JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering, Vol. 15, No. 1, March 2021, pp. 18-33

A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images

Sunaina, Ramanpreet Kaur*, and Dharam Veer Sharma

Department of Computer Science, Punjabi University, Patiala, India

sunaina175@gmail.com, ramanpreet.star@gmail.com, dveer72@hotmail.com

Abstract

Due to the rising demand for automatic interpretation of visual relationships in several domains, human-object interaction (HOI) detection and recognition have also gained more attention from researchers over the last decade. This survey paper concentrates on human-centric interactions, which can be categorized as human-to-human and human-to-objects. Although an extensive amount of research work has been done in this area, real-world constraints like the domain of possible interactions make the research a challenging task. This paper provides an analysis of conventional hand-crafted representation-based methods and recent deep learning-based methods, ongoing advancements taking place in the field of HOI recognition and detection, and challenges faced by the researchers. Moreover, we present a detailed picture of publicly available datasets for HOI evaluations. At the end, the future scope of the study is discussed.

Category: Human-Computer Interaction

Keywords: Human-object interactions; Action recognition; Visual relationships; Deep learning; Hand crafted; Computer vision

I. INTRODUCTION

The computer vision, an imitation of the human vision system, has witnessed significant development recently. Visual relationship recognition [1-4], is a fundamental problem in this area. The task of visual scene understanding in- volves detection of object instances and recognizing interactions among them. The technical advancements in the computer vision era help to achieve the great potential for many vision tasks including activity mining in social networks [5], activity detection in surveillance [6], activity analysis [7], video understanding [8], image captioning [9] and visual question answering [10-13]. The human-object interaction (HOI) [14-24], is an important facet of visual relationship recognition. It localizes humans and objects and then identifies the relationships among them to answer various questions like “What is happening in a particular visual scene?”, “What is the objective of an interaction?” and “Which of the objects involve in the interaction?” HOI is an understanding of how humans interact with the surrounding objects. Fig. 1 shows some of the examples of HOI activities.

Earlier work on vision-based recognition mainly focused on action recognition. Many survey papers are available [25-30], which present a detailed analysis of action recognition methods. The authors of [25] review pose recognition and tracking human motion in the videos. In [26], the analysis of the action and activities has been discussed according to their complexity. In [27], a survey on local representation and deep learning methods on

Open Access http://dx.doi.org/10.5626/JCSE.2021.15.1.18 http://jcse.kiise.org

Received 04 January 2021; **Accepted** 12 March 2021

*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

Fig. 1. Illustrations of human-object interactions, cleaning floor with a mop (left), playing guitar (middle), and a group of children playing football (right).

action recognition is presented. The authors of [28] focus on the challenges encountered while recognizing human action and activities. In [29], the authors divide action recognition into four categories; gestures, actions, interactions, and group activities. Out of these categories, the interaction can be further classified into different classes like interaction with the objects and interaction with humans. However, there are a large number of survey papers regarding action recognition in images and videos, but nowadays more attention is paid to recognize more fine-grained actions to completely understand what is going on in the particular scene as there is the difference between "person holding the guitar" and "playing the guitar". The motivation behind the paper is to provide a comprehensive overview of research work specific to interaction between human and object recognition and detection from still images. The remaining part of the paper is organized as follows; challenges specific to HOI recognition in still images are listed in Section II; a summarized view of publicly available datasets for evaluations is presented in Section III; and Section IV covers the techniques used for HOI detection and recognition. The paper covers both the handcrafted and learning-based methods for HOI. The conclusion and future scope are discussed in Section V.

II. CHALLENGES

While dealing with HOI recognition and detection many challenges come in the way which makes this work more difficult. Following are some of the challenges encountered while dealing with HOI.

A. Occlusion

To carefully detect the HOIs, the model must identify the object. But sometimes the part of the object gets occluded by the actor itself while performing the interaction. This issue affects the performance of the model. For example, when a person makes a call on a mobile phone, the major portion of the mobile phone is occluded by the hand and the face of the person, in such cases it becomes challenging to detect the object and interaction.

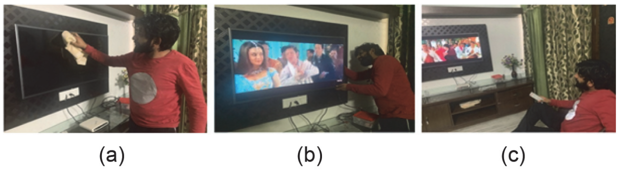

Fig. 2. (a) Cleaning the television, (b) Repairing the television, (c) Watching the television.

B. Inter-class Similarity

If the model successfully identifies the object and actor in the image or video, it does not imply that it can correctly identify the interactions because there may be multiple interactions with the same object that causes inter-class similarity. Fig. 2 demonstrates three different interactions with the same object. In Fig. 2(a), a person is cleaning the television; in Fig. 2(b), the person is repairing the television; and in Fig. 2(c), the person is watching the television. In this case, even after detecting the human and object, the model needs to be trained for different interactions.

C. Shape Variation

When an object comes in contact with the actor for any interaction then the shape of the object changes, which poses another challenge to correctly detect an object. For example, when a man eats an apple, in the image of in-between the interaction, it becomes hard to correctly identify the object and hence interaction.

D. Intra-class Variability

Many actions and activities are such that there is a similar name for the interaction but the objects and subjects are entirely different. For example, "man-eating apple" and "cow eating grass," here the interaction class is similar i.e., eating, but similar class subject and object are varied.

E. Limited Dataset

As there a large number of interactions exist between human and object pairs in the real world, no such dataset exists to cover those interactions. Manually creating and annotating the dataset is entirely impossible.

F. Background Variation

The background environment of the recorded HOI is a very important source. In old datasets, the recorded videos and images are unrealistic and staged where the background is static. But this is not possible in real life. In realistic environments, it becomes even harder to detect humans and objects in cluttered or dynamic backgrounds. For example, if the color of an object is alike to the background color then the object localization will become more complex.

G. Varying Speed

In videos, the speed of approaching objects for interaction varies as per the actors. So, it is difficult to generalize the interaction among different actors.

H. Lighting Conditions and Viewpoint Variation

The appearance of humans and objects can be influenced by lighting conditions and also the same object can be observed differently from diverse viewpoints. In some datasets, only one viewpoint is fixed and the recordings are done by static cameras. But this seems very constrained and opposite to real-life environments. By the use of multiple cameras, the static viewpoint problem gets reduced to some extent and the occlusion can also be alleviated. However, the use of multiple viewpoints also increases the complexity.

I. Multi-label Classification

After identifying any interaction in the image, it does not need to be the only interaction that a human is performing. Humans can perform multiple interactions with different objects at the same time as cutting the vegetables while sitting on the sofa and watching the television. Such recognitions make the task more challenging.

