EG UK Computer Graphics & Visual Computing (2019)
G. K. L. Tam and J. C. Roberts (Editors)

Short Paper

Recognising Human-Object Interactions Using Attention-based LSTMs

M. Almushyti and F. Li

Durham University, Department of Computer Science, UK

Abstract

Recognising Human-object interactions (HOIs) in videos is a challenge task especially when a human can interact with multiple objects. This paper attempts to solve the problem of HOIs by proposing a hierarchical framework that analyzes human-object interactions from a video sequence. The framework consists of LSTMs that firstly capture both human motion and temporal object information independently, followed by fusing these information through a bilinear layer to aggregate human-object features, which are then fed to a global deep LSTM to learn high-level information of HOIs. The proposed approach applies an attention mechanism to LSTMs in order to focus on important parts of human and object temporal information.

CCS Concepts

• ***Computing methodologies*** → *Human-object interactions (HOIs); LSTM; CNN; Hierarchical design; Temporal information; Attention;*

1. Introduction

Recent research has highlighted the importance of understanding human behavior in videos for a wide range of applications, including video indexing, health support and surveillance [PSF12]. Videos offer useful information and cues about human behaviour that can help researchers recognise particular human activities and actions [KW16]. The terms activity and action are not interchangeable. In this paper, these terms are defined as follows. An activity involves more than one action. For example, an activity in a basketball game may involve two actions, namely, "running" and "jumping". An action is defined as any motion that a human body can perform [KW16]. Human motions refer to movements of human body parts. Most activities are performed by individuals who come into contact with different objects or with other people. This research analyses the interactions between humans and objects, referred to as human-object interactions (HOIs). HOI identification is important in a variety of scenarios. For example, at the AmazonGo grocery store, doing a physical checkout is unnecessary. Instead, customers simply scan the AmazonGo app at the entrance. These customers are then tracked; so when they obtain items, their virtual cart is updated [PB18]. In this kind of system, HOI identification is imperative.

Distinguishing human actions is a main challenge in computer vision. The ability of machine learning algorithms to recognise HOIs aids in addressing this challenge [SYH*18]. HOI identification generally involves localising human and the corresponding object for interaction. In the case of videos, both the human and the object are required to track over time, and their relationships are also needed to model. This phase is essential in recognising HOIs [CLL*18]. To enhance the successful rate of HOI recognition, researchers have used many features, such as the appearance of human or object [PSF12, SYH*18, GGDH18], human pose [XLW*18, YFF12], human gaze [XLW*18] and the relative location of an object with respect to the human [PSF12]. In existing work, HOIs have been widely studied in terms of images, but having only limited studies in examining HOIs from video streams. This paper investigates HOIs in videos by proposing a deep learning framework that use hierarchical LSTMs in order to capture spatio-temporal information of an interactions. HOIs are recognised based on only RGB frames from videos without using skeleton or depth information. Because training a detector from scratch to find human and object in each video frame, needs extensive human and object annotations, such as their bounding boxes, which is time-consuming. We avoid this by using a pre-trained detection model to localize humans and objects in videos. We also use LSTMs and an attention mechanism to highlight important parts of human and object temporal information. Each human and object in videos are represented by LSTMs and the second-level global LSTM captures high level information of object and human interaction. The contributions of this paper are as follows:

- We propose an LSTM-based framework with attention mechanism for recognising human-object interaction in videos. HOIs is modeled by using hierarchical LSTMs to capture the dynamics of Human and objects in a video sequence.
- We investigate the use of a bilinear layer which can handle the

DOI: 10.2312/cgvc.20191269

features of human and object, generating a discriminative feature representation from human and object information.
- Experiments were performed based on a subset of UCF101 dataset (UCF101-20) that related to HOIs [SZS12]. We show that using a bilinear layer can produce more discriminative feature for recognising HOIs with a 5% improvement than just performing a standard way of feature fusion, e.g., concatenating human-object (H-O) features.

