Detection of Object Interactions in Video Sequences

Ali Al-Raziqi, Mahesh Venkata Krishna and Joachim Denzler
Computer Vision Group
Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2 07743 Jena, Germany
Email: {ali.al-raziqi, mahesh.vk, joachim.denzler}@uni-jena.de

***Abstract*—In this paper, we propose a novel framework for unsupervised detection of object interactions in video sequences based on dynamic features. The goal of our system is to process videos in an unsupervised manner using Hierarchical Bayesian Topic Models, specifically the *Hierarchical Dirichlet Processes* (HDP). We investigate how low-level features such as optical flow combined with Hierarchical Dirichlet Process (HDP) can help to recognize meaningful interactions between objects in the scene, for example, in videos of animal interaction recordings, kicking ball, standing, moving around etc. The underlying hypothesis that we validate is that interactions in such scenarios are heavily characterized by their 2D spatio-temporal features. Various experiments have been performed on the challenging JAR-AIBO dataset and first promising results are reported.**

I. Introduction

Application fields such as video-based surveillance systems, animal monitoring systems etc., often require us to distinguish the interactions between objects or the interactions between objects and their surroundings. Figure 1 shows an example scenario where various objects in a scene are interacting with each other. The meaningful interactions in a scene are characterized by the spatio-temporal dynamics of the objects within the scene.

Detecting interactions between objects in scenes is a challenging problem in computer vision. The challenge is compounded by various aspects such as occlusions, variations in objects sizes, illumination variations, noisy recordings etc. It is important that any system tackling the problem is robust with respect to such factors.

Further, in many of these application scenarios, the interactions are not well-known beforehand, and preparation of a well-labeled data-set covering all possible interactions for the purpose of training a machine learning algorithm may not be possible. For example, in the scenario where we observe interactions between animals, all the interactions the animals might be involved in can not be determined beforehand, and sometimes, even the exact number of possible interactions is impossible to predict. In such situations, use of unsupervised methods becomes imperative.

For unsupervised scenarios, as the kind of interactions are not known beforehand, interactions are defined as co-occurring actions from multiple actors or actors performing actions using some inanimate objects in the scene.

In the literature, *Hierarchical Dirichlet Processes* (HDP) and their derivatives have been used for unsupervised activity perception and analysis [1]–[3]. While they have been demonstrated for activity perception and detection for crowded scenes or individual actors, it is not clear whether HDP can be extended to analyze specific interaction between actors, or between actors and objects, in a scene. Further, determining the correct representation schemes for the current task remains a challenge.

According to our knowledge, most of the current object interactions modeling systems rely on supervised learning methods and some features such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), shape/appearance feature matching etc. [4]–[12]. These frameworks typically start with the localization of an object in the frames and then determining the relevant action. Some of these works done on learning the interactions applied on static images [8], [9], [13], [14]. However, object segmentation and localization are often error prone steps, leading to performance deterioration. They suffer from problems such as camouflage, noisy recording process, occlusions, or poor visibility.

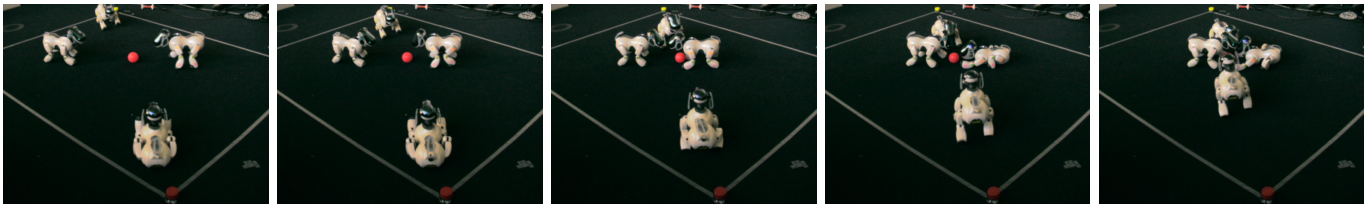
Another interesting line of approaches are based on recognizing objects, actions and human poses [6], [13], and then detecting/recognizing interactions from static images of single object without using feature matching and motion analysis.

