Unsupervised Framework for Interactions Modeling between Multiple Objects

Ali Al-Raziqi and Joachim Denzler
Computer Vision Group, Friedrich Schiller University of Jena, Jena, Germany

Keywords: Interaction Detection, Multiple Object Tracking, Unsupervised Clustering, Hierarchical Dirichlet Processes.

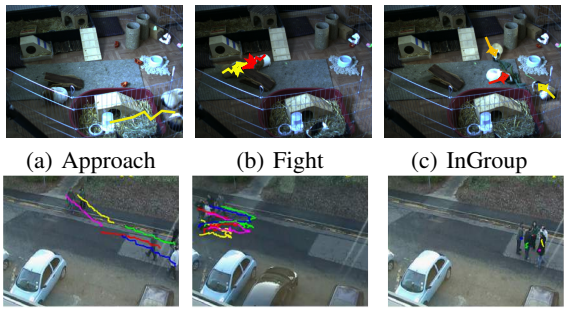
Abstract: Extracting compound interactions involving multiple objects is a challenging task in computer vision due to different issues such as the mutual occlusions between objects, the varying group size and issues raised from the tracker. Additionally, the single activities are uncommon compared with the activities that are performed by two or more objects, e.g., gathering, fighting, running, etc. The purpose of this paper is to address the problem of interaction recognition among multiple objects based on dynamic features in an unsupervised manner. Our main contribution is twofold. First, a combined framework using a tracking-by-detection framework for trajectory extraction and HDPs for latent interaction extraction is introduced. Another important contribution is the introduction of a new dataset, the Cavy dataset. The Cavy dataset contains about six dominant interactions performed several times by two or three cavies at different locations. The cavies are interacting in complicated and unexpected ways, which leads to perform many interactions in a short time. This makes working on this dataset more challenging. The experiments in this study are not only performed on the Cavy dataset but we also use the benchmark dataset Behave. The experiments on these datasets demonstrate the effectiveness of the proposed method. Although the our approach is completely unsupervised, we achieved satisfactory results with a clustering accuracy of up to 68.84% on the Behave dataset and up to 45% on the Cavy dataset.

1 INTRODUCTION

Activity recognition is a very important task in computer vision and has many applications such as video surveillance and animal monitoring systems. Computer vision can help the biologists to understand and recognize the behavior of animals in the videos. Activity recognition can be roughly divided into three categories. The first category is single activity, in which the activity is performed by only a unique object without interacting with any other objects (Ohayon et al., 2013; Guha and Ward, 2012; Delaitre et al., 2011). In many situations, single activities are uncommon compared with the activities performed by several active objects *e.g. gathering, chasing, fighting, running, etc*. The second category is pair activity which includes the interaction between two objects. Pair-activity methods can be classified into two approaches. The first approach performs segmenting and tracking the body parts (heads, hands, legs, etc.) of two objects to discover the interactions between them *e.g. high five, kiss, hand shake, etc* (Patron-Perez et al., 2010; Dong et al., 2011; Kong and Jia, 2012; Li et al., 2011). This may be unfeasible for low image resolution and occlusions in surveillance videos. In the Cavy dataset which is introduced for the first time in this study, it is difficult to segment cavy parts to discover the interactions between the cavies. The second category is characterized by tracking the whole body of objects to extract the interactions between objects, *e.g. gathering, scattering, leaving, etc* (Zhou et al., 2011; Sato and Aggarwal, 2004; Blunsden et al., 2007). The third category is group activity which refers to the interaction among multiple objects (two or more objects) within a specific distance. In group activity methods, the scene has to be divided into subgroups, interaction in each group are then analyzed and recognized, *e.g. In-Group, Approach, WalkTogether, Fight, etc*, (Blunsden and Fisher, 2009; Cheng et al., 2014; Kim et al., 2014; Ni et al., 2009; Lin et al., 2010; Yin et al., 2012; Münch et al., 2012; Zhang et al., 2012). Generally, activities involving multiple active objects are considered as a group activity. As an example, *scattering* activity consists of multiple *running* individuals.

Figure 1 shows some scenarios where various objects in a scene are interacting with each other in Cavy

Al-Raziqi, A. and Denzler, J.
Unsupervised Framework for Interactions Modeling between Multiple Objects.
DOI: 10.5220/0005680705090516
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 509-516
ISBN: 978-989-758-175-5



(a) Approach (b) Fight (c) InGroup

(a) Approach (b) Fight (c) InGroup

Figure 1: Interactions between multiple persons on the Cavy and the Behave datasets (Blunsden and Fisher, 2010). For better visibility, refer to the web version.