III. DATASETS FOR HOI

Since the initiation of work on human action recognition, many datasets were published from time to time. Initially, video datasets like KTH [31] and Weizmann [32] were published which contained videos of different human actions, but the videos were recorded in a very controlled environment and represented very limited action categories, which are not very useful in real-world applications. With the advent of techniques some more video datasets were created like Hollywood [33], UCF101 [34], and HMDB51 [35] that contained more realistic actions. But again, these

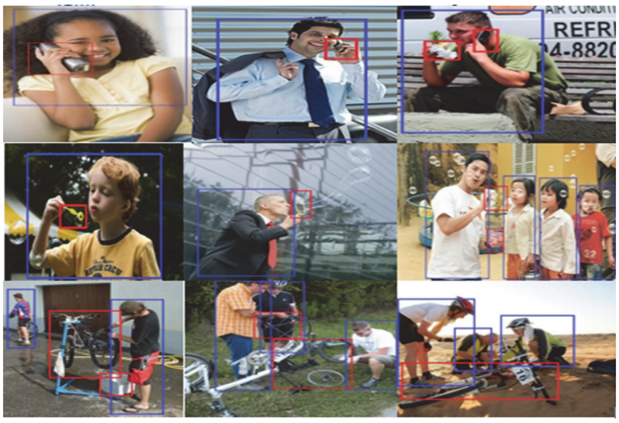

Fig. 3. Visualization of human-object interactions in the Standford40 dataset. It shows the humans in the blue boxes and objects in the red boxes. Multiple humans may involve in a single interaction. The three rows show; talking over the phone, blowing bubbles, and fixing bikes.

Table 1. Summary of hoi datasets with number of images, interaction classes, examples, and year

Dataset name	Images	Interactions	Example activities	Year
PPMI [16]	-	7	Playing violin, playing guitar, playing a flute	2010
TBH [50]	341	3	Playing trumpet, wearing hat	2010
Stanford40 Action [51]	9,532	40	Brushing the teeth, cleaning the floor	2011
89 Action [52]	2,038	89	‘Sitting on a chair, ‘drinking from a bottle’	2013
TUHOI [42]	10,805	2,974	Playing ping-pong, using a laptop, holding a computer mouse	2014
MPII [43]	40,522	823	Playing violin, riding a bus, horse grooming	2014
HICO [45]	47,774	600	Feeding a giraffe, sailing a boat, talking on a cellphone	2015
V-COCO [1]	10,346	26	Laying on the bed, reading a book, working on a laptop, kicking sports ball	2015
VRD [2]	5,000	70/37,993 (predicates/relationships)	Person kicking ball, a person on top of the ramp	2016
HCVRD [47]	52,855	9,852	Man holding surfboard, a man wearing kneepad	2017
HICO-DET [45]	47,776	600	Tying a boat, feeding a bird, riding an airplane	2018
HOI-A [24]	38,668	10	Reading document, talking on a mobile phone, kicking sports ball	2020

datasets contained videos of different types of action categories like the person acting without any object for example running, walking, the action between two persons (fighting, handshaking), action among multiple persons (group dance, eating in a restaurant), the action between the human and object, etc. Fig. 3 shows some examples of interaction activities from Stanford 40 datasets.

There are datasets [36-40] available for the recognition and detection of objects. But only detecting the objects in the scene is not enough to get a complete idea of what is going on in the image. So, some datasets have been introduced further, which consists of not only humans or objects but also the relationships among them. Table 1 represents the summary of datasets related to human-object interactions. The introduction of these datasets is given in the following sections.

A. PASCAL Visual Object Classes (PASCAL VOC)

The first version of this dataset was published in 2005 [41]. Initially, it included only four classes, i.e., bicycles, cars, motor bikes, and people, having 1,578 images and 2,209 annotated objects. Subsequently, the dataset evolved and included more classes and images. The 2012 version contains 20 classes of objects and 11,530 images containing 27,450 annotated objects.

B. Trento Universal Human Object Interaction (TUHOI)

TUHOI [42] dataset is designed to represent the dispersal of HOIs in real life as some interactions like riding a bike are more common and occur frequently than washing a bike. It also describes each action with different verbs and different usage of the same verb.

C. MPII Human Pose

In MPII Human Pose [43] dataset, a wide variety of activities are included from different fields like skating, fishing, watching the bird, etc., covering both common and rare human poses. Out of 20 action categories in the dataset, 12 categories are related to HOIs.

D. Visual Genome

Krishna et al. [44] presented the visual genome dataset to model interactions between objects to characterize cognitive understanding of the image. It consists of 108,249 images from MS-COCO and YFCC100M datasets. The annotations involve multiple regions and their description, objects and their bounding boxes, attributes, relationships, region graphs, scene graphs, and question-answer pair of the image.

E. Human Interacting with Common Objects (HICO)

Chao et al. [45] presented the HICO dataset in which a list of interactions with the objects is carefully selected. It contains sense-based HOI categories and multi-labeling of co-occurring categories is done.

F. Verbs in Common Objects in Context (V-COCO)

Gupta and Malik [1] created the dataset for HOI detection by augmenting MS-COCO [46] dataset by annotating the verb besides the objects named V-COCO. Where the MS-COCO dataset contains only persons and objects annotation, the V-COCO dataset describes the action annotation that each person is performing, and the different semantic roles of the object like riding a bicycle, sitting on a bicycle, etc.

G. Visual Relationship Dataset (VRD)

Lu et al. [2] proposed the dataset based on the visual relationship of objects in the images. The dataset contains 100 object categories and 70 predicates having 24.25 predicates per object category. The dataset also represents a long list of rare actions that are learned by language priors.

H. Human-Centric Visual Relationship Detection (HCVRD)

Zhuang et al. [47] released the HCVRD dataset consisting of the visual relationships which comprise humans and object relationships. The human subject is further fine-grained in man, woman, boy, and girl to capture the better relationship details like age and gender significantly affect the relationship. However, it contains a very large number of relationships but many relationships have only fewer images (less than 10) in the dataset, which causes the long tail distribution problem.

I. Human Interacting with Common Objects-Detection (HICO-DET)

HICO-DET [48] is the further augmentation of the HICO dataset, where HICO contains the image level HOI recognition. HICO-DET is designed not only to determine the presence of HOI but also to estimate the location of the action. In it, the annotation consists of a bounding box of objects, a bounding box of humans, and the linking of each person to an object.

J. Human-Object Interaction Applications (HOI-A)

Recently published HOI-A dataset [24], which contains HOI categories is frequently used and has some practical applications like smoking, talking over the phone, playing on the phone, etc. While shooting for the dataset both indoor and outdoor environments are used for each activity to include the intra-class variation.

