Relative Dense Tracklets for Human Action Recognition

Piotr Bilinski Etienne Corvee Slawomir Bak Francois Bremond

INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, BP93, 06902 Sophia Antipolis, France
{Piotr.Bilinski,Etienne.Corvee,Slawomir.Bak,Francois.Bremond}@inria.fr

Abstract— **This paper addresses the problem of recognizing human actions in video sequences for home care applications. Recent studies have shown that approaches which use a bag-of-words representation reach high action recognition accuracy. Unfortunately, these approaches have problems to discriminate similar actions, ignoring spatial information of features. As we focus on recognizing subtle differences in behaviour of patients, we propose a novel method which significantly enhances the discriminative properties of the bag-of-words technique. Our approach is based on a dynamic coordinate system, which introduces spatial information to the bag-of-words model, by computing relative tracklets. We perform an extensive evaluation of our approach on three datasets: popular KTH dataset, challenging ADL dataset and our collected Hospital dataset. Experiments show that our representation enhances the discriminative power of features and bag-of-words model, bringing significant improvements in action recognition performance.**

I. Introduction

Automatic recognition of human actions has gained tremendous interest in recent years. Video surveillance, video data indexing, video retrieving, human-computer interaction or sport event analysis are just a few of many applications, in which action recognition plays the main role. Recently, recognition of human behaviour is also becoming more and more important and frequently used in medicine, especially for purposes of health care monitoring applications. In particular, more and more emphasis is placed on systems that allow for early detection of upcoming or existing physical and health disorders. Identifying changes in human everyday behaviour such as food preparation, walking, housekeeping, exercise or sleeping, allows medical scientists to propose strategies related to diet, exercise and medication adherence. This is important for elderly people, for which such systems allow them to have a longer, healthier and safer life within the comforts of home.

In this paper, we focus on assisting elderly people, proposing an approach for recognizing subtle differences in behaviour of analysed humans for purposes of health care monitoring applications. Recent studies have shown, that approaches which use a bag-of-words representation, reach high action recognition accuracy [16], [19], [24]. Unfortunately, these approaches have problems to discriminate similar actions, ignoring spatial information of features. A common way to overcome these limitations is to use either spatio-temporal grids [16] or multi-scale pyramids [17]. However, these methods are still limited in terms of a detailed description providing only a coarse representation. To differ from these ideas, we propose novel descriptors based on a dynamic coordinate system. Our suitable design descriptors introduce important spatial information to the bag-of-words model enhancing its discriminative power. Our main idea is that action recognition ought to be performed using a dynamic coordinate system corresponding to an object of interest. Computing relative tracklets, we are able to keep spatial information in a bag-of-words framework. We propose to use a head position as a center of our dynamic coordinate system, providing description invariant to changes in camera viewpoint. Our novel descriptors improve the discriminative power of features and help to distinguish similar features detected at different positions (e.g. to distinguish similar features appearing on hands and feet). We perform an extensive evaluation on three datasets: popular KTH dataset, challenging ADL dataset and our locally collected Hospital dataset. Consistently, performed experiments show that our representation enhances the discriminative power of features and bag-of-words model, bringing significant improvements in action recognition performance.

The main contributions of this paper are summarized as follows:

- We offer a novel action recognition approach based on a dynamic coordinate system. We propose to compute relative tracklets, introducing spatial information to the bag-of-words model. The tracklets computation is based on their relative positions according to the central point of our dynamic coordinate system. As this central point, we choose the center of a head to provide camera invariant description.
- We report experimental results on three action recognition datasets (KTH, ADL and our collected Hospital dataset), showing that our representation enhances the discriminative power of features and bag-of-words model, bringing significant improvements in action recognition performance.

The rest of the paper is organized as follows. Section II looks at the most relevant state-of-the-art work in the literature. In section III, we present our novel action recognition approach. In section IV, we present the obtained results from the conducted experiments. Finally, in section V, we conclude with future directions of work.

II. Related Work

Over the last few years, many different action recognition techniques have been proposed. However, due to appearance variations of both people and actions, camera view point changes, occlusions, noise, and enormous amount of video data, action recognition still remains a challenging problem.