Figure 2: A set of frames taken from different views and different time with changing the illumination.

and Behave datasets (Blunsden and Fisher, 2010).

It can be observed that most of the previous methods share two common characteristics. First, at a high level, most of them implement the same framework according to which motion/appearance features are extracted. Second, supervised machine learning method are used to classify the interactions. In many of the activity recognition categories, the number of involved objects cannot be determined beforehand. Furthermore, the exact number of activities is usually a prerequisite for classification, which is often unavailable especially for new videos to be analyzed. Hence, using an unsupervised method is a necessity in such situations to extract the interactions.

Our proposed approach incorporates the capabilities of the Hierarchical Dirichlet Processes (HDP) with spatio-temporal dynamic features based on the trajectories to tackle the problem of interactions between objects. The main contribution of this paper is twofold:

1. A combined framework using a tracking-by-detection method for trajectory extraction and HDPs for latent interaction extraction is introduced.
2. The introduction of the Cavy dataset[1], which contains six dominant interactions performed several times by two or three cavies at different locations as shown in Figure 2 and table 1.

Table 1: The dominant interactions performed by two or three cavies at different locations on the Cavy dataset.

Interaction	Description
Approach	One object approaches to another(s) object(s)
RunTogether	Objects walking together
Split	Object(s) split from one another
Ingroup	Several objects are close to each other and with small moving
Fight	Objects fighting each other
Follow	Object(s) following other

The Cavy dataset can be useful in many disciplines, in addition to computer vision, since the dataset is taken at various time, it may help the biologists to study and monitor the cavies behavior in specific periods.

For unsupervised clustering tasks, HDP has been widely used in many fields such as text analysis (Teh et al., 2006), traffic scene analysis and action recognition (Kuettel et al., 2010; Krishna and Denzler, 2014; Krishna et al., 2013) and yielded significant results. In this paper, we apply for the first time HDP to the group activity recognition problem.

The rest of this paper is organized as follows. In Sect. 2 we provide a brief overview of the existing literature on interaction recognition. Sect. 3 describes the interaction modeling. Sect. 4 discusses the applied HDP model, and the corresponding inference procedure. The experiments conducted on the Cavy and the Behave datasets are described in Sect. 5 along with results.

2 RELATED WORK

In this work, we focus on the interaction detection between multiple objects. The related work can be divided into two categories, supervised and unsupervised learning methods.

Supervised Learning, In (Yang et al., 2013), the authors used a graph framework to analyze the interaction between parts of an object. The body parts and objects are represented as nodes of graphs, the parts are tracked to extract the temporal features and the

[1]Available at http://www.inf-cv.uni-jena.de/Group/Staff/M_Sc_++Ali+M_+Al_Raziqi.html

network analysis provide the spatial features. They then use Support Vector Machine (SVM) and a Hidden Markov Model to classify the interactions of the object's parts. In (Zhou et al., 2011), the authors analyzed the interaction between objects based on Granger Causality Test. The GCT causality measures the effect of the objects on each other. In (Ni et al., 2009), the authors divided the individuals into subgroups and cluster them using k-means algorithm. The causality is analyzed with respect to individual, pair, and inter-group activity. Finally, classifiers such as Nearest Neighbor (NN) and SVM are used for group activity classification.

Another relevant approach is introduced in (Kim et al., 2014), where the authors recognize group activities by detecting meaningful groups. This is done by defining Group Interaction Zone (GIZ). Group activities in each GIZ can be illustrated by attraction and repulsion properties which are represented by the relative distance during k frames. Furthermore, the study in (Cheng et al., 2014) presented a new approach in different semantic layers: individual, pair and group. Motion and appearance features are extracted from those layers. For the appearance features, Histograms of Oriented Gradients (HoG) are extracted for each object in the group and Delaunay triangulation is used to extract the whole group features.

Unsupervised Learning, In the work presented in (Al-Raziqi et al., 2014), the authors have developed an HDP-based interaction extraction approach in which the optical flow is extracted in the whole image without object localization or trajectories motion analysis. Another interesting method which tried to tackle this problem is described in (Zhu et al., 2011). The authors extracted features such as appearance, causality and feedback based on GCT and leant an extended probabilistic Latent Semantic Analysis (pLSA). Then, pLSA used to categorize new sequences.

Unlike many of the approaches described above, our approach integrates an unsupervised clustering method, namely HDP, with optical flow based on motion trajectories to identify the interactions of multiple objects without further knowledge.

