Egocentric Activity Recognition using Histograms of Oriented Pairwise Relations

Ardhendu Behera, Matthew Chapman, Anthony G. Cohn and David C. Hogg
School of Computing, University of Leeds, LS2 9JT, Leeds,U.K.

Keywords: Egocentric Activity Recognition, Histogram of Oriented Pairwise Relations (HOPR), Spatio-temporal Relationships, Pairwise Qualitative Relations, Bag-of-visual-words.

Abstract: This paper presents an approach for recognising activities using video from an egocentric (first-person view) setup. Our approach infers activity from the interactions of objects and hands. In contrast to previous approaches to activity recognition, we do not require to use an intermediate such as object detection, pose estimation, etc. Recently, it has been shown that modelling the spatial distribution of visual words corresponding to local features further improves the performance of activity recognition using the bag-of-visual words representation. Influenced and inspired by this philosophy, our method is based on global spatio-temporal relationships between visual words. We consider the interaction between visual words by encoding their spatial distances, orientations and alignments. These interactions are encoded using a histogram that we name the Histogram of Oriented Pairwise Relations (HOPR). The proposed approach is robust to occlusion and background variation and is evaluated on two challenging egocentric activity datasets consisting of manipulative task. We introduce a novel representation of activities based on interactions of local features and experimentally demonstrate its superior performance in comparison to standard activity representations such as bag-of-visual words.

1 INTRODUCTION

In this work, we address the problem of recognising activities using video from a wearable camera (egocentric view). Several approaches have been proposed in the past to address the problem of generic activity recognition (Moeslund et al., 2006; Turaga et al., 2008; Aggarwal and Ryoo, 2011). These approaches use various types of visual cues and compare them using some similarity measure. In the course of the last decade or so, activity recognition has received increasing attention due to its far-reaching applications such as intelligent surveillance systems, human-computer interaction, and smart monitoring systems. Researchers are now advancing from recognising simple periodic actions such as 'clapping', 'jogging', 'walking' (Schuldt et al., 2004; Blank et al., 2005) to more complex and challenging activities involving multiple persons and objects (Laptev et al., 2008; Kuehne et al., 2011; Liu et al., 2009a; Gupta and Davis, 2007). Even more recently, there has been growing interest in activity recognition from an egocentric approach using first-person wearable cameras (Fathi et al., 2011b; Kitani et al., 2011; Fathi et al., 2011a; Aghazadeh et al., 2011). Most real-

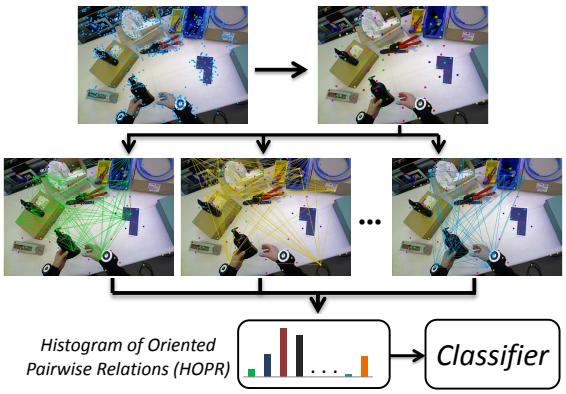


Figure 1: Hierarchical framework for activity recognition: 1) detected keypoints representing visual features (*e.g.* SURF (Bay et al., 2006)) in an image (top-left), 2) filtered keypoints based on their strength with assigned codewords using K-means clustering (top-right), 3) extraction of pairwise relations between keypoints belonging to the same codewords (middle row), 4) histogram of oriented pairwise relations (HOPR) representation of these extracted relations, which is used for framewise classification of activity using a classifier. The wrist marker in images are used for the detection and tracking of wrist in the existing method (Behera et al., 2012b).

Behera A., Chapman M., G. Cohn A. and C. Hogg D..
Egocentric Activity Recognition using Histograms of Oriented Pairwise Relations.
DOI: 10.5220/0004655100220030
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 22-30
ISBN: 978-989-758-004-8

world activity recognition systems utilize a bag-of-visual words paradigm, which use spatio-temporal features (Schuldt et al., 2004; Dollár et al., 2005; Blank et al., 2005; Ryoo and Aggarwal, 2009). These features are shown to be robust to the changes in lighting and invariant to affine transformations. These approaches are designed to classify activities after fully observing the entire sequence assuming each video contains a complete execution of a single activity. However, such features alone are often not enough for modelling complex activities as the same action pattern can produce a variety of different movement patterns. For example, while cooking pasta, one can pour water using one hand while the other hand is used for stirring and perform actions sequentially using one hand. In order to improve the recognition performance of complex activities, recently there is a growing interest in modelling the spatial distributions of the above-mentioned spatio-temporal features (Matikainen et al., 2010; Ryoo and Aggarwal, 2009; Sun et al., 2009; Gilbert et al., 2009). Such ideas are inherited from object recognition approaches (Savarese et al., 2006; Shechtman and Irani, 2007; Liu et al., 2008; Deselaers and Ferrari, 2010).

