Egocentric Activity Monitoring and Recovery

Ardhendu Behera, David C Hogg and Anthony G Cohn

School of Computing, University of Leeds, Leeds, LS2 9JT, UK
{A.Behera, D.C.Hogg, A.G.Cohn}@leeds.ac.uk

Abstract. This paper presents a novel approach for real-time egocentric activity recognition in which component atomic events are characterised in terms of binary relationships between parts of the body and manipulated objects. The key contribution is to summarise, within a histogram, the relationships that hold over a fixed time interval. This histogram is then classified into one of a number of atomic events. The relationships encode both the types of body parts and objects involved (e.g. wrist, hammer) together with a quantised representation of their distance apart and the normalised rate of change in this distance. The quantisation and classifier are both configured in a prior learning phase from training data. An activity is represented by a Markov model over atomic events. We show the application of the method in the prediction of the next atomic event within a manual procedure (e.g. assembling a simple device) and the detection of deviations from an expected procedure. This could be used for example in training operators in the use or servicing of a piece of equipment, or the assembly of a device from components. We evaluate our approach ('Bag-of-Relations') on two datasets: 'labelling and packaging bottles' and 'hammering nails and driving screws', and show superior performance to existing Bag-of-Features methods that work with histograms derived from image features [1]. Finally, we show that the combination of data from vision and inertial (IMU) sensors outperforms either modality alone.

1 Introduction

Automatically recognising human activities from videos is one of the fundamental research problems in computer vision and has generated a rich literature [2–4]. During the last decade or so, it has received increasing attention due to its far-reaching applications such as intelligent surveillance systems, human-computer interactions, robotics, and smart monitoring systems. Most of the earlier work in this area has been focused on recognizing periodic actions such as 'clapping', 'jogging', 'walking' etc. on relatively simple datasets [5, 6]. Lately, attention has moved to more realistic, complex and challenging datasets [1, 7–9]. These datasets incorporate videos collected from YouTube, movies, or by an amateur using a hand-held camera. Even more recently, there has been growing interest in activity recognition from an egocentric approach using first-person wearable cameras [10–13].

Most real-world activity recognition systems only classify activities after fully observing the entire sequence, but this is unsuitable for recognition of atomic-level, incomplete and/or ongoing activity. Such systems usually expect that the same number of people or objects are observed over the entire activity whilst in realistic scenarios often people and objects enter/leave the scene while activity is going on. In order to approximate prediction from partial observation, traditional sequential models such as hidden Markov models (HMMs) are often used [3]. However, from our experience they are unlikely suitable for 1) a varying number of objects observed at different times and 2) the high dimensional discontinuous sparse features such as histograms typically used today.

In this paper, we address the above-mentioned challenges while designing a fast but competitively accurate computer vision system for monitoring industrial activities. An industrial activity is assumed here to be a temporally ordered set of procedural steps or atomic events for accomplishing a task in which people and tools are involved in each step of the process. In such environments, the aim of activity monitoring is to recognize atomic events using video from wearable cameras in order to assist users/operators by providing *on-the-fly* instructions from an automatic system. This enables continual interaction between users and the system while performing a task. The system provides instructions and/or messages via augmented reality in the form of video clips and/or text using a see-through Head Mounted Display (HMD)[14]. There are three main objectives of the proposed monitoring system: 1) to recognise the current event from a short observation period (typically two seconds); 2) to anticipate the most probable event that follows on from the current event; 3) to recognise operator deviations from the correct activity (which may lead to quality and/or health and safety problems).

In our approach, events are represented as object-object and object-wrist spatiotemporal relations. For example, in everyday activities such as 'making coffee or tea' and 'cooking pasta', the interactions between the person's hands and objects involved (cups, teapots, pans, etc.) are key to representing events. These interactions contain important cues for recognizing such manipulative activities. This is in contrast to traditional approaches where often configurations and movements of the human body are the main cues. In our system, the instantaneous positions of objects and wrists, in a world frame of reference, are provided through detection and tracking using visual SLAM (Simultaneous Localization and Mapping) [15]. Spatiotemporal relations r between objects and wrists are generated by considering their separation and its first derivative with respect to time. We generate a codebook for these spatiotemporal relations using an unsupervised K-means clustering algorithm on a training dataset. In order to handle noisy object detection and tracking of the objects in our domain, we characterise atomic events by the frequency of occurrence of codewords occurring within a short temporal window (typically two seconds). This *bag-of-relations* (BoR) approach contrasts with logical inference from the set of relations occurring within the window. A key aspect of our approach is that the spatiotemporal relations are learned instead of being predefined by hand as is common in other

