Multi-Type Activity Recognition from a Robot's Viewpoint

Ilaria Gori[1], **J. K. Aggarwal**[1], **Larry Matthies**[2] and **Michael. S. Ryoo**[3]
[1]Computer and Vision Research Center, The University of Texas at Austin, Austin, TX, USA
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[3]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
ilary.gori@gmail.com, aggarwaljk@utexas.edu, lhm@jpl.nasa.gov, mryoo@indiana.edu

Abstract

The literature in computer vision is rich of works where different *types* of activities – single actions, two persons interactions or ego-centric activities, to name a few – have been analyzed. However, traditional methods treat such types of activities separately, while in real settings detecting and recognizing different types of activities simultaneously is necessary. We first design a new *unified* descriptor, called Relation History Image (RHI), which can be extracted from all the activity types we are interested in. We then formulate an optimization procedure to detect and recognize activities of different types. We assess our approach on a new dataset recorded from a robot-centric perspective as well as on publicly available datasets, and evaluate its quality compared to multiple baselines.

1 Introduction

Activities can be categorized based on how many people participate (one person, two persons, a group of people), and the point of view from which they are recorded (third person, ego-centric). The combination of these two characteristics generates 6 possible *types* of activities. Traditional computer vision methods address the recognition of only one *type* of activity at a time. In such cases, videos to be classified belong to a specific *type* of activities. However, in an environment where for example a service robot provides directions and recommendations in public places, people who surround the robot usually perform several different types of activities, and the robot is required to recognize them all. This paper summarizes our previous work [Gori *et al.*, 2016] on classifying complex scenes where multiple people perform activities of different types concurrently or in a sequence. We provide a new video dataset recorded from a robot-perspective where multiple activities and interactions happen at the same time, and we present a novel approach to appropriately classify such videos.

In activity recognition tasks, videos are usually encoded based on one or more *visual descriptors*, which are a collection of variables often fed as *features* to a classifier. Designing (or learning) good descriptors is very important, as robust descriptors lead to better accuracy of the entire system. In particular, if we want to discriminate different types of activities, a *unified* descriptor that can be extracted from activity videos regardless of their type is necessary. This paper presents a new unified descriptor, called *Relation History Image* (RHI) (Fig. 1), which is built as the variation over time of relational information between pairs of local joints/patches, belonging to one or a pair of subjects. The fact that it can be extracted from different types of activities while maintaining the same format and dimensionality, allows to compare videos directly. This direct comparison enables the system to detect ongoing activities even in the scenario where multiple types of activities are present in the same video.

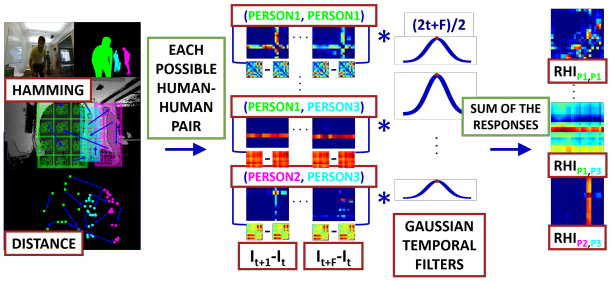

Figure 1: This picture represents how RHIs on skeleton and depth are extracted from a single person (e.g., person1) and pairs of persons (e.g., {person1, person3} and {person2, person3}).

The main peculiarity of multi-type activity videos is that it is impossible to know a priori who is performing what types of activities. In traditional videos with single person actions, it is assumed that each person is doing something singularly. In traditional videos of two-person interactions, it is expected that there are two persons and that they are interacting. In our case, if there are three persons in the scene, there is no way for the system of knowing if two of those are interacting, or if some of them are interacting with the robot. To solve this problem, we propose a new method based on optimization.

Our main contribution is the idea of recognizing concurrent activities belonging to different *types*, which is crucial in real-world scenarios. Our technical contributions are: 1) a new *unified* descriptor, RHI, which can be extracted from multiple types of activities and thus simplifies the classification step; 2) a new method to identify what persons are involved in what activity; 3) a systematic evaluation to compare the proposed

approach to several baselines on a newly collected dataset as well as on public ones. The full version of this work can be found in [Gori *et al.*, 2016].

