-3 relations are graphically shown in Figure 2 and they are denoted as $R = \{\text{D, PO, P}\}$.

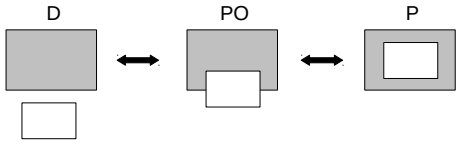

Fig. 2. Region Connection Calculus-3 showing the three distinct relations between a pair of objects

We first compute these pairwise RCC relations from the tracks of the entities, producing sequences of RCC relations for each pairwise combination of entities. A low-pass filter is then applied across the sequences of relations to suppress any jitter caused due to objects and skeleton detection error. An important aspect in QSR activity recognition is to model the relation changes that occur between entities, as these represent the discriminative stages of an activity. In related work [19, 13, 14], this is achieved by aggregating repeated consecutive occurrences of a relation for each pairwise combination of entities. In other words, by parsing individually every sequence row $S_{e_i,e_j} \in S$ which are the RCC chains (sequences) between any two entities e_i and e_j, each chain is suppressed while $S_{e_i,e_j,t} = S_{e_i,e_j,t-1}$. By only locally focussing on how the RCC relations of a specific pair of objects is evolving, it limits the representative strength of the spatial feature, as it ignores how the changes in S_{e_i,e_j} are affecting the changes in the spatial relations of the rest of the entities; i.e. local segmentation ignores the holistic picture.

We propose an alternative approach to suppress the spatial relation chains in S when only all chains are the same as the ones before. Again in simple terms, instead of looking individually at every S_{e_i,e_j} which are the rows of S, we suppress the RCC relations while $S_t = S_{t-1}$, i.e. if there is a change between column t and $t-1$ of S. An example of local and propagated segmentation is illustrated in Figure 3 and the benefits of propagated segmentation over local segmentation are demonstrated in the experimental results section. Finally, we compute the number of occurrences of sub-sequences of length 1, 2, 3 and 4 in the propagated segmented RCC sequence, as these sub-sequences represent the minimal blocks that describe an activity in terms of qualitative spatial relations. The histogram of these sub-sequences is our qualitative spatial feature F_1.

Formally this procedure is described as follows. Let E be the set of entities in the scene. Then at each frame, t, we compute between entities $e_i, e_j \in E$

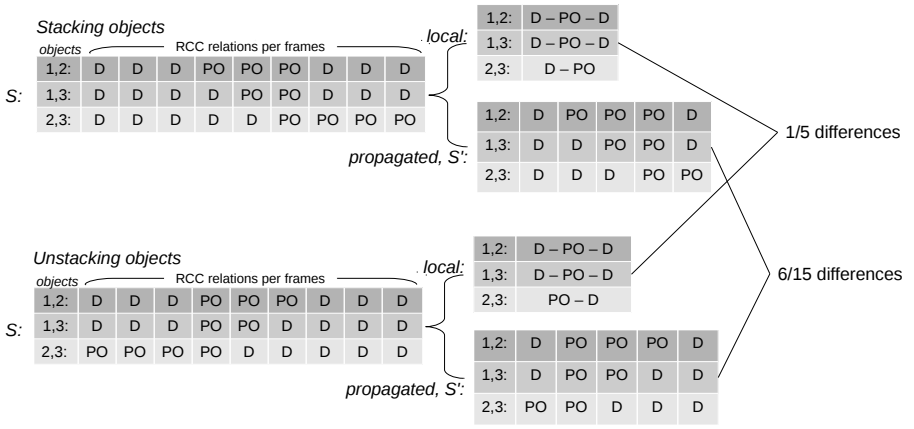

Fig. 3. Example showing the expressive power of propagated segmentation over local segmentation in two activities (stacking and unstacking objects)