2. Related work

2.1. Shallow learning

Traditional methods used in HOIs recognition involve the use of hand-crafted features and machine learning algorithms for classification. In order to recognise HOIs, primitive features, such as position of joints and distances between joints, were extracted from RGB-D video sequences. The authors of [YLY14] used a new middle-level representation, called orderlet. In addition to skeleton features, object features (e.g. object position in each frame) were detected. Meng et al. [MDDB15] examined the impact of using the distance between an object and human rather than using the position of the object as cue to model HOIs in both videos and images. The results illustrate that accuracy improved from 71.4%, as reported in [YLY14], to 75.8% for the same dataset [MDDB15]. However, the accuracy of these approaches was not high, implying that some temporal or contextual information is either missing or handled improperly. In [PFS13], HOIs were explicitly modelled by tracking people and action objects in a video sequence. HOIs were described as the relative position, area and motion of an object with respect to a human. Gupta et al. [GKD09] investigated how to recognise HOIs by employing a Bayesian model. The advantage of Bayesian models is their ability to use contextual information, which is perceived as the information surrounding the detected object, to identify other parts of the HOIs. However, the proposed method considered only the trajectory of the hand and ignores the whole body pose. Yao and Fei-Fei [YFF10] used the mutual context of object and human pose when modelling HOIs in images. The model is computed by considering the co-occurrence frequency of a pair of variables (e.g. object and human pose) in training images. A random field model was used to encode the connectivity between the object and human pose by using structure learning. The drawback of this approach is that it is limited to handling only one interaction per image [YFF10].

2.2. Deep learning

In contrast to shallow learning approaches, deep learning models rely on feature learning in such a way that prior knowledge of features is not required because the features are directly derived from the data by the models. In [JZSS16], the Structural-RNN (SRNN) design was proposed; in this design, HOIs are modelled using spatial-temporal graph with RNN. This method can capture high-level information and perceive the sequence of an interaction. Truong et al. [TY17] extended the SRNN by modelling object-object relations wherein spatial and temporal information between objects was observed to recognise HOIs. The results illustrate that accuracy improved from 83.2%, as reported in [JZSS16], to 90.4% for the same dataset [TY17]. Furthermore, a number of studies have attempted to solve the problem on HOI recognition in still images. In [CLL*18], human-object proposals were generated and fed to multi-stream deep neural networks. The first two streams encode the local features of objects and human whereas the final stream captures the spatial relation between human and object bounding boxes. The output score of these streams was then summed to produce the final score for each HOI class. In [GGDH18], the object is again detected in the first branch of the network, but the design differs from that of [CLL*18] in the manner by which HOIs are detected. HOIs are recognised in the form of "human, action, object" triplets. In the second branch of the network, the action is classified based on the human's pose, making this a human-centric branch, and the location of the target object is predicted. For instance, if the individual is sitting, the object is located below. The appearance of the object was also considered in order to make the model more discriminative.

Moreover, attention network design is used for HOIs recognition where in these networks, research has focused on parts of features that play an important role in boosting recognition accuracy. For example, instead of relying on the whole human appearance, it is more important to concentrate only on human parts involved in an interaction. In this direction, a pair-wise attention model that focuses on important parts of the human body and their relationships was proposed in [FCTL18] in order to recognise HOIs in still images. The authors proposed an ROI-pairwise pooling layer, which encodes the relative spatial location between a pair of body parts. The feature maps of all body part pairs are then fed to an attention model that produces the most important body part pairs. The selected body part features, which are local features, are combined with global features, including the appearance of the human, object and scene, to recognise HOIs. However, this method is only applicable when skeleton data is available. The design of the framework in this paper is inspired by the work of [IMD*16] where a hierarchical designed is used for recognising group activities. However, our work differs in the way we handle the temporal information by using attention mechanism with bilinear layer and the task is human-object interaction recognition rather than group activity.

3. Human-object interaction Recognition

3.1. Preliminary

Recurrent neural networks (RNN)s are a powerful network architecture for recognising sequential data of different lengths, such as sentences and sequences of images in a video. In neural networks, layers behave independently. However, in RNNs, all inputs are related to one another in a way that the same task is performed to all elements of a sequence. The type of RNNs using in this paper is Long short-term memory LSTM [HS97]. In LSTM, two vectors are involved at each time step, namely the hidden vector and the cell state vector. LSTM can add or remove information to the cell state by using different gates, namely input, forget and output gates. This design helps maintain the long dependency of a sequence in LSTM.