Existing techniques can be divided into four categories. The first group of techniques uses silhouette or body contour information to represent an action [1], [12], [18]. Such techniques usually require precise segmentation, which is often difficult to achieve, especially in real-world videos. The second category of techniques uses local spatio-temporal features [7], [13], [15], [20], [26], [29]. The local spatio-temporal features are able to capture both visual and motion appearance. They are robust to viewpoint and scale changes, they are easy to implement and fast to process. Moreover, they do not require object localization and in addition they are robust to background clutter. Over the last few years, many different local interest point detectors (like Harris3D [15], Cuboid [5], Hessian [34] or Dense sampling [33]) and many spatio-temporal descriptors (like HOG [16], HOG3D [13], HOF [16], Cuboid [5] or ESURF [34]) have been proposed. One of the most commonly used descriptors in the literature showing a high performance over the various datasets [3], [33] are: Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) descriptors [16]. The former describes the local visual appearance and the latter characterizes the local motion appearance of an interest point. The third category contains methods analysing motion trajectories [9], [23], [28], [32]. This group of techniques usually requires tracking of feature points or objects [6], [21], [27], [35]. Recent techniques, based on feature tracking, have shown high action recognition rate, especially when combining trajectories with local spatio-temporal features. Becha *et al.* [9] have proposed to track corner points using HOG tracker, and then represent feature trajectories by angle descriptors. Raptis *et al.* [28] have proposed spatio-temporal feature descriptors (named average of gradient orientation and average of optical flow) that capture the local structure of an image around trajectories tracked over time. Messing *et al.* [23] have proposed to track Harris feature points using KLT tracker, and then represent trajectories by temporal velocity histories. Recently, especially dense trajectories have drawn a lot of attention and have shown to obtain high performance for action recognition in videos. Wang *et al.* [32] have proposed to use dense short trajectories together with HOG, HOF and MBH (Motion Binary Histograms) features. Wu *et al.* [35] have proposed to use Langrangian particle trajectories which are dense trajectories obtained by advecting optical flow over time. Raptis *et al.* [27] have proposed to extract salient spatio-temporal structures by forming clusters of dense optical flow trajectories. Then, the assembly of these clusters into an action class is governed by a graphical model.

Most of the recent techniques, based on local spatio-temporal features and trajectories, use the bag-of-words model. The bag-of-words model have shown to achieve high recognition rate across various datasets [16], [19], [24]. It simplifies the structure of 3D video data assuming conditional independence across spatial and temporal domains. It encodes global statistics of features computing histogram of feature occurrences in a video. However, the bag-of-words model has limitations. The main drawback of this technique is that it ignores important spatial position of features. A common way to overcome this limitation is to use either spatio-temporal grids [16] or multi-scale pyramids [17]. However, these methods are still limited in terms of a detailed description providing only a coarse representation.

In contrary, we design a novel approach based on short relative dense tracklets, local spatio-temporal features and bag-of-words model. We propose novel descriptors based on a dynamic coordinate system, which introduce important spatial information to the bag-of-words model. Our novel descriptors improve the discriminative power of features and bag-of-words model, and help to distinguish similar features detected at different positions (e.g. to distinguish similar features appearing on hands and feet).

III. Our Approach

The bag-of-words model has achieved high action recognition rate across various datasets. It encodes global statistics of features ignoring important information on their spatial position. A common way to overcome this limitation is to use either spatio-temporal grids or multi-scale pyramids. Unfortunately, these methods are still limited in a detailed description providing only a course representation. Instead, we propose relative tracklets introducing spatial information of features to the bag-of-words approach. We focus on home care applications thus human head can be used as a reference point for computing our relative trajectories. Using the head estimation framework, we create a dynamic coordinate system, which allows us to compute our relative tracklets.