2 Related Work

Most of the approaches in the literature tackle the problems of third-person activity recognition [Wang *et al.*, 2012b], two persons interactions [Ryoo and Aggarwal, 2009; Kong *et al.*, 2014] and group activities [Khamis *et al.*, 2012]. In the last few years, also first-person activities [Lu and Grauman, 2013] and first-person interactions [Ryoo and Matthies, 2013] have been analyzed. However, previous works deal with only one type of activities at a time. In contrast, our method explicitly addresses situations where the goal is recognizing different types of activities performed concurrently.

Fathi et al. [Fathi *et al.*, 2012] analyzed the problem of detecting who is interacting with whom: given a set of persons, their goal is to classify monologues, dialogues, and discussions. Nonetheless, it is assumed that only one social interaction happens in each scene. Differently, we consider cases where there are multiple activities performed at the same time.

Lan et al. [Lan *et al.*, 2012] analyze videos with multiple people and formulate a hierarchical classifier to recognize single activities at a low level, while inferring their social roles at an intermediate level. However, they only focused on multi-person group activities and simple actions composing those.

3 Method

We introduce a new *unified* feature descriptor to handle multiple different types of human activities: Relation History Image (RHI). Our descriptor is unified as it is able to describe different types of human activities while maintaining the same format and dimension. In this paper, we focus on interactions, single person activities and ego-centric interactions. If we had three distinct classification systems (one for interactions, one for single person actions and one for ego-centric ones), it would be difficult to infer whether an interaction between two persons p1 and p2 is more likely than person p1 and p2 each performing individual actions or activities targeting the robot. Instead, thanks to the unified descriptor, we can train all the activities within the same classification framework and directly select the most likely activity.

Our approach takes a set of unlabeled RGB-D frames, where multiple people perform activities of different types, and returns a set of activities along with who is involved in which. We consider a pair formed by the same person selected twice as an acceptable pair to enable single-person action representations. Relation History Images (RHIs) from **each possible pair of appearing persons** is extracted over a sliding window of frames. During training, we know who are the persons that are interacting, who is alone and who is interacting with the robot; the related descriptors model *valid* actions. All the other descriptors extracted from pairs that are not modeling a real activity (non-*valid* actions) are associated to the 'null' class. A one-vs-one SVM is trained based on all the RHIs, including those assigned to the 'null' class, regardless of their types. When testing, we do not use any annotation on the pairs who are interacting and the pairs who are not. In order to simultaneously recognize the activities and identify the persons performing them, we solve a linear optimization problem, exploiting the SVM responses for each pair in the scene.

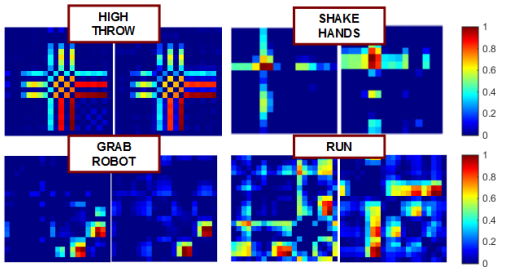

Figure 2: Examples of the RHI descriptors on human joint locations extracted from multiple datasets. Notably, RHIs extracted from single persons are symmetric.

3.1 Relation History Image

Given a set of m local regions $(r_1,...,r_m)$ identified on a person, we extract a local feature vector from each of them $(x_1,...,x_m)$. For example, if the local region corresponds to a certain part of the person's body, the local feature vector can be the position of that body part with respect to the camera. We then build a $m \times m$ matrix $\mathbf{R_t}$ for each frame t, for each pair in the scene. Each element of $\mathbf{R_t}$ is equal to $R_t^{i,j} = K(x_t^i, x_t^j)$, where $K(\cdot,\cdot)$ is a function that measures the relation between the low-level descriptors extracted from the i-th and the j-th local regions at time t. Notably, $\mathbf{R_t}$ can describe the relation between local regions on two different persons, as well as the relation between local regions on the same person. In the above example, where the descriptor is built on body parts, then relations can be simply the Euclidean distances between pairs of joints. If the RHI is built on depth information, then we build a bounding box around each person in the scene, and we split it in cells, which represent the locally fixed regions. Then, we extract from each cell the modified τ test proposed in [Lu *et al.*, 2014], and the relation between pairs of cells is represented using the Hamming distance between pairs of strings (see [Gori *et al.*, 2016] for details).

