
Article:
Al Harbi, N. and Gotoh, Y. (2015) A unified spatio-temporal human body region tracking approach to action recognition. Neurocomputing, 161. 56 - 64. ISSN 0925-2312

https://doi.org/10.1016/j.neucom.2014.11.072

A Unified Spatio-temporal Human Body Region Tracking Approach to Action Recognition

Nouf Al Harbi, Yoshihiko Gotoh

Department of Computer Science, University of Sheffield, United Kingdom

Abstract

There are numerous instances in which, in addition to the direct observation of a human body in motion, the characteristics of related objects can also contribute to the identification of human actions. The aim of the present paper is to address this issue and suggest a multi-feature method of determining human actions. This study addresses the matter by applying a sturdy region tracking method, instead of the conventional space-time interest point feature based techniques, demonstrating that region descriptors can be attained for the action classification task. A cutting-edge human detection method is applied to generate a model incorporating generic object foreground segments. These segments have been extended to include non-human objects which interact with a human in a video scene to capture the action semantically. Extracted segments are subsequently expressed using HOG/HOF descriptors in order to delineate their appearance and movement. The LLC coding is employed to optimise the codebook, the coding scheme projecting every one of the spatio-temporal descriptors into a local coordinate representation developed via max pooling. Human action classification tasks were used to assess the performance of this model. Experiments using KTH, UCF sports and Hollywood2 dataset show that this approach achieves the state-of-the-art performance.

Keywords: spatio-temporal segmentation, human body volume, object tracking, regions of interest, action recognition

1. Introduction

It is not a difficult task for us to comprehend the actions occurring in a video clip, regardless of the scene context, individuals in the scene or the camera angles with which the scene is presented. Furthermore, viewers can follow an extensive series of actions no matter how complex they are. From the computational point of view however, action representation poses considerable challenges. To provide a solution to this problem, most existing approaches are geared towards the expression of motion information within a scene. Descriptors for motion information are highly important; in recent years methods used to garner space-time interest point (STIP) features have been greatly improved [1]. It has been demonstrated that high-level models, which functioned on representations developed based on tracked objects, their features and/or interaction, were capable of identifying complex actions [2, 3]. Such high-level model, relying on interaction primitives, was proven to be highly effective when extracting appearance and STIP primitives.

STIPs are extricated from video data using the Bag of Words (BoW) model. They provide the basis for current action recognition research, which depends mostly on the ability that differentiates unique local space-time descriptors. STIP primitives are outlined on person and object trajectories, and attained via a flexible part-model detector [4]. However, despite their efficiency, such models remain incapable of tracking all object types or of functioning in a variety of observation conditions. It disregards information about the spatio-temporal organisation of the interest points which could be important for various computer vision tasks. This paper moves away from point-feature-based approaches, instead examines a spatio-temporal 'region-based' approach to interpret motion extracted from a video stream. In order to process complex actions which are challenging to track efficiently using conventional descriptors, this paper investigates a new model for action representation that relies on detecting spatio-temporal person-object interaction regions [5, 6].

Email addresses: nmalharbi1@sheffield.ac.uk (Nouf Al Harbi), y.gotoh@dcs.shef.ac.uk (Yoshihiko Gotoh)

The argument brought forth is that video ought to be perceived as an assemblage of three-dimensional volumes. The integrated analysis of the temporal and spatial dimensions of a video presents a number of advantages. First of all, it facilitates the preservation of spatial and temporal consistency. Secondly, higher-level algorithms are able to concentrate on extensive, sparse regions rather than undertaking a multiple frame analysis of pixels, thus enhancing efficiency. Thirdly, joint modelling of object appearance and movement leads to improved recognition results.

To this end the approach builds on a segmentation of human and non-human objects, where a human body volume is detected along a video stream [7]. These segments are extended to accommodate an non-human objects to form up final key-segment regions. They formulate a descriptor that encompasses the static and dynamic features of detected key segments. The KTH, UCF sports and Hollywood2 dataset are employed to assess these representations [1]. It is demonstrated that, in comparison to conventional methods, action representation is substantially optimised, and that the codebook enhancement based on the locality constrained linear coding (LLC) technique [8] conveys the highest performance. The contributions of this work can be summarised as follows:

- Extraction of a spatio-temporal human body volume by incorporating generic object foreground segments guided by the state-of-the-art of human detection and segmentation approaches as well as extension of these regions to accommodate an interacted non-human objects regions;
- Extension of the existing two-dimensional (2D) image LLC scheme to a spatio-temporal video signal;
- Development of an efficient and robust schema to represent a human action signal;
- Application of the spatio-temporal region-based approach to the action classification task with Hollywood2, one of the most challenging real-world dataset, demonstrating that the approach outperforms the state-of-the-art, interest point-based techniques by a clear margin.

The idea of using objects to improve the recognition rate for actions was proposed by [9]. However that work used a traditional approach to detecting humans and objects [10], suffering from the large size of 'space' not belonging to a human body or an object. To the contrary, as illustrated in Figure 1, this paper presents an approach that segments a human body and object regions at a frame level and tracks them over a sequence of video frames, thus creating a carefully trimmed spatio-temporal human body volume. Outcomes from the experiment indicate that the availability of an exact object region results in more accurate action representation.