K. Some Other Datasets

Along with the above-discussed datasets, some researchers worked on different datasets to experiment on HOI recognition. Initially, Sadeghi and Farhadi [49] published a dataset to represent visual relationships between objects. It consists of 2,769 images having 17 visual phrases, out of which 9 are HOIs like "person sitting on a chair," "drinking from a bottle," etc. In [50], the authors used the Trumpets, and Bikes and Hats (TBH) dataset for experiments which include only three activities, i.e., "playing trumpet," "riding a bike," and "wearing a hat." Stanford40 dataset [51] contains images of 40 varied daily real-life human action classes. Each class has 180-300 images which represent a rich set of variations in human poses, cluttered background, and appearance. Le et al. [52] presented 89 action datasets containing 19 objects and 36 verbs, a total of 89 actions. The dataset is prepared from PASCAL VOC 2012. A real-world scene graph dataset [53] is published, built upon COCO and YFCC100M [54] datasets containing 5,000 images which represent 93,000 object instances, 110,000 attribute instances, and 112,000 relationship instances. The examples are like "the man wearing a black helmet riding the bike" and "the man wearing a red shirt on a skateboard." Willow dataset [55] represents seven human actions out of which 5 are HOIs. The authors of [15] generated their own video and image dataset for HOI where the image dataset consists of 6 interactions from sports like "cricket bowling" and "volleyball smashing." In [56], the authors used the dataset containing 6 interaction classes with four objects like pouring from the cup, making a phone call, etc. Some scene and situation-based databases [57] are also published which provide a rich set of contextual information to aid action and interaction recognition. The dataset in [58], contains a large number of different types of scenes like the zoo, church, mall, etc.

Many video-based datasets for simple human action recognition [31, 32], for complex human action recognition [33-35, 59] and HOI [60, 61] also exist. But discussing those datasets is beyond the scope of the present work. In [62], the authors provide a detailed survey of existing video datasets for action recognition.

IV. TECHNIQUES USED FOR HUMAN-OBJECT INTERACTION DETECTION AND RECOGNITION

The visual relationship is an old concept; researchers have developed several methods for visual relationship detection. The set of visual relationships includes spatial relations (e.g., "Dog is under the table"), actions (e.g., "Boy is running"), verbs (e.g., "Girl wears red shirt"), prepositional phrases (e.g., "Man hit the ball with a bat"), etc. A general idea behind visual relationship detection is identifying a triplet <*object1*, *predicate*, *object2*>.

Human action recognition and pose estimation methods [6-12] focus on individual actions (e.g., "playing," "running," and "walking") and human body structures. Object detection tasks [1-4] detect objects and persons present in the scene (e.g., "piano", "computer", and "boy") by generating bounding boxes. Various applications of visual relationship recognition require a deep understanding of the scenes, by concentrating on all humans, objects, and relationships present.

Human-object interaction is a kind of visual relationship, which concentrates on the interactions between humans as sub- jects and other objects present in the scene (e.g., "Girl playing piano"). A human can play out a few activities at the same time (e.g., "eating bread and watching TV while sitting on a sofa"). So, recognition of human-object interactions is a difficult assignment. This section presents a survey of techniques used for HOI recognition and detection from still images.

A. Hand Crafted Representation based Approaches

The earliest approaches for recognition of HOIs from visual scenes are capturing hand-crafted local features and then classifying features into particular interaction classes. Most of the models consider appearance features, pose features, spatial features, temporal features, language features, etc. The development of these models initiated with the arrival of algorithms like Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transformation (SIFT), Histogram of Oriented Flow (HOF), Motion History Image (MHI), etc. Hand-crafted features-based models are considered simple and unambiguous to understand. Dalal and Triggs [63] developed HOG for detecting humans and identifying the poses them from images. A technique that uses motion and shape information for action recognition is called Motion History Image (MHI) and was introduced by Bobick and Davis [64]. The action recognition rate of the MHI method depends upon the silhouette information. It requires complete silhouette information in advance. Lu and Little [65] presented a template-based algorithm that uses a PCA-HOG descriptor for tracking and recognizing the actions. PCA-HOG descriptor is robust under various challenging situations such as pose, illumination, and view-points variations, etc. The use of Principal Component Analysis (PCA) [66] increases the speed of tracking and recognition, without any loss in accuracy because PCA reduces the dimensionality of the feature vector by selecting a few high-energy components of the transformed image and increases the processing speed. SIFT [67] has proven to be very effective and as a successful descriptor for recognition and classification systems. The features are invariant to image scale and rotation, also partially invariant to changing 3D viewpoint, the addition of noise, and change in illumination. MOSIFT [68] not only detects interest points and encodes their local appearances but also explicitly models the local motion. For detection of distinctive local features, local appearance and motion are used. The feature point is represented as an aggregation of the HOG and HOF [33]. The HOG algorithm is used to describe spatial appearance and HOF to indicate motions of the feature point.

Scovanner et al. [69] presented a 3D SIFT descriptor, which can reliably capture the spatio-temporal nature of the video data as well as 3D images such as MRI data. The SIFT descriptor is a popular descriptor due to its properties. They have improved performance on the task of action recognition in a bag-of-words paradigm. The SIFT features are equally likely to be found on human-objects and in the background. Moreover, the SIFT uses only gray scale information and misses important appearance information, such as color. A 3D SIFT overcomes the problems of SIFT and provides better results. SIFT has been extended in other ways also. Ke and Sukthankar [70] presented PCA-SIFT, by applying PCA to reduce the dimensions of SIFT. It stated that PCA-SIFT is not only compact but also more robust since PCA may help to reduce noise in the original SIFT descriptors. For recognizing actions, the maximum likelihood estimation (MLE) was executed based on observations using HMM classifier. An improved version of SIFT, called Gradient Location and Orientation Histogram (GLOH), was proposed in [71] to use a log-polar location grid instead of the original rectangular grid in SIFT. Van De Sande et al. [72] studied color descriptors that incorporated color information into the intensity-based SIFT for improved object and scene recognition. Further, to improve the computational efficiency, Bay et al. [73] developed SURF as a fast alternative descriptor using 2D Haar wavelet responses. The authors of [55] tried to improve the performance of action recognition for still images by combining the bag-of-features model with the latent SVM model and also merged the scene background information into the classifier for good results.