A. Dense Multi-Scale Tracklet Extraction

The amount of data retrieved from a video content usually depends on the action-video parameters such as: a length of the action taking place and a video resolution. As certain daily living actions like walking or sitting could only last a few seconds, information provided by commonly used tracking algorithms such as KLT and SIFT might not be enough for recognizing these actions. Similarly to [32], we cope with this problem by employing dense tracklets extracted on multiple spatial scales. For each scale, we sample feature points on a grid with a step size of W pixels and track densely sampled feature points using optical flow and median filtering. Using dense tracklets, we are able to distinguish similar and short actions. Moreover, limiting the length of tracklets to L frames, we avoid a drifting problem and enhance the discriminative properties of tracklets. As tracklets themselves do not contain spatial-temporal information, we propose to introduce relative positions of trajectories, computed using a dynamic coordinates system.

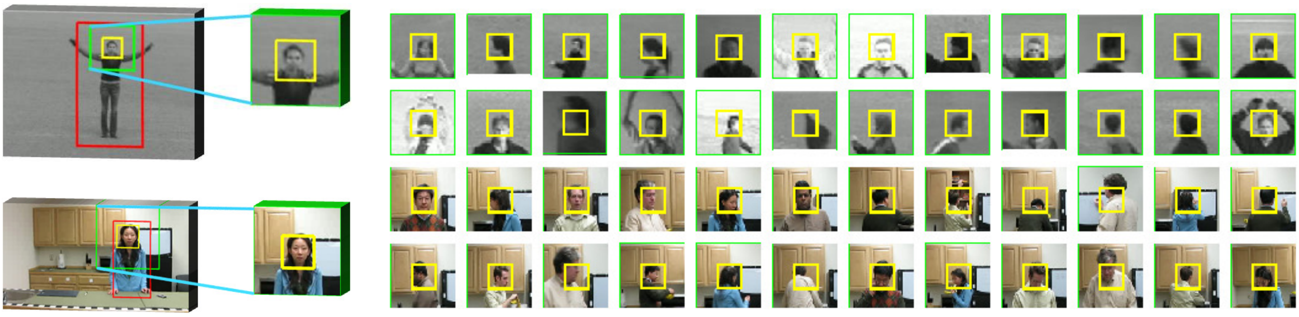

Fig. 1: Samples of estimated head positions for the KTH (first row) and ADL (second row) datasets.

The central point of this system is selected using a head position, which is computed by applying our robust head detection algorithm.

B. Head Detection

Head detection is of particular interest in our recognition framework, thus we have to ensure robust localization of this body part. As head is a common pattern, we needed to combine several techniques for estimating head position. Cues provided by object detectors (people, head, and face) are combined with motion information (background subtraction) and object tracking results. We employ: (1) histogram of oriented gradients for people detection [4], (2) LBP patterns for head detection [25], and (3) haar-like features for face detection [4]. Refining object detection results by background subtraction step, we produce preliminary detection-based tracking results. Then, detection results are used as an input to the tracking algorithm (applied for forward and backward tracking in a video) [10] to overcome missed detections. Both detection-based tracking and TLD tracking algorithm [10] provide multiple hypothesis along which we select the most likely one. This selection is based on our probability framework $\mathcal{P}$, which smooths trajectories by replacing rapid object displacements with the interpolated results. The final position of the head is obtained by maximizing the trajectory-depended probability:

$$\mathcal{P}(l_h|t_{h,i}) = \sum_{z\in\{f,b\}} \mathcal{P}(l_z \mid t_{h,i} \propto t_{z,j})\ \mathcal{P}(t_{h,i} \propto t_{z,j}), \quad (1)$$

where l_h is a head location and $t_{h,i}$ is their corresponding trajectory. $\propto$ describes proportional variance of trajectory $t_{h,i}$ w.r.t. trajectory of other body part $t_{z,j}$. f and b refer to the face detection and the full body detection, respectively. Sample head positions estimated by this method are shown in Figure 1.

C. Combined Multi-Scale Tracklet (CMST) Descriptors

In this section, we introduce our novel descriptor, which contains both shape characteristics of a tracklet and relative positions of tracklet's elements according to the central point of the dynamic coordinate system. We focus on home care applications, thus we use head as a reference point to provide camera invariant description. Our novel descriptors improve the discriminative power of features and bag-of-words model, and help to distinguish similar features detected at different positions (e.g. to distinguish similar features appearing on hands and feet).