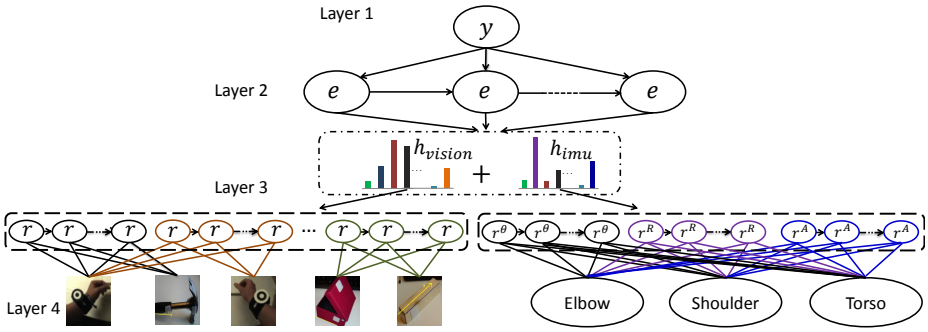

Fig. 1. Overview of our hierarchical framework: atomic events e are inferred using spatiotemporal pairwise relations r from observed objects and wrists, and relations r^θ, r^R and r^A between body parts (*elbow-shoulder* and *shoulder-torso*) using inertia sensors. Activities y are represented as a set of temporally-consistent e.

work [16–18] and which may not best fit the particular domain under consideration. While creating the BoR, we replicate the histogram for each pair of object categories, accumulating only those counts that apply to the relations between instances of the respective categories. Thus, the final histogram differentiates between events that involve the same relational codewords but different object categories, as in for example *pick up hammer* and *pick up screwdriver* where the action involving a hand is the same, but the category of object manipulated is different.

The overview of our hierarchical framework for the *hammering nails* activity is shown in Fig. 1. We model activities at layer 1 as a set of temporally-consistent atomic events in layer 2. Similarly, each atomic event is described using a BoR, which is composed of two histograms h_{vision} and h_{imu} (layer 3). h_{vision} is extracted from object (and wrist) interactions over a sliding window. The system observes objects and wrists, which are the leaf nodes of the model (layer 4). Similarly, h_{imu} captures the relative movements between body parts such as elbow w.r.t. shoulder and shoulder w.r.t. torso using a biomechanical upper-body model, over the same sliding window. These relative movements are captured by inertial sensors (IMU) as Euler angles $\boldsymbol{\theta}$, Euler rates $\boldsymbol{R}$ and Euler accelerations $\boldsymbol{A}$ [19, 20]. We represent these movements as two pairwise relations such as *elbow-shoulder* and *shoulder-torso*. In each pair, there are three possible relations of r^θ, r^R and r^A which correspond to the respective $\boldsymbol{\theta}$, $\boldsymbol{R}$ and $\boldsymbol{A}$. In this example, activity y refers to *hammering nails* which contains $e = \{$*pick up hammer*, *pick up nail*, *hammer nail*, *put down hammer*$\}$ atomic events.

The paper makes the following principal contributions: 1) a learnt representation for the spatial and kinematic relationship between pairs of objects, 2) a histogram-based representation that summarises the relational structure between sets of (categorised) objects within a temporal window, and provides the basis for atomic event classification, 3) an experimental evaluation that demonstrates the viability of the approach within an industrially motivated setting and that compares the contribution of different wearable sensors.

2 Related Work

Several different approaches for activity recognition can be identified in the vast literature on computer vision [2–4]. We categorize the existing work into three main classes based on the complexity involved. The first class of research considers recognition of body movements of a single person such as jogging, walking, clapping, etc. [21, 22, 5, 6]. The second considers object context for realistic actions such as skipping, drinking, smoking, etc. [9, 1]. The third kind involves interaction of multiple objects and/or persons [23, 24].

In this work, we focus on the activity recognition involving manipulations and/or interaction of multiple objects which forms the focus of this work. For example, Ryoo [23] presented a probabilistic approach for human-human activity prediction from streaming video using both dynamic and integral bag-of-words representation of spatiotemporal features. Gupta et al. [9] introduced a Bayesian model for recognizing human-object interactions with a likelihood model, which is based on hand trajectories. Shi et al. [25] presented propagation networks (P-nets) to describe glucose model calibration by applying temporal and logical constraints. Veres et al. [26] proposed a method for monitoring workflow activities in a car assembly line using a global motion descriptor. Sridhar et al. [18] described an unsupervised approach for finding event classes from videos by considering spatial and temporal relations between tracklets. Workflow activities were monitored by Behera et al. [27] using HMM and pLSA (probabilistic Latent Semantic Analysis). Simon et al. [28] presented a method that uses Petri-Nets in order to recognize workflow activities.