1) Use of Mutual Context for Human-Object Interaction

Object and action recognition activities provide mutual context and facilitate the recognition of each other while the objects are difficult to be recognized based on shape/appearance features alone. Some methods of HOI and recognition rely on this idea of mutual context [14, 15, 56, 74]. The use of contextual information plays a vital role to accelerate the performance of visual relationship recognition. Oliva and Torralba [74] have discussed the contextual association of objects with other objects used for object localization and recognition activities when the local features are not enough due to the presence of challenges. Gupta and Davis [56] used the Bayesian model for HOI recognition by concentrating on objects, human motions, and object reactions. The contextual information from object type and object reaction is exploited for action recognition and is difficult to achieve by using the motion features alone. They train HOG-based object detectors and HMM for interactions. Gupta et al. [15] provide a fully supervised learning-based framework for classifying HOIs using spatial and functional constraints for object and action recognition from images and videos. They worked on the idea that objects having similar shape and appearance features may have different functionality and the same human poses may have different purposes depending on the context in which they occur. The context helps to differentiate between the confusing actions.

Yao and Li [14] exploited human pose and objects for HOI recognition. They have addressed challenges involved in both human pose estimation and object detection and how these can help to recognize each other for successful recognition of HOIs. Desai et al. [18] formulated a model based on the same idea of contextual interaction between human body parts and objects for recognizing human actions from still images by using the HOG template and latent SVM classifier.

The authors of [17, 49] modeled spatial relationships between objects and humans based on relative geometric information for HOI recognition. Sadeghi and Farhadi [49] show that a combination of visual phrases and object detectors can improve the detector results. Hu et al. [17] propose a model for HOI recognition based on exemplars for learning. They exploited the spatial pose-object interaction exemplars to describe the interactions between objects and humans. Yao and Li [16] performed spatial grouping of low-level SIFT features for detecting the interactions between humans and objects rather than the modeling of human poses, worked for the interactions with small variations like “playing a musical instrument” and “holding a musical instrument.” They used both discriminative and generative classifiers for classification purposes. They also experimented with their model for object recognition on Caltech 101 [75] dataset. Another work by Yao and Li [21] exploited spatial context between objects and human poses to recognize HOI by extracting SIFT features and SVM for classification tasks. They trained the activity classifier by using Spatial Pyramid Matching (SPM).

2) Use of Human Poses for Human-Object Interaction

An action may involve several pose variations. So, methods using global feature representations are not suitable due to the inflexibility of global templates for representing action variations. A single global template cannot capture all the pose variations. Some models [19, [76-80] focus on the feature representations based on human poses for successful recognition of human-object interactions. Zheng et al. [77] developed poselet activation vector-based model for action recognition from still images by focusing on both contextual classifier and pose-based classifier. They utilized SIFT descriptor and SVM classifier for context-based action classification. Ikizler et al. [76] extracted and represented features of human body structure using circular histogram of rectangles (CHORs), feature reduction using linear discriminative analysis (LDA), and classification using SVM for action recognition from the still images. Maji et al. [19] relied on body part detectors and concentrated only on action-specific poselets for 3D pose estimation. The algorithm is based on HOG and linear SVM for learning action-specific poses and appearances. Poselet activation vector-based models outperform the bag-of-words methods.

Similarly, Delaitre et al. [20] performed action recognition by learning human-object interactions of only action-specific body parts and objects rather than complete body pose configuration. Local object and human body part detectors were used rather than low-level HOG/SIFT features. Some models treat pose estimation and action recognition separately. The output of the pose estimation phase acts as input for the action recognition phase. Yang et al. [80] provided a model that integrates pose estimation and action recognition. They used latent information of human poses to recognize actions from static images. Desai and Ramanan [81] used phraselets for the detection of actions, human poses, and objects. The authors of [50], [82] did not use pose and object annotations but worked on weakly-supervised models. Prest et al. [50] formulated a probability-based model for HOI recognitions. They also compared a fully supervised version of their model with existing models for Sports, TBH, and PASCAL Action 2010 datasets. Ikizla and Sclaroff [82] proposed a multiple instance learning model by combining multiple feature channels such as a person, object, and scene. They combined the HOG features and optical flow for action recognition from video datasets. Table 2 represents the accuracy rate of some of these methods on the sports dataset.

Table 2. Recognition rates of methods experimented on sports dataset

Study	Year	Accuracy rate (%)
Gupta et al. [15]	2009	78.67
Yao and Li [14]	2010	83.30
Desai et al. [18]	2010	82.50
Prest et al. [50]	2012	81.00
Prest et al. [50]	2012	83.00
Yao and Li [21]	2012	87.00
Hu et al. [17]	2013	92.50

An image may have several objects/persons surrounding a person, known as interactees. Some of them are not manipulated/contacted during interactions but still act as central to the interactions. Several human/object relationships such as “standing on the floor,” “looking at an object from a distance,” “carrying a bag on the shoulder,” etc., may exist in an image. Chen and Grauman [83] tried to identify the locations and sizes of the interactees independently. Most of the existing methods considered the dependency of action and object classes for identifying the interactions. Other systems recognize the fine-grained nature of action recognition and exploit discriminative templates [84, 85], and color [86]. Table 3 represents the summary of various handcrafted methods for HOI.

HOI recognition based on the mutual context of human pose estimation and object detection does not produce good results due to the problem of pose estimation in occlusion and object detection is also complex when objects are very small. Hand-crafted feature representation methods do not require a large amount of data for training

Table 3. hand-crafted techniques to recognize and detect HOI

Study	Year	Technique	Datasets
Gupta et al. [15]	2009	Generative probabilistic models (fully supervised model)	Sports
Yao and Li [14]	2010	Discriminative grouplets (low-level SIFT features)	PPMI
Desai et al. [18]	2010	Discriminative Model (HOG descriptor+Latent SVM)	Sports
Yao and Li [14]	2010	Structure learning and max-margin algorithm (fully supervised model)	Sports
Yang et al. [80]	2010	Latent poselets (HOG+SVM)	YouTube dataset
Delaitre et al. [20]	2011	Replaced local SIFT/HOG features with trained object and body part detectors	Pascal VOC 2010
Yao and Li [21]	2012	Mutual context model	Sports, PPMI
Prest et al. [50]	2012	Weakly supervised learning and fully supervised learning	Sports, TBH, and PASCAL Action Challenge 2010
Hu et al. [17]	2013	Exemplar Based Modeling	Sports, PPMI

purposes. But these techniques are computationally less expensive and less robust.

B. Deep Learning-based Approaches

The performance of visual relationship recognition surged with the formulation of deep learning-based methods. These methods tend to achieve good accuracy rates for a huge amount of data. By extracting features from very large sets of training data, these methods make use of more information available in visual scenes rather than being limited to a small set of features like conventional hand-crafted methods.