Also, in [11], the authors used network graphs framework to analyze the interaction between parts of an object. The body parts and objects are represented as nodes of social network graphs, the parts are tracked to extract the temporal features and the social network analysis features provide the spatial features. They then, used SVM and a Hidden Markov Model to classify the interactions of the object's parts. However, an approach free from object localization requirement and using features that better characterize the interactions in the scene is called for. As a solution, some methods focus on background subtraction [4], [8].

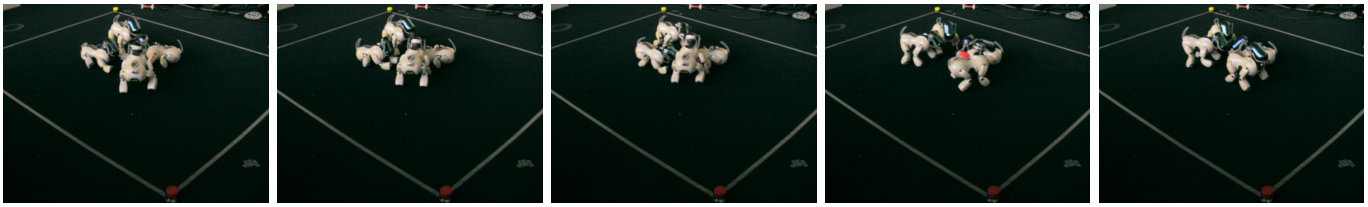
In contrast, to tackle the task of interaction detection in an unsupervised manner and without object localization / pose estimation, we combine the HDP model presented in [15], and low level features such as optical flow using [16]. Since, to the best of our knowledge, no such work has been done in the past, we evaluate the advantages and drawbacks of our HDP-based algorithm on the challenging JAR-AIBO dataset [17] and present the results.

II. Optical Flow and HDP

Due to their wide applicability, clustering techniques are applied commonly in many areas of computer vision. Unlike supervised classification methods, in clustering, class labels are not supplied. There are two categories of clustering algorithms: partitioning and hierarchical. Most of the partitioning based



(a) Coming together



(b) Playing with the ball

Fig. 1: Examples of interactions between objects. In (a), in successive frames, the dogs are coming together from the corners of the marked area. (b) shows the four dogs playing with the ball in the middle. Images are from the JAR-AIBO dataset [17].

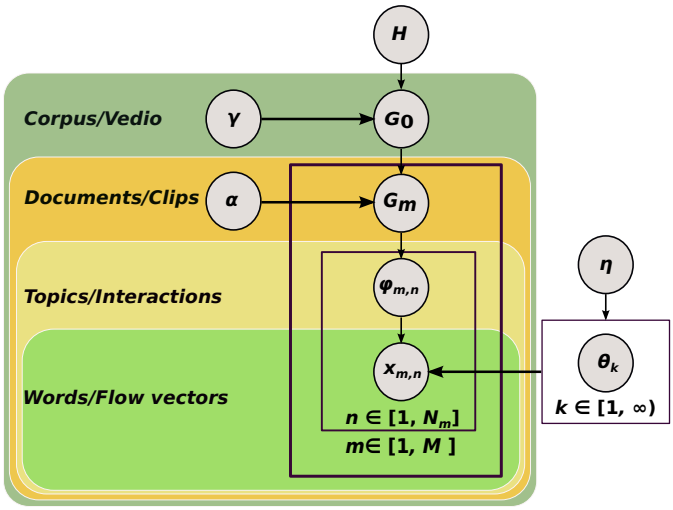


Fig. 2: HDP Model.

clustering techniques such as k-means and Latent Dirichlet Analysis (LDA), require a set of parameters, such as the number of clusters to be provided, which limits their applicability in many situations where such information is not available. In HDP, the number of clusters is deduced automatically from the data and hyper-parameters. As will be formally shown later (*cf.* 3), the number of resulting clusters in HDP can be controlled by the hyper-parameters α, γ and η. The hyper-parameters, especially η, determine the number of extracted clusters, in our case interactions.

HDP has been originally designed for clustering words in documents based on word co-occurrences. Figure 2 shows the basic HDP model. Suppose we are given an input data corpus, which is divided into M documents and each document consists of a set of words $x_{m,n}$, where $n \in [1, N_m]$. The goal of the HDP model is to cluster these words into meaningful latent structures, or *topics*.