3 INTERACTIONS MODELING

The interaction is an activity performed by several objects within a specific region. Figure 3 shows the main steps of our approach. In order to perform object interaction modeling, as a preliminary step, a reliable and accurate tracker is required. Since the objects in the Cavy dataset are not annotated by bounding

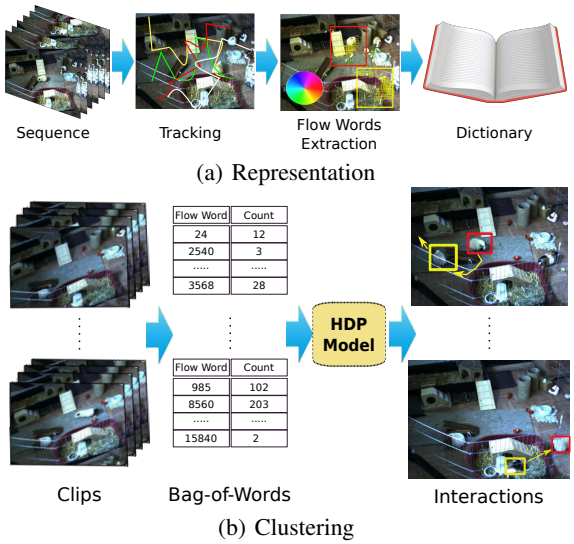


Figure 3: Our framework for interaction detection: (a) Objects tracking to extract the trajectories, low-level visual features inside bounding boxes and build the dictionary with all possible flow words. (b) Divided the video sequence into short clips. Local flow motions are computed for each clip. Each clip is represented by a Bag-of-Words. HDP is used to extract the interactions between the objects.

boxes (BB), we cannot start with ground truth object positions and trajectories, but have to compute this information from the data itself. This makes the Cavy data set a very challenging one since the consecutive interaction detection step must deal with errors in the previous tracking step. The set of detections (BB) are generated by background subtraction method using Gaussian Mixture Model (GMM) presented in (Zivkovic, 2004).

Due to this simple detection method, errors in the detection cannot be avoided, such as missing, false, merging or splitting objects. Examples are shown in Figure 6. To mitigate the effect of wrong or missing detections, we apply a two-stage graph method presented in (Jiang et al., 2012). The result of the tracking algorithm are trajectories of all objects, e.g. the $i-th$ object trajectory is represented from time 1 to k as: $T_i^k = [x_i^1, x_i^2, ..., x_i^k]$, where T_i^k is a sub-trajectory of the object $i's$ trajectory in k frames and x_i is a center of mass coordinate (x, y) of a particular object. The average distance is computed using Euclidean distance between the sub-trajectories which consists of the largest value k for which the length of the trajectories of object i equals that of object j.

$$D_{i,j} = \frac{1}{k}||T_i^k - T_j^k|| \quad (1)$$

Subsequently, optical flow inside the BBs regions is computed using the TV-L^1 algorithm (Zach et al., 2007), if the $D_{i,j}$ is below a user defined threshold

. This threshold depends on the kind of interactions to be identified by the system and is application depended. The video sequence is then divided into short and equally sized clips without overlap. In each clip, optical flow is quantized into eight directions (flow words). The optical flow features can be defined as X=(x, y, u, v), where (x, y) is the location of a particular pixel in the image, and (u, v) are the flow values. Following the approaches are described in (Kuettel et al., 2010; Krishna and Denzler, 2014), all clips are represented by accumulated flow words over their frames. Finally, a dictionary is built with all possible flow words.

4 HIERARCHICAL DIRICHLET PROCESSES

HDP was first presented for clustering words in documents based on words co-occurrence to infer the latent topics (Teh et al., 2006). Specifying the number of topics beforehand is impractical. So, in HDP the number of clusters (*topics*) is extracted automatically from the data and a set of hyper-parameters (α, γ and η). Figure 4 shows the basic HDP model.

For text analysis, the corpus is divided into M separated documents where each document contains a set of unordered words N_m, denoted as $x_{m,n}$, where $m \in [1, M]$ and $n \in [1, N_m]$. Hence, each document is represented by its words.

In our case, we follow (Krishna and Denzler, 2014; Al-Raziqi et al., 2014; Kuettel et al., 2010), where the corpus, documents and words correspond to the video sequence, short equal sized clips, and optical flow respectively. Generally, for a given input video, optical flow features are extracted from each pair of successive frames. Then, the video sequence is divided into short clips. Each clip is represented by accumulated a Bag-of-Words (see Section 3).