Convolution neural networks (CNN) acts as a powerful tool for object detection and semantic segmentation issues. Initially, CNN was used for handwritten digit classification [87]. Krizhevsky et al. [88] exploited CNN for the classification of images, trained on ILSVRC-2010 and ILSVRC-2012 datasets to outperform the existing methods for the object recognition task. Girshick et al. [89] accelerated the accuracy rate of object detection and semantic segmentation by combining region proposals with CNN (R-CNN). It worked by generating region proposals at the initial phase. CNN extracts features from each region and class-specific linear SVMs used for classification. Girshick [90] improved the accuracy rate and speed for object detection by proposing a Fast region-based CNN (Fast R-CNN) as a single-stage framework. It outperformed R-CNN and SPPnet [91] for object detection. In [92], authors proposed a two-stage framework called Faster R-CNN, which comprises of region proposal network (RPN) to generate object bounding boxes and the second stage is the same as Fast RCNN, which tried to achieve real-time object detection. Gkioxari et al. [93] further explored the idea that contextual cues lying in the image are helpful to recognize correct action and proposed R*CNN which uses more than one reason for the classification of the action.

After the successful application of neural network techniques in object classification and recognition, deep learning is also applied for action recognition in images and videos [94-100]. The efficiency of these methods spurred the researchers to work in more specific categories. After that many applications came into existence like visual relationship detection and human interaction detection.

Gupta and Malik [1] used R-CNN to detect the object and the agent and then a regression model was applied to find the semantic role of the object. Girdhar et al. [101] modified the base architecture ResNet-101 and plugged the attention module in its last layer so that it can learn the attentional maps around human body pose and object bounding box which further improved the recognition of action. In images, visual relation is the key that offers scene understanding. Considering this key, Zhang et al. [102] proposed a VTransE network consisting of Faster R-CNN for object detection module, having a feature extraction layer and translation embedded to predict the relation between the objects. Graph R-CNN [103] also focuses on the relationships between objects. It detects the objects and then generates the graph by linking the objects that likely have a relation among them. To add contextual information, attentional graph convolutional network (aGCN) is used. To use the information of different branches of object and relationship detection and message sharing among these in [104], the author proposed the Zoom Net which is the stack of Spatiality Context Appearance module. Li et al. [105] converted the visual relationship detection into three inter-linked detection problems and introduced visual phrase convolutional neural network (ViP-CNN), which connects the three problems jointly. Dai et al. [106] presented a deep relation network that focuses on the statistical dependency between the detected objects and relationships. In [107], the authors propose the weakly supervised discriminative clustering model which is trained by image-level labels to detect the relations. Liang et al. [108] introduced a deep variation structure reinforcement learning (VRL) model which captures the global context information to identify the visual relationships among objects. Zhuang et al. [109] gave the concept of using the context information while recognizing and introducing a context-aware interaction recognition model. As some scenes may contain a large number of objects and it is challenging to analyze the relationship combination among all of them, in [110] authors introduced relationship proposal network (Rel-PN) which used pairs of related regions in the image to detect the relationship. Zhang et al. [111] presented a parallel pairwise region-based fully convolutional network (PPR-FCN) using R-FCN [112] to apply weakly supervised recognition of objects in the image to avoid the expensive method of annotating each object in the image.

While a lot of datasets for human-object interaction evaluation are available, in the real world the list of interactions is very long and many interactions are rare enough to collect a good amount of data. Considering this problem [113] introduced zero-shot scaling of interactions. The thought behind this is that however it is very uncommon to have numerous interactions, the object and the verb associated with the interaction may be very common. For example, “riding the elephant,” is a very uncommon interaction, but both riding and elephant are very common. So, the model can detect such interactions by leveraging the learning of verbs and object differently. To address the long tail distribution problem of some rare interaction classes, [114] proposed the idea of using an action co-occurrence matrix that modeled the correlation between HOI and utilized it for better training. The [113] contains VGG19 [4] to extract feature map and HOI detection is done using two different networks, i.e., object detection network and verb detection network. For improvement in learning, [3] formulated a scalable and generic model GPNN (graph parsing neural network) by integrating graphical structure and neural networks. GPNN can be applied for both images and video datasets for HOI detection and recognition. Chao et al. [48] proposed human-object region-based CNN (HO-RCNN), a multi-stream architecture that detects the HOI using three streams, i.e., human stream to extract the local features of humans, object stream to extract the local features of objects, and pairwise stream to encode spatial relation between humans and objects.

Mallya and Lazebnik [11] utilized the local and global context information to predict human activity. The deep convolution model consists of Fast RCNN, VGG16, and ROI Pooling layer. To tackle the lack of full person instance-level super- vision multiple instance learning is used. While performing any task, the intention and body part movement of the human provides a good amount of information to detect the correct interaction. Xu et al. [118] explored this idea by fusing the human gaze, human pose, and relative distance between human joints and object into HOI recognition. Gao et al. [117] also extended this idea and proposed instance centric attention network (iCAN) which focuses on the regions near detected human and object instances and generated an

Table 4. Table of deep learning-based techniques to recognize and detect hoi

Study	Year	Technique	Data sets
Gupta and Malik [1]	2015	RCNN	V-COCO
Lu et al. [2]	2016	RCNN for detecting object pairs + visual module and language module	VRD
Mallya and Lazebnik [11]	2016	Fast RCNN + VGG16	HICO and MPII
Zhuang et al. [47]	2017	Faster RCNN and deep metric learning module	HCVRD
Zhang et al. [102]	2017	Proposed VTransE to perform object detection and predicate classification.	VRD, VG
Chao et al. [48]	2018	HO-RCNN	HICO-DET
Gkioxari et al. [115]	2018	Faster RCNN for object detection + action classification branch + interaction branch	V-COCO and HICO-DET
Yang et al. [103]	2018	Graph RCNN (Relation Proposal network (RePN) + attentional Graph convolutional Network(aGCN))	VG
Yin et al. [104]	2018	Zoom Net (Spatiality Context Appearance module)	VG
Qi et al. [3]	2018	GPNN having four functions Link function, Message function, Update function, and readout function	HICO-DET and V-COCO
Fang et al. [116]	2018	VGG+ROI pooling+attention module	HICO and MPII
Gao et al. [117]	2018	Faster RCNN+object iCAN+human iCAN+pairwise stream	HICO-DET and V-COCO
Shen et al. [113]	2018	Zero-shot learning using VGG19 + Faster RCNN + verb detection network.	HICO-DET
Li et al. [23]	2019	ResNet is used for feature extraction and multi-stream CNN is used for interactiveness knowledge	V-COCO and HICO-DET
Xu et al. [118]	2019	Faster RCNN for object detection and pose and gaze network for intention feature extraction.	V-COCO and HICO-DET
Wan et al. [119]	2019	PMFNet (Pose aware Multi-level Feature Network)	HICO-DET
Gupta et al. [22]	2019	Factor graph consists of human and object appearance,box-pair configuration, and human-pose factors	V-COCO and HICO-DET V-COCO and HICO-DET
Zhou et al. [120]	2019	RPNN (object body part graph + human body part graph)	V-COCO and HICO-DET
Liao et al. [24]	2020	Faster RCNN + ResNet50	HOI-A and HICO-DET
Zhou et al. [121]	2020	Instance localization Network + interaction recognition network	V-COCO
Ulutan et al. [122]	2020	VisualSpatial Graph network	V-COCO, HICO-DET
Li et al. [123]	2020	Part state network+Activity2Vec model	HICO, V-COCO
Li et al. [124]	2020	2D-3D joint learning	HICO-DET