In the proposed model, our goal is to recognize activities from the egocentric viewpoint using a wearable camera and is quite different from the above-mentioned approaches. Starner and Pentland were one of the first to use an egocentric setup to recognize American sign language in real-time [29]. More recently, Fathi et al. [12] presented a hierarchical model of daily activities by exploring the consistent appearance of objects, hands and actions from the egocentric viewpoint. Aghazadeh et al. [13] extracted novel events from daily activities and Kitani et al. [11] identified ego-action categories from first-person viewpoint. Ward et al. [30] proposed a method to recognize wood workshop assembly activities by using on-body sensors of microphones and accelerometers. Reiss et al. [20] described a method for activity recognitions based on a biomechanical upper-body model using on-body sensors of IMUs.

Our proposed approach initiates a framework in which activities and atomic events are recognized in real-time using streaming data from wearable sensors and the system should be able to provide required feedback to the user. Moreover, the goal is to describe events and activities with semantically meaningful spatiotemporal relations based on object-object and object-hand interactions.

3 Activity Monitoring Model

The main goal of the proposed activity monitoring model is to analyze the live streaming of an image sequence and assign atomic event and activity labels to

it. The real-time detection and tracking algorithm processes each image and then provides 3D positions of each detected object with respect to workspace coordinates and its class type for further analysis [15]. Our model uses a sliding window approach over the image sequence. Therefore, a complete activity sequence contains a set of windows $\mathbf{w} = \{w_1, w_2, \ldots, w_W\}$, where each window w_i comprises F_i images $w_i = \{I_1, I_2, \ldots, I_{F_i}\}$. As a result, each image I_j contains N_j objects $I_j = \{o_1, o_2, \ldots, o_{N_j}\}$ and each object o_k will have class type $c \in \mathcal{C}$ and a 3D XYZ position, where $\mathcal{C} = \{$*hammer*, *screwdriver*, *wrist*, *bottle*, etc.$\}$ is the set of object categories, which are known in advance.

Inference involves assigning an atomic event label $e \in \mathcal{E}$ to each window w_i and an activity label $y \in \mathcal{Y}$, where $\mathcal{E} = \{$*pick up hammer*, *write address*, etc.$\}$ is a set of atomic events and $\mathcal{Y} = \{$*hammering nails*, *labelling and packaging bottles*, etc$\}$ is the set of possible activities.

3.1 Pairwise Spatiotemporal Relational Features

The proposed spatiotemporal relational feature jointly contains two types of information in a view-invariant fashion: 1) spatial configuration of objects in 3D space and 2) the kinematics between them over time. Although, the proposed relational feature is not scale-invariant, it has little or no effect in an egocentric setup since a user only manipulates objects within his/her reach. Suppose the system observed an image I_t at time t, then the detector provides category $c_m, c_n \in \mathcal{C}$ and locations of the detected respective objects $o_m, o_n \in I_t$. The spatiotemporal relation between the objects o_m and o_n is represented by

$$\mathbf{r} = (d_{m,n}, \frac{\dot{d}_{m,n}}{d_{m,n} + \epsilon}) \in \Re^2 \tag{1}$$

where $d_{m,n}$ is the Euclidean distance between object o_m and o_n, and ϵ is a small positive value to avoid division-by-zero errors.

We describe the relational feature $\mathbf{r}$ with K possible relational words $\alpha_1 \ldots \alpha_K$. In order to achieve this, the spatiotemporal features are grouped into K clusters that constitute the relational vocabulary. We use a standard unsupervised K-means clustering algorithm. Let an input $\mathbb{F} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ be the set of relational features. If we denote the center of the jth cluster as $mean_j$, then each feature $\mathbf{r} \in \mathbb{F}$ is now mapped into the nearest relational word via:

$$\alpha_i(\mathbb{F}) = \{\mathbf{r} | \mathbf{r} \in \mathbb{F} \wedge i = argmin_j ||\mathbf{r} - mean_j||^2\} \tag{2}$$

where $||\mathbf{r} - mean_j||$ denotes the Euclidean distance between feature $\mathbf{r}$ and $mean_j$. As a result, we have decomposed the set $\mathbb{F}$ into K subsets, $\alpha_1(\mathbb{F}), \ldots, \alpha_K(\mathbb{F})$, based on their spatiotemporal relations.

3.2 Category-specific Bag-of-relations (BoR)

We now discuss the creation of a BoR from the relational vocabularies.The proposed spatiotemporal relational feature is based on the available object category.

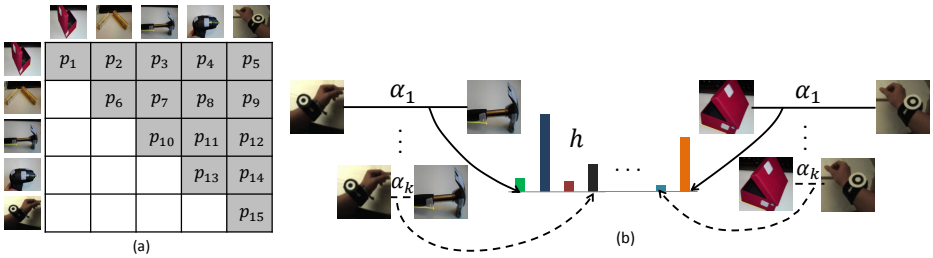

Fig. 2. a) Category-specific object pairs for the 'hammering nails and driving screws' scenario and b) category-specific bin assignment to BoR histogram.