The HDP in this work uses Dirichlet Process (DP) to infer the interactions at two levels. The global list of interactions G_0 is generated in the first DP level, where G_0 is a prior distribution over the video. In the second DP, specific interactions G_m are drawn from the global list G_0 for each clip. The interactions might be shared among different G_m. Formally, we write the generative HDP formulation

$$\begin{aligned} G_0 \mid \gamma, H &\sim DP(\gamma, H) \\ G_m \mid \alpha, G_0 &\sim DP(\alpha, G_0) \quad \text{for } m \in [1, M]. \end{aligned} \tag{2}$$

In Eq 2, the hyper-parameters α and γ are called the concentration parameters and the parameter H is called the base distribution. Therefore, the observed words $x_{m,n}$ are seen as being sampled from the mixture priors $\phi_{m,n}$, which can be interpreted as being drawn from a DP G_0. The values of mixture components are drawn from θ_k. Consequently, this model can be written as

$$\begin{aligned} \theta_k &\sim P(\eta) \qquad \text{for } k \in [1, \infty) \\ \phi_{m,n} \mid \alpha, G_m &\sim G_m \qquad \text{for } m \in [i, M], n \in [1, N_m] \\ x_{m,n} \mid \phi_{m,n}, \theta_k &\sim F(\theta_{\phi_{m,n}}) \end{aligned} \tag{3}$$

where M is the number of clips in the video, N_m is the number of words in clip m, $P(\cdot)$ and $F(\cdot)$ are the prior distribution over topics and words respectively.

Consequently, given the observed flow words, HDP *infers* the latent topics (interactions) which is called Bayesian inference. Following the formulation of (Krishna et al., 2013), the conditional probability of the topic-word association for each iteration step is evaluated as

$$p(\phi_{m,n} = k, \alpha, \gamma, \eta, \theta, H) \propto (n_{m,k}^{\neg m,n} + \alpha\theta_k) . \frac{n_{k,t}^{\neg m,n} + \eta}{n_k^{\neg m,n} + V.\eta} \tag{4}$$

where $n_{m,k}$, $n_{k,t}$, and n_k represent statistics of the word-topic, topic-document and the topic-wise word counts, respectively. The current word $x_{m,n}$ must be excluded from that topic. The size of the dictionary is represented by V. The probability of assignment of the current flow word $x_{m,n}$ to a particular topic is relative to the number of words previously associated with that topic as shown in the first term of equation 4. The second term shows the effects of the hyper-parameters α, γ and η on determining the number of extracted topics and the possibility of creating a new topic. In this paper, interactions are interpreted as flow words which co-occur in the same clip. The idea is that optical flow measured within the area of a tracked object represent fine-grained details of the activity which in combination with the respective activity of the other object identifies the interaction.

5 EXPERIMENTS AND RESULTS

For evaluating the performance of our proposed framework, we performed several experiments on two different datasets, the Cavy dataset and the benchmark dataset Behave (Blunsden and Fisher, 2010) to illustrate the effectiveness and capability of the HDP in interaction extraction. Both datasets provide various challenging interactions of multiple objects as shown in Figure 1.

As the Cavy dataset does not contain ground truth in terms of interactions among objects, we marked

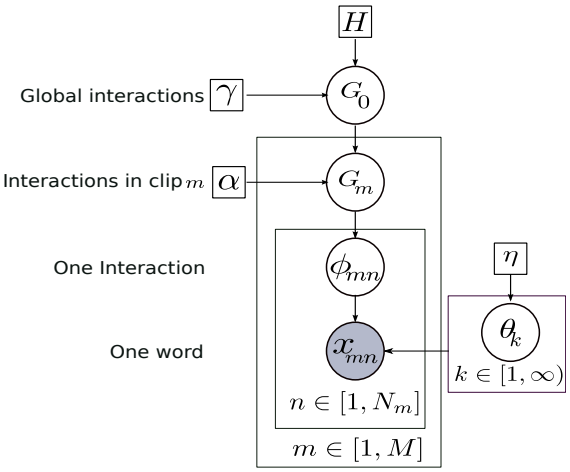


Figure 4: HDP model. Dirichlet Processes are used to generate the global interactions G_0 and G_m which are drawn from the global G_0.

the semantically meaningful interactions in the scene (clip-wise annotations). Then, similar to the procedure in (Kuettel et al., 2010; Krishna and Denzler, 2014; Al-Raziqi et al., 2014), the output of our system is manually mapped to the ground truth labels and the performance measures are calculated. For the performance evaluation, we use the accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

where TP, FP, FN, and TN are True Positives, False Positives, False Negatives, and True Negatives respectively.