In this paper, we address the problem of recognizing activities in an egocentric setting. Our approach considers the interactions between feature descriptors as a discriminating cue for recognising activities. The framework of the proposed approach is presented in Fig. 1. The proposed approach is in contrast to traditional approaches where interaction between objects and wrists are often used for recognising activities (Fathi et al., 2011a; Gupta and Davis, 2007; Behera et al., 2012b; Behera et al., 2012a). Such approaches use pre-trained object detectors. Moreover, our approach can recognise activities using a single frame and can make a decision before observing the entire video. This is very helpful for real-time monitoring systems. There also have been previous approaches which are successful for recognising activities using single frames (Niebles and Li, 2007; Fathi et al., 2011a). However, they are limited to either simple activities or require pre-trained object detectors.

In this work, we introduce a new descriptor called Histogram of Oriented Pairwise Relations (HOPR) for recognising activities in egocentric settings. The proposed descriptor captures the interactions between the extracted features/patches such as SIFT (Lowe, 2004), STIP (Laptev and Lindeberg, 2003), SURF (Bay et al., 2006) and summarises the pairwise relationships structure between them within an image. This provides the basis for activity classification and does not require any object detector. We demonstrate the advantages of our representation by evaluating it on challenging egocentric datasets, which are publicly available namely GTEA (GeorgiaTech Egocentric Activities) consisting of kitchen activities (Fathi et al., 2011b) and Leeds' egocentric dataset ('labelling and packaging bottles') for manipulative tasks (Behera et al., 2012b). In order to recognise activities, the proposed method captures the wrist-object interactions using pairwise relationships between visual words. Therefore, we evaluate our method on egocentric datasets because poses and displacements of manipulated objects are consistent in workspace coordinates with respect to an egocentric view.

2 PREVIOUS WORK

Several different approaches for activity recognition can be found in the computer vision literature (Moeslund et al., 2006; Turaga et al., 2008; Aggarwal and Ryoo, 2011). In this work, we mainly concentrate on activity recognition involving spatial distribution of visual words in an egocentric setup, which is the focus of our work. To our knowledge, there is no existing previous work which uses the distribution of visual words for recognising egocentric activities. However, they do appear in a different context in the literature. Therefore, we discuss both the approaches.

Pairwise relationships in the form of correlograms, constellations, star topologies and parts model have been used frequently in static image analysis (Savarese et al., 2006; Crandall and Huttenlocher, 2006; Carneiro and Lowe, 2006). Practical limitations have prevented transitioning of these methods into video (Matikainen et al., 2010). Therefore, different approaches have been adopted for recognising activities in videos using pairwise relationships. Matikainen *et al.* (2010) proposed a method for activity recognition by encoding pairwise relationships between fragments of trajectories using sequencing code map (SCM) quantisation. Ryoo and Aggarwal (2010) presented a spatio-temporal relationships match for recognising activities that uses relationships between spatio-temporal cuboids. Sun *et al.* (2009) proposed a method for recognising actions by exploring the spatio-temporal context information encoded in unconstrained videos based on the SIFT-based trajectory, in a hierarchy of three abstraction levels.

In this work, the main objective is to recognise activities from the egocentric viewpoint and is quite different from the above-mentioned approaches. Real-time recognition of American sign language is the first to use an egocentric setup and is proposed by (Starner and Pentland, 1995). Lately, Behera *et al.* (2012b) described a method for real-time monitor-

ing of activities using bag-of-relations in an industrial setup. Fathi *et al.* (2011a) presented a hierarchical model of daily activities by exploring the consistent appearance of objects, hands and actions from the egocentric viewpoint. Aghazadeh *et al.* (2011) extracted novel events from daily activities and Kitani *et al.* (2011) identified ego-action categories from a first-person viewpoint.

Most of the above-mentioned approaches are designed to perform after-the-fact classification of activities after fully observing the activities. Furthermore, they often require object detectors for detecting wrists and objects as object-wrist interactions have been used as cue for discriminating activities. Our proposed approach initiates a framework in which complex activities can be recognised using a single frame in real-time without using any object detector. The proposed novel Histogram of Oriented Pairwise Relations (HOPR) captures the interaction between visual descriptors (SURF) and represents them as a relational structure that encodes the pairwise relationships.

3 PROPOSED MODEL

A video sequence $v_i = \{I_1 \dots I_T\}$ consists of T images. Every image $I_{t=1\dots T}$ is processed to extract a set of keypoints $\mathcal{S}_t = \{f\}$. Each keypoint $f = (f^{desc}, f^{loc}, f^{st})$ is represented by a feature descriptor f^{desc}, its xy position f^{loc} in the image plane and its strength f^{st} representing the quality of the keypoints. Here, keypoints refer to the detection and description of local features such as SIFT (Lowe, 2004), SURF (Bay et al., 2006) and STIP (Laptev and Lindeberg, 2003). However, STIP requires more than a frame in order to extract the feature descriptors.