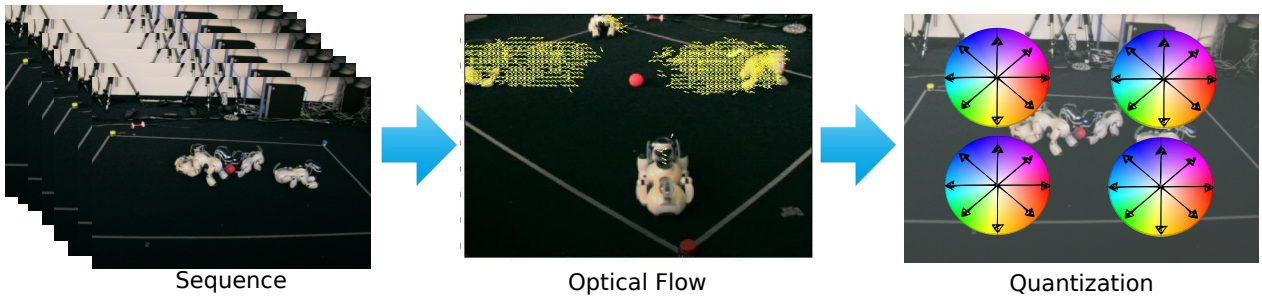
In our case, given an input video, optical flow features are extracted from each pair of successive frames using TV-L^1 algorithm [16]. The resulting optical flow is threshlolded to remove noise such as changing illumination or camera motion, and only significant motion is used for feature extraction. Subsequently, the optical flow vectors are quantized into eight directions. 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 which represent the vector of optical flow. Based on the flow values, the magnitude and direction of the optical flow can be represented as $P = \sqrt{u^2 + v^2}$ and $\theta = \tan^{-1}\left(\frac{v}{u}\right)$ respectively. Figure 3 illustrates the complete procedure.

Then a dictionary or codebook is built with all possible flow words (flow words are four-tuples, x-y co-ordinates and associated flow values). The video is divided into small equally sized clips (e.g. 10 sec) without overlapping, and each clip is represented by a bag-of-words based on the dictionary. In our framework, clips and optical flow words correspond to documents and words, respectively.

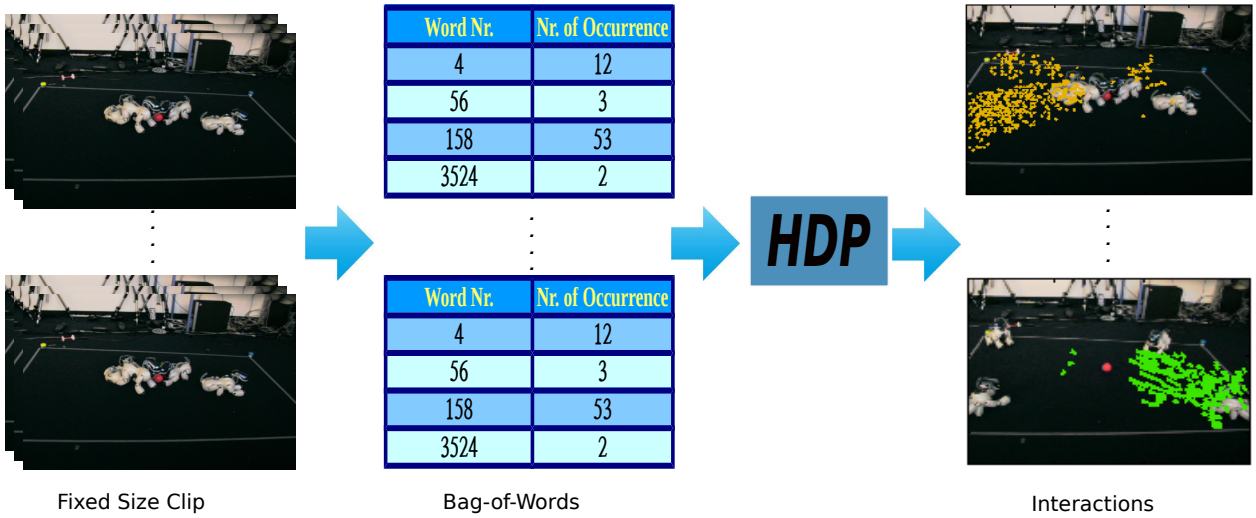
The HDP model generates the global list of interactions using a top level Dirichlet Process (DP) G_0. Then, Dp generates specific interactions G_m which are drawn from the global list G_0 for each clip. Formally, we write the generative HDP formulation as shown in 1:

$$\begin{aligned} G_0 \mid \gamma, H &\sim DP(\gamma, H) \\ G_m \mid \alpha, G_0 &\sim DP(\alpha, G_0) \quad \text{for } m \in [0, M] \end{aligned} \tag{1}$$

where the hyper-parameters α and γ are called the concentration parameters and the parameter H is called the base distribution (Dirichlet distribution). Therefore, the observed words $x_{m,n}$ are seen as being sampled from the mixture priors $\phi_{m,n}$, which in turn are seen as being drawn from a Dirichlet Process G_0 The values of mixture components drawn from θ_k