We define our CMST descriptor as:

$$\Phi = \left[(\vartheta_X)^T \quad (\vartheta_Y)^T \quad (X_{TR}-X_H)^T \quad (Y_{TR}-Y_H)^T\right]^T \quad (2)$$

where the first two elements of the vector $((\vartheta_X)^T, (\vartheta_Y)^T)$ are refereed to as Shape Multi-Scale Tracklet (SMST) descriptor, which represents shape characteristics of a tracklet based on the displacement vector descriptor [32], and the two remaining parts correspond to our Relative Multi-Scale Tracklet (RMST) descriptor.

1) Shape Multi-Scale Tracklet (SMST) Descriptor: Encoding a local motion pattern of a given tracklet $t_i = \{(x_1, y_1), ..., (x_L, y_L)\}$ of length L, we compute displacement vectors θ_X and θ_Y:

$$\theta_X = \Delta(X - \overline{X}) \quad (3)$$

$$\theta_Y = \Delta(Y - \overline{Y}) \quad (4)$$

where $X = [x_1, x_2, ..., x_L]^T$ and $Y = [y_1, y_2, ..., y_L]^T$. Symbols $\overline{X}$ and $\overline{Y}$ represent mean of the vector X and Y, respectively. Then, we normalize displacement vectors by the sum of their magnitudes:

$$\vartheta_X = \frac{\theta_X}{\sum_{i=1}^{L}\sqrt{\theta_{X_i}^2 + \theta_{Y_i}^2}} \quad (5)$$

$$\vartheta_Y = \frac{\theta_Y}{\sum_{i=1}^{L}\sqrt{\theta_{X_i}^2 + \theta_{Y_i}^2}} \quad (6)$$

where θ_{X_i} and θ_{Y_i} represent i-th elements of the vector X and Y, respectively. Finally, we obtain our shape characteristic tracklet representation by defining a vector ψ, which is the result of the concatenation of the vector ϑ_X and ϑ_Y:

$$\psi = \left[(\vartheta_X)^T \quad (\vartheta_Y)^T\right]^T \quad (7)$$

2) Relative Multi-Scale Tracklet (RMST) Descriptor: Encoding a local motion pattern of a given tracklet $t_i = [(x_j, y_j), ..., (x_{j+L}, y_{j+L})]$ (where L is the length of the tracklet and j is the frame number, where the tracklet occurred for the first time) with respect to the head trajectory

$t_h = [(x'_k, y'_k), ..., (x'_m, y'_m)]$ (where $k \leq j$ and $j + L \leq m$), we define the RMST descriptor by:

$$\phi = \left[(X_{t_i} - X_{t_h})^T \quad (Y_{t_i} - Y_{t_h})^T\right]^T \tag{8}$$

where $X_{t_i} = [x_j, ..., x_{j+L}]^T$, $Y_{t_i} = [y_j, ..., y_{j+L}]^T$, $X_{t_h} = [x'_j, ..., x'_{j+L}]^T$, and $Y_{t_h} = [y'_j, ..., y'_{j+L}]^T$.

Our CMST descriptors introduce the relative positions of features to the bag-of-words approach. Our novel descriptors improve the discriminative power of features and bag-of-words model, and help to distinguish similar features detected at different positions. Fusing the discriminative power of both SMST and RMST descriptors, we significantly improve action recognition accuracy. Our final descriptor (CMST) allows classifier to recognize an action even in the case when the estimation of the head is not perfect or head detection is missing. This is obtained by the discriminative power of the SMST descriptor.

D. Action Recognition using CMST features

Additionally, to increase the discriminative power of tracklets, we compute the HOG (Histogram of Oriented Gradients) and HOF (Histogram of Oriented Flow) features along space-time neighbourhood of each tracklet [32]. The former feature describes the local visual appearance and the latter characterizes the local motion appearance of a tracklet.

The tracking algorithm, used to compute the SMST and HOG-HOF descriptors, was selected based on its use in the literature, and provide a good baseline for comparison with the state-of-the-art techniques. However, our action representation method can be also used together with any other tracking algorithm.