First, we select a subset of keypoints by considering their strength f^{st} (Fig. 1 top). All the keypoints in the set $\mathcal{S}_t$ are sorted with decreasing f^{st} and iterated over each keypoint from the highest to the lowest strength. In each iteration, the keypoints F which are within a radius p (image plane) w.r.t. to the given keypoint f_i are removed from the set $\mathcal{S}_t$ i.e.

$$\begin{aligned} F &= ||f_i^{loc} - f_j^{loc}|| < p, \forall f_i, f_j \in \mathcal{S}_t \text{ and } i \neq j \\ \mathcal{S}_t &= \mathcal{S}_t - F \end{aligned} \tag{1}$$

where $||f_i^{loc} - f_j^{loc}||$ is the Euclidean distance between the locations of a pair of keypoints f_i and f_j. In our experiment we set p = $.05 \times image\_height$.

Secondly, we encode a keypoint f with K codewords $\alpha_1 \dots \alpha_K$ using only the descriptor f^{desc} part of the keypoints. In order to achieve this, we generate a codebook of size K using a standard K-means clustering algorithm. If we denote the center of the jth cluster as $mean_j$, then each keypoint $f \in \mathcal{S} = \{S_1 \dots S_T\}$ is mapped into the nearest codeword via

$$\alpha_i(\mathcal{S}) = \{f | f \in \mathcal{S} \wedge i = argmin_j ||f^{desc} - mean_j||\} \tag{2}$$

where $||f^{desc} - mean_j||$ denotes the Euclidean distance between feature descriptor f^{desc} and $mean_j$. As a result, we have decomposed the set $\mathcal{S}$ into K subsets, $\alpha_1(\mathcal{S}), \dots, \alpha_k(\mathcal{S})$ based on the keypoints descriptor. This is the quantisation step of the standard *bag-of-word* approach used in literature (Behera et al., 2012b; Ryoo and Aggarwal, 2009; Laptev et al., 2008).

In the third step, we extract relations between all possible pairs of keypoints within a subset $\alpha_1(\mathcal{S}), \dots, \alpha_k(\mathcal{S})$ i.e. the relations between keypoints assigned to the same codewords within an image. This relation is represented as $\overrightarrow{r}_{m,n} = (d_{m,n}, \theta_{m,n})$ between m^{th} and n^{th} keypoints (f_m and f_n), where $d_{m,n} = ||\overrightarrow{r}_{m,n}||$ and $\theta_{m,n}$ is the orientation w.r.t. the x-axis of the image plane i.e.

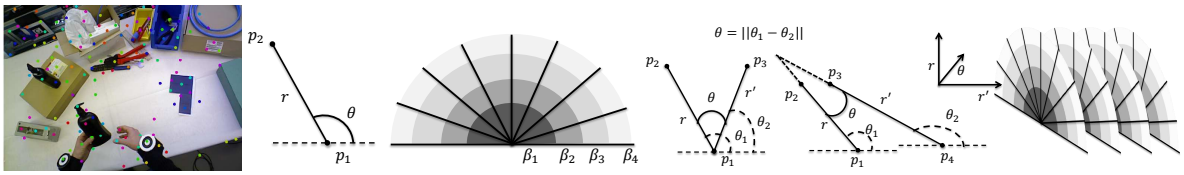
$$\begin{aligned} d_{m,n} &= ||f_m^{loc} - f_n^{loc}||, \forall f_m, f_n \in \mathcal{S}_k \text{ and } n > m \\ \theta_{m,n} &= \begin{cases} acos\left(\dfrac{\overrightarrow{r}_{m,n} \cdot \overrightarrow{x}}{\|\overrightarrow{r}_{m,n}\|}\right), \text{ if } (\overrightarrow{r}_{m,n} \cdot \overrightarrow{y}) > 0 \\ \pi - acos\left(\dfrac{\overrightarrow{r}_{m,n} \cdot \overrightarrow{x}}{\|\overrightarrow{r}_{m,n}\|}\right) \end{cases} \end{aligned} \tag{3}$$

where $\overrightarrow{x}$ and $\overrightarrow{y}$ are the orthogonal unit vectors defining the image plane. We extract all possible pairwise relations from all the subsets $S_{1\dots K}$.

Finally, the magnitude $d = \{d_{m,n}\}$ of the spatial relation is described with R possible codewords $\beta_1 \dots \beta_R$ using a K-means clustering algorithm. Each element in d is assigned to the nearest codeword using (2). The codewords in this codebook are sorted i.e. $\beta_1 < \beta_2 < \dots < \beta_R$ because we apply the smoothing over the histogram bins which represent the distribution of spatial relations. We discuss this further in the next section while generating our histogram of oriented pairwise relations.

3.1 Histogram of oriented pairwise relations (HOPR)

In this section, we explain the generation of histograms of pairwise relations (HOPR) from the extracted relations between all possible pairs of keypoints assigned with the same codewords. In our representation, the average distance between visual words in an image is represented with R possible codewords which are learned from the training set.