For example, there are $P = 15$ possible category-specific pairs from $\mathcal{C} = \{box, baton, hammer, screwdriver, wrist\}$ categories (Fig. 2a). Both left-wrist and right-wrist are separately detected and tracked. Our model considers 'left-wrist' and 'right-wrist' as a single category of 'wrist'. This is due to often the same manipulative tasks are carried out by a left-handed person and the system might have not seen a left-handed person during training; further discussion on this point is provided in the experiments and discussion section. We also includes intra-category pairs (p_1, p_6, p_{10}, p_{13} and p_{15}) as shown in Fig. 2a. The main reasons are 1) We believe that multiple objects belonging to the same category are often observed in an activity; for example, two batons are observed at the same time in this activity. 2) Misclassification, i.e. if a hammer is classified as box, and we do not consider box-box pair then the relations between actual hammer-box pair will not be considered at all; our model also adds another category called *unknown* to the existing list of categories. This category is for those objects whose types cannot be decided by the object detector due to a low confidence measure. This could be due to occlusion or very little information being observed by the detector. The spatiotemporal relations of this *unknown* category with other known categories might provide vital cues to discriminate between events/activities and therefore our system also considers this for the computation of pairwise relations. Moreover, our system could provide feedback to the detector about the possible types for *unknown* based on previously observed activities.

Next, we represent the sliding window $w_i \in \mathbf{w}$ as a histogram h of relational words with encoded object categories. In the conventional bag-of-words method, the number of bins in the histogram is equal to the number of relational words in the codebook, i.e. K. Each relational word α_k in the codebook is mapped to a unique bin $b_i \in h$ of the histogram. Each b_i represents the frequency of the corresponding relational word α_k appears in the sliding window w_i. In our representation, we encode the object category along with relational word i.e.

$$b_i = |\mathcal{A}_i|, \text{where } \mathcal{A}_i = \{(\alpha_k, p)\}, k \in \{1 \dots K\} \text{ and } p \in \{1 \dots P\} \tag{3}$$

$\mathcal{A}_i$ is the set of co-occurrences of relational word α_k and the category pair p of the related objects. Therefore, there are $P \times K$ bins in the histogram h for our category-specific representation instead of K. To illustrate, consider the example

of the *pick up hammer* and *take nail* atomic events. Both of these events may have the same temporal transitions $\alpha_1 \rightarrow \cdots \rightarrow \alpha_k$ i.e. in the case of *pick up hammer* the wrist moves towards the hammer (Fig. 2b). Similarly, for *take nail* the wrist moves towards to the box. Both have similar relative spatiotemporal movement patterns and the only difference is that different objects are involved. Therefore, encoding object category should provide discriminative cues for activity recognition.

The approach generates a histogram h_i for each sliding window $w_i \in \mathbf{w}$ after processing all images $F_i \in w_i$ and considering all possible category specific pairwise relations $r \in \mathbb{F}$. This histogram is the input feature vector for our learning and inference procedure, which is presented in the next section.

3.3 Learning and Inference

During supervised learning, the task is to infer a function mapping from sequences $\bar{\mathbf{x}}^{(n)}$ to atomic events $\bar{e}^{(n)}$ and activities $\bar{y}^{(n)}$, given training examples of ($\mathbf{x}^{(n)}$, $e^{(n)}$,$y^{(n)}$). Joint learning of these variables requires an unmanageably large training set and very expensive approximate inference as pointed out by Fathi et al. [12]. Therefore, an alternative approach is to exploit the independence structure of the proposed hierarchical framework (Fig. 1). First, the model learns to predict the atomic events from object-object and object-wrist interactions (layers 2-4 of the framework) and then envisage activities (layer 1) from a set of atomic events by exploiting their temporal structure. Consider the *driving screws* example to have a better understanding of the temporal structure. The atomic event 'drive screw' is possible only after events 'pick up screwdriver' and 'pick a screw'.

We first solve the sub-problem of learning atomic events from the object-object and wrist-object interaction is modelled by removing the activity layer (layer 1) from the hierarchical framework. The goal is to learn a discriminative function $e = f_1(h)$ by training a classifier on histogram h for each atomic event e without considering the adjacent horizontal links (temporal) between atomic events. We use a probabilistic multi-class Support Vector Machine (SVM) [31] with a χ^2 kernel using 'one-vs-one' methodology. Recently, Vedaldi and Zisserman [32] reported that the χ^2 kernel performs better than other additive kernels such as intersection and Hellinger's for histogram-based classifications.