attentional feature map. In [22], the authors encoded the appearance and layout information generated from the object and human pose detector and pass ed to the factor graph. Fang et al. [116] considered the fact that only some parts of the human body are involved in an interaction with objects rather than the entire body. To incorporate this, the model used ROI pooling to select the joint feature maps of pair-wise body parts and an attention module to find the pairwise correlation between the body parts. Wan et al. [119] extended the idea by exploring relational reasoning on three semantic levels for each human-object interaction, i.e., interaction, visual objects, and human body parts. They proposed the pose aware multi-level feature network (PMFNet) consisting of four modules named backbone module, holistic module, zoomed-in module, and fusion module to detect humans, objects, and part-level attention to improve part cues. Gkioxari et al. [115] again hypothesized that the appearance of individuals, their clothes, and pose gave the cue of the location of the target object, human interacting with. So rather than looking at the whole image, the model learns to predict the density over the possible location and the output is combined with the location of actually detected objects to finally localize the target object. Following the idea, Zhou et al. [120] introduced relation parsing neural network (RPNN) consisting of two attention-based graphs where object location is estimated from refined human body part features and interaction is detected from refined object body part features. Li et al. [23] improved the performance of HOI detection by introducing an interactiveness network that can be combined with any existing HOI detection method. An interactiveness network is a transferable knowledge learner that learns the interactiveness from several HOI datasets and applies to a specific dataset. In their two-stage model after performing the non-interactive suppression (NIS), the remaining interactive pairs are sent for classification. Low-grade object detection is suppressed and high-grade objects are rewarded by low-grade instance suppression (LIS). Faster R-CNN [92] with ResNet 50 [125] is used for HOI detection. Table 5 shows the performance of some of the techniques on the V-COCO and HICO- DET dataset.

Liao et al. [24] formulated the HOI detection with a different approach by detecting the points rather than the human-object proposal. They designed single-stage parallel point detection and matching (PPDM) framework. Both the branches run parallel rather than sequentially. PPDM outperforms the existing two stage methods. Zhou et al. [121] proposed the multistage cascaded architecture to detect HOI where each stage consists of two network instance localization networks and interaction recognition networks. At each stage, the network is connected to the network of the previous stage for cross information propagation and the framework can perform pixel-wise and fine grain segmentation of relation. Li et al. [126] created the human activity knowledge engine (HAKE) by annotating the part state of each active person in each image of the existing datasets so that the information of the part state of objects and person can be considered while recognizing the activity. Li et al. [123] considered the knowledge requirement of states of human body parts to recognize the activities from images. For that, they designed the Activity2Vec model to extract features of body parts and PaStaNet to deduce the activity. Li et al. [124] considered the detailed 3D body shapes to understand the complex interactions and proposed a detailed 2D 3D joint representation learning method. For that, they captured detailed 3D body, face, and hand shapes of the human body and 3D object location and size. Ulutan et al. [122] used the structural connection between objects and spatial configuration to understand the HOIs. They designed VSGNet which utilizes the structural connection between the human object pairs using graph convolutions. Table IV shows the summary of various deep learning methods for HOI detection.

Table 5. Performance table of some of the techniques on V-COCO and HICO-det dataset

Technique	mAP	
	HICO-DET	V-COCO
GPNN [3]	13.11	44.0
iCAN [116]	14.84	45.3
InteractNet [119]	9.94	40.0
No-Frills model [22]	17.18	-
Interactiveness network [23]	17.22	48.7
iHOI [115]	13.39	45.79
PMFNet [118]	17.46	52.0
RPNN [120]	17.35	47.53
DJRN [125]	21.34	-
Cascaded [122]	-	48.9
VSGNet [126]	19.80	51.76
PPDM [24]	21.73	-

V. CONCLUSION AND FUTURE SCOPE

HOI is central to both action recognition and visual relationship detection. Automatically understanding and analyzing the activities and interactions of the human is a very intricate task. In this paper, we have discussed the challenges, datasets dedicated to HOI, and a summary of handcrafted and deep learning-based approaches used in HOI. Recent approaches appear to be taking over the field and are successful in detecting HOIs but they still need to unlock the full potential to address various challenges hindering them to achieve the required precision to apply it to real-life applications. Some of the challenges like occlusion and shape variation can be handled by increasing the size of the dataset by including the images containing different shapes of objects and occluded objects. The challenges like inter-class and intra-class variation can be resolved by increasing the number of classes however it can cause overlapping between classes. To cope up with background variation and illumination changes, transfer learning can be used where first the model can be trained on interactions of simple background images and after that, the weights can be used to train on complex background and images of varying illumination. In the following section, we have discussed some future directions that need to be explored further.

However, many datasets available for HOI detection consist of a large number of classes but still, the domain of the possible HOIs is myriad. Some authors tried zero-shot learning to detect the interactions for which there are no examples but it still needs to be explored further.

Even after detecting the interaction between humans and objects, further research work is required to understand the real motive of the interaction. For example, if a person is touching some object lying on the floor, it creates ambiguity whether the person is picking or putting something on the floor. It can only be done when context information is included while detecting the interaction. In some techniques, frameworks are designed to include context information but still, there is room for further research.

Many interactions with the objects become clearer when the pose of the person is taken into the account. A lot more work is required to be done in this direction to get better results on HOI.