(a) (b) (c) (d) (e) (f)

Figure 2: (a) Assigned codewords $\alpha_1 \ldots \alpha_K$ to the filtered keypoints in an image, (b) pairwise relationship feature $\overrightarrow{r}$ for all possible pairs (2^{nd} order) of keypoints assigned with the same codewords, (c) creation of histogram of oriented pairwise relations (HOPR) using 8-bin orientations (unsigned) and codebook size of 4 i.e. $\beta_1 \ldots \beta_4$ for $||\overrightarrow{r}||$, (d) relationship features for 3^{rd} order, (e) relationship features for 4^{th} order and (f) HOPR for relationship features for 3^{rd} and 4^{th} order.

A similar approach is proposed by Savarese *et al.* to represent these relationships using the correlogram which is a function of kernel radius (Savarese et al., 2006). Our method is different from the above-mentioned approach as we characterise the spatio-temporal distribution of keypoints associated with the same visual words. The relationships between the different visual words can be extracted using higher order spatial features. For example, 4^{th} order relationship features (Fig. 2e) can be used to represent the spatial relationships between two different visual words, where pairs $(p_1, p_2) \in S_i$, $(p_3, p_4) \in S_j$ and $i \neq j$. The features originating from local keypoints (bag-of-visual words) are called 1^{st} *order features*. Similarly, the features that encode spatial relationships between a set of two, three or N keypoints are called as 2^{nd}, 3^{rd}, and N^{th} order features, respectively (Liu et al., 2008). These are analogous to *N-grams* used in statistical language modelling.

Fig. 2 shows our systematic approach for extracting 2^{nd}, 3^{rd} and 4^{th} order relationship features. Fig. 2a shows the distribution of keypoints over an image. These keypoints are filtered based on their strength f^{st} (step 1) and assigned respective codewords $\alpha_1 \ldots \alpha_k$ (step 2). Keypoints with identical color are assigned with the same codewords. Fig. 2b represents the extraction of relationships between pairs of keypoints having the same codewords. Each relationship r is represented with a distance and angle pair (d, θ). The distance d is assigned with the corresponding distance codewords $\beta_1 \ldots \beta_R$ and is the last step of our extraction process. The HOPR for the 2^{nd} order relationships features is shown in Fig. 2c and its dimension is $O \times R$. O represents the number of orientation bins and R describes the pairwise distance bins i.e. the distance codebook $\beta_1 \ldots \beta_R$ (in Fig. 2c, $O = 8$ and $R = 4$). One HOPR per descriptor codewords $\alpha_1 \ldots \alpha_K$ per frame is generated. Our approach considers the contribution from the adjacent bins before normalising the HOPR. These contributions are assigned a fixed weight of 0.6 for the current bin and 0.2 for the previous and following bins. The process is essentially a smoothing of the HOPR with predefined 1-D centered filter kernels of $[0.2, 0.6, 0.2]$ and $[0.2, 0.6, 0.2]^T$. Due to this, the distance codewords $\beta_1 < \beta_2 < \ldots < \beta_R$ are sorted as mentioned before. We use the L2-norm for normalising the HOPRs. The normalised HOPR from each descriptor codeword $\alpha_1 \ldots \alpha_K$ is concatenated to produce a final 2^{nd} order relationships feature vector that consists of $O \times R \times K$ elements, and will be used by a classifier for activity recognition.

Fig. 2d depicts the extraction of 3^{rd} order relationship features using a sets of three keypoints. In this setting, there are two pairwise relationships r and $r^{'}$ with keypoint p_1 appearing in both the relations (junction keypoint). We consider all possible configurations consisting of these three keypoints i.e. in the other two configurations p_2 and p_3 will be the respective junction point. During the computation of the 2^{nd} order relationship features, we have already extracted the distance angle pair (d, θ_1) and $(d^{'}, \theta_2)$, and assigned distance codeword $\beta_1 \ldots \beta_R$ for the respective r and $r^{'}$ relationships. While generating the HOPR for the 3^{rd} order relationship features, the relative angle $\theta = ||\theta_1 - \theta_2||$ between the relationships r and $r^{'}$ is used for orientation bins O and their respective pairwise distance for the distance codewords bins R. Fig. 2f shows the HOPR for 3^{rd} order relationship features with dimension of $O \times R \times R$. As in the 2^{nd} order HOPR, the same smoothing and normalisation steps are applied and the (smoothed) HOPR from each descriptor codeword is concatenated to represent the final 3^{rd} order relationship feature vector that consists of $O \times R \times R \times K$ elements.