(a) Representation



(b) Clustering

Fig. 3: Illustration of the process of extracting optical flow features and arranging them according to a bags-of-words representation scheme.

Thus, the formulation of this construction can be written as,

$$\begin{aligned} \theta_k &\sim P(\eta) \quad \text{for } k \in [1, \infty) \\ \phi_{m,n} \mid \alpha, G_m &\sim G_m \quad \text{for } m \in [i, M], n \in [1, N_m] \\ x_{m,n} \mid \phi_{m,n}, \theta_k &\sim F(\theta_{\phi_{m,n}}) \end{aligned} \tag{2}$$

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

In our problem, we perform the Bayesian inference, where given the observed words, we *infer* the latent interactions. As a closed form solution for the inference process is not available for our case, we use the *Markov Chain Monte Carlo* (MCMC) approximation, specifically Gibbs sampling, using the Chinese Restaurant Franchise-based formulation. Following the formulation of [2], the conditional probability of the topic-word association for each iteration step evaluates to:

$$p(\phi_{m,n} = k, \alpha, \gamma, \eta, \theta, H) \propto (n_{m,k}^{\neg m,n} + \alpha\theta_k).\frac{n_{k,t}^{\neg m,n} + \eta}{n_k^{\neg m,n} + V.\eta} \tag{3}$$

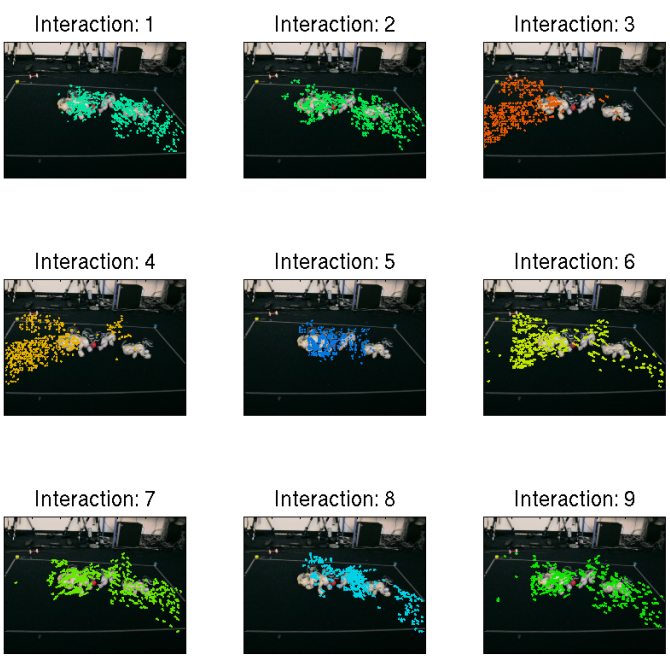
where $n_{m,k}$; $n_{k,t}$; and n_k represent count statistics of the word-topic, topic-document and the topic-wise word counts, respectively.

The superscript $\neg m, n$ means that the current word $x_{m,n}$ must be eliminated from these statistics. V is the size of the dictionary. The first part of the equation 3 reveals that the probability of assigning the current word to a topic is proportional to the number of words already assigned to that topic. This forms the basis of the clustering property of the HDP model. The second part (the probability of creating a new topic) shows that the hyper-parameters α, γ and especially η can be used to determine the number of extracted topics We also perform hyper-parameter sampling to make our framework completely data-driven. For further details on the sampling procedure, we refer to [15].

