Exploration of Spatial and Temporal Modeling Alternatives for HOI

Rishabh Dabral Srijon Sarkar Sai Praneeth Reddy Ganesh Ramakrishnan
Department of Computer Science, IIT Bombay
{rdabral, srijon, praneeth20, ganesh}@cse.iitb.ac.in

Abstract

Human-Object Interaction detection from a video clip can be considered as a special case of video-based Visual-Relationship Detection wherein the subject must be a human. Specifically, it involves detecting the humans and objects in the clip as well as the interactions between them. Conventionally, the problem has been formulated as a space-time graph inference problem over the video clip features. In this work, we explore alternate spatial approaches for detecting Human-Object Interactions. We consider a hierarchical setup that decouples spatial and temporal aspects of the problem and analyse the impacts of a variety of design choices for the spatial networks. Particularly, to capture spatial relationships in the scene, we analyze the effectiveness of the traditionally used Graph Convolutional Networks against Convolutional Networks and Capsule Networks. Unlike current approaches, we avoid using ground truth data like depth maps or 3D human pose during inference, thus increasing generalization across non-RGBD datasets as well. We demonstrate a comprehensive analysis of the exploration, both quantitatively and qualitatively, while achieving state-of-the-art results in human-object interaction detection (88.9% and 92.6%) and anticipation tasks of CAD-120 and competitive results on image based HOI detection in V-COCO dataset, setting a new benchmark for visual features based approaches.

1. Introduction

Visually understanding a scene as depicted in an image or video is one of the fundamental problems of Computer Vision. It builds on top of existing sub-problems like object detection, activity recognition, saliency estimation, etc. Humans are, arguably, one of the most important entities to understand. As such, understanding human activities and the way humans interact with the surrounding environment becomes a crucial and interesting problem to solve. In this work, we investigate this problem of identifying Human-Object Interactions from videos. Given a video stream, the goal is to identify the objects interacting with the humans while also estimating the kind of interaction, *eg.*, holding the cup, placing the bowl, moving the furniture, *etc.* The availability of such information can be crucial in higher order tasks, such as human-motion prediction, scene generation, etc. Furthermore, such information has the potential to facilitate downstream applications such as unmanned supermarkets, surgery documentation, robotics, *etc.*

In this work, we investigate the possible solutions to the HOI-from-video problems, with special focus on spatial model design - the relative ordering of the subjects and objects in the scene and its effects on interaction detection. There have been a significant number of works that model the spatial relationships in the form of Graphs. The subjects, object and relationships, typically act as nodes while the edges correspond to the potentials indicating the strengths of associations. The graphs may be processed using message-passing alorithms [21] or Graph Convolutional Networks [47, 38] (GCNs). As an alternative, CNNs have also been used for spatial models in prior works [52] with the model being fed the object/human regions as inputs.

Recently, Capsule Networks [11, 43, 42], with multiple variations, have been proposed as models capable of inherently being able to reason about the spatial information in the scene. Furthermore, past works have demonstrated their ability to learn the part-to-whole relationships in the scene without having to memorize the same from thousands of data points. We hypothesize that this ability makes Capsule Networks a potentially suitable spatial network to capture the relationships between the objects and humans in the video effectively.

There has been a significant amount of research on HOI with images [56, 38, 26, 52, 18], thanks to the availability of V-COCO [13] and HICO [2] datasets. However, learning human-object interactions within videos is challenging and relatively less explored owing to multiple reasons. *Firstly*, it requires the model to account for the changing orientations of objects in the scene with respect to the humans. This makes it difficult to extend the image-based approaches that use the RoI features of the union of human and object to the video setting. Secondly, the unavailability of large scale

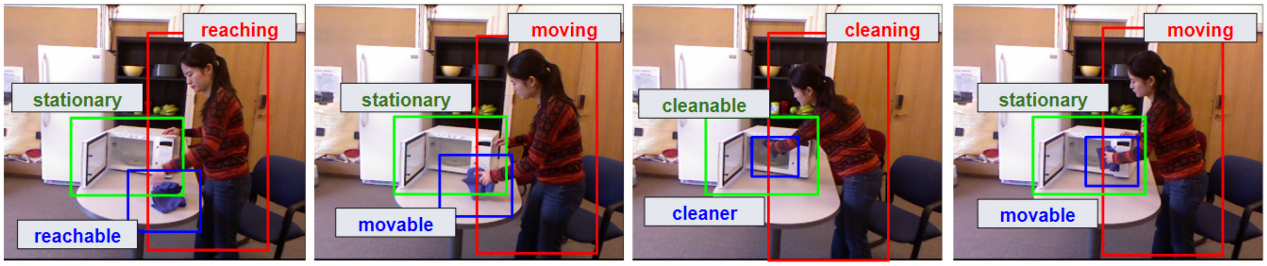


Figure 1. Illustration of human-object interaction detection in video (CAD-120) setting

video datasets (except CAD-120 [20]) makes it difficult to train an HOI model that is generic, and performs well for in-the-wild videos. *Finally*, the interaction definitions tend to become confusing when defined for a video, *e.g.*, *placing* vs. *moving* vs. *reaching*, *opening* a jar vs. *closing* a jar, etc. In spite of these challenges, videos allow for exploiting temporal visual cues that are, otherwise, absent in images.