Similarly, we extract the HOPR for the 4^{th} order relationship feature set as depicted in Fig. 2e. In this case, there is no junction keypoint as in the 3^{rd} order HOPR (Fig. 2d). However, if the relationships r and $r^{'}$ are not parallel then there is a point in the image plane where these relationships are joined. The extraction process and the dimension of this HOPR is the same as in the 3^{rd} order HOPR. It is worth mentioning that although the order of the

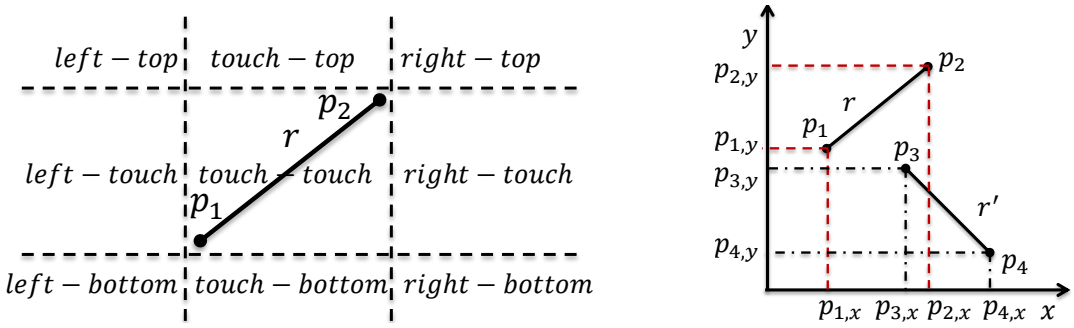


Figure 3: (a) Simplified RA relations in the image plane (left), (b) example projection of pairwise relationships in the image co-ordinates (right) - in this case the relationship is *touch-top*.

relationship feature set is increased from 3 to 4, the size of the HOPR descriptor is the same. In this work, we limit ourselves up to 4^{th} order relationship features. So far, our HOPR encodes the distance d between the positions f^{loc} of keypoints and their orientations θ w.r.t. the image plane. However, one further piece of important information is the spatial alignment of the relationships r w.r.t. $r^{'}$ can be used for further discriminating the activity pattern. We include this information in our HOPR representation as it encapsulates the relation between a pair of pairwise relations. We incorporate this information by simplifying Allen's temporal relations (Allen, 1983). We incorporate this information by a coarsened version of the rectangle algebra (RA) (Balbiani et al., 1999) which is a cross product of the Allen interval algebra (IA). Whereas the IA has 13 jointly exhaustive and pairwise disjoint (JEPD) relations, the RA at 13×13 = 169. By collapsing *before* and *meets* to a single relation (and correspondingly their inverses) and all the remaining nine relations *o, oi, s, si, f, fi, d, c* then we obtain a calculus with 3 JEPD relations in the 1D case and with 9 JEPD relations in the 2D case which has been called DIR9 (Liu et al., 2009b); originally DIR9 was conceived as a calculus for the bounding rectangles planar regions rather than line segments, but it is clear that once axis-aligned bounding rectangles have been computed the two cases are identical. The calculus with our names for the relations is depicted in Fig. 3. There are 3 x-relations (*left*, *right* and *touch*) for the x-axis and another 3 y-relations (*top*, *bottom* and *touch*) for the y-axis of the image plane. A total combination of 9 possible relations are extracted (Fig. 3). These relations are extracted using the positions f^{loc} of the keypoints in the image plane. The spatial alignment between a pair of pairwise relationships r (keypoints p_1 and p_2) and $r^{'}$ (keypoints p_3 and p_4) is computed as:

$left : p_{1,x} < p_{3,x} \wedge p_{1,x} < p_{4,x} \wedge p_{2,x} < p_{3,x} \wedge p_{2,x} < p_{4,x}$, $right : p_{1,x} > p_{3,x} \wedge p_{1,x} > p_{4,x} \wedge p_{2,x} > p_{3,x} \wedge p_{2,x} > p_{4,x}$ and $touch : \neg right \wedge \neg left$

$top : p_{1,y} < p_{3,y} \wedge p_{1,y} < p_{4,y} \wedge p_{2,y} < p_{3,y} \wedge p_{2,y} < p_{4,y}$, $bottom : p_{1,y} > p_{3,y} \wedge p_{1,y} > p_{4,y} \wedge p_{2,y} > p_{3,y} \wedge p_{2,y} > p_{4,y}$, and $touch : \neg top \wedge \neg bottom$