The next task is to estimate the activity label y from the complete probability distribution $P(e_1 \dots e_T | h_1 \dots h_T)$ of sequence of atomic events for a given sequence of histograms. One could learn another discriminative function $y = f_2(e_1 \dots e_T)$ from a sequence of atomic events e and an activity label y. However, our model is targeted for live recognition of activities and atomic events. Therefore, we learn the transition probability $P(e_t|e_{t-1})$ between atomic events, starting probability of an atomic event $P(e_1)$ and the distribution of atomic events given activities $P(e_t|e_1 \dots e_{t-1}, y)$ from the training examples. Thus, it establishes the horizontal links i.e. transition probability between atomic events and these links along with the $P(e_t|e_1 \dots e_{t-1}, y)$ distribution decide the most likely activity label y (Fig. 1).

Fig. 3. Snapshots from the 'hammering nails and driving screws' (1^{st} row) and the 'labelling and packaging bottles' (2^{nd} row). The first two images of each row are from the top view (RGB-D) and the rest are from the chest-view fisheye camera.

During prediction at time t, the most probable on-going atomic event $\bar{e}_t$ and activity $\bar{y}_t$ label corresponding to the observed histogram $\bar{h}_t$, and the most likely next atomic events $\bar{e}_{t+1}$ are computed as:

$$\begin{aligned} P(\bar{e}_t|\bar{h}_t) &\propto P(\bar{e}_t)P(\bar{e}_t|f_1(\bar{h}_t))P(\bar{e}_t|\bar{e}_{t-1}) \\ \bar{e}_t &= \arg\max\{P(\bar{e}_t|\bar{h}_t)\} \\ \bar{y}_t &= \arg\max\{P(\bar{y}_t|\bar{e}_t)P(\bar{e}_t|\bar{h}_t)\} \\ \bar{e}_{t+1} &= \arg\max\{P(\bar{e}_{t+1}|\bar{e}_1 \ldots \bar{e}_t, y)\} \end{aligned} \tag{4}$$

4 Experiments and Discussion

In this section, we present results to validate the performance of our proposed hierarchical framework.

4.1 Datasets

In order to test our hierarchical framework, we have obtained two datasets using an egocentric setup. These datasets consist of non-periodic manipulative tasks in an industrial context. All the sequences were captured with on-body sensors consisting IMUs, a backpack-mounted RGB-D camera for top-view and a chest-mounted fisheye camera for front-view of the workbench (Fig. 3).

The first dataset is the scenario of 'hammering nails and driving screws' (Fig. 3). In this dataset, subjects are asked to hammer 3 nails and drive 3 screws using prescribed tools (not necessarily in this order and there is inter-subject variation in this dataset). There are 9 different types of atomic events: 1) take and put box, 2) take and put baton, 3) pick hammer, 4) take nail/screw, 5) hammering nail, 6) put down hammer, 7) pick screwdriver, 8) driving screw and 9) put down screwdriver. A total of 27 sequences were captured, consisting of 3

people performing the activity 5 times each and 2 people performing the activity six times each. In this dataset, there are five types of object categories which consists of hammer, screwdriver, baton, wrist and box.

The second dataset is a 'labelling and packaging bottles' scenario. In this dataset, participants asked to attach labels to two bottles, then package them in the correct positions within a box. This requires opening the box, placing the bottles, closing the box, and then writing on the box as completed using a marker pen. There are 9 different atomic events: 1) pick and put bottle, 2) stick label, 3) pick and put box, 4) remove cover, 5) put bottle inside box, 6) take and put cover, 7) write address, 8) take and put sticky tape dispenser, and 9) seal the box with sticky tape. In this dataset, there are 23 sequences, consisting of 3 performing the activity 5 times each, and 2 people performing it 4 times each. Five types of object categories namely, bottle, box, sticky tape dispenser, wrist and marker pen are used in this dataset.

The duration of each activity is about 2500 frames recorded at 25 frames per second. For evaluation purposes, these datasets were manually annotated by providing the start and end frames of each atomic event. To our knowledge, there are no other available datasets that are captured in an egocentric setup with on-body sensors of camera and IMUs in an industrial context. Our dataset is available at http://www.engineering.leeds.ac.uk/computing/research/vision/cognito/. At the moment, GTEA (GeorgiaTech Egocentric Activities) is the only available egocentric video dataset. This contains 7 kinds of daily activities [10].

4.2 Evaluation Methods

For all of our evaluations, we used a sliding window $w_i \in \mathbf{w}$ of duration 2 seconds with 50% overlap and 'one-vs-all-subject' evaluation strategies. In 'one-vs-all-subject', all sequences from one subject are used for validation and the rest are used to train the model. We used this strategy on the assumption that experts are involved in demonstrating activities from which the system learns, and subsequently the system would be required to guide a naïve worker to perform those activities in an industrial scenario (thus, those subjects used for training and those used for testing are distinct from one another).