REFERENCES

1. S. Gupta and J. Malik, “Visual semantic role labeling,” 2015 [Online]. Available https://arxiv.org/abs/1505.04474.
2. C. Lu, R. Krishna, M. Bernstein, and F. F. Li, “Visual relationship detection with language priors,” in *Computer Vision - ECCV 2016*. Heidelberg, Germany: Springer, 2016, pp. 852-869.
3. S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, “Learning human-object interactions by graph parsing neural networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 401-417.
4. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014 [Online]. Available https://arxiv.org/abs/1409.1556.
5. X. Yang, T. Zhang, and C. Xu, “Deep-structured event modeling for user-generated photos,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2100-2113, 2017.
6. W. Chu, H. Xue, C. Yao, and D. Cai, “Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos,” *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 246-255, 2018.
7. F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, “Activitynet: a large-scale video benchmark for human activity understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 961-970.
8. H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, “Video paragraph captioning using hierarchical recurrent neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4584-4593.
9. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 4651-4659.
10. L. Yu, E. Park, A. C. Berg, and T. L. Berg, “Visual Madlibs: fill in the blank image generation and question answering,” 2015 [Online]. Available https://arxiv.org/abs/1506.00278.
11. A. Mallya and S. Lazebnik, “Learning models for actions and person-object interactions with transfer to question answering,” in *Computer Vision - ECCV 2016*. Heidelberg, Germany: Springer, 2016, pp. 414-428.
12. W. Norcliffe-Brown, S. Vafeias, and S. Parisot, “Learning conditioned graph structures for interpretable visual question answering,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 8334-8343, 2018.
13. L. Li, Z. Gan, Y. Cheng, and J. Liu, “Relation-aware graph attention network for visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 10313-10322.
14. B. Yao and F. F. Li, “Modeling mutual context of object and human pose in human-object interaction activities,” in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 17-24.
15. A. Gupta, A. Kembhavi, and L. S. Davis, “Observing human-object interactions: using spatial and functional compatibility for recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775-1789, 2009.
16. B. Yao and F. F. Li, “Grouplet: A structured image representation for recognizing human and object interactions,” in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 9-16.
17. J. F. Hu, W. S. Zheng, J. Lai, S. Gong, and T. Xiang, “Recognising human-object interaction via exemplar based modelling,” in *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp. 3144-3151.
18. C. Desai, D. Ramanan, and C. Fowlkes, “Discriminative models for static human-object interactions,” in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, 2010, pp. 9-16.
19. S. Maji, L. Bourdev, and J. Malik, “Action recognition from a distributed representation of pose and appearance,”

in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 3177-3184.

20. V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1503-1511, 2011.
21. B. Yao and F. F. Li, "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*, vol. 34, no. 9, pp. 1691-1703, 2012.
22. T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 9677-9685.
23. Y. L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, pp. 3585-3594.
24. Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 482-490.
25. T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90-126, 2006.
26. P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473-1488, 2008.
27. S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: a survey," *Image and Vision Computing,* vol. 60, pp. 4-21, 2017.
28. I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "Visionbased human action recognition: An overview and real world challenges," *Forensic Science International: Digital Investigation*, vol. 32, article no. 200901, 2020. https://doi.org/10.1016/j.fsidi.2019.200901
29. J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1-43, 2011.
30. R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.
31. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 2004, pp. 32-36.
32. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 12, pp. 2247-2253, 2007.
33. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.
34. K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," Center for Research in Computer Vision, Orlando, FL, *Report No. CRCV-TR-12-01*, 2012.
35. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of 2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2556-2563.
36. F. F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *Proceedings of 2004 Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, 2004, pp. 178-178.
37. G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Pasadena, CA, *Report No. CNS-TR-2007-001*, 2007.
38. S. A. Nene, S. K. Nayar, H. Murase et al., "Columbia object image library (coil-100)," Department of Computer Science, Columbia University, New York, NY, *Report No. CUCS-006-96*, 1996.
39. A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large data set for nonparametric object and scene recognition,*" IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958-1970, 2008.
40. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "ImagNet: a large-scale hierarchical image database," in *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 248-255.
41. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.
42. D. T. Le, J. Uijlings, and R. Bernardi, "TUHOI: Trento universal human object interaction dataset," in *Proceedings of the 3rd Workshop on Vision and Language*, Dublin, Ireland, 2014, pp. 17-24.
43. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: new benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 3686-3693.
44. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, et al., "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision,* vol. 123, no. 1, pp. 32-73, 2017.
45. Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: a benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1017-1025.
46. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV 2014*. Heidelberg, Germany: Springer, 2014, pp. 740-755.
47. B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. V. d. Hengel,

"Care about you: towards large-scale human-centric visual relationship detection," 2017 [Online]. Available https://arxiv.org/abs/1705.09892.

48. Y. W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018, pp. 381-389.
49. M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 1745-1752.
50. A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601-614, 2011.
51. B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and F. F. Li, "Human action recognition by learning bases of action attributes and parts," in *Proceedings of 2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1331-1338.
52. D. T. Le, R. Bernardi, and J. Uijlings, "Exploiting language models to recognize unseen actions," in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, Dallas, TX, 2013, pp. 231-238.
53. J. Johnson, R. Krishna, M. Stark, L. J. Li, D. Shamma, M. Bernstein, and F. F. Li, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 3668-3678.
54. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. J. Li, "YFCC100M: the new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64-73, 2016.
55. V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *Proceedings of the British Machine Vision Conference*, Aberystwyth, UK, 2010.
56. A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.
57. M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: visual semantic role labeling for image understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 5534-5542.
58. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 3485-3492.
59. M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.
60. J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1728-1743, 2011.
61. H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14-29, 2015.
62. J. M. Chaquet, E. J. Carmona, and A. Fernandez-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633-659, 2013.
63. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 886-893.
64. A. Bobick and J. Davis, "An appearance-based representation of action," in *Proceedings of 13th International Conference on Pattern Recognition*, Vienna, Austria, 1996, pp. 307-312.
65. W. L. Lu and J. J. Little, "Simultaneous tracking and action recognition using the PCA-HOG descriptor," in *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision*, Quebec, Canada, 2006, pp. 6-6.
66. M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Lahaina, HI, 1991, pp. 586-587.
67. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
68. M. Y. Chen and A. Hauptmann, "MoSIFT: recognizing human actions in surveillance videos," School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, *Report No. CMU-0CS-09-161*, 2009.
69. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*, Augsburg, Germany, 2007, pp. 357-360.
70. Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, 2004, pp. II-II.
71. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, 2005.
72. K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582-1596, 2009.
73. H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision - ECCV 2006.* Heidelberg, Germany: Springer, 2006, pp. 404-417.
74. A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520-527, 2007.
75. T. Kinnunen, J. K. Kamarainen, L. Lensu, J. Lankinen, and H. Kaviainen, "Making visual object categorization more challenging: randomized caltech-101 data set," in *Proceedings of 2010 20th International Conference on Pattern Recognition*,

Istanbul, Turkey, 2010, pp. 476-479.