An example of the process of extracting such relations ($touch - top$) using pairwise relationships r and $r^{'}$ is shown in Fig. 3b. For convenience, we represent these relations as x-relation followed by y-relation *e.g.* for the spatial alignment of *touch-top*, the projection of the pairwise relationships r and $r^{'}$ on the x-axis are *touched*. Whereas on the y-axis, the projection of the relationship r is on the top of the relationship $r^{'}$. For a given order of relationship feature sets i.e. $2^{nd}, 3^{rd}$ or 4^{th}, we have already extracted all the involved pairwise relations r between all possible pairs of keypoints assigned with the same descriptor codeword (step 3). Let $\mathcal{R} = \{r\}$ be a set containing all pairwise relations r for a given order of relationships feature set in the image plane. The spatial alignment is computed by considering all possible pair of relations $(r_i, r_j) \in \mathcal{R}, i \geq j$ within the set $\mathcal{R}$. The relative orientation $\theta = ||\theta_i - \theta_j||$ between the pair (r_i, r_j) is used for the orientation bin O of the HOPR. The relative spatial alignment (9 relations) between the pair (r_i, r_j) is then added to the extracted HOPR. The final dimension of the HOPR for the 2^{nd} order relationships feature is $O \times (R+9) \times K$. Similarly, for the 3^{rd} order and above, the dimension of the HOPR is fixed and is of $O \times (R \times R + 9) \times K$. This is due to the fact that from the 3^{rd} order and above, we use the compute the relationships between a pair of lines as mentioned earlier and for N^{th} order relationship features, the respective dimension is $O \times (R^{N-1} + 9) \times K$.

Table 1: Framewise performance comparison for the experiment one-vs-rest-subject (Leeds dataset without video stabilisation).

	H_1	H_2		H_3		H_4		$\hat{H}_2$
	-	O=6	O=8	O=6	O=8	O=6	O=8	O=6
s_1	22.4	26.6	27.4	24.4	24.6	21.2	22.2	30.9
s_2	25.4	34.5	35.1	32.8	33.7	24.5	25.4	41.2
s_3	30.0	36.2	38.3	34.7	35.5	31.3	32.0	40.3
s_4	28.8	38.6	39.8	33.0	33.1	29.2	29.9	36.1
s_5	29.0	29.8	30.3	27.6	27.9	25.2	25.2	31.4
Avg.	27.1	33.1	34.2	30.5	31.0	26.3	26.9	36.0

3.2 Learning and Inference

We use a standard Support Vector Machine (SVM) to solve our multi-class classification problem in a supervised fashion. Every frame in a video is processed and the corresponding relationships feature vector HOPR is extracted and is used by the SVM for training and prediction. The activity label for each frame is provided by manual annotation. We use the χ^2-kernel which is given by $k(x,y) = 2(xy)/x+y$ and is named after the corresponding additive squared metric $D^2(\mathbf{x},\mathbf{y}) = \chi^2(\mathbf{x},\mathbf{y})$ which is a χ^2 distance between HOPR $\mathbf{x}$ and $\mathbf{y}$. The χ^2-kernel performs better than other additive kernels such as intersection and Hellinger's for histogram based classifications (Vedaldi and Zisserman, 2010). Due to the large dimensionality of the HoPR, we use the linear approximation of the χ^2-kernel in order to reduce the computational complexity which is one of the most important requirement for the real-time prediction of ongoing activity. This linear approximation is presented in (Vedaldi and Zisserman, 2010). We use the order $N = 2$ for the approximation i.e. if L is the dimension of the HOPR then after approximation the final dimension will be $L \times (2N+1)$. We use this approximation as an input feature vector for the linear SVM for the classification of activities (Fan et al., 2008).

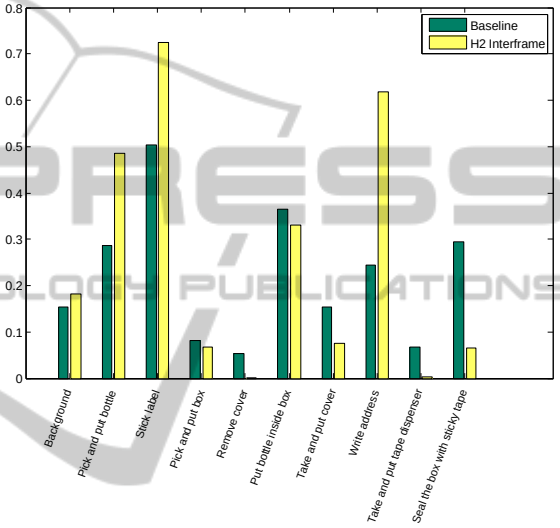


Figure 4: Action recognition results for Leeds dataset (Behera et al., 2012b) are compared with the baseline approach: SURF (Bay et al., 2006) (green) 27.1% and our approach (yellow) 36.0%.

4 EXPERIMENTS

In order to validate our novel representation of pairwise relationships using Histogram of Oriented Pairwise Relationships (HOPR), we use two publicly available egocentric datasets: 1) GTEA (GeorgiaTech Egocentric Activities) dataset consisting of kitchen activities (Fathi et al., 2011b) and 2) Leeds egocentric dataset ('labelling and packaging bottles') for manipulative tasks (Behera et al., 2012b). All evaluations are presented as a framewise classification accuracy.