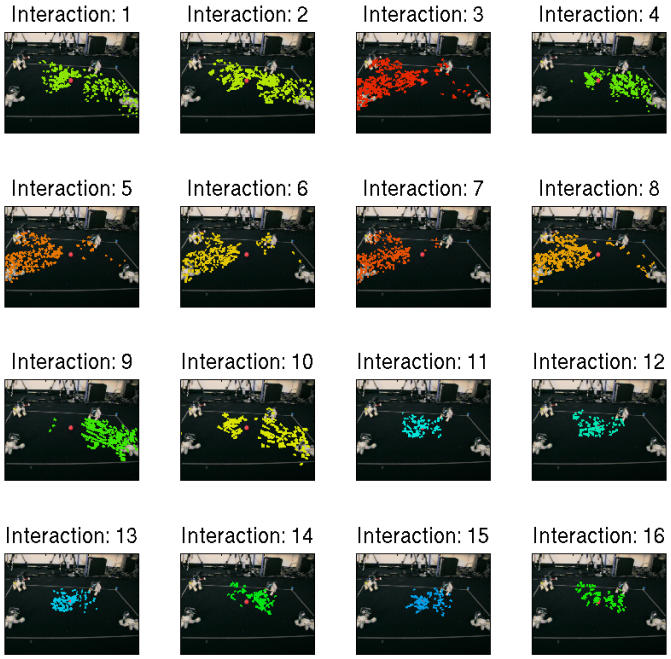
III. Experiments

A. Data-set

We use the challenging JAR-AIBO dataset [17] to evaluate our system (*cf.* Fig. 1). JAR-AIBO dataset enables us to test our system in the face of many issues such as changing illumination, changing object view and occlusions. It contains 5 sequences taken of four SonyAIBO robot dogs performing



(a) Hyper-Parameters $\alpha = 0.5$, $\eta = 0.5$



(b) Hyper-Parameters $\alpha = 1.5$, $\eta = 1.5$

Fig. 4: Some qualitative results with various extracted interactions, coded by different colors. Note the impact of varying the value of the HDP's hyper-parameters α and η on the number of extracted interactions.

actions autonomously, which are captured by six cameras at 17 fps with a resolution of 640 x 480 pixel. The camera feeds are synchronized to a frame level. In our experiments, we utilized the sequences of two cameras.

In all, we have approximately 15 interactions involving four dogs. Interactions include the dogs "converging" from all corners of the frame to the center, "playing with the ball", one or more dogs "leaving the group", one dog "walking around" the others, one dog "kicking the ball" as other dogs walk around, etc. Figure 1 shows some example frames of "coming together" and "playing with the ball" interactions. In all these, more than one dog is involved and the challenge is to detect these interactions without any prior knowledge about them.

B. Experimental setup

The optical flow extraction is performed as follows. As we mentioned above, optical flow is computed using [16]. Each frame is divided into grid cells of size 8 x 8 pixels, and quantized into eight directions. Hence, the size of the dictionary is 80 x 60 x 8. In our experiments, in order to study the effects of clip lengths on performance, the video is divided into clips of various sizes ranging from 100 to 400 frames each – corresponding to approximately 5 to 23 seconds in the videos – and constructed bags-of-words representations for them.

Though the HDP model provides possibility of assigning multiple topics per word based on its context. In this paper, we also study the effect of changing the hyper-parameters α, η where their values ranging from 0.1 to 1.5. Further, as it gets re-sampled depending on the data and the initial value does not significantly affect performance, we initialize γ=1 in all experiments.

For quantitative performance evaluation, we use the true positive rate (TPR) and the false positive rate (FPR), defined as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}; \qquad \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{4}$$

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

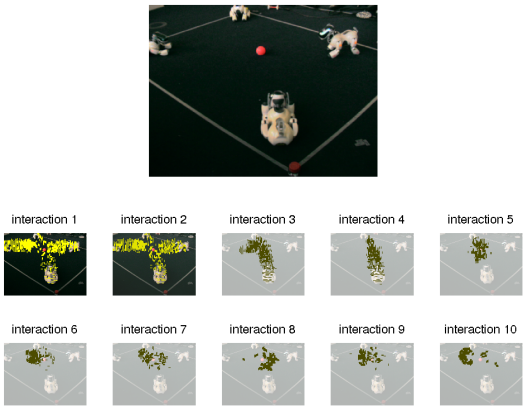
As the data-set does not contain ground truth in terms of object interactions, the video sequences were marked with clip-wise annotations regarding the interactions contained within them[1]. Then, following the procedure similar to [1], [3], the output of our system is manually mapped to the ground truth labels and the performance measures are calculated.