Most existing methods are designed to work in either the image setting [56, 26, 52], or the video setting [21, 16] but not both. Recently, Qi *et. al.* [38] proposed a graph-parsing based method that caters to both the settings. While the method indeed achieves state-of-the art results in video setting, it does so by using carefully designed and pre-computed hand-crafted features such as SIFT [36] transforms, object centroids, 3D poses, object depths, *etc.*, which were originally proposed in [20]. It is worth noting that these features were derived from the ground-truth data provided in the CAD-120 dataset. Thus, it is expected that using ground-truth based features for estimating HOI would not allow the method to perform equally well on in-the-wild videos because such features may either not be available (3D pose) or may be noisy and inconsistent across frames (object bounding boxes, centroids, *etc.*).

With these caveats in mind, we work on a hybrid approach that argues about the spatio-temporal relationships between the humans and the objects at multiple levels of hierarchy. The method is designed to infer from videos and *does not* rely on hand-crafted features. We use pure visual features derived from a re-trainable off-the-shelf network to represent the inputs to the network and demonstrate strong performance on the CAD-120 dataset. Specifically, we use a two-level architecture which, i) performs spatial embedding extraction from the video and learns temporal reasoning functions at the frame level, followed by ii) a segment level temporal network which learns inter-segment temporal cues from previous segments, for regressing the human subactivities and object affordances. This choice of using de-coupled networks for spatial and temporal modeling allows us to experiment with two spatial models: Capsule Networks and Graph Convolutional Networks. Both the networks have the potential to argue about complex spatial relationships, when provided with suitable inputs. The temporal functions rely on sequence models such as RNNs and LSTMs which are designed to learn the temporal relationships between human-object pairs across the video.

Despite not using the ground truth based pre-computed features and in spite of the small amount of data available for training from videos, our visual input based model achieves state-of-the-art performance on subactivity, affordance detection tasks, setting a strong baseline for the future of such methods. When used with the segment level pre-computed features, the segment-level temporal model of our proposal performs at par with the state-of-the-art methods. Finally, despite being designed for video-based tasks, our method also demonstrates competitive performance on the V-COCO dataset that corresponds to the image setting. In the supplementary material, we qualitatively illustrate the improved performance of our trained models, vis-a-vis state-of-the-art spatio-temporal model [38] on several 'in-the-wild' videos and images. As anticipated, use of ground-truth based features does not help [38] generalize to settings that are significantly different from the training data.

In summary, we make three contributions in this paper: *First*, we propose a generalizable, multi-level method for identifying Human-Object Interactions from videos. *Second*, We analyze multiple architectures for modeling the spatial relationships between the objects and the humans in the scene. *Third*, we show how our method naturally lends itself to static, image-based settings.

2. Related Work

A key element of scene understanding is human perception and human cognition. Human perception involves inferring the physical attributes about the humans such as in the case of detection [7, 60, 54, 41], pose recognition [33, 6, 30, 12, 51, 5], shape identification [17, 34], clothing recognition [35, 25], *etc.*. On the other hand, human cognition seeks to reason about the finer details relating to human behaviour [40, 31], human activity [8, 58, 29, 32, 59], human-object visual relationship detection [45, 39, 48, 28, 50, 44], and human-object interactions [49, 38, 52, 45, 39, 48, 28, 50, 44]. Human-Object Interaction detection has been a well researched problem.

In this section, we discuss the existing literature from two broad viewpoints: static (images) and dynamic (videos).

HOI from images: A significant amount of work [58, 15, 57, 8] in this area pre-dates the deep learning advent. However, deep learning based methods [46, 1, 61, 3, 14, 53, 10, 37], bolstered by the availability of large amounts of in-the-wild training data [13, 2] have lead to significantly improved performance in HOI detection. Among such methods, Li *et. al.* [26] proposed to learn the knowledge about the *interactiveness* between the humans and object categories from HOI datasets and use this knowledge as a prior while performing HOI detection. Several methods have attempted to leverage the human pose information in their pipelines. Wan *et. al.* [52] propose a pose-aware network architecture that employs a multi-level feature strategy. Likewise, Xu *et. al.* [55] use the human pose features in conjunction with the gaze estimates to discover human intentions, which are then used for HOI detection. Since the HOI problem is well-suited for graph-based representations, Graph Convolutional Networks have been regularly used to model the interactions. In this line of work, Xu *et. al.* [56] propose to deal with long-tail HOI categories by modeling underlying regularities among verbs and objects. They do so by constructing a knowledge graph and enforcing similarity of graph embeddings derived from a GCN with visual feature embeddings derived from a CNN using a triplet-loss. Qi *et. al.* [38] propose GPNN, a method that uses an iterative message passing framework on a parse graph comprising of verbs and objects as nodes. Our work is inspired by graph based methods in that we represent humans and objects as graph nodes and learn their interactions based on the image-based node features.