$(i \neq j)$, the RCC relation $R_{e_i,e_j,t}$, from the Cartesian positions of these two entities at this frame. This produces a $m \times n$ matrix S of r relations for all m possible pairwise combinations of the entities for each one of the n frames in the video. Propagated segmentation on S is computed by aggregating relation changes that occur in any of the pairwise relations sequences $S_{[t:t+1]}$, producing a new matrix S'. Then, we form a histogram of all possible RCC relation changes with length $l \in [1:4]$ that occur in the whole video represented by S' for all m pairwise combination of entities. The total number of these histogram bins is $m \times \sum_{l=1}^{4} \text{len}(R)^l$, where as previously described $\text{len}(R) = 3$. For example, consider the propagated segmentation S' for *stacking objects* activity shown in Figure 3. The bins of the histogram for object 1 and 2 would be $\langle D; D\text{-}PO; \ldots; D\text{-}PO\text{-}PO\text{-}D \rangle_{1,2}$ with counts $\langle 2; 1; \ldots; 1 \rangle$, and all other possible RCC combinations filled in with zeroes so that the length of the histogram feature has the same length for all activities. This is repeated for the remaining pairwise combinations $(1,3$ and $2,3)$ and the resulting histograms for each pairwise combination are joined together to form the complete feature representing the activity video. This becomes the qualitative spatial representation feature set F_1, with length of $\mathbb{R}^{120 \times m}$.

3.2 Qualitative Temporal Representations (F_2)

Spatial changes in F_1 alone do not capture the notion of time in an activity which, for some activities, are maybe important. For example, there might be similarity of the spatial relations between the *talking on the phone* and *biting an apple* activities; capturing the time dependencies between spatial changes can help in discerning such situations. One commonly used method for describing temporal relations is Allen's Interval Algebra [18]. Since Allen's temporal relations do not encode quantitative duration relations, we expand the *meets* relation, by adding

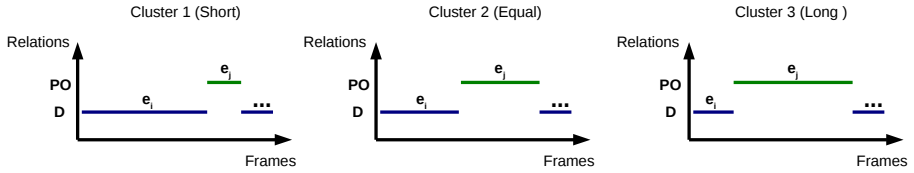

Fig. 4. Ratios computed between relative lengths of two consecutive different spatial relations e_i and e_j, then discretised into one of three clusters representing *short*, *equal* or *long* durations.

a qualitative measure of relative duration between two consecutive different spatial relations. More formally, from the non-segmented spatial relations sequences of S we calculate the relative duration ratio d:

$$d = \frac{\text{len}(e_j)}{\text{len}(e_i)},\ e_i \text{ meets } e_j \tag{2}$$

To obtain a qualitative measure of the relative duration, the quantitative ratios are clustered together using k-means. In practice, we found that discretisation of the continuous space of ratio values can be sufficiently captured with three clusters ($k = 3$). As illustrated in Figure 4, these three clusters give the notion of duration ratios as either being *short*, *equal* or *long*[3]. Once the temporal relations are in a qualitative form it is possible to compute their histogram which is our qualitative temporal feature F_2, with a length of $\mathbb{R}^{9\times m}$. The total number of features in F_2 is given by $m \times (\text{len}(R)^2 - \text{len}(R)) \times k$ where m is the number of all pairwise combination of objects, $\text{len}(R)^2 - \text{len}(R)$ denotes the total number of pairwise combinations of spatial relation changes with no repetition. For example for objects 1 and 3 in S of activity *stacking objects* shown in Figure 3, the bin $\langle (D\text{-}PO)_{short}; (D\text{-}PO)_{equal}; \ldots; (PO\text{-}D)_{long} \rangle_{1,3}$ provides the counts $\langle 1; 0; \ldots; 1 \rangle$.

3.3 Quantitative Spatial Representations (F_3)

Qualitative spatial representations successfully abstract a good representation of a video scene through capturing interactions. However, due to the coarse representation of space and time, it is often not possible to discern similar looking activities that are performed at a different scale or speed. Quantitative spatial representations, on the other hand, are able to encode such finer motions in an activity as seen in previous work [35]. In our approach we make use of various quantitative spatial representations to aid our model in the recognition problem.