5.1 Results on Cavy Dataset

The Cavy dataset is a new dataset introduced in this work. The Cavy dataset contains a variety of conditions that have been taken from a stationary camera. As can be observed in Figure 2, sequences are recorded from different views with changing illumination and in different periods. It contains 16 sequences with 640×480 resolutions recorded at 7.5 frames per second (fps) with approximately 31621506 frames in total (272 GB). The sequences are recorded non-synchronously and stored in *ppm* format. The Cavy dataset contains six dominant interactions performed several times by two or three cavies at different locations in the scene. Table 1 shows the types of the interactions. Some interactions are easy to distinguish, while others only differ a bit in execution period, velocity and the number of involved cavies. In these experiments, we used eight sequences with a total number of 159358 frames.

Results: As baseline experiments on the Cavy dataset, we first extract trajectories of objects and dynamic features. As next step, the optical flow is computed inside the bounding boxes. Then, HDP is used to extract the global interactions in the video. As next step, the optical flow is computed inside the bounding boxes.

Table 2: Confusion matrix representing the performance of the HDP on the Cavy dataset.

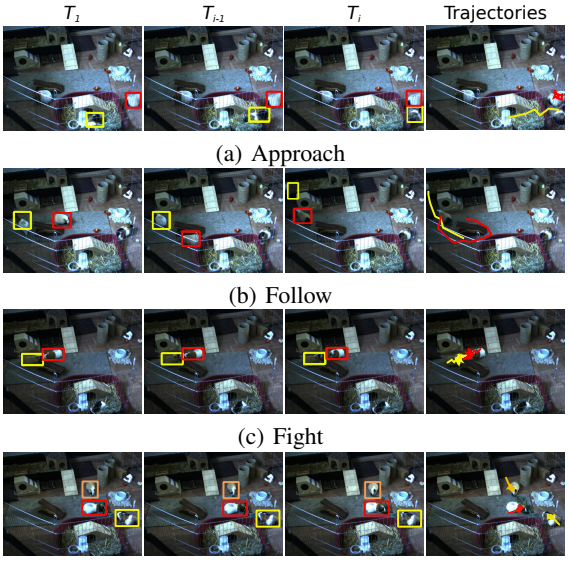
A	0.5	0.03	0.05	0.00	0.00	0.00	0.41	61
S	0.01	0.28	0.03	0.00	0.01	0.00	0.67	75
I	0.03	0.01	0.4	0.00	0.02	0.00	0.54	373
FO	0.00	0.25	0.00	0.63	0.00	0.00	0.13	8
F	0.02	0.00	0.1	0.00	0.35	0.00	0.52	48
R	0.00	0.17	0.00	0.00	0.00	0.50	0.33	6
N	0.06	0.01	0.14	0.00	0.05	0.03	0.71	403
	A	S	I	FO	F	R	N	#

For qualitative analysis of our method, Figure 5 shows the interactions extracted by the HDP model. In Figure 5(a), interaction interpreted as one cavy approaches another one. Figure 5(b) represents one cavy follows another one. Figure 5(c) shows two objects are fighting each other. This wrong result is caused by the detection error (split bounding box). Figure 5(d) shows two objects close to each other (InGroup). The interactions are performed by two or three cavies in *k* frames and represented by flow words co-occurring in the same clip.

Also we studied the effect of the hyper-parameters of HDP (α and η) on the number of the extracted interactions as depicted in Figure 7 (a). The hyper-parameters (α and η) values ranging from 0.1 to 2, and the clip length is 150 frames. As mentioned, the hyper-parameter values control the number of obtained topics (in our case interactions). As notice from Figure 7 (a), the number of extracted interactions has fluctuated significantly with increasing the hyper-parameter values especially η. This is due to the fact that the hyper-parameter η controls the probability of generating new interactions. It is crucial mentioning that increasing the values of η does not always lead to the generation of new interactions. This is likely due to the randomness in the Bayesian inference step.

Table 2 shows the quantitatively evaluation for the selected interactions, *Approach (A), Split (S), RunTogether (R), Fight (F), InGroup (I), Follow (FO), No (N)*. We add the field (No) which represents the false positive and false negative of the particular interaction, where the ground truth does not contain that interaction. The last column in Table 2 represents the number of instances of each category in the ground truth.