HOI from Video: The HOI labels predicted in this task are typically indicative of an activity spanning over a period of time. Therefore, utilizing temporal cues in a video setting is naturally expected to provide important insights on the interactions and thereby benefit the HOI detection. Albeit less, there have also been significant contributions towards research on HOI detection in videos, mostly on the CAD-120 dataset. Koppula *et. al.* [20] proposed the dataset and introduced an MRF base formulation for handling spatio-temporal requirements. The authors hand-crafted a set of features for humans (pose, displacement of joints, *etc.*) and objects (3D centroids, transforms of SIFT matches between adjacent frames, *etc*). Instead of being used at the frame-level, these features, put together, represented a video segment as a whole. Since then, most existing methods (deep learning and traditional methods alike) work on the same segment level features. Qi *et. al.* [38] extend their GPNN method for videos and construct a parse graph for every video segment using the segment level features to initialize the node and edge features in their parse graph. Likewise, Jain *et. al.* [16] design a spatio-temporal graph for performing structured predictions on a video consisting of multiple segments. Kopulla *et. al.* [21] present ATCRF - a CRF based approach that models anticipatory trajectories of objects and humans.

While there have been remarkable improvements over the years, we submit that there are two major areas for improvement. Firstly, avoiding the usage of such hand-crafted features, since the above approaches limit the scope for in-the-wild HOI detections. More often than not, the 3D poses or 3D centroids of objects (used as features) are either not available or are too erroneously estimated to be simply plugged into a model trained on hand-crafted features. Secondly, all the existing methods model temporal relations only between multiple *segments* of a video. This may be, partly, because the hand-crafted features discussed above are defined for a segment as a whole. We believe that there is scope for exploring temporal cues even at a more fine-grained level, *viz.*, frame-level. Using image-based features facilitates the same.

We, therefore, propose an approach to model HOI relevant spatial-structures from every frame of a segment and further design a temporal aggregation regime using these frame level structures. Again, such aggregation strategies have proved to be extremely effective for problems such as image labelling [23], entity-linking [24, 22] and text classification [4]. Deep-learning based computer vision models have enough representation power to be able to extract meaningful visual features from images or videos. Thus, our primary intent is to construct a model which can effectively learn hierarchical HOI embeddings at a fine-grained frame level as well as at a coarser segment level, using only visual attributes, and set a new baseline for human-object interaction detection in videos.

3. Our Approach

In this section, we present our approach for HOI detection on video. The HOI information in the videos is dealt with at two levels of granularity. The first, and the coarser, granularity corresponds to viewing the video as a sequence of segments, with each segment representing an atomic interaction. For example, a video may include a sequence of segments such as: *reaching* for a jar, *opening* the jar, and *placing* the jar back. The second, and finer, granularity corresponds to dissecting each segment into its constituent frames. Lastly, the visual features at the frame level provide crucial spatial cues about the possible interactions. Our architecture leverages these constructs and is outlined in Figure 2.