3.2. Human-object interaction model

As discussed in the previous section, limited research has addressed the problem of HOIs in videos. Learning the global description of a video's temporal information is important to classify videos accurately. Inspired by the success of employing hierarchical architecture in modelling the temporal dynamics of group activities [IMD*16], a hierarchical design for handling HOIs is proposed. Firstly, the inputs to our framework are human and object tracklets which are a sequence of bounding boxes of human and object in a video, which are extracted by CNN networks [LBB*98]. The outputs of CNNs will then be fed to LSTM layers. Particularly, each tracklet will be fed into a LSTM layer to capture intensive temporal information between frames in a sequence. Specifically, the model can be divided into three part: input pipeline, modeling H-O interactions and classification procedure.

- **Input pipeline**: To model human-object interactions, it is essential to have information about the parts involved in an interaction. Therefore, spatial and appearance information of the human and objects in video is very important. These information includes the shape and texture of human and objects during the video. The input of the model is object and human tracklets in a video sequence. The spatial features of these tracklets is extracted via convolution neural networks (CNNs).Here, we used a pertained model for feature extractions which includes series of convolution layers with kernels that are used to extracts different features, such as edges, color, gradient orientation, etc, from the bounding box around the person and object in each video frame.
- **Modeling H-O interactions:** As shown in Figure 1, each of human and object tracklet features is fed to LSTMs. In other word, the human tracklet is fed to an LSTM to capture the temporal information in human movement (e.g. motion). Also, the object's tracklet is fed to another LSTM to learn object's motion during the video. A soft attention mechanism [LPM15] is applied to the output of both LSTMs that are related to human and object. In soft attention, a soft alignment score between the last hidden state and each hidden state in LSTM layer is computed through multiplication. This score is then fed to a softmax layer where the output represents the attention distribution. This output considers as alignment weights with the size being equal to the number of time steps in the LSTM layer. Finally the context vector is computed by multiplying alignment weights and LSTM hidden states. The vectors that are generated after applying attention mechanism over LSTMs are fused using a bilinear layer. The purpose of this layer is to aggregate features from human and object which can imply the pairwise interactions between these H-O features [YYX*18]. The bilinear layer operation can be formulated by:

$$Y = WA_h \otimes A_o \tag{1}$$

  where A_h, A_o are the human and object features after attention layer is applied, respectively. W is the learnable weights and $\otimes$ indicates the outer product. This can produce a representation of human and object interactions. Y is then fed to a deep LSTM to learn high level information of HOIs. This is followed by Softmax layer for classification. Figure 1 illustrates the proposed framework.
- **Classification procedure**: In order to predict HOI label of each video, a Softmax layer is used where the probability for each class is computed. The cost function used during training is Cross-Entropy that for a training example can be formulated as:

$$H(y_i, \hat{y}_i) = -\sum_{i=1}^{c} y_i \log \hat{y}_i \tag{2}$$

  where y_i and $\hat{y}_i$ are the actual and predicted probability distributions for HOIs classes, respectively. c indicates the number of HOIs classes. The loss over the whole dataset is formulated as:

$$loss = \frac{1}{m} \sum_{i=1}^{m} H(y_i, \hat{y}_i) \tag{3}$$

  where m is the number of training examples. The goal during training is to minimize this loss by gradient descent algorithm.