[3]Note that this is similar to the INDU calculus [34] which extends the interval calculus by discretising whether intervals in a before, meets or overlaps relationship are shorter, equal or longer than each other.

Euclidean distances: We compute the Euclidean distances between the centroid of the bounding boxes of each pair of elements in the scene. An element could be any of the following skeleton parts: head, hands, torso, shoulders, hips, and any of the objects in the scene. Lower body parts are not used due to their high rate of occlusion. For a compact and generic representation we compute the descriptive statistics of this distances distribution, namely we compute the mean (μ), standard deviation (σ), kurtosis (κ) and skewness (γ). We use these statistical measures as quantitative spatial features.

Relative direction of motion: A problem with some qualitative spatial relations is that in some cases they are unable to distinguish mirror activities, e.g. pushing and pulling. To resolve this issue we calculate the relative direction of motion between two objects, i.e. whether two objects are approaching or departing from each other. We calculate this for every possible pair of objects using their timed minimum and maximum Euclidean distances. By knowing how the relative direction of the motions change for pairs of objects in the scene it is possible to distinguish between mirror activities.

The descriptive statistics of the Euclidean distances combined with the relative direction of motion of the entities, form the quantitative spatial representation F_3. The number of total features in F_3 is given by $\mathbb{R}^{5\times m}$ where m denotes all pairwise combinations of entities and 5 represents the number of statistical metrics of Euclidean distances plus relative directions of motion.

3.4 Feature Selection and Learning

The feature template F is the most generic representation of an activity. However, the importance of each of the features in it is determined by the nature of the activities. We employ a feature selection step that automatically identifies from F the feature set $F^{'}$ that is more discriminating for activity classes c. We apply this by using the *Minimum-Redundancy Maximum-Relevance* (MRMR) feature selection method [33], which is based on mutual information between two random variables, α and β as shown in Equation 3. Specifically, MRMR is based on two criteria, namely maximum-relevance and minimum-redundancy, which are described below.

$$I(\alpha;\beta) = H(\alpha) + H(\beta) - H(\alpha,\beta) \tag{3}$$

Maximum-Relevance: This criterion approximates Max-Dependency $D(F,c)$, by searching for features using the mean of all mutual information values between feature x_i and classes c, that satisfy equation 4.

$$max\ D(F^{'},c), D = \frac{1}{\mid F^{'} \mid} \sum_{x_i \in F^{'}} I(x_i;c). \tag{4}$$

Minimum-Redundancy: If any two features, x_i and x_j, have a high dependency between them, one of them is redundant in the feature set, $F^{'}$, and

is therefore removed while still preserving the discriminative-class power of the feature set. Therefore, by using the Minimal-Redundancy criterion, as shown in Equation 5, mutually exclusive features are chosen.

$$min\ R(F^{'}), R = \frac{1}{\mid F^{'} \mid^2} \sum_{x_i, x_j \in F^{'}} I(x_i; x_j). \tag{5}$$

By defining the operator $\phi(D, R)$ to combine these two criteria, as shown in Equation 6, the minimal most discriminative feature set, $F^{'}$, for a given set of activities classes is obtained.

$$max\ \phi(D, R), \phi = D - R \tag{6}$$

We use the feature set $F^{'}$ to train a multi-class SVM [36] to recognise the high-level activities using a polynomial kernel of degree $d = 2$ and $\gamma = 0.75$.

4 Results and Discussion

We evaluate our framework on the Cornell activity dataset (CAD-120)[4] which we describe in section 4.1. We then describe our experimental setup and evaluation method in section 4.2. In section 4.3 we compare our approach against a state of the art benchmark. Section 4.4 provides an in depth analysis and discussion of the strengths of our approach.