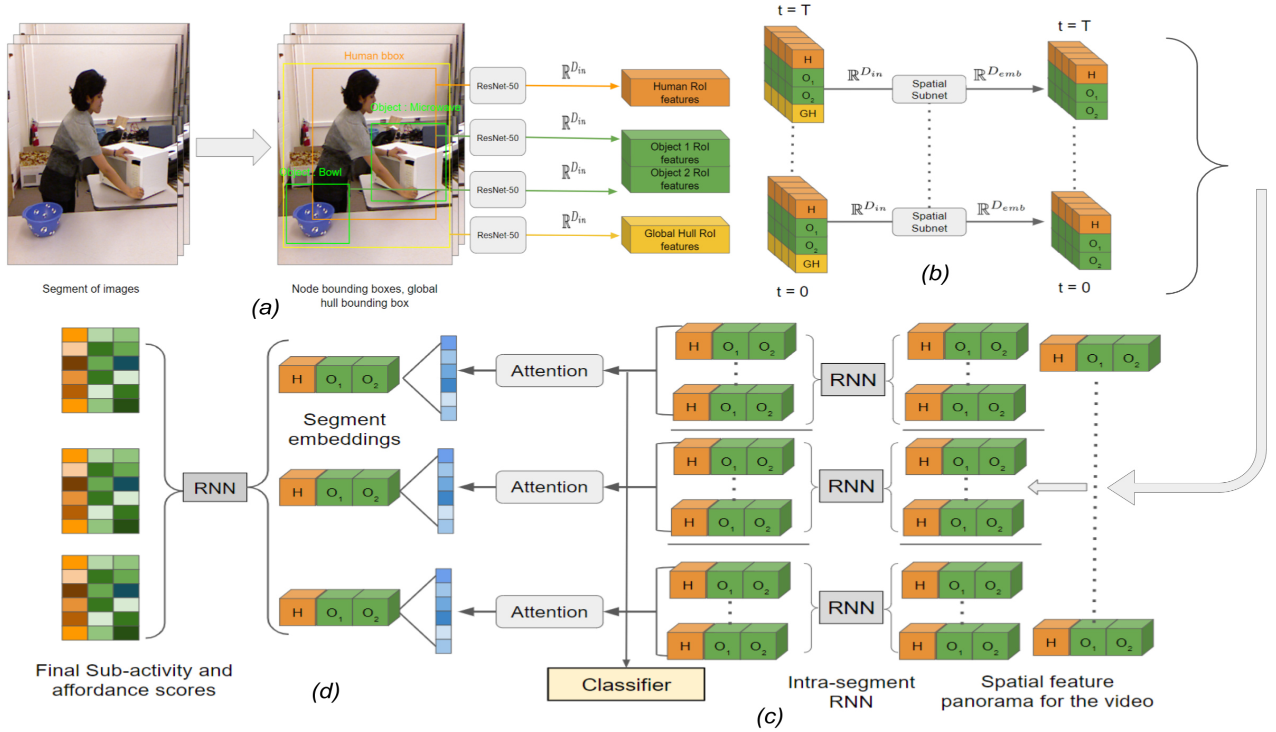


Figure 2. Overall pipeline of our network. Given an input video segment with T frames and bounding box coordinates of the humans and objects in every frame, we (a) first extract the visual features from ResNet-50. (b) These features are then processed in a per-frame fashion by a Spatial Subnet. (c) The graph structure is disentangled and temporal cues between frames in a segment are learnt from spatial features. (d) The frame-wise features are summarised into segment embeddings using attention mechanism and refined using inter-segment relations, to regress the human subactivities and object affordances. Best viewed in colour and/or digitally with zoom.

3.1. The Proposed Learning Framework

Given an input video $\mathcal{I} = \{I_1, I_2, \ldots, I_T\}$ consisting of T frames such that the video includes a single human and N objects, our task is to regress human subactivities (placing, opening, *etc.*), $H = \{H_0, H_1, \ldots, H_M\}$ for the human and object affordances (placable, openable *etc.*), $O = \{O_{0,0}, O_{0,1}, \ldots, O_{N,M}\}$ for each of the N objects and M segments in the video. To this end, we propose a pipeline consisting of three stages: (i) the spatial subnet, (ii) the frame-level temporal subnet, and (iii) the segment-level temporal subnet.

The spatial subnet feeds on an input frame I_t and learns a set of embeddings $\phi_t \in \mathbb{R}^{D_{emb}}$ for each human and $\theta_{n,t} \in \mathbb{R}^{D_{emb}}$ for each object. These per-frame, spatial embeddings are then fed to the *frame-level* temporal subnet that churns out the corresponding spatio-temporal embeddings, $\Phi_t \in \mathbb{R}^{D_{emb}}$ and $\Theta_{n,t} \in \mathbb{R}^{D_{emb}}$, while also providing initial estimates of H_m and $O_{n,m}$, where m corresponds to the segment index, and n corresponds to the object index. The frame-level spatio-temporal embeddings are then consolidated for each segment using an attention mechanism to produce A^{Φ}_{m} and $A^{\Theta}_{n,m}$, and passed on to *segment-level* temporal subnet that produces the final outputs for the subactivity and affordance estimates.

Traditionally, previous works have derived spatial features not from the raw images, but from the ground-truth data like depth of the objects, pose of the human and objects, *etc.* It is easy to see that such a construction prohibits its use on any video for which depth information is unavailable. In this work, we do not use the depth-based features and only rely on RGB inputs. Next, we now elaborate on each step of the pipeline.

3.2. Spatial Subnet

As just discussed, the sole job of the spatial subnet is to learn features relevant to the spatial ordering of the objects and the human. Formally, the Spatial Subnet, S transforms the features corresponding to the t^{th} frame as $\phi_t = S(x_{v,t})$ if v is a human node and $\theta_t = S(x_{v,t})$ if v corresponds to an object node. At the end of the Spatial Subnet, the network produces an intermediate feature set in $\mathbb{R}^{T\times(N+1)\times D_{emb}}$ space. To this end, we investigate three variants of the spatial subnet based on their ability to effectively model the spatial relationships.

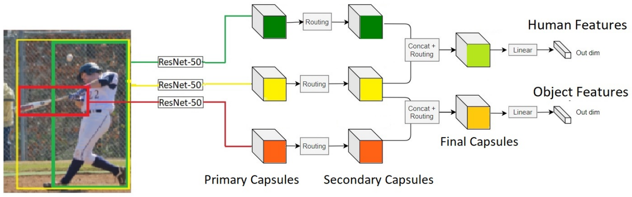


Figure 3. Architecture of Capsule Spatial Subnet. The object and human features, derived from primary and secondary capsules, are concatenated with the features of the global hull I_{gh} (yellow bbox). The subnet outputs spatial embeddings which are then processed by the temporal subnet.

Capsule Network: Capsule Networks [11, 43] have been proposed as an alternative to conventional CNNs for inherently reasoning about the spatial organization and rotation invariance of the scenes without having to memorize the same across a large dataset. This, arguably, fits in the requirements of an ideal spatial subnet. To this end, we propose a variant which uses capsule networks for spatial subnet. The schema of the Capsule Spatial Subnet is described in Fig. 3. For each object O (and human) in frame t, the capsule net is subjected to two inputs: the object (or human) bounding box RoI features $I_{o,t}$, and features of the global hull $I_{gh,t}$. The global hull is the super-bounding-box that includes the human and all the objects. This design choice is motivated by the requirement that network must be provided with enough image context. We do the same for the human node. The input features are extracted by passing the corresponding image crops through a ResNet network upto the third last layer.