Figure 1: A sample segmentation from the Hollywood2 dataset: the original frame from a video clip 'sceneclipautoautotrain00319' (left), a segmentation using Felzenszwalb *et al.* [10] (middle) and a segmentation using the approach presented in this paper (right).

2. Related Work

Action representations incorporating low-level track point features in a video have been embraced by a large number of research works [11, 12, 13, 14, 15]. However these types of representations can incur tracking errors, particularly when there is background clutter present. On the other hand such representations circumvent the onerous task of object and person identification.

The use of representations for human motion to identify human actions has received a fair amount of attention. One of the first to investigate this phenomenon was by Bobick and Davis [16], who managed to capture view-dependent motion, as well as Yacoob and Black [17], who developed parameter-based motion models. For action recognition, Ali *et al.* [18] suggested the use of kinematic flow features. A different approach proposed by Efros *et al.* [19], and later by Zelnik-Manor and Irani [20], was the correlation-based categorisation of human motion. Schechtman and Irani [21] have implemented this approach to associate correspondences and self-similarities between images and videos.

Space-time interest point (STIP) methods have attracted an increasing amount of attention recently. By employing local STIPs, a number of studies have generated representations on the basis of visual vocabularies outlined with the help of gradient-based descriptors obtained either at determined points of interest [14, 15, 22, 23, 24] or from the actual point locations [11, 25]. The positive implications of associating static and dynamic descriptors have been emphasised as well [1, 12]. Compound neighbourhood-based features were originally developed for static images and object identification [26, 27], but have been recently expanded to video processing [14, 15, 23]. A wide range of ap-

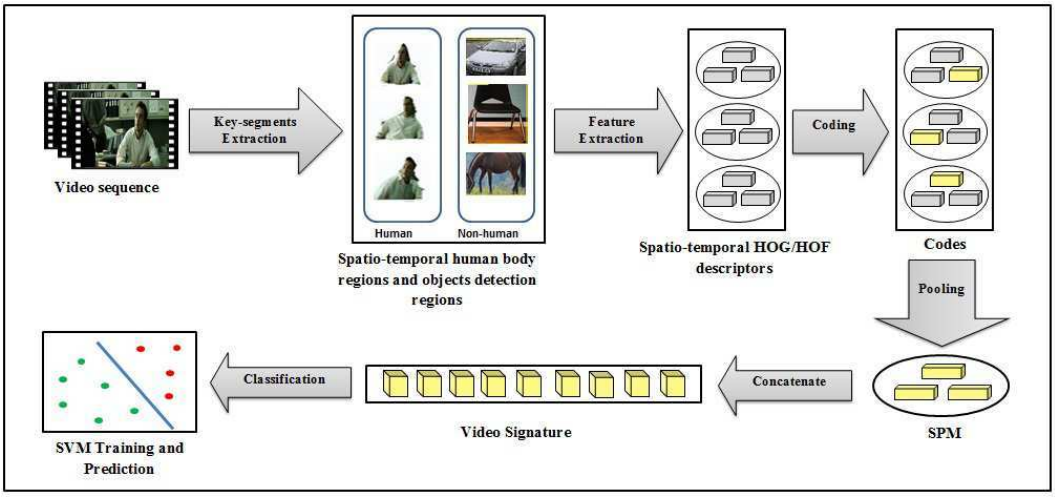

Figure 2: Processing flow of the 'human body region tracking' approach with visual object recognition (HBRT/VOC).

proaches are available, aiming to apply a coarse grid of histogram bins to subdivide the space-time volume globally [14, 15, 23] or else to position grids around the raw points of interest in order to generate new representations based on the location of the interest points which are included in the grid cells encompassing a central point [11].

A dataset derived from Hollywood films was presented by Laptev *et al.* [15]. Due to the fact that it consists of a wider variety of viewing angles, background clutter and scenarios, this dataset is considerably more difficult to process than the earlier datasets. This new dataset was referred to as 'Hollywood2: Human Actions and Scenes Dataset', and developed further by Marszalek *et al.* [1], who incorporated contextual information in their method. Recently another line of work was proposed by Bilen *et al.* [28] and assessed using Hollywood2 data. They attempted to describe actions by extracting salient local regions, applying motion segmentation, then tracking them with optical flow.

3. Human Action Representation

Despite relying on local information as well, the approach in this paper is different from existing works in that it focuses on a human body region-based feature representation. It is more concerned with the temporal continuity (or tracking) of regions than with isolated spatio-temporal regions. Figure 2 illustrates the processing flow of the technique presented in this paper, which is later on referred to as the 'human body region tracking' approach with visual object recognition (or HBRT/VOC).

3.1. Detecting and Tracking Human Body Regions

Our goal is to segment human body volume in an unlabelled video. The approach consists of two main stages (Figure 3). Firstly, human body objects are segmented at a frame level by combining low-level cues with a top-down part-based person detector, formulating grouped patches. Secondly, detected segments are propagated along the time line of video frames, exploiting the temporal consistency of detected foreground objects using colour models and local shape matching [5]. The final output is a spatio-temporal segmentation of the human body in a video stream. Figure 4 presents sample segmentations of a 'SitDown' action from the Hollywood2 dataset, forming a three-dimensional (3D) human body volume with two spatial and one temporal domains. In the following each stage is described in turn.