In order to compare our approach with the state of the art, we have implemented a baseline 'bag-of-features' (BoF) method using STIP (Space-Time Interest Points) descriptors as described in [1]. In this, a visual vocabulary is generated by randomly sampling a subset of 100,000 STIP descriptors from the training set and using K-means clustering to generate 4000 visual words. Descriptors are assigned to their closest visual word using the Euclidean distance and the histogram h_{stip} of visual word occurrences is computed over each sliding window $w_i \in \mathbf{w}$. Results for both approaches are achieved using a χ^2 kernel and multi-class classifications using the 'one-vs-one' approach. We fix the histogram normalization to the $L1$-norm and optimize the parameters of SVM classifier using 10-fold cross-validations on training set.

We evaluated three variations of our proposed *bag-of-relations* (BoR) method. The first is as described above. In the second, we ignore object categories and

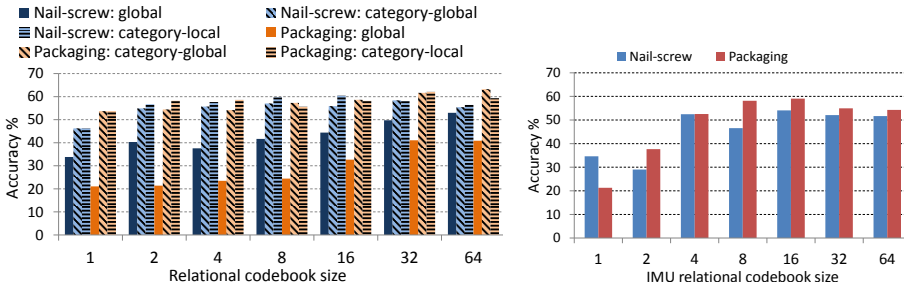

Fig. 4. Average performance with varying relational codebook size for 'one-vs-all-subjects' experiments using only h_{vision} (left) and h_{imu} (right).

simply accumulate codewords, ignoring the categories of the objects involved. The second kind is a category-specific BoR with a global codebook (category-global). The third variation is like the first, except that the codebook is specific to each category pairs (category-local). Therefore, there are $|P|$ relational codebooks and each codebook is attached to a particular category-pair.

In a similar fashion, we generate the histogram h_{imu} for each window $w_i \in \mathbf{w}$ by considering the relational codebook created from IMU data. There are six different codebooks (3 each for *elbow-shoulder* and *shoulder-torso* relations) for IMU data. The 3 codebooks for each relation representing a pair of body parts e.g. *shoulder-torso* correspond to the respective Euler angles $\boldsymbol{\theta}$, Euler rates $\boldsymbol{R}$ and Euler accelerations $\boldsymbol{A}$. In total, we have 6 category-specific-codebook representing *elbow-shoulder* and *shoulder-torso* relations for both left and right hand. We have also evaluated all possible combinations (e.g. $[h_{vision}h_{imu}]$, $[h_{vision}$ $h_{stip}]$, $[h_{imu}h_{stip}]$, and $[h_{vision}h_{imu}h_{stip}]$) of the different modalities/approaches by concatenating the generated histograms.

4.3 Results and Discussion

We first compared the performance of our proposed method (BoR) with a varying size of relational codebook for both the above-mentioned datasets. In this experiment, we selected $K = \{1, 2, 4, 8, 16, 32, 64\}$. The mean performance over all subjects (one-vs-all-subject) is shown in Fig. 4. All of our results are presented as the classification accuracy over all windows. For $K = 1$, it just implies the presence or absence of one or more objects for the global codebook and object-pairs for both the category-specific codebooks (local and global). The performance using category-specific BoR is better than the generic one for all values of K (except in packaging dataset for $K = 64$). This justifies our representation of activities by considering object-object and object-wrists interactions. From Fig. 4, it is evident that most often the performance using a category-specific codebook (category-local) is better than the other two representations. This suggests the kinematic variations in the way we use different objects, are important. The optimal category-specific relational codebook size is 32 and 16 for 'labelling and packaging' and 'nails and screws' datasets respectively. This

Table 1. Performance comparison for the experiment one-vs-rest-subject (bold text represents the best performance out of the individual modalities).