The network consists of a primary capsule layer followed by a secondary capsule layer. The first layer creates capsules out of the visual features individually, before the global hull features are appended. At this point, the global hull features and features from human/object hulls are appended in the following way. The human node capsules input for the final layer are concatenation of human RoI features and global hull RoI features. The object node capsules are concatenation of object RoI features, human RoI features and features from human-object hull. We use a 1×1 conv layer to preprocess the ResNet features of dimension 1024×14×14 into embeddings of dimension 256×14×14. These embeddings are individually converted into primary capsules. Routing is done on these capsules to produce secondary capsules. The final layer of capsules are flattened and passed through a linear layer to get the output of dimension 1024. We perform capsule routing using the Variational Bayes Routing algorithm [42].

ConvNets: A direct substitute of the proposed Capsule Network architecture in Fig.3 is by drop-in replacement of the capsule layers in the netwok by convolutional layers. Specifically, we subject the incoming image features corresponding to objects, humans and the global hull to multiple convolutional layers which are then fused together, followed by more convolutional layers. The embedding size, D_{emb}, remains the same.

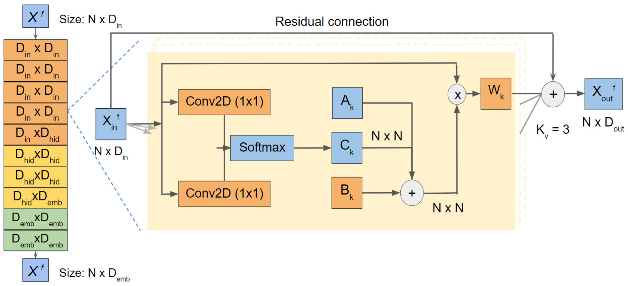


Figure 4. Architecture of GCN Spatial Subnet. Each block augments the adjacency matrix by a learnable correction, B, and a data-dependent course-correction, C. A residual connection is added to facilitate faster training of the model

Graph Convolutional Network: The spatial subnet can also be modeled by a Graph Convolutional Network (GCN) which lends itself naturally to the task at hand. We define the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the nodes $\mathcal{V} = \{1, 2, \ldots, N+1\}$ correspond to N objects and one human and $\mathcal{E} = (p, q) \in \mathcal{V} \times \mathcal{V}$.

We extract the node features $x_{v,t} \in \mathbb{R}^{D_{in}}$ corresponding to the v^{th} node (human/object) of the t^{th} frame by feeding the corresponding image crop $I_{v,t}$ to an off-the-shelf feature extractor F. Formally, $x_{v,t} = F(I_{v,t})$. The edge weights are initialized to be 1 for human-object edges and 0 for the rest. The adjacency matrix is dynamically learnt while training the Spatial Subnet. Unlike Capsule Networks, a major challenge in GCN based formulation is to account for variability in the number of nodes across segments in a video. For example, a video may include the following segments: picking a bowl (1 object), moving the bowl (1 object), putting the bowl in the microwave (2 objects). Typically, this number varies from two nodes to six nodes.

To alleviate this issue, the network is designed to learn course-corrections to the adjacency matrix. As depicted in Figure 4, every graph-convolution layer is followed by an update of the adjacency matrix which involves addition of the following two refinement components to the base adjacency matrix A. The first is a learnable additive matrix, B, that is learnt during the training process. The second is a data-driven additive matrix, C, that is estimated uniquely for every input. Note, that this formulation has overlaps with a parallel proposal in [47]. However, unlike [47], we do not operate in time dimension at the GCN level.

3.3. Frame-level Temporal Subnet

Once the per-frame spatial features for the graph are extracted, (in the case of video data such as CAD-120) we process the graph features in time dimension, thus providing a feature-panorama of the entire segment. As discussed earlier, temporal reasoning occurs in two granularities - frame level and segment level. It is at this stage that we disintegrate the graph structure of the network and construct individual feature sets for each node, aggregated over time. These frame-level embeddings are subjected to a bidirectional Recurrent Neural Network (RNN) which produces two outputs for every frame:

For human nodes, given the input embeddings $\phi_t \in \mathbb{R}^{T \times N \times D_{emb}}$, the frame-level bidirectional-RNN outputs the estimates of human subactivity, $H_{m,t}$, and updates the recurrent embedding, $\Phi_t \in \mathbb{R}^{D_{emb}}$ for frame t in segment m. Note, that while the learnt embeddings are further fed into the segment-level subnet, we also use them to classify subactivities and affordances for each frame to facilitate stronger supervision. For object nodes, we concatenate human node features along with the object node features and feed it to the frame-level RNN which outputs the estimates of object affordances $O_{n,m,t}$ and updates the corresponding recurrent embeddings, $\Theta_{n,t} \in \mathbb{R}^{D_{emb}}$

The aggregated activity and affordance classification scores at frame level are computed by taking a summation of the sequential frame-wise scores output by the RNN. Formally, the frame-level subactivity prediction can be written as: $H_m = softmax(\sum_t H_{m,t})$

Loss Functions: Both the classifiers are subjected to standard Cross-Entropy losses $\mathcal{L}_h$ and $\mathcal{L}_o$ corresponding to human subactivities and object affordances, respectively. The overall loss is a weighted sum of the two losses and can be written as:

$$\mathcal{L} = \mathcal{L}_h + \lambda \mathcal{L}_o$$

3.4. Segment-level Temporal Subnet

The previous subnet learns intra-segment temporal relations, but does not utilize the temporal information from the previous segments of the video, thus lacking wider context. The segment-level subnet learns inter-segment temporal cues by leveraging the context from previous segments of the video. We use another RNN to model these relations.