In this experiment, the video divided into clips of 150 frames which achieved an average clustering of up to 45 %. It is clear that there are different factors



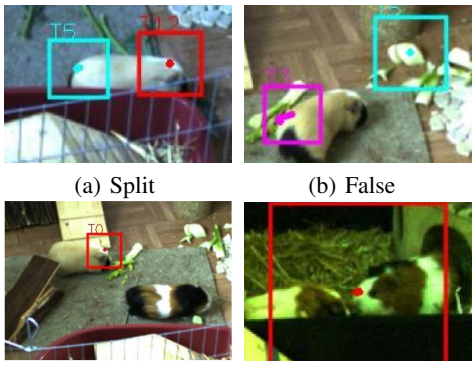
(a) Approach

(b) Follow

(c) Fight

(d) InGroup

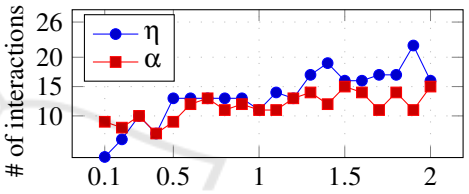
Figure 5: The Cavy dataset. Illustration of different interactions occur in k successive frames. Each row represents one interaction extracted by HDP. The trajectories correspond to the interactions. For better visibility, refer to the web version.



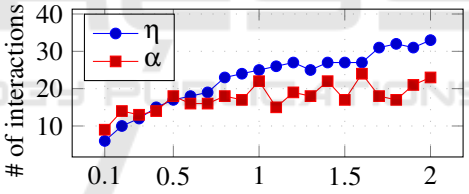
(a) Split (b) False

(c) Missing (d) Merge

Figure 6: Illustration of the main tracking issues. (a) Split object, which leads to discover interaction interpreted as following. (b) Non-cavies object detected as cavy (false), HDP discovers an interaction as gathering or as leaving. (c) Missing of detecting and (d) Merge objects, in these cases HDP will not be able to discover the interaction.



(a) Cavy dataset



(b) Behave dataset

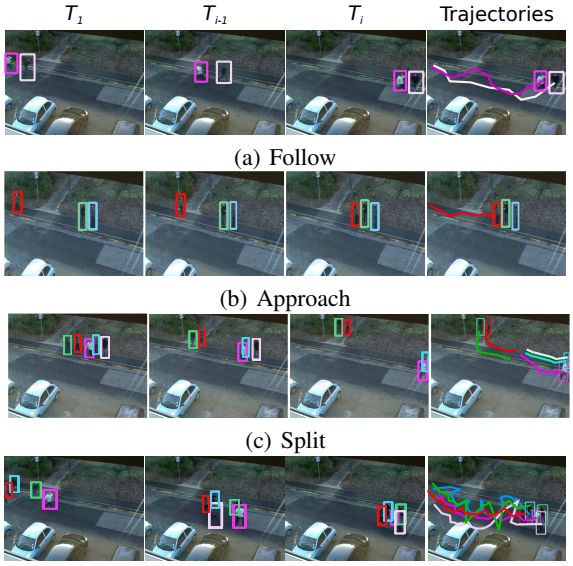
Figure 7: Effects hyper-parameters α, η on number of extracted interactions. In each experiment, we change one of the hyper-parameters while the other is held constant at 0.5 and vice versa. The clip size is 150 frames.

that have an effect on the results, such as errors raised from detector and tracker (missing, false, merged or splitted objects) as shown in Figure 6. More precisely, missing and merged object(s) issues lead to decrease the TP, whereas the false detection increases the FP. For instance, as can observed from Table 2, the highest false positive ratio is detected for the Split interaction, while the interaction is not found in the ground truth. Consequently, our method showed lower performances for the Split interaction. Additionally, the optical flow is probably not helpful in case of the fixed objects which leads to increase the false negative. All of these factors lead to degraded the performance of our approach.

5.2 Results on Behave Dataset

Also we used the Behave dataset (Blunsden and Fisher, 2010). Behave dataset consists of four video sequences, and 76,800 frames in total and recorded at 25 frames per second with a resolution of 640 × 480 pixels. The Behave dataset provides different challenging interactions include: *InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase , Fight, RunTogether,* and *Meet*. The number of objects involved in the interaction ranging from two to five. Due to the limited number of annotated frames, (Kim et al., 2014; Yin et al., 2012; Zhang et al., 2012; Münch et al., 2012) used subsets of the categories to demonstrate the performance of their methods. However, we use the same subsets to compare our approach with their methods. In this study, we divided the sequences into clips with size 150 frames and only analyze clips for which ground truth are available. It must be mentioned that *Meet* and *Ignore* categories found just once and twice in the ground truth respectively. Hence, these categories are excluded.