To represent videos, we apply the bag-of-words model for each feature class (SMST-RMST, HOG-HOF) independently. We construct visual vocabularies from training videos clustering computed features. Then, we assign each feature to its closest visual world. The concatenated histograms of visual word occurrences over video forms the final representation.

To classify a new video sequence, we use multi-class non-linear Support Vector Machines (SVM). We apply a χ^2 distance to compare two n-bins histograms $H_i = [H_i(1), ..., H_i(n)]^T$ and $H_j = [H_j(1), ..., H_j(n)]^T$:

$$\chi^2(H_i, H_j) = \frac{1}{2}\sum_{k=1}^{n}\left(\frac{(H_i(k) - H_j(k))^2}{H_i(k) + H_j(k)}\right) \tag{9}$$

This distance is then converted into SVM multi-channel χ^2 kernel using a multi-channel generalized Gaussian kernel:

$$K(H_i, H_j) = exp(-\frac{1}{A}\chi^2(H_i, H_j)) \tag{10}$$

where the parameter A is the width of the kernel.

IV. Experiments

We perform an extensive set of experiments on multiple datasets to demonstrate the effectiveness of the proposed descriptors. We evaluate our approach on three datasets for human action recognition: popular KTH dataset, challenging ADL dataset and locally collected Hospital dataset. Sample images from video sequences of these datasets are presented in Figure 2.

A. Implementation Details

We compute HOG and HOF descriptors on a spatio-temporal grid of size $n_x \times n_y \times n_t$, where: $n_x = 2$, $n_y = 2$ and $n_t = 3$. For each individual cell of the grid, we compute a 8-bins histogram of orientation for the HOG and 9-bins histogram for the HOF. We normalize both descriptors with the L_2 norm.

During the quantization process of calculated features, we use the k-means clustering technique and nearest neighbour algorithm. To compute the bag-of-words representation, features are quantized to the codebook size of 1000, which has shown empirically to give good results. As a metric to calculate a distance between features and visual words, we use the L_2 norm.

In all our experiments, we apply the cross-validation technique to both gauge the generalizability of the proposed approach, and select the most discriminative parameters. We use the Leave-One-Out Cross-Validation (LOOCV) technique, where videos of one person are used as the validation data, and the remaining videos as the training data. This is done repeatedly so that videos of each person are used once as the validation data.

B. KTH Dataset

The KTH dataset [31] does not contain real home care videos. However, we have decided to evaluate our approach on it due to its popularity and possibility to compare our approach with most of the state-of-the-art techniques.

The KTH dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action is performed several times by 25 different subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). In total, the dataset contains 599 video files. All sequences were recorded with 25 fps frame rate.

The dataset contains a set of challenges like: scale changes, illumination variations, shadows, different scenarios, cloth variations, inter and intra action class speed variations and low resolution (160 × 120 pixels spatial resolution).

There are two commonly used experimental setups to evaluate an approach on the KTH dataset: splitting-based scheme and LOOCV technique. Therefore, to compare our approach with all of them, we evaluate our approach on both experimental setups.

1) LOOCV evaluation scheme: We follow the recent evaluations [8], [35], [36], [38] on the KTH dataset using LOOCV scheme. The experimental results are presented in Table I. Comparison of our approach with state-of-the-art methods using LOOCV technique is presented in Table II. The detailed comparison for each scenario separately is presented in Table III. We observe that overall and for