76. N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, “Recognizing actions from still images,” in *Proceedings of 2008 19th International Conference on Pattern Recognition*, Tampa, FL, 2008, pp. 1-4.
77. Y. Zheng, Y. J. Zhang, X. Li, and B. D. Liu, “Action recognition in still images using a combination of human pose and context information,” in *Proceedings of 2012 19th IEEE International Conference on Image Processing*, Orlando, FL, 2012, pp. 785-788.
78. C. Thurau and V. Hlavac, “Pose primitive based human action recognition in videos or still images,” in *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.
79. Y. Wang, H. Jiang, M. S. Drew, Z. N. Li, and G. Mori, “Unsupervised discovery of action classes,” in *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, 2006, pp. 1654-1661.
80. W. Yang, Y. Wang, and G. Mori, “Recognizing human actions from still images with latent poses,” in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 2030-2037.
81. C. Desai and D. Ramanan, “Detecting actions, poses, and objects with relational phraselets,” in *Computer Vision - ECCV 2012*. Heidelberg, Germany: Springer, 2012, pp. 158-172.
82. N. Ikizler-Cinbis and S. Sclaroff, “Object, scene and actions: combining multiple features for human action recognition,” in *Computer Vision - ECCV 2010*. Heidelberg, Germany: Springer, 2010, pp. 494-507.
83. C. Y. Chen and K. Grauman, “Predicting the location of “interactees” in novel human-object interactions,” in *Computer Vision - ACCV 2014*. Heidelberg, Germany: Springer, 2014, pp. 351-367.
84. B. Yao, A. Khosla, and F. F. Li, “Combining randomization and discrimination for fine-grained image categorization,” in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 1577-1584.
85. G. Sharma, F. Jurie, and C. Schmid, “Expanded parts model for human attribute and action recognition in still images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, Portland, OR, pp. 652-659.
86. F. S. Khan, R. M. Anwer, J. Van De Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, “Coloring action recognition in still images,” *International Journal of Computer Vision*, vol. 105, no. 3, pp. 205-221, 2013.
87. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989.
88. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.
89. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 580-587.
90. R. Girshick, “Fast R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1440-1448.
91. K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
92. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 91-99, 2015.
93. G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R*CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1080-1088.
94. S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2012.
95. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 568-576, 2014.
96. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 2625-2634.
97. Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA 2015, pp. 1110-1118.
98. A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, “Deepproposals: Hunting objects and actions by cascading deep convolutional layers,” *International Journal of Computer Vision*, vol. 124, no. 2, pp. 115-131, 2017.
99. G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510-1517, 2017.
100. L. Wang, L. Ge, R. Li, and Y. Fang, “Three-stream CNNs for action recognition,” *Pattern Recognition Letters*, vol. 92, pp. 33-40, 2017.
101. R. Girdhar and D. Ramanan, “Attentional pooling for action recognition,” *Advances in Neural Information Processing Systems,* vol. 30, pp. 34-45, 2017.
102. H. Zhang, Z. Kyaw, S. F. Chang, and T. S. Chua, “Visual translation embedding network for visual relation detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 5532-5540.
103. J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 670-685.

104. G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy, “Zoom-net: Mining deep feature interactions for visual relationship recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 322-338.
105. Y. Li, W. Ouyang, X. Wang, and X. Tang, “VIP-CNN: visual phrase guided convolutional neural network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 1347-1356.
106. B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 3076-3086.
107. J. Peyre, J. Sivic, I. Laptev, and C. Schmid, “Weakly-supervised learning of visual relations,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 5179-5188.
108. X. Liang, L. Lee, and E. P. Xing, “Deep variation-structured reinforcement learning for visual relationship and attribute detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 848-857.
109. B. Zhuang, L. Liu, C. Shen, and I. Reid, “Towards context-aware interaction recognition for visual relationship detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 589-598.
110. J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, “Relationship proposal networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 5226-5234.
111. H. Zhang, Z. Kyaw, J. Yu, and S. F. Chang, “PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 4233-4241.
112. J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: object detection via region-based fully convolutional networks,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 379-387, 2016.
113. L. Shen, S. Yeung, J. Hoffman, G. Mori, and F. F. Li, “Scaling human-object interaction recognition through zero-shot learning,” in *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018, pp. 1568-1576.
114. D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, “Detecting human-object interactions with action co-occurrence priors,” in *Computer Vision - ECCV 2020*. Heidelberg, Germany: Springer, 2020, pp. 718-736.
115. G. Gkioxari, R. Girshick, P. Dollar, and K. He, “Detecting and recognizing human-object interactions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359-8367.
116. H. S. Fang, J. Cao, Y. W. Tai, and C. Lu, “Pairwise body-part attention for recognizing human-object interactions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 51-67.
117. C. Gao, Y. Zou, and J.-B. Huang, “iCAN: instance-centric attention network for human-object interaction detection,” 2018 [Online]. Available https://arxiv.org/abs/1808.10437.
118. B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, “Interact as you intend: Intention-driven human-object interaction detection,” *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423-1432, 2019.
119. B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, “Pose-aware multi-level feature network for human object interaction detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 9469-9478.
120. P. Zhou and M. Chi, “Relation parsing neural network for human-object interaction detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 843-851.
121. T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, “Cascaded human-object interaction recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 4263-4272.
122. O. Ulutan, A. Iftekhar, and B. S. Manjunath, “VSGNet: spatial attention network for detecting human object interactions using graph convolutions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 13617-13626.
123. Y. L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H. S. Fang, Z. Ma, M. Chen, and C. Lu, “PaStaNet: toward human activity knowledge engine,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 382-391.
124. Y. L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu, “Detailed 2D-3D joint representation for human-object interaction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 10166-10175.
125. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770-778.
126. Y. L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, M. Chen, Z. Ma, S. Wang, H. S. Fang, and C. Lu, “HAKE: human activity knowledge engine,” 2019 [Online]. Available https://arxiv.org/abs/1904.06539.

Sunaina

Sunaina received an undergraduate degree in 2012 and post graduate degree in 2015 in computer applications from Punjabi University, Patiala, Punjab, India. Currently she is pursuing Ph.D. degree in computer science. Her research interests include computer vision, deep learning and image processing.

Ramanpreet Kaur

Ramanpreet Kaur received an undergraduate degree in computer applications in 2010 from Panjab University, Chandigarh, India. She received her postgraduate degree in computer applications in 2013 and M.Phil. in 2016 from Punjabi University, Patiala, India. She started her career as an assistant professor in Khalsa College for Women, Sidhwan Khurd, India in 2018. She is currently pursuing Ph.D. degree in computer science from Punjabi University, Patiala. Her research interests include image processing and computer vision.

Dharam Veer Sharma

Dharam Veer Sharma received his Ph.D. in Computer Science from Punjabi University, Patiala, India. He has teaching experience of 22 years. He is currently working as a Professor and HOD of Department of Computer Science, Punjabi University, Patiala, India. His research interests are in Optical Character Recognition, Natural Language Processing, Green Computing and General Computing.