4.1 Description of the Benchmark Dataset

CAD-120 comprises of 120 RGB-D video sequences of four human subjects performing daily living activities which are recorded using a Microsoft Kinect camera. Out of these four subjects two are male and two are female; three are right-handed and one is left-handed. Each video is labelled with a single high-level activity name: *making cereal*, *taking medicine*, *stacking objects*, *unstacking objects*, *microwaving food*, *picking objects*, *cleaning objects*, *taking food*, *arranging objects* and *having a meal*. The dataset provides skeleton tracks of people in the scene, as well as auto and ground truth tracks of the objects present in each one of the videos. Figure 5 shows some sample images of the dataset.

4.2 Experimental Procedure

For validation and comparison, we follow the same evaluation procedure as the one presented in the current state of the art [10]. We adopt a 4-fold cross validation approach where we train on three subjects and test on the fourth new subject. In addition, on the training set we perform a 3-fold cross validation for the feature selection process where we train on two subjects and we use the third subject for feature selection with the method described in Section 3.4. We then

[4] http://pr.cs.cornell.edu/humanactivities/data.php

Fig. 5. CAD-120 dataset sample screen shots[4]

combine the extracted features of each of the three folds together and remove repetitions to form our final most-discriminative features set.

We then use the new feature set to compute the results of the main testing fold which we take an average across the four folds. We report the micro accuracy, macro precision and macro recall for the activity recognition. Micro accuracy is the average of the percentages of correctly classified labels across the four folds. Macro precision and recall are the averages of precision and recall respectively for all classes.

4.3 Activity Recognition Results

Table 1 shows the performance of our approach on high-level activity recognition of the CAD-120 dataset. It can be observed that we achieve an accuracy of 95.2%, precision of 95.2% and recall of 95.0%. This is a significant improvement of 12.1%, 8.2% and 15.0% in terms of accuracy, precision and recall when ground-truth temporal segmentation of sub-activities is not known, as well as an improvement of 1.7%, 0.2% and 1.7% when it is known. The assumption of knowing the temporal segmentation of the sub-level activities is not required by our method, but it is needed by the benchmark method. These results demonstrate that our approach efficiently and effectively captures the interactions between the human subjects and the objects without needing any prior knowledge about the types and the affordances of the objects in the scene or knowledge of sub-level activities. Figure 6 presents the confusion matrix obtained with ground truth bounding boxes. From the strong diagonal it is evident that there is nearly no confusion in discriminating different high-level activities.

The results presented so far are obtained using ground-truth object tracks. We evaluate our method on more realistic scenarios by using the noisy automatic object tracks provided by the CAD-120 dataset. We compare our results with [10, 32]. Rybok et al. requires no object tracks as their method is based on saliency and optical features. Table 1 shows that our method is robust and achieves comparable results to the other two methods. Specifically, it achieves an accuracy of 75.8%, which is only 2.4% lower than the highest performance by Rybok et al. Results also show marginal increase of 0.8% over results of Koppula et al.

Figure 7 illustrates the confusion matrix for our results using automated object tracks. We can observe that most activities obtain a high accuracy while

Table 1. Performance measurements with or without ground-truth temporal segmentation based on accuracy, precision and recall

Method	Accuracy	Precision	Recall
assuming ground-truth temporal segmentation			
Koppula et al. [11]	84.7 ± 2.4	85.3 ± 2.0	84.2 ± 2.5
Koppula, Saxena [10]	93.5 ± 3.0	95.0 ± 2.3	93.3 ± 3.1
assuming no ground-truth temporal segmentation			
Koppula et al. [11]	80.6 ± 1.1	81.8 ± 2.2	80.0 ± 1.2
Koppula, Saxena [10]	83.1 ± 3.0	87.0 ± 3.6	82.7 ± 3.1
QQSTR-gt-tracks	$\mathbf{95.2} \pm 2.0$	$\mathbf{95.2} \pm 1.6$	$\mathbf{95.0} \pm 1.8$
assuming no ground-truth temporal segmentation and no ground-truth object bounding boxes			
Koppula et al. [11]	75.0 ± 4.5	75.8 ± 4.4	74.2 ± 4.6
Rybok et al. [32]	**78.2**	-	-
QQSTR-auto-tracks	75.8 ± 6.8	$\mathbf{77.9} \pm 11.0$	$\mathbf{75.4} \pm 9.1$

there is confusion between the *cleaning objects* and *taking food* activities. We suspect that this confusion between these two activities occurs due to potentially high level of noise in the object tracks. This suspicion is supported by Figure 6 which shows that when the tracks are noiseless a high degree of separation is achieved.