C. Results and Discussion

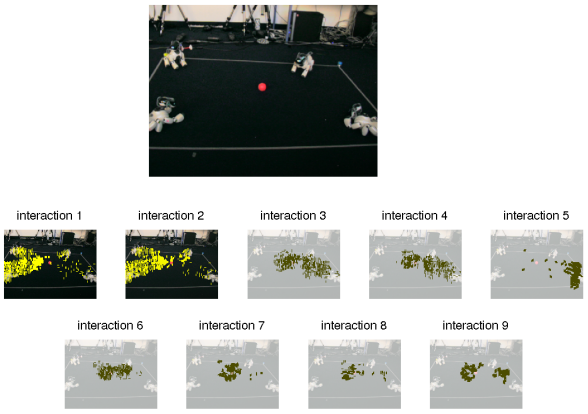
We can see some quantitative results in Fig. 4, for a video containing four dogs, where the dogs start from different corners of the frame, converge at the center, play with the ball, and finally one dog leaves the group to the bottom right corner of the frame. In Fig. 4a, interactions 1-4 and 7 represent the "converging" interaction, interactions 5 and 6 represent "playing with the ball" interaction, and interactions 8 and 9 represent the "dog leaving the group" interaction. Similar parallels can be seen in Fig. 4b.

The impact of varying the values of the HDP's hyper-parameters on the number of extracted interactions can be

[1]The ground truth will be made available as a part of the data-set



(a) View1



(b) View2

Fig. 5: Interactions extracted for multiple views. Note that despite the change of views, the interactions are still detected meaningfully.

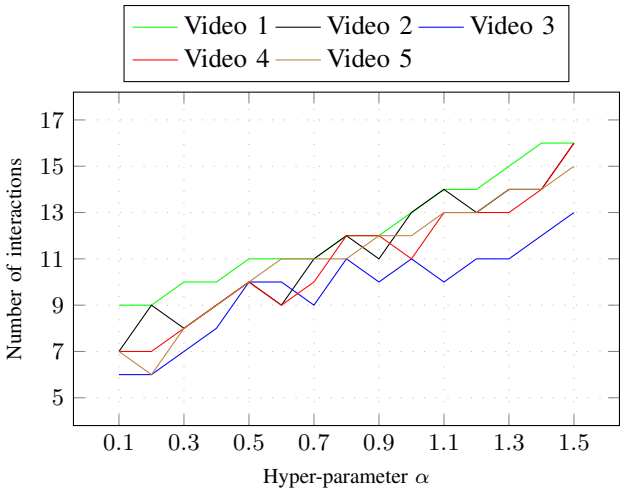


Fig. 6: Number of interactions extracted for each of the five videos as a function of the hyper-parameter α. η was held constant at 0.5.

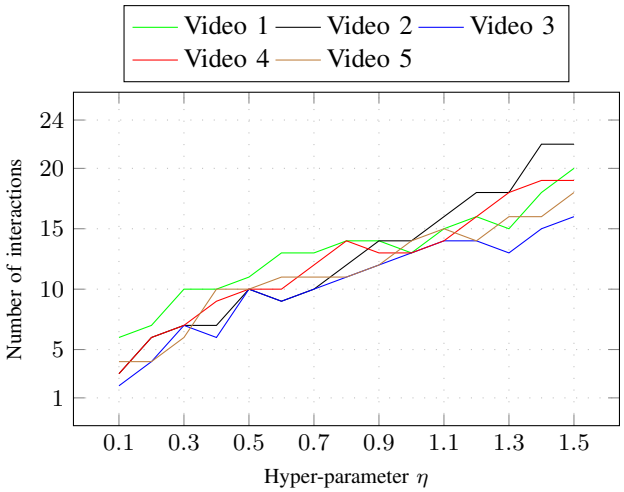


Fig. 7: Number of interactions extracted for each of the five videos as a function of the hyper-parameter η. α was held constant at 0.5.

clearly seen. The number of topics grows with increasing hyper-parameters values. Figures 4(a),(b) show that high values of hyper-parameters in situations with smaller number of interactions result in the creation of duplicate interactions. For example, in Fig. 4(a) interaction 4 is a duplicate of interaction 3, with only a few noisy flow vectors being the difference. In Fig. 4(b), this is more pronounced, where interaction 3, for example, is repeated four more times in interactions 5 to 8. Sometimes, due to high hyper-parameter values, a single interaction, such as interaction 5 in Fig. 4(a), is split into multiple smaller interactions, such as interactions 11 to 16 in Fig 4(b). This increase in the number of inferred interactions follows from the HDP inference process, where higher values of hyper-parameters imply a higher probability of drawing new interactions, and the presence of noisy features compounds the effect.