3.1.1. Estimating Human Body Region at Frame Level

This stage builds on the graph-based image segmentation technique by Maire *et al.* [7]. It produces a grouping of parts and pixels along the following idea:

- pixels are connected based on low-level cues in order to accomplish region consistency;
- detected parts are bound together when they belong to the same object;
- the regions belonging to a part are included in the foreground, whereas the remaining regions are pushed to the background.

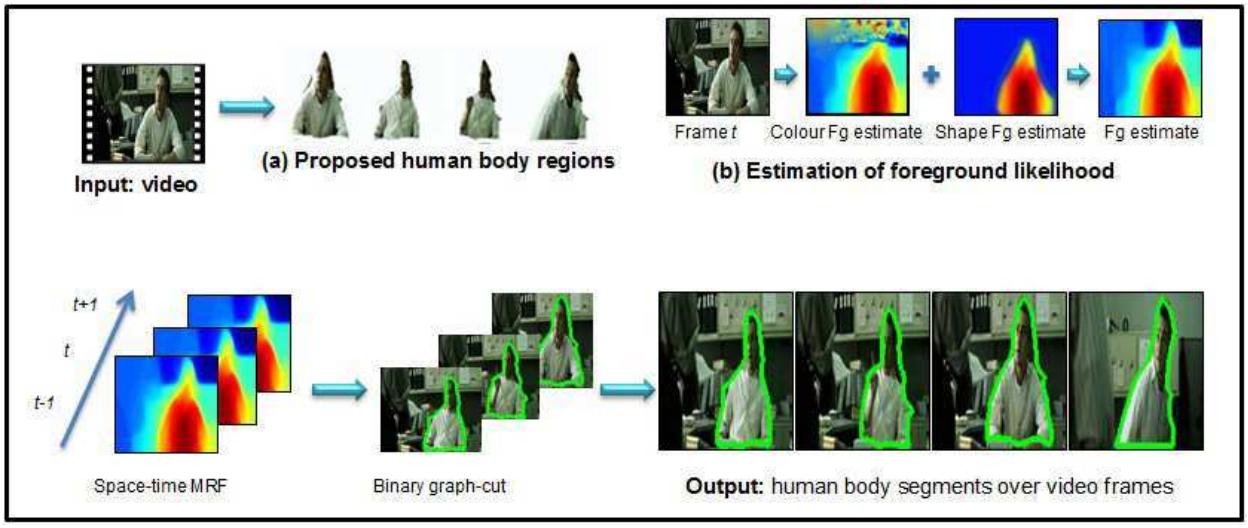

Figure 3: Two-stage approach to human volume segmentation. A human body detected in the first stage is propagated along video frames in the second stage.

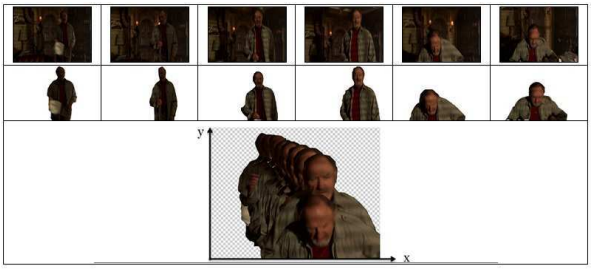

Figure 4: Sample segmentations of a human body volume of a 'Sit-Down' action from the Hollywood2 dataset. The first row shows key frames from a video clip, identified as '`actioncliptest00269`'. The second row presents the segmentation made by our approach of detecting and tracking human body regions. The montage in the third row is the human body volume created from the continuous video frames by localising the body regions at a frame-level (in 2D space) and tracking these regions over time. The furthest one is the oldest segmentation while the closet one is the most recent.

The key-segment extraction consists of several stages. A preliminary list of top-down regions for the unannotated video is generated and subsequently ordered on the basis of the appearance for each region of the human body. The development of the list relies on the segmentation of each frame with person detection [7]. To formulate several key-segment hypotheses, the detected regions across all frames were grouped together. Further detail should be referred to [5].

3.1.2. Spatio-Temporal Segmentation of Human from a Video Stream

Every hypothesis outlines a foreground colour probability model within the spatio-temporal context of Markov Random Field (MRF), in which each node takes the form of a pixel and each edge spatio-temporally links the adjoining pixels. Graph-cuts are used to subdivide this graph in order to generate a pixel-wise segmentation for the temporal movement of the identified human. The segmentations are then computed according to the hypothesis rank and non-overlap between the chosen key-segments is imposed so that a correlation between each hypothesis and a unique human in the video is achieved. Figure 5 shows two samples for human body segmentation using the approach. Ample discussion of this method was made in [6].

3.1.3. Object Detection Regions

In many instances, human activities can be effectively presented by collaboration and interaction between human and non-human objects. Eating action, for example, can be illustrated by describing a person who sits around a dining table and grasps the food. Consequently action classification will operate more effectively if non-human objects are incorporated into the zone of interest. A number of studies have been conducted for visual object recognition tasks[1] (VOC) that can be employed as a frontend processor for the human body region tracking

[1] *e.g.*, `pascallin.ecs.soton.ac.uk/challenges/VOC/` for the PASCAL visual object classes.