We embed the temporal information in our descriptor computing a series of temporal differences. We consider windows $[t, t+F]$ of F frames and we build a tensor $\mathbf{W_t}$ composed of the differences between all the matrices $\mathbf{R_{t+f}}$, $f \in \{2,...,F\}$ in the window and the first one:

$$\mathbf{W_t} = [\mathbf{R_{t+1}} - \mathbf{R_t} \quad \mathbf{R_{t+2}} - \mathbf{R_t} \quad ... \quad \mathbf{R_{t+F-1}} - \mathbf{R_t}]. \qquad (1)$$

We further convolute $\mathbf{W_t}$ with a set of 1D Gaussian filters over the temporal dimension. The responses are summed up in the final RHI descriptor:

$$\mathbf{RHI_t} = \sum_{j=1}^{s} \mathbf{W_t} * h(\mu, \sigma_j^2), \qquad (2)$$

Fig. 2 depicts some examples of RHI.

Figure 3: Sample images from the new Multy-Type dataset.

3.2 Learning and Classification

We evaluate temporal *segments*, where a *segment* is defined as a set of frames where each user is performing one activity that lasts for the entire segment. A new segment starts when a new person enters the scene, or someone leaves the scene, or someone starts performing a different activity. For each segment, there is no a priori information on who is interacting with whom, who is on their own and who is interacting with the robot. Therefore, we extract RHIs on skeleton and depth from all the possible pairs, including pairs of each person with themselves.

All the RHIs extracted from videos of a specific class are used as training data for that class. The descriptors that do not model valid activities are associated with the 'null' class. We train a one-vs-one Support Vector Machine (SVM), which builds a classifier to discriminate between each pair of classes. For each test RHI_t, the one-vs-one SVM returns a set of scores $\mathbf{s^t}$ such that $s^t_{k,l}$ is positive if the descriptor has been classified as belonging to class c_k, negative if belonging to class c_l. We then define a function that, given a SVM score, returns 1 if the score is positive, 0 if it is negative:

$$h(s^t_{k,l}) = \begin{cases} 1, & \text{if } s^t_{k,l} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Based on these scores, given the set $\mathbf{C}$ of the activity classes, for each pair of users (u_i, u_j) we build a vector of votes $\mathbf{v^{ij}} = [v^{ij}_1, v^{ij}_2, ..., v^{ij}_{|C|}]$ on the entire segment, where: $v^{ij}_k = \sum_{t=1}^{T} \sum_{l=1, l\neq k}^{|C|} h(s^t_{k,l})$ and T is the number of RHIs extracted from the segment. We, therefore, obtain a matrix $\mathbf{V}$, where the element v^{ij}_k indicates the votes obtained by class c_k on the test data extracted for the pair (u_i, u_j).

The simultaneous identification and classification is performed using an optimization procedure. For each segment, the system assigns activities to the pairs of users, to maximize the number of votes globally over the entire segment. The optimization is subject to the constraints that each pair must be performing exactly one activity (including the 'null' activity), and that if one person is performing a one-person activity, they cannot take part in any other interaction. In particular, we create a set of variables Φ, one for each element of the matrix $\mathbf{V}$. Then, given the set of users $\mathbf{U}$, the set of activities $\mathbf{C}$ and the matrix of votes $\mathbf{V}$ for one segment we solve:

$$\begin{aligned} \Phi^* = \underset{\Phi}{\text{argmax}} \quad & \sum_{k=1}^{|C|} \sum_{i=1}^{n} \sum_{j=1}^{n} v^{ij}_k \phi^{ij}_k \\ \text{s. t.} \quad & \sum_{k=1}^{|C|} \phi^{ij}_k = 1;\ i, j \in [1, |U|]; \\ & \sum_{k=1}^{|C|} \phi^{ii}_k + \sum_{k=1}^{|C|} \sum_{j=1, j\neq i}^{n} \phi^{ij}_k = 1;\ i \in [1, |U|]; \\ & \phi^{ij}_k \in \{0, 1\};\ i, j \in [1, |U|], k \in [1, |C|]. \end{aligned} \tag{4}$$

The objective function represents the fact that we want to choose classes and users that maximize the votes obtained from the classifier. The first constraint models the fact that each pair has to be assigned to only one class. The second constraint handles the assignment between pairs composed of the same user and pairs of different users.