Results: Qualitatively, Figure 8 shows the set of the probable interactions in one sequence. As observed in Figure 8(a), one object follows another one from the left corner to the right corner. Figure 8(b) interpreted as one object approaches from the left corner to join other objects (converge to center). In Figure 8(c), the group splitted into two groups each one



(a) Follow

(b) Approach

(c) Split

(d) Fight

Figure 8: Illustration of different interactions occur in k successive frames.Each row represents one interaction, last column represents the extracted interaction using HDP.

Table 3: Confusion matrix represents the performance of the HDP on the BEHAVE dataset.

I	0.54	0.13	0.16	0.03	0.03	0.00	0.10
A	0.11	0.68	0.16	0.05	0.00	0.00	0.00
W	0.03	0.14	0.75	0.06	0.00	0.03	0.00
S	0.13	0.13	0.00	0.67	0.00	0.00	0.07
FO	0.00	0.00	0.00	0.00	1	0.00	0.00
R	0.00	0.00	0.17	0.00	0.00	0.5	0.00
F	0.1	0.00	0.00	0.00	0.00	0.00	0.8
	I	A	W	S	FO	R	F

walking into different directions and a set of objects fighting each other as shown in Figure 8(d). It is worth mentioning that the interactions $a-d$ in Figure 8 represented by flow words based on their co-occurrence in the same clip. The spatial flow patterns is formed by flow words at different coordinates in the frames.

For the quantitatively evaluation for the selected categories, *Approach (A), Split (S), WalkTogether (W), RunTogether (R), Fight (F), InGroup (I), Follow (FO)*. Table 3 shows the confusion matrix of HDP performance. Our method demonstrated lower performances for InGroup and RunTogether interactions. Most likely due to that the optical flow is not worked precisely with fixed objects. Addition to the highest similarity and the spatial overlap between the interactions.

For the comparison with previous work, we used the same subsets from the Behave dataset and compare directly with their results. We compared our approach with (Kim et al., 2014; Yin et al., 2012; Münch et al., 2012) with the selected group activities *e.g. Approach, Split, WalkTogether, InGroup,Fight* as shown in Table 4. As can be seen from Table 4, despite the fact that our approach is completely unsupervised, we achieved a clustering accuracy close to (Münch et al., 2012) of up to 65.95%. Unlike (Münch et al., 2012), our approach extracted the interactions without prior knowledge.

Table 4: Interaction recognition comparison with (Kim et al., 2014), (Münch et al., 2012) and (Yin et al., 2012).

Category	Our	(Kim et al., 2014)	(Münch et al., 2012)	(Yin et al., 2012)
Approach	68.42	83.33	60	n/a
Split	66.42	100	70	93.10
WalkTogether	75.00	91.66	45	92.10
InGroup	53.73	100	90	94.3
Fight	80.00	83.33	n/a	95.10
Average	65.95	93.74	66.25	93.65

The essential benefits of our approach is that it is able to extract the interactions automatically for the new unseen videos without any further knowledge.

5.3 Implementation

The presented tracking framework was implemented in C++ using the OpenCV library while the optical flow computation and HDP modeling was realized in MATLAB using the standard toolboxes. The experiments are performed on a desktop computer Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz and 32 GB RAM. The implementation parameters were set as follows. The threshold distance for equation 1 was 35 pixels, see Sec. 3. For optical flow feature extraction, the trajectories interval k was 10, see Sec. 3. The hyper-parameters (α and η) values in equation 4 ranging from 0.1 to 2, and the clip length is 150 frames corresponding to approximately 20 seconds, see Sec. 4. The run time for inference process depends on the video size and number of objects interacting with each other in the scene (BB). The Euclidean distance is computed between sub-trajectories for every k frames, the time complexity for Euclidean distance is $O(n)$.

6 CONCLUSIONS AND FUTURE WORK

The aim of this paper was to address the problem of interaction among multiple objects. Our proposed approach incorporates the unsupervised clustering capabilities of the HDP with the spatio-temporal features to recognize the interactions of multiple objects without prior knowledge. Furthermore, the Cavy dataset is introduced in this work. The Cavy dataset is created by capturing the interactions between three cavies.

The Cavy dataset contains six dominant interactions performed several times by two or three cavies at different locations. The challenging aspect of the the Cavy dataset is that the cavies are behaving and interacting in complicated and unexpected ways. The experiments have been performed on the Cavy dataset and the Behave dataset. Extensive experiments on these datasets demonstrate the effectiveness of the proposed method. Our approach achieved satisfactory results with a clustering accuracy of up to 68.84% on the Behave dataset and up to 45% on Cavy dataset.

In the future, robust tracker needs to be developed to mitigate the tracker effects. Also the appearance-based and trajectory-based features beside optical flow could possibly be included.