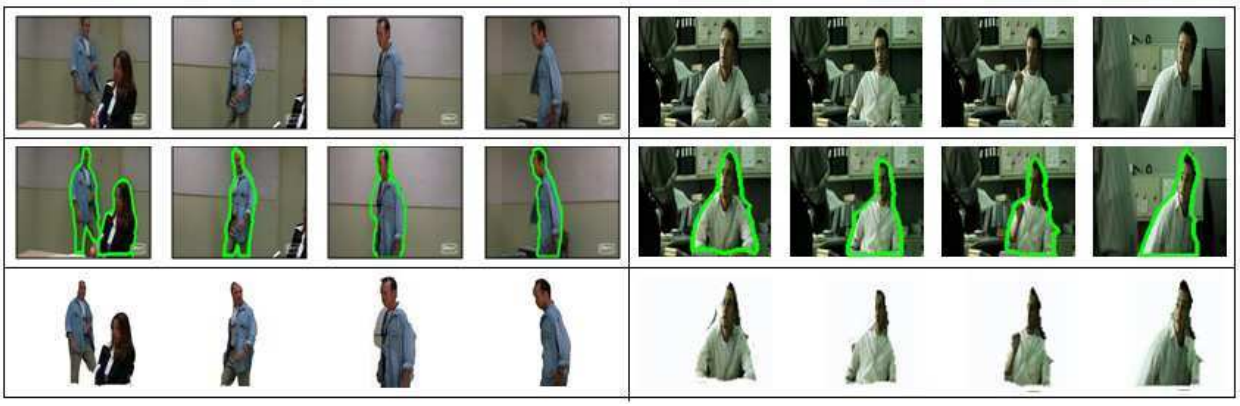

Figure 5: Two sample segmentations of action localisation, segmentation and recognition from the Hollywood2 dataset. They are identified as '`sceneclipautoautotrain00405`' (left) and '`sceneclipautoautotrain00319`' (right), selected based on the number of humans present. The first row shows key frames from two video clips. The second and the third rows respectively present the results of key segments and the corresponding segmentation using our approach of detecting and tracking human body regions.

(HBRT) approach. In this study a detector developed by Felzenszwalb *et al.* [29] is adopted, creating a store of the following twenty object classes: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train , bottle, chair, dining table, potted plant, sofa and tv/monitor. A window is tightly fitted to the identified object segment and a bounding box is drawn on the window. In order to guarantee the consistency of the object segments perceived with the human body regions in the video, certain spatial restrictions are imposed; if a confluence of human body and non-human regions exists, the segments are included as key-segments regions. See this in Figure 6.

3.2. Describing Detected Regions

Once key-segments are determined, they must be described as the identified hypotheses. In order to encompass the appearance and motion patterns of the regions of interest throughout a video clip, for each pixel contained in the segmented regions the HOG/HOF descriptor [15] is employed in this study. The 162-bin descriptor is composed of a histogram of oriented gradients (HOG) and a histogram of oriented flow (HOF). To outline the movement and appearance of selected features, the histogram descriptors of space-time volumes are positioned in the proximity of the identified points. Each volume is subsequently separated into a $n_x \times n_y \times n_t$ grid of cells[2]; for each cuboid a coarse HOG with 4-bin histogram and a HOF with 5-bin histogram are generated. Normalised histograms are integrated into HOG/HOF descriptor vectors, exhibiting certain similarities with the SIFT (scale invariant feature transform) descriptor by Lowe [30]. As an additional note, in the event that key-segment hypotheses are not produced for some video clip due to a failure of human detection, the HOG/HOF descriptor is augmented with space-time interest points.

[2]The parameter values employed in this study are $n_x = n_y = 3$ and $n_t = 2$, following the setup described in [15].

Figure 6: A sample clip from the Hollywood2 dataset: GetOutCar action from a video clip 'actioncliptest00108'. A region was detected using Felzenszwalb *et al.* [29] (red bounding box), while a human body was detected by using the approach presented in this paper (green contour). The car region was included in the action representation as there was an overlap between a car and a human.

4. Learning Feature Sets

A training set consists of M videos, and we define $X = \{x_1, x_2, \ldots, x_M\}$ where x_i represents a set of concatenated N-dimensional spatio-temporal descrip-

tors for each video. A sufficient number of features are randomly selected and grouped together using k-means in order to attain a preliminary codebook B of the fixed vocabulary size for spatio-temporal features. LLC is subsequently applied to enhance the codebook, which consists of the following three steps: representing video signals with spatio-temporal local descriptors X, generating the locality-constrained sparse code S, and finally optimising the codebook B. The codebook basis and LLC coefficients should efficiently approximate spatio-temporal descriptors. The following objective function [31] is employed:

$$\underset{S,B}{\operatorname{argmin}} \sum_{i=1}^{M} \{||x_i - Bs_i||^2 + \lambda ||d_i \odot s_i||^2\} \quad st. \quad 1^\top s_i = 1, \forall i \tag{1}$$

where $\odot$ denotes element-wise multiplication and λ is a weight parameter to control the locality constraint. The constraint, $1^\top s_i = 1$, meets the requirement of shift-invariance for the LLC coding scheme. The locality constrained parameter d_i represents every basis vector in codebook with different freedom on the basis of its similarity to the spatio-temporal descriptor x_i:

$$d_i = \exp\left(\frac{dist(x_i, B)}{\sigma}\right) \tag{2}$$
$$\text{with} \quad dist(x_i, B) = \{dist(x_i, b_1), \ldots, dist(x_i, b_M)\}^T$$

where $dist(x_i, b_j)$ represents the Euclidean distance between the spatio-temporal descriptor and the basis codebook B, and σ is a weight parameter to control the locality constraint. This is a convex problem in B only or in S but not in both together, and can be iteratively solved by the coordinate descent method:

1. Initialise the dictionary B with the codebook generated by clustering:
$$B \leftarrow B_{init} \tag{3}$$
2. For each spatio-temporal descriptor x_i, compute the new LLC coefficient s_i using the current B:
$$s_i \leftarrow \underset{s}{\operatorname{argmax}} \{||x_i - Bs||^2 + \lambda ||d \odot s||^2\} \quad st. \quad 1^\top s = 1 \tag{4}$$
3. Update the current dictionary, only if the computed LLC coefficient value is greater than a predefined threshold:
$$\triangle B_i \leftarrow -2\tilde{s_i}(x_i - B_i\tilde{s_i}) \tag{5}$$
$$\mu \leftarrow \sqrt{\frac{1}{i}} \tag{6}$$
$$B_i \leftarrow B_i - \frac{\mu \triangle B_i}{|\tilde{s_i}|_2} \tag{7}$$
4. Project the computed dictionary onto the output matrix:
$$B(:, id) \leftarrow proj(B_i) \tag{8}$$

The features are quantised on the basis of the vocabulary, with the purpose of creating a feature histogram which represents the vector for feature categorisation. A linear support vector machine (SVM) classifier from LIBLINEAR package is used to learn a model from the feature vectors for each action [32], where the regularisation parameter was set $C = 10$.

5. Experiments

This section assesses the effectiveness of the approaches, spatio-temporal HBRT and its extension HBRT/VOC, using three action recognition datasets. For all datasets we apply HOG/HOF descriptors with 162 dimensions. 100,000 features are randomly selected for initialisation of the codebook with a vocabulary size of 4000 words (the key parameter for dictionary training), and the number of neighbours is $K = 5$. A codebook size of 4000 is adopted in order to make a fair comparison with the standard action recognition studies, although a much larger codebook may be used to improve the modelling capacity. In Equation (1) $\lambda = 500$ is selected, and $\sigma = 100$ is set for Equation (2). These parameter values are adopted from the original work conducted by Wang *et al.* [31]. To evaluate the outcome of the action classification task, accuracy per class is calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{9}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positive and false negative, respectively.

5.1. Datasets and the experimental procedure

The body region tracking approaches (HBRT and HBRT/VOC) were comprehensively evaluated using three action datasets selected from the KTH, the UCF sports and the Hollywood2 video data. The datasets encompassed a variety of locations and scene settings shown in video clips, including controlled experimental settings, popular films and televised sporting events. The assessment incorporated a range of variations resulting from different resolutions, perspectives, lighting shifts, occlusion, background disorder, and irregular motion. Overall, more than 4000 video segments were assessed with 28 action classes. Some sample frames are presented in Figure 7.

Figure 7: Sample frames from the three action recognition datasets, the KTH (top row), the UCF Sports Action (middle) and the Hollywood2: Human Actions and Scenes (bottom), used for the experiments.

KTH Dataset[3] (Schuldt *et al.* 2004 [24])
2391 video segments made this dataset with six types of human action: 'walking', 'jogging', 'running', 'boxing', 'waving', and 'clapping'. Each action was carried out for a number of times by 25 people and filmed in a variety of settings: outside, outside with scale variation, outside in changed clothing and inside. In most of the segments the background was regular and still. Segments were resized to a spatial resolution of 160×120 pixels and the mean duration of video clips was four seconds. We adhered to the experimental format of the existing studies by splitting the samples into a test set (nine subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and a training set (the other 16 subjects). Emulating the original paper, we trained and assessed a multi-class classifier [24], and calculated the accuracy for each class and finally reported the average accuracy over all classes.

UCF Sports Action[4] (Rodriguez *et al.* 2008 [33])
This dataset contained ten human actions: 'swinging' (both on a pommel horse and on the ground), 'diving', 'kicking a ball from the front and the side', 'lifting weights', 'horse-riding', 'running', 'skateboarding', 'swinging from the high bar', 'gold swinging from the back, front and side', and 'walking'. Nearly 200 video segments were used with a resolution of 720×480, indicating significant intra-class variability. As with the KTH set, we employed a multi-class classifier and reported the average accuracy in all classes.

Hollywood2: Human Actions and Scenes Dataset[5] (Marszalek *et al.* 2009 [1])
This data has been collected from 69 different Hollywood movies. It consisted of the following 12 action classes to be identified from real-life film scenes: 'answering a phone', 'driving a car', 'eating', 'fighting', 'getting out the car', 'hand shaking', 'hugging', 'kissing', 'running', 'sitting down', 'sitting up' and 'standing up'. In total there were 1707 video sequences divided into a training set (823 sequences) and a test set (884 sequences), with the average length of ten seconds. Training and test sequences were mutually exclusive. The experiment was performed on this dataset with a spatial resolution of 360×288 pixels and a sample rate of 4.6 fps (frames per second) as suggested by [34]. A one-against-all SVM categorisation was applied where a binary classifier recorded every action [32].