For the baseline evaluation, we use the standard approach of a bag-of-visual words i.e. 1^{st} order feature H_1. In our experiment, we use SURF (Bay et al., 2006) feature descriptors as visual features. There is no specific reason for choosing SURF instead of SIFT (Lowe, 2004). We found that SURF is computationally less expensive than SIFT and prefer not to use STIP (Laptev and Lindeberg, 2003) due to the fact that the baseline performance of bag-of-visual word using STIP (14.4%) performed less well in comparison to SIFT (29.1%) on the GTEA dataset (Fathi et al., 2011a). It is worth to mention that the extraction of STIP features require more than a frame. In our baseline evaluation, we use a χ^2-kernel without any approximations and the size of the descriptor codebook is varied from 20 to 1000. We follow the same experimental setup i.e. 'leave-one-out' subject cross-validations presented in (Behera et al., 2012b; Fathi et al., 2011a). In the Leeds dataset, there are 5 subjects and a total of 26 video sequences having 9 different activities, whereas in the GTEA dataset, there are 4 subjects, 28 sequences and 10 verbs. The

Table 2: Framewise performance comparison for the experiment one-vs-rest-subject (GETA dataset without video stabilisation).

	H_1	H_2		H_3		H_4		$\hat{H}_2$	
	-	O=6	O=8	O=6	O=8	O=6	O=8	O=6	O=8
s_1	24.8	27.7	27.4	28.2	28.1	26.0	26.3	30.3	30.1
s_2	29.4	33.9	33.8	33.0	33.2	29.8	30.0	37.5	37.1
s_3	32.3	35.6	36.0	33.6	35.2	31.9	32.1	40.0	40.8
s_4	28.6	32.9	32.9	32.6	33.1	28.5	28.9	37.5	37.8
Avg.	28.8	32.5	32.5	31.8	32.4	29.1	29.3	36.3	36.5

Table 3: Framewise performance comparison for the experiment one-vs-rest-subject (GETA dataset with video stabilisation).

	H_1	H_2		H_3		H_4		$\hat{H}_2$	
	-	O=6	O=8	O=6	O=8	O=6	O=8	O=6	O=8
s_1	26.3	27.6	27.8	27.3	27.7	27.4	27.7	29.1	29.1
s_2	33.1	33.8	34.3	32.4	32.4	31.0	31.6	36.4	36.9
s_3	28.8	33.5	34.1	30.5	31.2	28.7	29.0	35.1	36.1
s_4	29.1	30.5	30.9	29.6	30.2	29.8	30.5	35.4	35.5
Avg	29.3	31.4	31.8	29.9	30.4	29.2	29.7	34.0	34.4

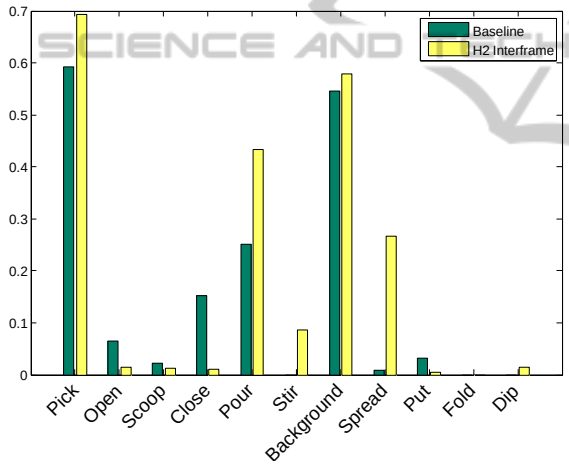


Figure 5: (a) Average action verb recognition results for GTEA datasets (Fathi et al., 2011b) over 'leave-one-out' subject, are compared with baseline approach: SURF (Bay et al., 2006) (green) 28.8% and our approach (yellow) 36.5% (left).

Leeds dataset does not provide any video stabilisation and the framewise recognition performance is presented in Table 1. We run the experiments on both the stabilized and unstabilised version of the GTEA datasets and the performance is provided in Table 2 and 3, respectively. The last row of all the Tables provides the average performance of 'leave-one-out' subject cross-validation.

The HOPR representation of the order 2^{nd}, 3^{rd} and 4^{th} is represented using H_2, H_3 and H_4, respectively. The extraction procedures for these histograms is explained in section 3.1. For this experiment, we have computed the HOPR for 2^{nd} order features sets between frames and is symbolised as $\hat{H}_2$. While computing $\hat{H}_2$, the current frame is compared with the previous 3 frames with a gap of 0.25 seconds between two consecutive frames. For this experiment, we keep the codebook size of 20 for visual words and a pairwise distance codebook size of 8. We compare the performance using two different orientations for HOPR i.e. $O = 6$ and $O = 8$.

From the performance tables of both the dataset (Table 1-3), it is evident that the performance of our representation i.e. HOPR is better than that of the bag-of-visual words approach. It is note-worthy that in the baseline, we use the full χ^2-kernel without any linear approximation and the best performance is selected using the varying size of the visual codebook. In both the datasets, the HOPR $\hat{H}_2$ gives best performance. For the GTEA dataset, it is 36.5% and 34.4% without using stabilisation and with stabilisation, respectively. For the Leeds it is 36% (without using stabilisation).