REFERENCES

Al-Raziqi, A., Krishna, M., and Denzler, J. (2014). Detection of object interactions in video sequences. *OGRW*, pages 156–161.

Blunsden, S., Andrade, E., and Fisher, R. (2007). Non parametric classification of human interaction. In *PRIA*, pages 347–354. Springer.

Blunsden, S. and Fisher, R. (2009). Detection and classification of interacting persons. *Machine Learning for Human Motion Analysis: Theory and Practice*, page 213.

Blunsden, S. and Fisher, R. (2010). The behave video dataset: ground truthed video for multi-person behavior classification. *BMVA*, 4:1–12.

Cheng, Z., Qin, L., Huang, Q., Yan, S., and Tian, Q. (2014). Recognizing human group action by layered model with multiple cues. *Neurocomputing*, 136:124–135.

Delaitre, V., Sivic, J., and Laptev, I. (2011). Learning person-object interactions for action recognition in still images. In *NIPS*, pages 1503–1511.

Dong, Z., Kong, Y., Liu, C., Li, H., and Jia, Y. (2011). Recognizing human interaction by multiple features. In *ACPR*, pages 77–81.

Guha, T. and Ward, R. K. (2012). Learning sparse representations for human action recognition. *IEEE Transactions on, Pattern Analysis and Machine Intelligence*, 34(8):1576–1588.

Jiang, X., Rodner, E., and Denzler, J. (2012). Multi-person tracking-by-detection based on calibrated multi-camera systems. In *Computer Vision and Graphics*, pages 743–751. Springer.

Kim, Y.-J., Cho, N.-G., and Lee, S.-W. (2014). Group activity recognition with group interaction zone. In *ICPR*, pages 3517–3521.

Kong, Y. and Jia, Y. (2012). A hierarchical model for human interaction recognition. In *ICME*, pages 1–6.

Krishna, M. and Denzler, J. (2014). A combination of generative and discriminative models for fast unsupervised activity recognition from traffic scene videos. In *Proceedings of the IEEE (WACV)*, pages 640–645.

Krishna, M., Körner, M., and Denzler, J. (2013). Hierarchical dirichlet processes for unsupervised online multi-view action perception using temporal self-similarity features. In *ICDSC*, pages 1–6.

Kuettel, D., Breitenstein, M. D., Van Gool, L., and Ferrari, V. (2010). What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, pages 1951–1958.

Li, B., Ayazoglu, M., Mao, T., Camps, O., Sznaier, M., et al. (2011). Activity recognition using dynamic subspace angles. In *CVPR*, pages 3193–3200.

Lin, W., Sun, M.-T., Poovendran, R., and Zhang, Z. (2010). Group event detection with a varying number of group members for video surveillance. *IEEE Transactions on CSVT*, 20(8):1057–1067.

Münch, D., Michaelsen, E., and Arens, M. (2012). Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering. In *KI 2012: Advances in Artificial Intelligence*, pages 233–236. Springer.

Ni, B., Yan, S., and Kassim, A. (2009). Recognizing human group activities with localized causalities. In *CVPR*, pages 1470–1477.

Ohayon, S., Avni, O., Taylor, A. L., Perona, P., and Egnor, S. R. (2013). Automated multi-day tracking of marked mice for the analysis of social behaviour. *Journal of neuroscience methods*, 219(1):10–19.

Patron-Perez, A., Marszalek, M., Zisserman, A., and Reid, I. (2010). High five: Recognising human interactions in tv shows.

Sato, K. and Aggarwal, J. K. (2004). Temporal spatio-velocity transform and its application to tracking and interaction. *Computer Vision and Image Understanding*, 96(2):100–128.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Yang, G., Yin, Y., and Man, H. (2013). Human object interactions recognition based on social network analysis. In *AIPR*, pages 1–4.

Yin, Y., Yang, G., Xu, J., and Man, H. (2012). Small group human activity recognition. In *ICIP*, pages 2709–2712.

Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition*, pages 214–223. Springer.

Zhang, C., Yang, X., Lin, W., and Zhu, J. (2012). Recognizing human group behaviors with multi-group causalities. In *WI-IAT*, volume 3, pages 44–48.

Zhou, Y., Ni, B., Yan, S., and Huang, T. S. (2011). Recognizing pair-activities by causality analysis. *ACM TIST*, 2(1):5.

Zhu, G., Yan, S., Han, T. X., and Xu, C. (2011). Generative group activity analysis with quaternion descriptor. In *Advances in Multimedia Modeling*, pages 1–11. Springer.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, volume 2, pages 28–31.