[3] www.nada.kth.se/cvap/actions/
[4] server.cs.ucf.edu/~vision/data.html

	bx	cl	hw	jg	rn	wk
boxing	1	0	0	0	0	0
clapping	0	1	0	0	0	0
handwaving	0	0.02	0.98	0	0	0
jogging	0	0	0	0.97	0.03	0
running	0	0	0	0.02	0.97	0.01
walking	0	0	0	0	0.01	0.99

Figure 8: **(KTH Dataset)** Confusion matrix between six action classes using the HBRT/VOC combination approach.

method	accuracy (%)
point feature based:	
Laptev *et al.* (2008) [15]	91.8
Le *et al.* (2011) [35]	93.9
Gilbert *et al.* (2009) [11]	94.5
Sadanand *et al.* (2012) [36]	98.2
local region tracking:	
HBRT	97.2
HBRT/VOC	98.5

Table 1: **(KTH Dataset)** Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature based methods.

5.2. Experimental Results

Table 1 presents the comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature based techniques using the KTH dataset [24]. To date, it is probably the most frequently-used dataset in assessment of action recognition. The region tracking approaches performed well; the HBRT method achieved 97.2%, and its integration with VOC reached 98.5%, outperforming the reported outcome of Sadanand *et al.* [36] by a small margin. Figure 8 illustrates the confusion matrix associated with the HBRT/VOC approach. It still made occasional, although rare, confusion between 'jogging', 'running' and 'walking' actions. The effectiveness of the HBRT/VOC resulted from the shrewd targeting of interest points represented by human body regions, which allowed the action to be pre-determined and eradicated superfluous and noisy background. Although KTH dataset only depict people, the VOC efficiently enhances the accuracy of results. This can be explained because the HBRT is failed to detect a person who is far from a camera.

Table 2 makes the same comparison, but this time using the UCF Sports Action Dataset [33]. For the region tracking approaches, the overall accuracy was

[5] `lear.inrialpes.fr/data`

method	accuracy (%)
point feature based:	
Le *et al.* (2011) [35]	86.5
Kovashka and Grauman (2010) [37]	87.3
Wu *et al.* (2011) [38]	91.3
Sadanand *et al.* (2012) [36]	95.0
local region tracking:	
HBRT	90.8
HBRT/VOC	96.2

Table 2: **(UCF Sports Action Dataset)** Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature based methods.

90.8% with the HBRT, which was further improved to 96.2% with the HBRT/VOC; the latter clearly outperformed the recent state-of-the-art (95.0%) by Sadanand *et al.* [36]. For the HBRT/VOC, the confusion matrix between ten action classes is presented in Figure 9. The figure shows some confusion pairs such as 'lifting'/'skating' and 'running'/'walking' as their action representation was quite similar. The outcome indicates that the local region tracking is an effective new approach to capturing human activity on video, and possesses the great potential to achieve consistent performance in realistic conditions.

The KTH and the UCF Sports Action were both relatively small datasets. Hollywood2: Human Actions and Scenes [1], on the other hand, was substantially more difficult dataset to process because of several reasons, *e.g.*, more classes, the larger number of videos, actions with more realistic background involving multiple objects, camera motions. Table 3 compares various approaches using the challenging Hollywood2 data. Laptev *et al.* [15] presented a technique where space-time interest points were identified by the Harris-Laplace detector and described with HOF. Another technique, based on the motion region, was proposed by Bilen *et al.* [28]. The table shows the significant improvement made by the HBRT and the HBRT/VOC, *i.e.*, the local region-based approaches. In particular the latter achieved improvement of more than 4% absolute over Laptev *et al.* [15]. Recently, Zhang *et al.* [39] introduced a novel simplex-based orientation decomposition descriptor to quantise and represent 3D spatio-temporal features. This approach decomposes every 3D visual cue in a features support region into three different angles and transforms the output decomposed angles into the simplex topological vector space. The proposed technique able to address the singularity and limited discrimination power issues. Then, quadrant decomposition is performed to improve our SOD

	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	1	0	0	0	0	0	0	0	0	0
golfing	0	1	0	0	0	0	0	0	0	0
kicking	0	0	1	0	0	0	0	0	0	0
lifting	0	0	0	0.92	0	0	0.08	0	0	0
riding	0	0	0	0	1	0	0	0	0	0
running	0	0	0	0	0	0.95	0	0	0	0.05
skating	0	0	0	0.06	0	0	0.94	0	0	0
swing-bench	0	0	0	0	0	0	0	1	0	0
h-swinging	0	0	0	0	0	0	0	0.10	0.90	0
walking	0.02	0	0	0	0	0.07	0	0	0	0.91

Figure 9: **(UCF Sports Action Dataset)** Confusion matrix between ten action classes using the HBRT/VOC combination approach.

method	accuracy (%)
point feature based:	
Laptev *et al.* (2008) [15]	44.4
Bilen *et al.* (2011) [28]	41.3
Zhang *et al.* (2014) [39]	50.9
local region tracking:	
HBRT	44.4
HBRT/VOC	48.6

Table 3: **(Hollywood2: Human Actions and Scenes Dataset)** Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature based methods.

descriptors discrimination capability, and a final feature vector is formed by combining decomposed histograms from all quadrants. The table 3 shows that Zhang *et al.* [39] approach improve the result significantly over the HBRT/VOC by 2%.