Quantitatively, Fig. 6 and 7 show the variations in number of interactions extracted as a function of the two hyper-parameters α and η respectively. Clearly, the number of interactions extracted increases with the hyper-parameter values. However, it is interesting to note that the range of the number of interactions is larger in the case of hyper-parameter η. This is due to the fact that, being the parameter controlling the probability of generation of new interaction directly, it has larger effect on the resulting number of interactions. Therefore, a user can provide prior knowledge about the number of interactions through setting the hyper-parameters accordingly.

Figure 5 shows the extracted interactions for two different views. It can be clearly observed that despite a change in viewpoint, the extracted interactions are stable.

Table I shows the quantitative evaluation of our experiments. As can be observed, View 1 with frame size 400 has

TABLE I: Results of the HDP algorithms for two views. The effects of clip-sizes on the performance can be clearly observed.

View	**View 1**			**View 2**		
Clip Size (Frames)	100	250	400	100	250	400
True Positive Rate%	77.14	78.60	**82.35**	77.14	78.57	76.47
False Positive Rate%	32.95	41.70	32.00	52.13	51.16	**31.81**

achieved the high value of TPR 82.35 % also lowest value of FPR 31.81 %, whereas the lower frames per clip values result in worse performance. This is likely due to the fact that, smaller clip sizes split the interactions into many sub-interactions, and consequently, performance suffers.

IV. Conclusions and Future Work

The aim of this paper was to show how low-level optical flow features combined with a Hierarchical Dirichlet Process can be used to extract meaningful interactions in video sequences in an unsupervised manner. We compared the effect of several values of HDP's hyper-parameters, and the qualitative results obtained from the various experiments performed on the challenging JAR-AIBO dataset were promising.

Future research topic will be a comparison of different features combined with Hierarchical Dirichlet Processes and other similar topic models. Furthermore, in order to reduce testing time during deployment, we can use a step-wise combination of generative and discriminative methods, following the approach of [3]. Use of other clustering schemes, such as DP-means of [18] also seems interesting.

Acknowledgments

The authors thank Golenur Khanam for her contribution in compiling the results presented in this work. The work was partially funded by Deutscher Akademischer Austauschdienst (DAAD) and ZEISS.

References

[1] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, March 2009.

[2] M. Venkata Krishna, M. Körner, and J. Denzler, "Hierarchical dirichlet processes for unsupervised online multi-view action perception using temporal self-similarity features," in *Seventh International Conference on Distributed Smart Cameras (ICDSC)*, 2013, pp. 1–6.

[3] M. V. Krishna and J. Denzler, "A combination of generative and discriminative models for fast unsupervised activity recognition from traffic scene videos," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014, pp. 640–645.

[4] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in tv shows," 2010.

[5] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 9–16.

[6] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 10, pp. 1775–1789, 2009.

[7] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1331–1338.

[8] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–16.

[9] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 601–614, 2012.

[10] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1593–1600.

[11] G. Yang, Y. Yin, and H. Man, "Human object interactions recognition based on social network analysis," in *Applied Imagery Pattern Recognition Workshop: Sensing for Control and Augmentation, 2013 IEEE (AIPR*. IEEE, 2013, pp. 1–4.

[12] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 17–24.

[13] B. Yao and L. FeiFei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1691–1703, 2012.

[14] S. Park and J. Aggarwal, "Semantic-level understanding of human actions and interactions using event hierarchy," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. IEEE, 2004, pp. 12–12.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[16] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Pattern Recognition*. Springer, 2007, pp. 214–223.

[17] M. Körner and J. Denzler, "Jar-aibo: A multi-view dataset for evaluation of model-free action recognition systems," in *New Trends in Image Analysis and Processing–ICIAP 2013*. Springer, 2013, pp. 527–535.

[18] B. Kulis and M. I. Jordan, "Revisiting k-means: New algorithms via bayesian nonparametrics." in *International Conference on Machien Learning (ICML)*, 2012.