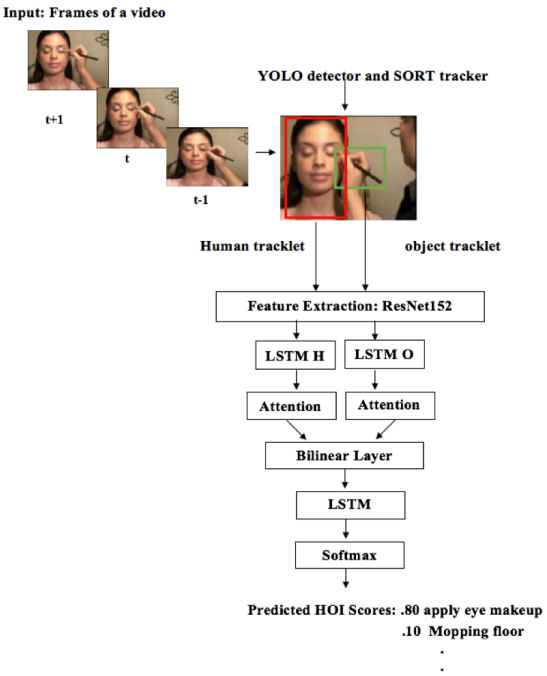

Figure 1: *The proposed hierarchical LSTM framework.*

4. Experiments

4.1. Dataset and evaluation metrics

- **Dataset:** The dataset used in this experiment is the UFC101 dataset [SZS12], which includes a variety of human actions, such as playing tennis and applying eye makeup. Since the scope of this study is recognising HOIs, 20 classes from the UCF101 dataset (split 1) relating to HOIs were used.
- **Evaluation metrics:** We use accuracy to measure the performance of our model which can be calculated by dividing the number of correctly recognised HOI videos by the total number of videos in the dataset.

4.2. Implementation details

We use PyTorch to implement our framework. In order to detect and track human and object in video frames, we used You only look once (YOLO) [RDGF16] object detection model which is pre-trained on COCO dataset [LMB*14] and Simple online and real-time tracking (SORT) tracker [BGO*16]. The detected object and human bounding boxes are then fed to ResNet152 [HZRS16] to extract spatial features.

Training details: The resolution of video frames is 224*224. The number of video frames that used from each video is 28 frames. We choose Adam optimizer [KB14], which is empirically shown that it is better than others in term of convergence speed, and the learning rate is set to 10^{-4}. To reduce overfitting, batch normalization and some of regularization techniques such as dropout are used. Since we are using pre-trained models for detection and feature extraction, our network is not following end-to-end training fashion. We only train all the hyper-parameters after Extracting features by the pre-trained model. The training is performed by using batch size of 32 videos and for 40 epochs. Figure 2 shows the training and validation accuracy against 40 epochs. We trained the model in different epochs such as 20, 30, and 40 epochs and the highest validation accuracy achieved when training the model with 40 epochs. All the experiments in this study are conducted on a single Nvidia GeForce RTX 2080 Ti GPU.

4.3. Results and discussion

As we can see in Table 1, the results show that including a bilinear layer can improve the encoding of human-object interaction than simply concatenating human and object features. Also, we evaluate the importance of different parts of our model. We train our model without the final global LSTM layer, showing a drop of accuracy to 63%, which reflects the important role of using the global LSTM layer for modeling HOIs. Also, we run the model without applying attention mechanism and we achieved very low accuracy of 46.14 %. This explains that giving more attention to the significant part of the video sequence improves the learning process. Also, we examine our model by using VGG-19 model [SZ14] as feature extraction instead of ResNet-152 and the results confirms that using residual mapping in the ResNet-152 leads to extracting more complex features than stacking convolutions in VGG-19. Also, this implies that the better models we use in extracting features and detection, the better results we can achieve in terms of accuracy. In fact, it is difficult to make a fair comparison of our results with existing methods because the results that reported on UCF101 are based on using all of the 101 categories in the dataset whereas in this study we use only 20 classes that are related to the scope of this research. Also, most of existing work only use UCF101 for action recognition tasks without considering human-object interaction. Moreover, their experiments only consider splitting the data into two sets, namely training and testing. Instead, our experiment further considers splitting a validation set for training by using 25% of the dataset. Due to time limitation, we could not apply our method to other datasets and we leave it as a future work. Figure 3 illustrates some false detection cases. They explain detecting most important object involved in human-object interaction is critical, where we will address this in our future work.

Table 1: *Results of the proposed framework.*

Architecture	Test accuracy
ResNet-152 + Our model with attention and concatenation	59.77
ResNet-152 + Our model with attention and bilinear layer	**64.50**
VGG-19 + Our model with attention and bilinear layer	55.66
ResNet-152 + Our model with attention and bilinear layer(w/o a global LSTM)	63.05
ResNet-152 + Our model with bilinear layer(w/o attention)	46.14