4.4 Discriminative strength of features types

We firstly investigate the strength of QQSRT (F) versus using individual and pairwise combinations of the different feature types. Figure 8 shows the accuracies for ground-truth and automatic tracks for F and all the different combinations of F_1, F_2 and F_3. It can be seen that F outperforms all other combinations. However, there are other interesting observations. To begin with, qualitative spatial representation (F_1) and qualitative temporal representation (F_2) seem to be robust to noisy automatic tracks. On the other hand, quantitative spatial representation F_3, although more prone to noisy tracks, achieves a higher performance in the case of smooth tracks. Furthermore it can be seen that F_1 and F_3 when combined together have a higher discriminating ability than when combined with F_2. This is confirmed by Figure 9 which shows post feature selection performance contributions of F_1, F_2 and F_3. It can be seen that the contribution of F_2 in F is much lower than those of F_1 and F_3. Despite the fact that F_2 is contributing less, it is still an important component of the overall feature set since its inclusion achieves the highest performance.

We next evaluate the benefit of propagated segmentation over local segmentation as described in the methodology section. Figures 10 and 11 show the confusion matrices obtained when using local and propagated segmentation re-

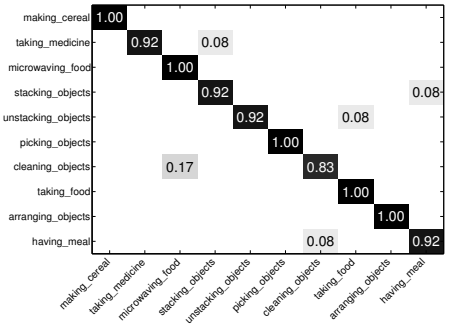

Fig. 6. Confusion matrix with ground-truth object tracks

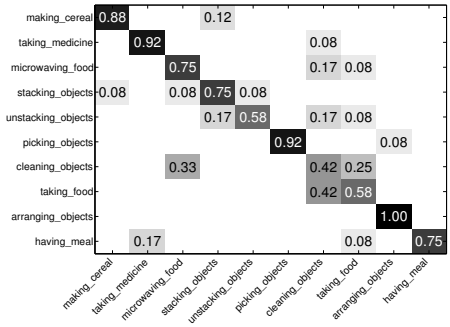

Fig. 7. Confusion matrix with automated object tracks

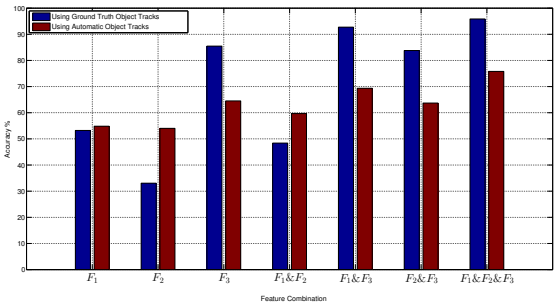

Fig. 8. Accuracy for different combinations of features types

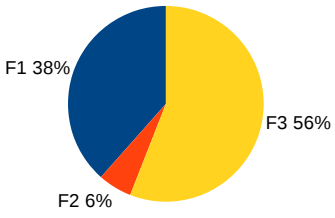

Fig. 9. Ratio of the number of each feature type over the total number of selected features

spectively. To objectively investigate the effect of propagated segmentation on the qualitative spatial relations F_1, we perform experimentation as before using F_1 alone. It can be seen that implementing the global propagation eliminates confusion between the mirrored activities of *stacking* and *unstacking*. These results validate our hypothesis that by taking into account the holistic picture and looking at how the spatial relations change at the global level yields much better results than a narrow focus on individual relational changes.