4 Experiments

This section presents our experimental results on multiple datasets. RHI performance is compared to other state-of-the-art descriptors on our new dataset as well as on three public datasets.

4.1 Experiments on our Multi-Type Dataset

We collected a new RGB-D dataset, called Multi-Type Dataset, which includes videos where multiple types of activities are performed concurrently and sequentially. We took advantage of Kinect 2.0 for the tracking. The sensor has been mounted on an autonomous non-humanoid robot, which is designed to move around in a building populated by students. We asked 20 participants divided in 5 groups of $4-5$ persons to perform 12 different sequences of activities. Each sequence is characterized by the presence of 2 to 5 persons performing actions, with different body orientations and at different scales. We defined 16 basic activities: 6 two-person interactions, *approach, hug, shake hands, push, talk, wave*, 6 first-person interactions, *approach robot, wave to the robot, point the robot, clear the path, talk to the robot, grab the robot*, and 4 single activities, *sit, stand up, walk, run*. We collected RGB, depth and skeletal data (see Fig. 3).

Given multiple continuous videos containing a total of 288 activity executions which may temporally overlap, 523 samples are extracted including action samples as well as non-valid action samples (i.e., 'null' class samples not corresponding to any activity) for the training/testing of our approach and baselines. The one-vs-one SVM is trained based on all these samples by using corresponding action samples as positive samples and all the others as negative samples. We performed a leave-one-group-out cross-validation treating 4 groups of people as training set and 1 group as test set.

We carried out two experiments. In the first one, we assume that there is no information on who are the pairs that perform valid actions. To evaluate the performance of our

Table 1: The table shows results achieved by our method and other state-of-the-art algorithms on publicly available datasets and on our new dataset.

Method	SBU	FP	MSR	MT-Acc	MT-F1	MT-F2	Real-time
[Yun *et al.*, 2012]	80.03%	-	-	-	-	-	Y
[Ji *et al.*, 2014]	86.9%	-	-	-	-	-	Y
HON4D [Oreifej and Liu, 2013]	77.0%	45.55%	88.36%	68.07%	0.7315	0.7870	N
DSTIP [Xia and Aggarwal, 2013]	42.69%	53.25%	37.76%	28.38%	0.3891	0.46	N
STIP [Laptev and Lindeberg, 2003]	66.28%	50.84%	-	38.59%	0.4519	0.5641	N
HOJ3D [Xia *et al.*, 2011]	41.88%	70.0%	78.97%	47.06%	0.4434	0.5075	Y
[Xia *et al.*, 2015]	-	83.7%	-	-	-	-	N
[Li *et al.*, 2010]	-	-	74.7%	-	-	-	?
[Wang *et al.*, 2012a]	-	-	86.5%	-	-	-	?
[Wang *et al.*, 2012b]	-	-	88.2%	-	-	-	?
[Wang *et al.*, 2013]	-	-	90.22%	-	-	-	?
[Evangelidis *et al.*, 2014]	-	-	89.86%	-	-	-	?
[Chaaraoui *et al.*, 2013]	-	-	91.8%	-	-	-	Y
[Vemulapalli *et al.*, 2014]	-	-	92.46%	-	-	-	?
[Luo *et al.*, 2013]	-	-	**96.70%**	-	-	-	?
RHI (ours)	**93.08%**	**85.94%**	95.38%	**76.67%**	**0.7954**	**0.8633**	Y

method, we use precision and recall. In the second experiment, we assume that the pairs who are performing valid activities are known. In this case, we evaluate the effectiveness of descriptors in terms of classification accuracy.

In order to establish a baseline, HON4D [Oreifej and Liu, 2013], STIP [Laptev and Lindeberg, 2003] and DSTIP [Xia and Aggarwal, 2013] are assessed on our dataset. Table 1, column **MT-F1** and **MT-F2**, summarizes the comparison between our descriptor and HON4D, STIP and DSTIP. The recognition accuracy, computed assuming that the true pairs are known, is reported in Table 1, column **MT-Acc**. As the table shows, RHI outperforms any other general purpose feature descriptor.