Additionally Table 4 presents the comparative analysis of the performance for individual classes by the region and point feature-based approaches. The complexity of the Hollywood2 data resulted in the low performance with several action classes, in particular with 'AnswerPhone', 'GetOutCar' and 'SitUp'. Some videos contained a variety of camera motions, peripheral actions as well as a multitude of viewing angles and action sequences. Even a human could fail the classification task when a subject was far from a camera position. The HBRT has proven to be highly effective particularly when processing subtle actions such as 'SitUp' and 'SitDown'. It is interesting to note that, according to Table 4, the HBRT result was improved by the HBRT/VOC, the latter extended the region of interest to accommodate non-human objects such as, car, dinning table and chair. The significant improvement was observed with classes such as, 'Eat', 'SitDown' and 'SitUp', indicating that the HBRT and VOC were complemented each other, especially when a human interacted with other objects in the video scene (*e.g.*, a human and a chair, a human and a dining table).

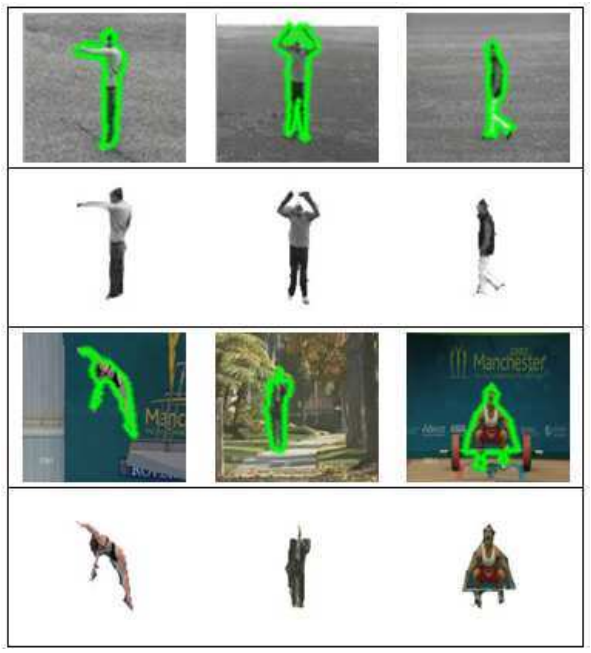

Figure 10: Samples for action localisation and segmentation. The 1st and 2nd rows respectively present the results of key segments and the corresponding segmentation using the HBRT/VOC on the KTH Dataset ('boxing', 'hand waving' and 'walking' actions). The 3rd and 4th rows show the results on the UCF Sports Actions Dataset ('diving', 'walking' and 'lifting' actions).

5.3. Discussion

The local region tracking schemes performed better than the point feature-based techniques for relatively small and well studied datasets such as the KTH and the UCF Sports Actions. Figure 10 presents several examples for action localisation and segmentation with

action class	Laptev	Bilen	Zhang	HBRT	HBRT/VOC
AnswerPhone	19.1	**21.9**	18.1	15.2	18.4
DriveCar	80.2	84.5	**88.1**	70.6	86.6
Eat	60.2	49.6	61.6	61.2	**72.4**
FightPerson	72.4	59.2	**76.2**	70.9	71.1
GetOutCar	25.6	24.0	**36.3**	18.2	28.7
HandShake	18.9	12.3	**55.9**	29.3	31.3
HugPerson	32.1	21.4	**48.3**	33.1	33.6
Kiss	47.8	49.3	**58.4**	50.3	52.3
Run	68.8	61.8	**72.1**	61.0	62.2
SitDown	49.2	40.9	51.9	50.9	**53.3**
SitUp	9.9	20.8	22.4	21.3	**23.5**
StandUp	49.0	**50.4**	21.6	50.2	50.3

Table 4: **(Hollywood2: Human Actions and Scenes Dataset)** Recognition accuracy for individual action classes. Units are in %. The best score for each class is highlighted by bold fonts. The numbers by Laptev *et al.* and by Bilen *et al.* were extracted from [28]. Zhang *et al.* is from [39].

these two datasets. Interestingly the accuracy by the HBRT/VOC improved over the HBRT even for datasets such as the KTH that did not contain any non-human objects. It was probably due to the person detector module of VOC, which contributed in successful human detection.

The local region tracking schemes showed its clear advantage when processing the complex and large dataset of Hollywood2, although the contribution of region tracking varied among action classes. It can be observed in Table 4 that, for 'FightPerson' and 'Run' actions, the space-time interest point features [15] performed better than the region based approaches. It was because the point features were able to provide more compact and abstract representation of video signals than the HBRT or the HBRT/VOC that relied on motion segmentation. The interest point features were useful when it was difficult to spatially localise the action using the region-based approaches.

The region tracking schemes clearly showed the state-of-the-art performance with some classes in the Hollywood2 dataset, in particular when the action could be fully identified using mainly human body regions interacting with some objects. The 'Eat' class from the Hollywood2 Dataset presented one such example, in which the regions of interest could be presented by multiple objects (*e.g.*, a human and a dining table).