Lastly, we evaluate the performance with and without employing feature selection. Table 2 shows these results. It can be seen that feature selection plays a significant role in achieving high performance of 95.2%.

Table 2. Performance measurements with and without feature selection

	Accuracy	Precision	Recall
QQSTR without feature selection	79.8 ± 1.5	82.45 ± 7.4	79.17 ± 7.8
QQSTR with feature selection	95.2 ± 2.0	95.2 ± 1.6	95.0 ± 1.8

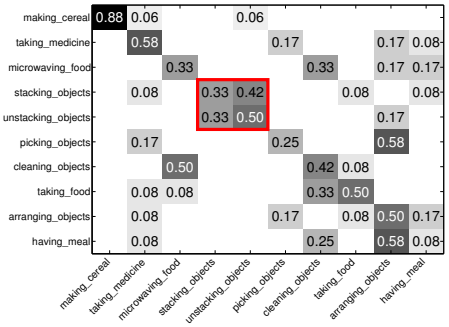

Fig. 10. High confusion between *stacking* and *unstacking* activities, highlighted in red, is evident when using local segmentation

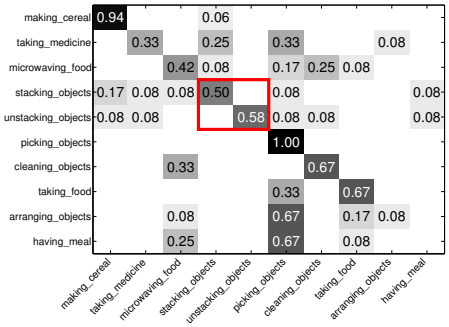

Fig. 11. There is no confusion between *stacking* and *unstacking* activities, highlighted in red, when using propagated segmentation

5 Conclusions and Future Work

In this paper we proposed a novel method of learning and recognising complex human high-level activities from video sequences. The method is based on qualitative and quantitative spatio-temporal features that capture the person-object interactions in the observed scene in a generic and effective manner. From these features we automatically selected the most discriminative ones and trained a multi-class SVM. We showed that the task of finding the most discriminative features from the original set is an important step. Unlike state of the art methods in activity recognition our method makes very few assumptions and does not need knowledge of object types, their affordances or sub-level activities that compose the high-level activity. We validated our method with extensive experiments over a challenging dataset, for which we significantly outperformed the state of the art approach. Specifically, we achieved an accuracy of 95.2%, precision of 95.2% and recall of 95.0%. This is a significant improvement of 12.1%, 8.2% and 15.0% in terms of accuracy, precision and recall when sub-level activities are not used, as well as an improvement of 1.7%, 0.2% and 1.7% when they are used by the state of the art approach; this assumption of knowing the temporal segmentation of the sub-level activities is not required by our method.

Although in this work our focus was in the recognition of high-level activities, recognition of the sub-level activities is also important. We plan to extend our work to recognise these sub-level activities using a top-down approach, where the recognition of the high-level activity helps to infer the sub-level ones. This is in contrast to a bottom-up approach used by the current state of the art where high-level activities are inferred from sub-level ones.

Acknowledgement. The financial support of RACE (FP7-ICT-287752) and STRANDS (FP7-ICT-600623) projects is gratefully acknowledged.