To show that RHI is more flexible than other joint-based methods, we extend HOJ3D [Xia *et al.*, 2011] such that when considering pairs constituted by different persons, we concatenated the two HOJ3D; when considering a pair composed of the same subject repeated twice, we concatenated a set of zeros at the end of the HOJ3D computed on the subject. Table 1 shows that the results achieved by this method on our dataset, as well as on the SBU dataset, are poor. This confirms that it is not trivial to find a way to extend conventional non-unified descriptors to recognize multiple types of activities.

4.2 Comparison on Public Datasets

We assess our new RHI descriptor on three publicly available datasets: SBU Interaction dataset [Yun *et al.*, 2012], which has been recorded for interaction recognition purposes, the non-humanoid robot dataset presented in [Xia *et al.*, 2015], which has been collected to recognize first-person interactions, and the MSRAction3D dataset [Li *et al.*, 2010], which contains video where a single person performs simple actions. In this case, we only need to evaluate its recognition accuracy. We compute the RHI descriptor on each set of F frames in every video, which constitute positive examples for the classifiers. We then train non-linear SVMs with Radial Basis kernel in a one-versus-all fashion.

Table 1 shows that RHI is not only the most effective descriptor on our new dataset, but also achieves excellent performance on all the datasets we have considered. Even though some of considered approaches obtained reasonable results on classifying activities of a single type, they do not have the ability to directly compare activities of different types, and thus are not suitable for classifying/detecting them. The only descriptors that may be suitable for our multi-type activity recognition setting are appearance-based descriptors such as HON4D, STIP and DSTIP. However, as the experiments confirm, they do not perform as well as the proposed approach in our multi-type scenarios.

5 Conclusion

Our work addresses the problem of labeling a complex robot-centric scene where multiple persons perform different types of activities concurrently and sequentially. We proposed a new unified descriptor that does not depend on any sensitive parameter and can be computed in real-time. We further propose a new optimization-based method that enables the simultaneous classification of activities and identification of the subjects involved. Experimental results confirm that RHI outperforms previous works on publicly available datasets as well as on the newly collected Multi-Type Dataset, a new RGB-D dataset which can be useful to the community for future benchmarking.

Acknowledgments

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016.

References

[Chaaraoui *et al.*, 2013] Alexandros Andre Chaaraoui, Jose Ramon Padilla-Lopez, and Francisco Florez-Revuelta. Fusion of skeletal and silhouette-based features

for human action recognition with rgb-d devices. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2013.

[Evangelidis *et al.*, 2014] Georgios Evangelidis, Gurkirt Singh, and Radu Horaud. Skeletal quads: Human action recognition using joint quadruples. *International Conference on Pattern Recognition (ICPR)*, 2014.

[Fathi *et al.*, 2012] Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Gori *et al.*, 2016] Ilaria Gori, J. K. Aggarwal, Larry Matthies, and Michael S. Ryoo. Multitype activity recognition in robot-centric scenarios. *IEEE Robotics and Automation Letters*, 1(1):593–600, Jan 2016.

[Ji *et al.*, 2014] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014.

[Khamis *et al.*, 2012] Sameh Khamis, Vlad I. Morariu, and Larry S. Davis. Combining per-frame and per-track cues for multi-person action recognition. In *European Conference on Computer Vision (ECCV)*, 2012.

[Kong *et al.*, 2014] Yu Kong, Yunde Jia, and Yun Fu. Interactive phrases: Semantic descriptions for human interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[Lan *et al.*, 2012] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Laptev and Lindeberg, 2003] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[Li *et al.*, 2010] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[Lu and Grauman, 2013] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[Lu *et al.*, 2014] Cewu Lu, Jiaya Jia, and Chi-Keung Tang. Range-sample depth feature for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[Luo *et al.*, 2013] Jiajia Luo, Wei Wang, and Hairong Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[Oreifej and Liu, 2013] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[Ryoo and Aggarwal, 2009] Michael S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[Ryoo and Matthies, 2013] Michael S. Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[Vemulapalli *et al.*, 2014] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[Wang *et al.*, 2012a] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *European Conference on Computer Vision (ECCV)*, 2012.

[Wang *et al.*, 2012b] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Wang *et al.*, 2013] Chunyu Wang, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[Xia and Aggarwal, 2013] Lu Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[Xia *et al.*, 2011] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2011.

[Xia *et al.*, 2015] Lu Xia, Ilaria Gori, J. K. Aggarwal, and Michael S. Ryoo. Robot-centric activity recognition from first-person rgb-d videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[Yun *et al.*, 2012] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.