6. Conclusion

The present study has put forward the human body region-based approach to action localisation, segmentation and recognition. The approach was further extended to accommodate non-human objects, resulting in the HBRT/VOC scheme. We showed that description of a human body volume with interacting objects regions using a spatio-temporal descriptor (HOG/HOF) generated stable representation for the appearance and motion patterns underpinning comprehension of the actions carried out. Three widely-used datasets were processed for evaluation and the region-based approach was able to outperform (or to perform at least as good as) the recent state-of-the-art, point feature-based techniques with all three datasets. It is hoped that this work stimulates further research on local region descriptors not only for action classification but also for many other video processing tasks.

Acknowledgements. The first author would like to thank Taibah University, Madinah, Saudi Arabia for funding this work as part of her PhD scholarship program.

References

[1] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Proceedings of CVPR, 2009.

[2] M. Sridhar, A. G. Cohn, D. C. Hogg, Unsupervised learning of event classes from video, in: Proceedings of AAAI, 2010, pp. 1631–1638.

[3] W. Brendel, A. Fern, S. Todorovic, Probabilistic event logic for interval-based event recognition, in: Proceedings of CVPR, 2011, pp. 3329–3336.

[4] B. Packer, K. Saenko, D. Koller, A combined pose, object, and feature model for action understanding, in: Proceedings of CVPR, 2012, pp. 1378–1385.

[5] N. Al Harbi, Y. Gotoh, Spatio-temporal human body segmentation from video stream, in: Computer Analysis of Images and Patterns, 2013, pp. 78–85.

[6] N. Al Harbi, Y. Gotoh, Action recognition: Spatio-temporal human body region tracking approach, in: Proceedings of the Second Workshop on Recognition and Action for Scene Understanding, CAIP, 2013.

[7] M. Maire, S. X. Yu, P. Perona, Object detection and segmentation from joint embedding of parts and pixels, in: Proceedings of ICCV, 2011.

[8] M. Al Ghamdi, N. Al Harbi, Y. Gotoh, Spatio-temporal video representation with locality-constrained linear coding, in: Computer Vision–ECCV 2012. Workshops and Demonstrations, 2012, pp. 101–110.
[9] N. Ikizler-Cinbis, S. Sclaroff, Object, scene and actions: Combining multiple features for human action recognition, in: Computer Vision–ECCV 2010, 2010, pp. 494–507.
[10] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.
[11] A. Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in: Proceedings of ICCV, 2009, pp. 925–931.
[12] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: Proceedings of CVPR, 2009, pp. 1996–2003.
[13] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Proceedings of ICCV, 2009, pp. 104–111.
[14] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: Proceedings of CVPR, 2009, pp. 2004–2011.
[15] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of CVPR, 2008, pp. 1–8.
[16] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.
[17] Y. Yacoob, M. J. Black, Parameterized modeling and recognition of activities, Computer Vision and Image Understanding 73 (2) (1999) 232–247.
[18] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 288–303.
[19] A. A. Efros, A. C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: Proceedings of ICCV, 2003.
[20] L. Zelnik-Manor, M. Irani, Statistical analysis of dynamic actions, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1530–1535.
[21] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1–8.
[22] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[23] J. Choi, W. J. Jeon, S. C. Lee, Spatio-temporal pyramid matching for sports videos, in: Proceedings of the ACM International Conference on Multimedia Information Retrieval, 2008, pp. 291–297.
[24] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of ICPR, Vol. 3, 2004, pp. 32–36.
[25] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space-time interest points, in: Proceedings of CVPR, 2009, pp. 1948–1955.
[26] D. Parikh, C. L. Zitnick, T. Chen, Unsupervised learning of hierarchical spatial structures in images, in: Proceedings of CVPR, 2009, pp. 2743–2750.
[27] T. Quack, V. Ferrari, B. Leibe, L. Van Gool, Efficient mining of frequent and distinctive feature configurations, in: Proceedings of ICCV, 2007, pp. 1–8.
[28] H. Bilen, V. P. Namboodiri, L. Van Gool, Action recognition: A region based approach, in: Proceedings of IEEE Workshop on Applications of Computer Vision, 2011, pp. 294–300.
[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.
[30] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision (2004) 91–110.
[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of CVPR, 2010, pp. 3360–3367.
[32] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, C. J. Lin, LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research 9 (2008) 1871–1874.
[33] M. D. Rodriguez, J. Ahmed, M. Shah, Action mach: a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of CVPR, 2008.
[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of BMVC, 2009.
[35] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proceedings of CVPR, 2011, pp. 3361–3368.
[36] S. Sadanand, J. J. Corso, Action bank: A high-level representation of activity in video, in: Proceedings of CVPR, 2012, pp. 1234–1241.
[37] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: Proceedings of CVPR, 2010, pp. 2046–2053.
[38] X. Wu, D. Xu, L. Duan, J. Luo, Action recognition using context and appearance distribution features, in: Proceedings of CVPR, 2011, pp. 489–496.
[39] H. Zhang, W. Zhou, C. Reardon, L. E. Parker, Simplex-based 3d spatio-temporal feature description for action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp. 2067–2074.