Attention Mechanism: The input to the segment-level RNN is a sequence of embeddings, A_m^{Φ}, corresponding to each segment for human nodes. We extract A_m^{Φ} by subjecting the frame-level embeddings, $\Phi_{m,t}$ to an attention network that produces a single embedding for a segment. Formally, $A_m^{\Phi} = \sum_t a_t * \Phi_{m,t}$, where a_t are the attention weights produced by a Multi-Layered Perceptron (MLP). Similar construction follows for the derivation of A_m^{Θ}.

The summarized sequence of segment embeddings is finally processed by an RNN, to leverage temporal dependencies from the previous segments for predicting human subactivity and object affordances for the current segment.

We use the same loss functions for classifiers at both frame-level and segment-level.

Table 1. A comparison of our approach with the existing methods. Note that unlike ours, all the methods that we compare with have been trained using hand-crafted features

	F1 Score in %	
Method	Sub-activity	Affordance
ATCRF [21]	80.4	81.5
S-RNN [16]	83.2	88.7
S-RNN (multi-task) [16]	82.4	91.1
GPNN [38]	88.9	88.8
Ours with Capsule Net	88.8	84.2
Ours with GCN	**88.9**	**92.6**

3.5. Implementation Details

We now discuss implementation details from two vantage points: model and training.

Model: Since the number of frames in a video segment may vary significantly, we uniformly sample a fixed number of frames, T, from the segment (for our experiments on CAD-120 dataset, we use T=20). We extract the RoI crops from each frame and reshape them to a fixed size of 224 × 224 × 3 (input dimension for ResNet). For our experiments, we use a pre-trained ResNet-50 backbone, which produces 2048 dimensional embeddings (for the GCN), and 14 × 14 × 1024 dimensional embeddings (for the CapsuleNet) for each node. In order to incorporate the information on positioning of humans and objects, we append normalized bounding box coordinates of human/objects to their respective visual node features. In the frame-level temporal subnet, we use a two-layered bidirectional RNN and three-layered unidirectional RNN network in the segment-level temporal subnet.

Training: We use the PyTorch deep learning framework for implementation. During training, we set $\lambda = 2$ for the overall loss. We use the Adam [19] optimizer with initial learning rate of 2×10^{-5}, learning rate decay factor of 0.8, and decay step size of 10 epochs. We train the network for a total of 300 epochs on Nvidia RTX 2080Ti GPU. We performed a hyper-parameter sweep to empirically obtain these configurations. Capsule Pooling [9] is performed for the CapsuleNet to be able to train it on a single GPU. The entire model is trained in two steps. Firstly, the model up to frame-level temporal subnet is trained by aggregating classification scores from the T frames of the segment. Finally, the entire model is trained in an end-to-end fashion, after initializing the parameters from the pre-trained frame-level model.

4. Experiments

We evaluate our model for the task of Human-Object Interaction detection on two datasets, *viz.*, i) CAD-120 [20] ,and ii) V-COCO [13].

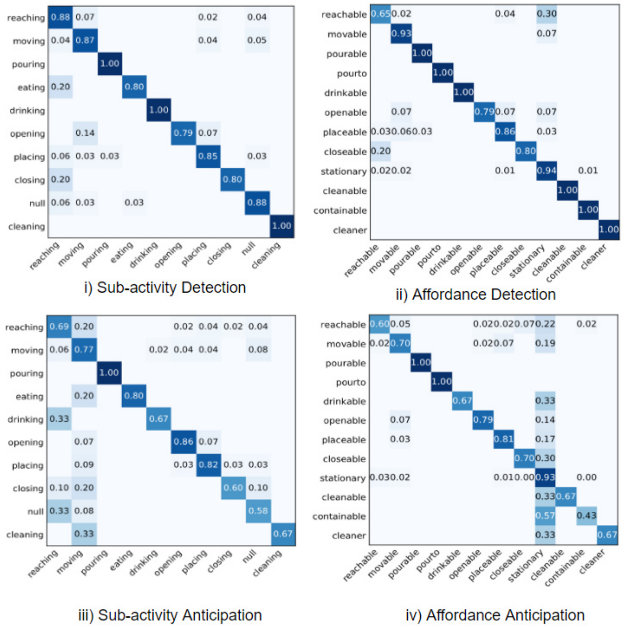
CAD-120 The CAD-120 dataset is a video dataset with

120 RGB-D videos of 4 subjects performing 10 daily indoor activities (*e.g., making cereal, microwaving food*). Each activity is a sequence of video segments involving finer-level activities. In each video segment, the human is annotated with an activity label from a set of 10 sub-activity classes *(e.g., reaching, pouring)* and each object is annotated with an affordance label from a set of 12 affordance classes (*e.g., pourable, movable*). The frame-length of each segment ranges from 22 to a little over 150 frames.

The metrics used for evaluating on the human-object interaction tasks of CAD-120 dataset are: i) sub-activity F1-score, and ii) object affordance F1-score computed for human sub-activity and object affordance classification.