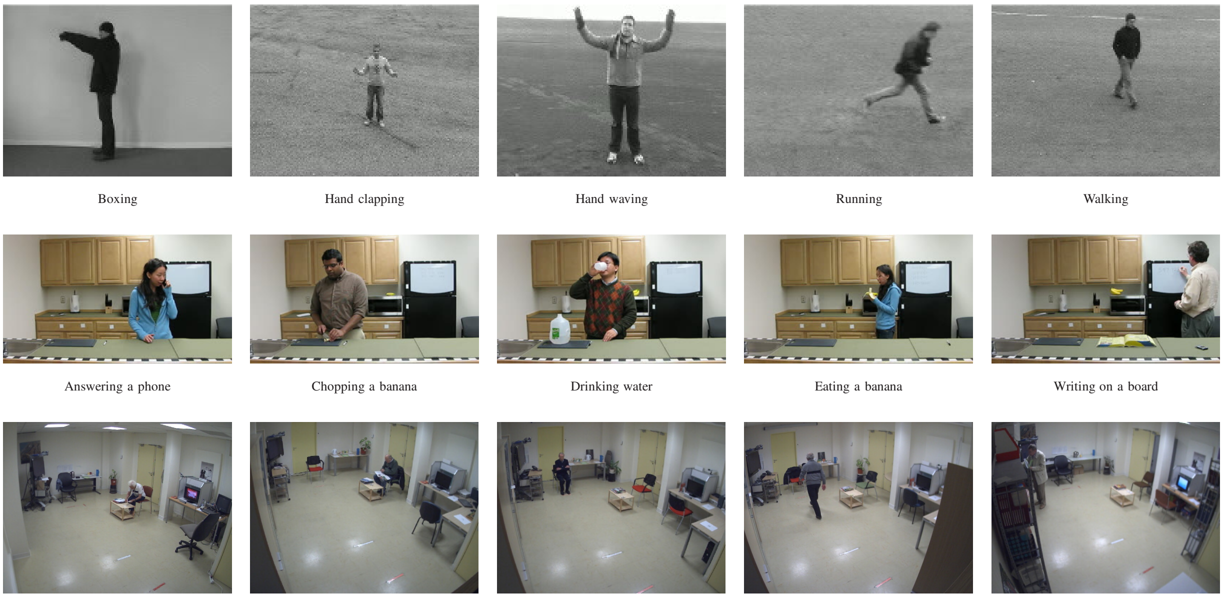
Boxing
Hand clapping
Hand waving
Running
Walking
Answering a phone
Chopping a banana
Drinking water
Eating a banana
Writing on a board
Playing cards
Reading
Sitting-standing
Up and go test
ABCD matching

Fig. 2: Sample frames from video sequences of the KTH (first row), ADL (second row) and Hospital (third row) datasets.

TABLE I: KTH dataset: Evaluation of SMST, RMST and CMST descriptors.

		recognition rate (%)				
	method	s1	s2	s3	s4	overall
Official Split	SMST	98.15%	88.89%	88.89%	90.74%	91.67%
	RMST	96.30%	88.89%	87.04%	92.60%	91.21%
	CMST	**98.15%**	**92.59%**	**92.59%**	**96.30%**	**94.91%**
LOOCV	SMST	98.00%	92.00%	93.29%	94.67%	94.49%
	RMST	98.67%	89.33%	95.30%	96.67%	94.99%
	CMST	**99.33%**	**93.33%**	**97.32%**	**98.67%**	**97.16%**

each scenario independently, our CMST descriptors outperform SMST descriptors, achieving 97.16%, 99.33%, 93.33%, 97.32% and 98.67%, respectively. We also observe that overall and for scenarios $s1$, $s3$ and $s4$, our RMST descriptors outperform SMST descriptors. Although continuous scale changes in scenario $s2$ cause time to time inaccurate and missing head estimations (what results in slightly lower accuracy of RMST descriptors compared to SMST features), our CMST descriptors still improve action recognition accuracy and outperform both SMST and RMST features. The results clearly show that our representation enhances the discriminative power of features and bag-of-words model, and outperforms state-of-the-art techniques. HOG-HOF features do not improve action recognition accuracy on this dataset (the accuracy 97.16% is already very high).

2) Splitting-based evaluation scheme: We also follow the original experimental setup, where video samples are divided into two parts: the training set and testing set. The testing set consists of 9 subjects (2, 3, 5, 6, 7, 8, 9, 10 and 22) and the training set consists of 16 remaining subjects.

The results from the experiments are presented in Table I. Comparison of our approach with state-of-the-art methods in the literature, using splitting-based evaluation scheme, is presented in Table II. Overall, our approach obtains 94.91% recognition rate. Also in this case, we observe that the CMST descriptors improve action recognition accuracy both overall and for each scenario independently. The results clearly show that our representation enhances the discriminative power of features and bag-of-words model, and outperforms state-of-the-art techniques.