	Hammering nails and driving screws							Labelling and packaging bottles						
	vision	IMU	STIP	vision IMU	vision STIP	IMU STIP	vision IMU STIP	vision	IMU	STIP	vision IMU	vision STIP	IMU STIP	vision IMU STIP
s_1	**65.7**	65.2	65.9	73.4	78.1	70.7	75.4	**61.3**	38.2	31.4	62.4	64.5	36.3	65.4
s_2	64.5	**67.5**	67.2	72.3	73.4	77.5	77.2	**53.5**	71.2	50.5	67.7	64.5	75.4	78.3
s_3	61.7	53.5	**73.1**	62.0	72.2	64.9	68.4	**63.0**	59.9	61.9	80.8	66.6	69.4	82.1
s_4	**38.0**	10.3	9.2	25.9	35.6	11.0	18.7	**76.1**	74.8	56.0	85.6	84.5	71.3	89.4
s_5	72.5	**74.0**	77.4	80.3	82.1	84.7	86.6	**56.5**	51.3	56.5	70.4	65.3	66.6	70.4
Avg	**60.5**	49.8	58.6	62.8	68.3	61.7	65.3	**62.1**	59.1	51.3	73.4	69.1	63.8	77.1

is due to the fact that there are different degrees of variation in the interactions between different categories of object. Whereas for inertial sensor data h_{imu}, the optimal codebook size for both the datasets is 16 which conveys the use of the same *elbow-shoulder* and *shoulder-torso* relations in both the datasets.

We present the complete evaluation of our framework using individual histograms i.e. h_{vision}, h_{imu} and h_{stip}, and their possible combinations in Table 1. The last row of Table 1 presents the mean performance over all subjects $s_1 \ldots s_5$. In this evaluation, we used the optimal size category-specific codebook as computed above. For both the datasets, vision (60.5%, 62.1%) performs better than the other two individual representations IMU (49.8%, 59.1%) and STIP (58.6%, 51.3%). However, the combined performance using IMU and vision (our proposed framework) is better for all individuals except subject s_4 in the 'nails and screws' dataset. In this dataset, subject s_4 alone is left-handed. Thus, the system has not seen a left-handed performer during training. However, the performance using vision (38.0%) is far better than using only STIP (9.2%), which demonstrates an invariance property of our BoR representation. Furthermore, the combined performance of vision and STIP (68.3%, 69.1%) is better than STIP (58.6%, 51.3%). Therefore, our pairwise relational approach can be used with existing *bag-of-features* approaches for further improvements.

We show the confusion matrix for both the datasets in Fig. 5. From both these matrices, it is evident that atomic events are often confused with the previous and next events. This is a typical synchronization error for sequential data. It is partly due to the manual assignment of labels while preparing ground-truth, since it is difficult for humans to assign boundaries consistently between consecutive events. Lets consider the atomic 'driving screw' (Fig. 5a). This is mostly confused with event 4 (take nail/screw) which comes before event 8. Similarly, event 4 is mostly confused with events 8 'driving screw' and 5 'hammering nail'. This is due to the fact that one should pick up the nails or screws, i.e. event 4, before hammering or driving it.

In the confusion matrix of the 'labelling and packaging' dataset (Fig. 5b), the atomic event 7 'write address' is mostly confused with event 9 'seal box with

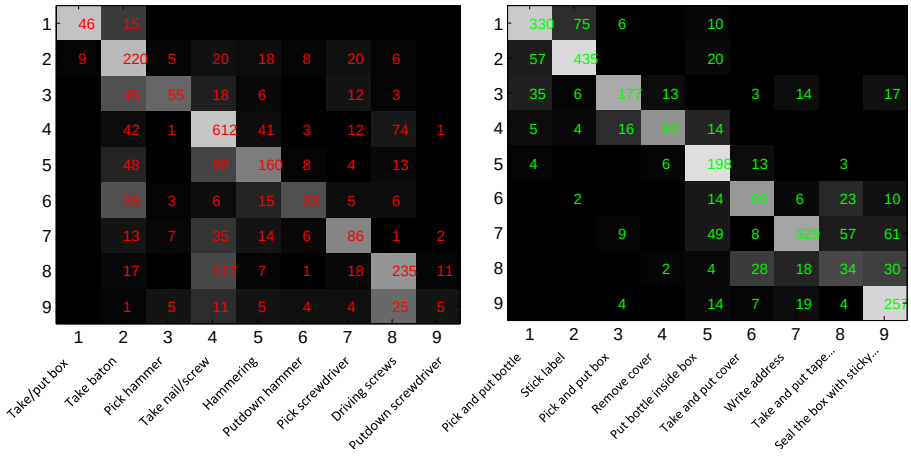

Fig. 5. Confusion matrix using $h_{vision+imu}$ for a) 'hammering nails and driving screws' (left) and b) 'labelling bottles and packaging' (right) dataset.

tape'. This is due to the fact that often the pen is not detected due to its size. Moreover, while writing on the box there is little kinematic variation between involved objects, which are wrist, pen and box. The system only observes the close proximity of these objects. In the case of event 9, the wrist takes a piece of tape (a deformable object, not detected by detector) from the sticky tape dispenser and attaches it to the box. While sticking the tape, the wrist is close to the box and stays there for sometime. Therefore, the system is unable to differentiate between writing and 'sticking tape' if the pen is not detected. The next most confused event is 'take and put sticky tape dispenser'. This is due to the same reason as while taking a piece of tape: the wrist comes closer to the dispenser stays for a while and then moves away. Similarly, while bringing the dispenser to the workbench, the wrist and the dispenser stay together for a while and then the wrist moves away. All three of these events exhibit similar spatiotemporal relations and therefore the model is unable to discriminate. The performance of event 2 'stick label' is very good given the fact that the model does not detect the label. However, while sticking the label both wrists come closer to the bottle and stay for a while. This information is discriminative enough to differentiate from the event 1 'pick and put bottle' in which one wrist is involved. This validates our category-specific representation of BoR.