V-COCO Crafted as a subset of the MS-COCO [27] dataset, V-COCO is an image dataset that provides annotations of Action labels for edges between human and object. There are 26 action classes.

4.1. Evaluation on the CAD-120 dataset:



i) Sub-activity Detection

ii) Affordance Detection

iii) Sub-activity Anticipation

iv) Affordance Anticipation

Figure 5. Confusion matrices for human-object interaction detection setting – (i), (ii) – and anticipation setting – (iii), (iv) – on CAD120 dataset. It is worth noting that most of the confusion occurs in visually similar categories, e.g. closing vs. reaching and opening vs. moving

We evaluate the performance of our model at both frame-level and segment-level, using both variants of spatial subnet. We tabulate the results of our approach in Table 1. As the numbers suggest, we achieve state-of-the-art performance with sub-activity detection F1 score of **88.9** and affordance detection F1 score of **92.6** with GCN and 88.8 and 84.2 for subactivity and affordance detection tasks with Capsule spatial net. This suggests that out of the two variants of spatial subnet, GCN module performs better than Capsules. We believe the reason to be the ability of graphs to better model the cases of multiple object scenarios in the scene. This comes from the fact that while the GCN, by construction, deals with multiple nodes (and objects), the same is not true for a vanilla Capsule network. The only context that the capsule receives is the global hull of the objects and human, which is not distinctive enough for the cases when there are more than one objects. This hypotheses is mildly corroborated by our results in V-COCO dataset wherein the Capsule networks perform better than GCN, possibly because every human deals with a single object most of the times.

Confusion Matrix: The confusion matrices for both detection and anticipation tasks using the GCN spatial net are displayed at Figure 5. Every row of a confusion matrix indicates the prediction distribution of various node samples of that ground truth class. From the confusion matrix for affordance detection, it is evident that most of the false predictions of object nodes are due to misinterpretation of object as stationary.

4.2. Evaluation on V-COCO dataset

Although our method is designed to leverage temporal cues within a video setting, we test our method on V-COCO dataset by setting T = 1. We observe the role mAP score of 47.26 while using Capsule spatial subnetwork for spatial learning, and role mAP of 38.28 using Spatial GCN, which, although not close to the state-of-the-art method [52] (52.0), achieves reasonable performance without bells and whistles. Moreover, we achieve a better performance using capsule networks for spatial learning than using a spatial graph convolution network. This might be attributed to the different formulations of HOI for CAD120 and V-COCO. HOI detection in V-COCO is done separately for each human-object pair, which implies that there are only two nodes, one for object and the other for human. This setting benefits the learning process of capsule networks, and is not well suited for graph convolutions, as there are very few nodes in the graph modeling. Finally, the ConvNet variant of the Capsule architecture achieves a role mAP of 44.8 compared to 47.26 of Capsule Nets.

4.3. Qualitative Evaluation

We provide some qualitative evaluation of our method using GCN spatial net on CAD-120 dataset in Figure 1. Figure 6 demonstrates some positive and negative cases of detection of edge action labels of human-object pairs for test images on V-COCO. The reader is referred to supplementary material for in-the-wild results on videos.

Table 2. Ablation experiments of the impact of design choices on subactivity and object affordance detection. Seg-RNN refers to segment-level RNN and vanilla GCN refers to GCN without adjacency matrix refinement.

Experiment	Human Subactivity	Object Affordance
Ours w/o spatial model	61.5	78.6
Ours w/o seg-RNN with MLP for frame-level temporal learning	84.1	85.0
Ours w/o seg-RNN w/o appending human node features to object nodes	85.2	84.6
Ours w/o seg-RNN	85.9	88.6
Seg-RNN on hand-crafted features	85.3	91.6
Ours with Capsules	88.8	84.2
Ours with GCN	88.9	92.6

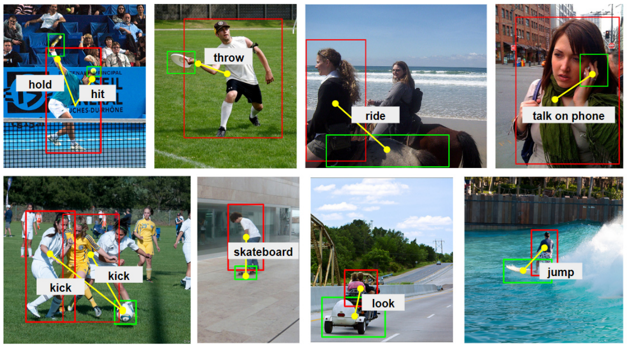


Figure 6. Detections of human-object action labels in test images of VCOCO. We report our failure cases on the last two images (bottom right). The rest are correct predictions.

4.4. Ablation Study

We now discuss the contributions of various components to the performance and their relevance to HOI detection.

Role of Spatial Models in Spatial Subnet: To verify the effectiveness of spatial graph convolution module, we designed an experiment where the image features from the backbone are directly passed to the frame-level model. We observed a significant degradation in performance in the absence of spatial model.

Role of human node features in affordance prediction: In the temporal subnet, we concatenate human node features along with object node features for the frame and segment level RNNs. We observed significant improvement in performance on object affordance detection (88.6% vs 84.6%) due to human node features. This improvement can be attributed to the high correlation between the human subactivity and affordances of active objects (objects which are not stationary).