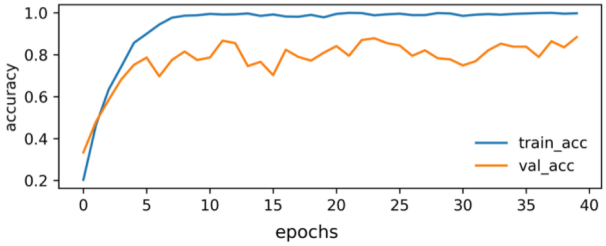

Figure 2: *Training our model with 40 epochs*

Figure 3: *Some of false detections cases: (Left) Drinking bottle is detected as the main object for interaction instead of lip brush. (Right) A Cup is detected as the main interacting object instead of the keyboard.*

5. Conclusion and future work

This paper introduced a new framework design to solve the problem of human-object interaction recognition with the use of attention and bilinear layer to model human and object temporal information. The results show that the importance of a hierarchical design that directs the network to learn high-level information of an interaction. Since we are using a subset of UCF101 dataset, which is related mostly to human object interactions, it is not fair to compare our results with other methods that have not reported results on the same subset. As a future work, we can use a better detection model that trained on more different kinds of objects. Also, we will consider the case where a human interaction with multiple objects in the video.

References

[BGO*16] BEWLEY A., GE Z., OTT L., RAMOS F., UPCROFT B.: Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)* (2016), IEEE, pp. 3464–3468. 4

[CLL*18] CHAO Y.-W., LIU Y., LIU X., ZENG H., DENG J.: Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), IEEE, pp. 381–389. 1, 2

[FCTL18] FANG H.-S., CAO J., TAI Y.-W., LU C.: Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 51–67. 2

[GGDH18] GKIOXARI G., GIRSHICK R., DOLLÁR P., HE K.: Detecting and recognizing human-object interactions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), IEEE, pp. 8359–8367. 1, 2

[GKD09] GUPTA A., KEMBHAVI A., DAVIS L. S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 10 (2009), 1775–1789. 2

[HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780. 2

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 4

[IMD*16] IBRAHIM M. S., MURALIDHARAN S., DENG Z., VAHDAT A., MORI G.: A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 1971–1980. 2, 3

[JZSS16] JAIN A., ZAMIR A. R., SAVARESE S., SAXENA A.: Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 5308–5317. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 4

[KW16] KANG S. M., WILDES R. P.: Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906* (2016). 1

[LBB*98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P., ET AL.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. 3

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *European conference on computer vision* (2014), Springer, pp. 740–755. 4

[LPM15] LUONG M.-T., PHAM H., MANNING C. D.: Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015). 3

[MDDB15] MENG M., DRIRA H., DAOUDI M., BOONAERT J.: Human-object interaction recognition by learning the distances between the object and the skeleton joints. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (2015), vol. 7, IEEE, pp. 1–6. 2

[PB18] POLACCO A., BACKES K.: The amazon go concept: Implications, applications, and sustainability. *Journal of Business & Management 24*, 1 (2018). 1

[PFS13] PREST A., FERRARI V., SCHMID C.: Explicit modeling of human-object interactions in realistic videos. *IEEE transactions on pattern analysis and machine intelligence 35*, 4 (2013), 835–848. 2

[PSF12] PREST A., SCHMID C., FERRARI V.: Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 3 (2012), 601–614. 1

[RDGF16] REDMON J., DIVVALA S., GIRSHICK R., FARHADI A.: You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 779–788. 4

[SYH*18] SHEN L., YEUNG S., HOFFMAN J., MORI G., FEI-FEI L.: Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), IEEE, pp. 1568–1576. 1

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 4

[SZS12] SOOMRO K., ZAMIR A. R., SHAH M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012). 2, 3

[TY17] TRUONG A. M., YOSHITAKA A.: Structured lstm for human-object interaction detection and anticipation. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2017), IEEE, pp. 1–6. 2

[XLW*18] XU B., LI J., WONG Y., KANKANHALLI M. S., ZHAO Q.: Interact as you intend: Intention-driven human-object interaction detection. *arXiv preprint arXiv:1808.09796* (2018). 1

[YFF10] YAO B., FEI-FEI L.: Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), IEEE, pp. 17–24. 2

[YFF12] YAO B., FEI-FEI L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 9 (2012), 1691–1703. 1

[YLY14] YU G., LIU Z., YUAN J.: Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision* (2014), Springer, pp. 50–65. 2

[YYX*18] YU Z., YU J., XIANG C., FAN J., TAO D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 99 (2018), 1–13. 3