References

1. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine Recognition of Human Activities: A Survey. IEEE Transactions on Circuits and Systems for Video Technology **18** (2008) 1473–1488
2. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28** (2010) 976–990
3. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding **115** (2011) 224–241
4. Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., Qiu, Y.: Exploring techniques for vision based human activity recognition: methods, systems, and evaluation. Sensors (Basel, Switzerland) **13** (2013) 1635–50
5. Collins, R., Lipton, A., Kanade, T.: Introduction to the special section on video surveillance. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 745–746
6. Gowsikhaa, D., Abirami, S., Baskaran, R.: Automated human behavior analysis from surveillance videos: a survey. Artificial Intelligence Review (2012) 1–19
7. Ko, T.: A survey on behavior analysis in video surveillance for homeland security applications. In: 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, IEEE (2008) 1–8
8. Chen, J., Cohn, A.G., Liu, D., Wang, S., Ouyang, J., Yu, Q.: A survey of qualitative spatial representations. The Knowledge Engineering Review **FirstView** (2013) 1–31
9. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64** (2005) 107–123
10. Koppula, H., Saxena, A.: Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation. In: Proc. of the International Conference on Machine Learning (ICML). (2013)
11. Koppula, H., Gupta, R., Saxena, A.: Learning Human Activities and Object Affordances from RGB-D Videos. International Journal of Robotics Research **32** (2013)
12. Sridhar, M., Cohn, A. G., Hogg, D. C.: Learning Functional Object-Categories from a Relational Spatio-Temporal Representation. In: European Conference on Artificial Intelligence. (2008)
13. Sridhar, M., Cohn, A. G., Hogg, D. C.: Unsupervised Learning of Event Classes from Video. In: AAAI. (2010)
14. Sridhar, M., Cohn, A. G., Hogg, D. C.: Discovering an Event Taxonomy from Video using Qualitative Spatio-temporal Graphs. In: European Conference on Artificial Intelligence. (2010)
15. Randell, D., Zhan, C., Cohn, A. G.: A Spatial Logic based on Regions and Connection. In: Third Int. Conf. on Knowledge Representation and Reasoning. (1992)
16. Cohn, A. G., Hazarika, S.: Qualitative spatial representation and reasoning: An overview. Fundamenta Informaticae **46** (2001) 1–29
17. Cohn, A. G., Renz, J.: Qualitative Spatial Representation and Reasoning. In van Harmelen, F., Lifschitz, V., Porter, B., eds.: Handbook of Knowledge Representation. Elsevier B.V. (2008) 551–596
18. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM (1983) 832–843
19. Dubba, K., Cohn, A. G., Hogg, D. C.: Event Model Learning from Complex Videos using ILP. In: European Conference on Artificial Intelligence. (2010) 93–98

20. Galata, A., Johnson, N., Hogg, D.: Learning Behaviour Models of Human Activities. In: British Machine Vision Conference. (1999)
21. Galata, A., Johnson, N., Hogg, D.: Learning Variable-Length Markov Models of Behavior. Computer Vision and Image Understanding **81** (2001) 398–413
22. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, IEEE (2005) 65–72
23. Xia, L., Aggarwal, J.: Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2013) 2834–2841
24. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2011) 2044–2049
25. Forbus, K.: Qualitative Modeling. In van Harmelen, F., Lifschitz, V., Porter, B., eds.: Handbook of Knowledge Representation. Elsevier B.V. (2008) 361–393
26. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-Temporal Phrases for Activity Recognition. In: European Conf. on Computer Vision (ECCV). (2012)
27. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional Learning of Spatio-temporal Features. In: European Conference on Computer Vision (ECCV). (2010) 140–153
28. Chen, B., Ting, J.A., Marlin, B., de Freitas, N.: Deep Learning of Invariant Spatio-Temporal Features from Video. In: NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop. (2010)
29. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011, IEEE (2011) 3361–3368
30. Ryoo, M.S., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: Computer Vision, 2009 IEEE 12th International Conference on. (2009) 1593–1600
31. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: Proceedings of the 2011 International Conference on Computer Vision. ICCV '11, IEEE Computer Society (2011)
32. Rybok, L., Schauerte, B., Al-Halah, Z., Stiefelhagen, R.: "Important Stuff, Everywhere!" Activity Recognition with Salient Proto-Objects as Context. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014)
33. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1226–1238
34. Pujari, A., Vijaya Kumari, G., Sattar, A.: INDu: An interval & duration network. In: Advanced Topics in Artificial Intelligence. Volume 1747 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (1999) 291–303
35. Behera, A., Hogg, D. C., Cohn, A. G.: Egocentric activity monitoring and recovery. In: The 11th Asian Conference on Computer Vision. (2012) 519–532
36. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27