In our model, the performance of the object detector has an effect on overall accuracy of activity recognition. The more accurate object detection and tracking, the better the accuracy. However, given the egocentric setup with a wearable camera, it is a difficult task. The object detector used [15] is evaluated on a subset of the above-mentioned datasets. It is a huge task to annotate the complete dataset of 50 sequences. Therefore 5 sequences from each dataset are selected and evaluated. The average performance for the 'nails and screw' dataset is 62.52% (recall) and 94.91% (precision). For the other dataset, the respective average

is 69.31% and 80.14%. Given these accuracies, the performance of our activity monitoring framework is good.

5 Conclusion

We present a novel approach for real-time monitoring of activities using on-body sensors in a egocentric environment. Our approach represents activity as a set of temporally-consistent atomic events within a hierarchical framework in which bottom-up propagation of evidence for atomic events is used to predict activity. This evidence is represented using a novel category-specific *bag-of-relations* (BoR) representation. The BoR is extracted using object-object and object-wrist interactions. We evaluated our approach on two new challenging datasets: 'hammering nails and driving screws' and 'labelling and packaging of bottles', which are captured in a egocentric setup with on-body sensors of camera and IMU in an industrial context. On these datasets, we demonstrate that the performance of our approach is superior to the existing 'bag-of-features' for activity monitoring.

In the near future, we hope to improve the performance by exploring the temporal structure in our *bag-of-relations* and feature selection method to select the most useful relations to represent activities.

Acknowledgement. This research work is supported by EU FP7 (ICT Cognitive Systems and Robotics) grant on COGNITO (www.ict-cognito.org, ICT-248290) project. We also thank our collaborators at the COGNITO partners.

References

1. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
2. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104** (2006) 90–126
3. Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Techn. **18** (2008) 1473–1488
4. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. **43** (2011) 1–16
5. Schldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR. (2004) 32–36
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV. (2005) 1395–1402
7. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. (2011) 2556–2563
8. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR. (2009) 1996–2003
9. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)

10. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR. (2011) 3281–3288
11. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR. (2011) 3241–3248
12. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV. (2011) 407–414
13. Aghazadeh, O., Sullivan, J., Carlsson, S.: Novelty detection from an ego-centric perspective. In: CVPR. (2011) 3297–3304
14. Wanstall, B.: HUD on the Head for Combat Pilots. In: Interavia 44. (1989) 334–338
15. Damen, D., Bunnun, P., Calway, A., Mayol-Cuevas, W.: Real-time learning and detection of 3d texture-less objects: A scalable approach. In: BMVC. (2012)
16. Pinhanez, C., Bobick, A.: Human action detection using pnf propagation of temporal constraints. In: Proc. of IEEE CVPR. (1998)
17. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. (2009) 1593–1600
18. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised learning of event classes from video. In: AAAI. (2010)
19. Bleser, G., Hendeby, G., Miezal, M.: Using egocentric vision to achieve robust inertial body tracking under magnetic disturbances. In: ISMAR. (2011) 103–109
20. Reiss, A., Hendeby, G., Bleser, G., Stricker, D.: Activity recognition using biomechanical model based pose estimation. In: EuroSSC. (2010) 42–55
21. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001) 257–267
22. Efros, A.A., Berg, A.C., Berg, E.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV. (2003) 726–733
23. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV. (2011) 1036–1043
24. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS. (2010) 1216–1224
25. Shi, Y., Huang, Y., Minnen, D., Bobick, A., Essa, I.: Propagation networks for recognition of partially ordered sequential action. In: CVPR. (2004) 862–869
26. Veres, G., Grabner, H., Middleton, L., Gool, L.V.: Automatic workflow monitoring in industrial environments. In: ACCV. (2010)
27. Behera, A., Cohn, A.G., Hogg, D.C.: Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In: MMM. (2012) 196–209
28. Worgan, S.F., Behera, A., Cohn, A.G., Hogg, D.C.: Exploiting petri-net structure for activity classification and user instruction within an industrial setting. In: ICMI. (2011) 113–120
29. Starner, T., Pentland, A.: Real-time American sign language recognition from video using hidden Markov models. In: Proc. of Int'l Symposium on Computer Vision. (1995) 265 – 270
30. Ward, J.A., Lukowicz, P., Troster, G., Starner, T.E.: Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans. PAMI **28** (2006) 1553–1567
31. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001)
32. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: CVPR. (2010) 3539–3546