C. ADL Dataset

The ADL (University of Rochester Activities of Daily Living) dataset [23] contains ten types of human activities of daily living, selected to be useful for an assisted cognition task. The full list of activities is: answering a phone, dialling a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. Each action is performed three times by five different people. In total, the dataset contains 150 video sequences recorded with 30 fps frame rate and 1280 × 720 pixels spatial resolution. The videos were down-sampled to the 640 × 360 pixels spatial resolution.

The dataset contains a set of challenges like: different

TABLE II: KTH dataset: Comparison of our approach with state-of-the-art methods in the literature using both official splitting-based evaluation scheme and LOOCV technique.

Official Split			LOOCV		
Method	Year	Accuracy (%)	Method	Year	Accuracy (%)
Laptev *et al.* [16]	2008	91.8%	Liu *et al.* [19]	2009	93.8%
Yuan *et al.* [37]	2009	93.3%	Ryoo *et al.* [29]	2009	93.8%
Zhang *et al.* [38]	2012	94.1%	Wu *et al.* [36]	2011	94.5%
Wang *et al.* [32]	2011	94.2%	Kim *et al.* [11]	2007	95.33%
Gilbert *et al.* [7]	2011	94.5%	Zhang *et al.* [38]	2012	95.5%
Kovashka *et al.* [14]	2010	94.53%	Wu *et al.* [35]	2011	95.7%
Becha *et al.* [9]	2012	94.67%	Lin *et al.* [8]	2011	95.77%
Our method		**94.91%**	**Our method**		**97.16%**

TABLE III: KTH dataset: Comparison of our approach with state-of-the-art methods in the literature for each scenario separately using LOOCV technique.

	recognition rate (%)				
method	s1	s2	s3	s4	avg.
Wu *et al.* [36]	96.7%	91.3%	93.3%	96.7%	94.5%
Lin *et al.* [8]	98.83%	94.00%	94.78%	95.48%	95.77%
Our method	**99.33%**	**93.33%**	**97.32%**	**98.67%**	**97.16%**

TABLE IV: ADL dataset: Evaluation of SMST, RMST, CMST, and CMST with HOG-HOF descriptors using LOOCV technique.

Method	Recognition Rate (%)
SMST	76.67%
RMST	78.67%
CMST	**88.00%**
CMST + HOG-HOF	**92.00%**

shapes, sizes, genders and ethnicities of people, and difficulty to separate activities on the basis of a single source of information (e.g. eating banana and eating snack or answering a phone and dialling a phone).

Results from the experiments are presented in Table IV. Comparison of our approach with state-of-the-art methods in the literature using LOOCV technique is presented in Table V. We observe that SMST descriptors overall achieves 76.67% and our RMST descriptors improve recognition rate up to 78.67%. Moreover, our CMST descriptors improve action recognition rate up to 88.0%, which means that our descriptors improve accuracy by 11.33%. We also fuse HOG-HOF features with our CMST descriptors and achieve 92.00% recognition rate, which means that our approach improves the accuracy by 15.33% compared to SMST descriptors. All these results clearly show that our representation enhances the discriminative power of features and bag-of-words model, bringing significant improvements in action recognition performance.

TABLE V: ADL dataset: Comparison of our approach with state-of-the-art methods in the literature using LOOCV technique.

Method	Year	Recognition Rate (%)
Matikainen *et al.* [22]	2010	70%
Satkin *et al.* [30]	2010	80%
Banabbas *et al.* [2]	2010	81%
Raptis *et al.* [28]	2010	82.67%
Messing *et al.* [23]	2009	89%
Our method		**92.00%**

D. Hospital Dataset

Most of the existing public action recognition datasets can be divided into a few categories: (a) low resolution videos of relatively simple actions (like Weizmann and KTH datasets) which do not include object interactions, (b) video sequences from broadcast television channels, YouTube, and personal cameras (like UCF Sports, YouTube, and UCF50 datasets) where often a person is not fully visible, videos are recorded in a significant distance from people, videos are often pixelated, blurred, and contain significant camera motion and background clutter, (c) video samples from movies (like Hollywood and Hollywood2 datasets) where often only parts of people and actions are visible, and camera view point is constantly moving, and (d) videos of activities of daily living (like ADL dataset) where the camera is set in front of the actor and background does not significantly change between videos. Therefore, a new dataset is needed for recognition of realistic human activities of daily living.