The other valuable observation in the GTEA dataset is that by using video stabilisation the average baseline performance increased from 28.8% to 29.3% whearas the performance only decreases slightly when using the HOPR. This provides evidence for our HOPR representation for the recognition of egocentric activities. This also explains the robustness of our pairwise relational structure to the uncontrolled movement of cameras in an egocentric setup.

In both the datasets $O = 8$ orientation bins gives slightly better (0.1 % - 0.6 %) performance than $O = 6$. The 2^{nd} order relationship features (H_2) encodes the spatial distribution and is more sparse than the 1^{st}

order features (bag-of-visual words i.e. H_1). Therefore, the performance of the 2^{nd} order HOPR is better than the bag-of-visual words. However, when we increase the relationships feature order to 3 or 4, the performance decreases. This can be explained by the fact that 3^{rd} and 4^{th} order features are more sparse than 2^{nd} order features and hence, statistically less reliable.

5 CONCLUSIONS AND FUTURE WORKS

We present a novel approach to egocentric video activity representation based on the relationships between visual words. These pairwise relations are encoded using Histogram of Oriented Pairwise Relations (HOPR). The movement and interaction between objects and hands are captured by observing the spatial relationships between features in video frames. This representation does not require the detection of objects or hands in comparison to other common approaches. In addition, it can be used for real-time activity detection which requires the recognition of partial observations i.e. single frame to few frames. In this work using egocentric data, we show that by encoding the spatiotemporal relationships between local features in activity representations improves performance over state-of-the-art activity representation approaches such as the bag-of-visual words. We would like to further investigate on the hierarchical relationships structure using local visual features.

ACKNOWLEDGEMENTS

This research work is funded by the EU FP7-ICT-248290 (ICT Cognitive Systems and Robotics) grant COGNITO (www.ict-cognito.org), FP7-ICT-287752 grant RACE (http://project-race.eu/) and FP7-ICT-600623 grant STRANDS (http://www.strands-project.eu/).

REFERENCES

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):1–16.

Aghazadeh, O., Sullivan, J., and Carlsson, S. (2011). Novelty detection from an ego-centric perspective. In *CVPR*, pages 3297–3304.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843.

Balbiani, P., Condotta, J.-F., and del Cerro, L. F. (1999). A new tractable subclass of the rectangle algebra. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 442–447.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). SURF: Speeded up robust features. In *ECCV*, pages 404–417.

Behera, A., Cohn, A. G., and Hogg, D. C. (2012a). Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In *MMM*, pages 196–209.

Behera, A., Hogg, D. C., and Cohn, A. G. (2012b). Egocentric activity monitoring and recovery. In *ACCV*, pages 519–532.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *ICCV*, pages 1395–1402.

Carneiro, G. and Lowe, D. (2006). Sparse flexible models of local features. In *ECCV*, pages 29–43.

Crandall, D. J. and Huttenlocher, D. P. (2006). Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV (1)*, pages 16–29.

Deselaers, T. and Ferrari, V. (2010). Global and efficient self-similarity for object classification and detection. In *CVPR*, pages 1633–1640.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Fathi, A., Farhadi, A., and Rehg, J. M. (2011a). Understanding egocentric activities. In *ICCV*, pages 407–414.

Fathi, A., Ren, X., and Rehg, J. M. (2011b). Learning to recognize objects in egocentric activities. In *CVPR*, pages 3281–3288.

Gilbert, A., Illingworth, J., and Bowden, R. (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, pages 925–931.

Gupta, A. and Davis, L. S. (2007). Objects in action: An approach for combining action understanding and object perception. In *CVPR*, pages 1–8.

Kitani, K. M., Okabe, T., Sato, Y., and Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, pages 3241–3248.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *ICCV*, pages 432–439.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *CVPR*, pages 1–8.

Liu, D., Hua, G., Viola, P. A., and Chen, T. (2008). Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, pages 1–8.

Liu, J., Luo, J., and Shah, M. (2009a). Recognizing realistic actions from videos "in the wild". In *CVPR*, pages 1996–2003.

Liu, W., Li, S., and Renz, J. (2009b). Combining rcc-8 with qualitative direction calculi: Algorithms and complexity. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 854–859.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Representing pairwise spatial and temporal relations for action recognition. In *ECCV (1)*, pages 508–521.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126.

Niebles, J. C. and Li, F.-F. (2007). A hierarchical model of shape and appearance for human action classification. In *CVPR*, pages 1–8.

Ryoo, M. S. and Aggarwal, J. K. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, pages 1593–1600.

Savarese, S., Winn, J. M., and Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlations. In *CVPR (2)*, pages 2033–2040.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *ICPR*, pages 32–36.

Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *CVPR*.

Starner, T. and Pentland, A. (1995). Real-time American sign language recognition from video using hidden Markov models. In *Proc. of Int'l Symposium on Computer Vision*, pages 265 – 270.

Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T.-S., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, pages 2004–2011.

Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488.

Vedaldi, A. and Zisserman, A. (2010). Efficient additive kernels via explicit feature maps. In *CVPR*, pages 3539–3546.