Role of RNN in frame-level temporal subnet: As a baseline for classification at frame-level subnet, we experimented with alternative temporal aggregation models. Specifically, we built an MLP network to obtain classification scores from spatial features concatenated across temporal dimension for each node separately. However, due to higher parameter count in MLP network, the model is prone to over-fitting, and thereby has a lower performance, as is evident from Table 2.

Role of segment-level temporal learning: Even though subactivity and affordance labels are predicted for every single segment, there are significant inter-dependencies between the activity in a segment and activities in previous segments. Using a temporal sequence processing network like an RNN after the frame-level aggregation step leverages these inter-segment dependencies and achieves a significant improvement in performance as compared to prediction at frame-level temporal subnet.

Evaluating the feature learning process: To measure the effectiveness of the hierarchical learning mechanism, we design an experiment where we feed the hand-crafted, segment-level features to segment-level RNN, instead of the visual embeddings learnt by the attention mechanism. The learnt visual features achieve a better performance than the hand crafted features, particularly for the more difficult case of human subactivity detection (85.3% vs 88.9%), thereby justifying the effectiveness of the proposed method in capturing the spatio-temporal relations from RGB video data.

5. Conclusion

In this paper, we investigated the spatial modeling approaches for identifying Human-Object Interaction in videos. We followed a two-step pipeline that decouples spatial modeling from temporal reasoning. We explore the usage of capsule networks, convnets and graph convolutions for spatial relation learning. Our approach is easily extendable to other videos for the task of HOI, where depth information and 3D pose information is not available. Our approach sets a new benchmark for Human-Object Interaction detection in videos with visual information.

Acknowledgements

We are grateful to IBM Research, India (specifically the IBM AI Horizon Networks - IIT Bombay initiative) for their support and sponsorship. Rishabh Dabral has also been supported by Qualcomm Innovation Fellowship 2019.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020.

[2] Y.W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[4] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Robust data programming with precision-guided labeling functions. In *AAAI*, 2020.

[5] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain. Multi-person 3d human pose estimation from monocular images. In *3DV*, 2019.

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011.

[9] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. In *NeurIPS*, 2018.

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[11] Nicholas Frosst Geoffrey Hinton, Sara Sabour. Matrix capsules with em routing. In *ICLR*, 2018.

[12] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018.

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. In *arXiv preprint arXiv:1505.04474*. 2015.

[14] T Gupta, A Schwing, and D Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. In *ICCV*, 2019.

[15] J.F. Hu, W.S. Zheng, J. Lai, S. Gong, and T. Xiang. Recognising human-object interaction via exemplar based modelling. In *ICCV*, 2013.

[16] A. Jain, A.R. Zamir, S. Savarese, and A Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[18] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*, 2020.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimizations. In *ICLR*, 2015.

[20] H.S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. In *The International Journal of Robotics Research*, 2013.

[21] H.S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *TPAMI*, 2016.

[22] Ashish Kulkarni, Kanika Agarwal, Pararth Shah, Sunny Raj Rathod, and Ganesh Ramakrishnan. Learning to collectively link entities. In *Proceedings of the 3rd IKDD Conference on Data Science, CODS*, 2016.

[23] Ashish Kulkarni, Narasimha Raju Uppalapati, Pankaj Singh, and Ganesh Ramakrishnan. An interactive multi-label consensus labeling model for multiple labeler judgments. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*. AAAI Press, 2018.

[24] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Acm sigkdd. 2009.

[25] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019.

[26] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[28] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*, 2020.

[29] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020.

[30] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. 2020.

[31] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *CVPR*, 2020.

[32] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *CVPR*, 2020.

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[34] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018.

[35] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR*, 2020.

[36] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *ECCV*, 2008.

[37] J Peyre, I Laptev, C Schmid, and J. Sivic. Detecting rare visual relations using analogies. In *ICCV*, 2019.

[38] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[39] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *ACM Multimedia*, 2019.

[40] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Rahul Sheth, Kurt Gray, and Dinesh Manocha. Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In *ACM Symposium on Applied Perception*, 2019.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*. 2015.

[42] Fabio De Sousa Ribeiro, Georgios Leontidis, and Stefanos Kollias. Capsule routing via variational bayes. In *AAAI*, 2019.

[43] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *CVPR*, 2017.

[44] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, 2019.

[45] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM Multimedia*, 2017.

[46] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018.

[47] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[48] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. Video visual relation detection via multi-modal feature fusion. In *ACM Multimedia*, 2019.

[49] Sai Praneeth Reddy Sunkesula, Rishabh Dabral, and Ganesh Ramakrishnan. Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos. In *ACM Multimedia*, 2020.

[50] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*, 2019.

[51] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[52] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[53] T Wang, R.M Anwer, M.H Khan, F.S Khan, Y Pang, L Shao, and Laaksonen J. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019.

[54] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*, 2018.

[55] Bingjie Xu, Junnan Li, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. Interact as you intend: Intention driven human-object interaction detection. In *arXiv preprint arXiv:1808.09796*, 2018.

[56] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019.

[57] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.

[58] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[59] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, 2020.

[60] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *ECCV*, 2018.

[61] P Zhou and M. Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.