We have locally collected dataset, created with the help of medical scientists. The new dataset contains 8 types of real human activities of daily living. The full list of activities is: (a) playing cards, (b) matching ABCD sheets of paper, (c) reading, (d) sitting down and standing up, (e) turning back, (f) standing up and moving ahead, and (g) walking back and forth (2 activities). These activities were selected and annotated by medical doctors.

The experiments have been approved by the national official committee, the Committee for the Protection of Patients in Biomedical Research. Once people have been selected and have agreed (with their relatives) to participate in the studies, videos were recorded during regular consultations of patients at hospital. The videos were recorded over a period of several months, for every recording slight changes were made to the positioning of the camera and objects in the room. As a result, we have obtained a dataset of 55 patients recorded at 640×480 pixels spatial resolution.

Our dataset contains a set of challenges like: different shapes, sizes, genders and ethnicities of people, occlusions, and multiple people (sometimes both patient and doctor are visible).

Our proposed CMST descriptors combined with HOG-HOF features achieve high action recognition rate (92.96% accuracy) improving the recognition rate by 6.67% compared to the SMST descriptors. Experiments on this dataset again confirm the effectiveness of our method and show that our representation enhances the discriminative power of features and bag-of-words model, bringing significant improvements in action recognition performance.

V. Conclusion and Future Work

We proposed a novel action recognition approach based on a dynamic coordinate system. Our approach employs head detection for computing relative tracklets. These relative tracklets enhance the discriminative power of features and bag-of-words model introducing important information on their spatial position. The proposed approach was evaluated on three benchmark datasets for human action recognition. Obtained results clearly show that our approach improves action recognition performance and outperforms existing state-of-the-art techniques. In future work, we intend to examine more efficient learning algorithms (like Multiple Kernel Learning) to combine features from the bag-of-words model. We also intend to examine different human body parts as reference points for computing our relative tracklets.

Acknowledgements

This work was supported by the Région Provence-Alpes-Côte d'Azur. However, the views and opinions expressed herein do not necessarily reflect those of the financing institution.

References

[1] M. Ahad, J. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *MVA*, 2010.

[2] Y. Benabbas, A. Lablack, N. Ihaddadene, and C. Djeraba. Action recognition using direction models of motion. In *ICPR*, 2010.

[3] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *ICVS*, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS workshop, in conjunction with ICCV*, 2005.

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[7] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *TPAMI*, 2011.

[8] Z. Jiang, Z. Lin, and L. S. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *TPAMI*, 2011.

[9] M. Kaaniche and F. Bremond. Recognizing gestures by learning local motion signatures of hog descriptors. *TPAMI*, 2012.

[10] Z. Kalal, J. Matas, and K. Mikolajczyk. P-n learning: Bootstrapping binary classifiers by structural constraints. In *CVPR*, 2010.

[11] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, 2007.

[12] T.-S. Kim and Z. Uddin. *Silhouette-based Human Activity Recognition Using Independent Component Analysis, Linear Discriminant Analysis and Hidden Markov Model*. InTech, 2010.

[13] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[14] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[15] I. Laptev. On space-time interest points. *IJCV*, 2005.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009.

[19] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[20] J. Liu and M. Shah. Learning human action via information maximization. In *CVPR*, 2008.

[21] W.-L. Lu, K. Okuma, and J. J. Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *IVC*, 2009.

[22] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *ECCV*, 2010.

[23] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[24] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[25] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002.

[26] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR*, 2009.

[27] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012.

[28] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *ECCV*, 2010.

[29] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[30] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010.

[31] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[32] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.

[33] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[34] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[35] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, 2011.

[36] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.

[37] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009.

[38] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *ECCV*, 2012.