Journal of Engineering and Applied Sciences 14 (19): 7223-7233, 2019
ISSN: 1816-949X

Computer Vision Methods for Looking at Peopleinteracting with Objects: A Taxonomy and Survey

Sultan M. Almotairi
Department of Natural and Applied Science, Majmaah University Majmaah,
Kingdom of Saudi Arabia, almotairi@mu.edu.sa

Abstract: Human-object interaction recognition is a challenging problem as it is a combination of three challenging tasks in computer vision, namely human-action recognition, object detection and the scene understating. These tasks share many challenges such as the appearance of a human performing a specific action can be a rich source of information and indication about the type of the performed action. other challenges such as occlusions, the layout of the scene, variation of body pose, and object appearance make it very important to understand to distinguish between two similar actions. The scope of this study is limited to actions were humans interacting with objects. Therefore, we introduce a new taxonomic classifications for human-object interaction. Also, we present a number of approaches that have been introduced recently that can be applied to a real-word applications. Finally, we present a number of human-object interaction datasets that are publicly available.

Key words: Human-object interaction methods, human-object interaction datasets, object recogni-tion, human action, human pose estimation1, publicly, taxonomic classifications

INTRODUCTION

Human-Object Interaction (HOI) recognition is a challenging problem as it is a combination of three challenging tasks in computer vision, namely human-action recognition, object detection and the scene understating. These tasks share many challenges such as the appearance of a human performing a specific action can be a rich source of information and indication about the type of the performed action. other challenges such as occlusions, the layout of the scene, variation of body pose and object appearance make it very important to understand to distinguish between two similar actions (Fig. 1). Despite the challenges of HOI, a number of approaches have been introduced recently that can be applied to a real-word applications (Filipovych and Ribeiro, 2011).

Among many practical applications of HOI recognition, the automatic surveillance of public places is of particular interest. For example, HOI recognition can be used for the surveillance of airports and underground transportation stations where detecting of the relationship between a person and their belongings are important as in the case of a person leaving a briefcase unattended. In this survey, we classify the representative methods for HOI recognition into four categorizes, namely, human-centric, object-centric, content-based and grasp-based (Table 1).

Fig. 1: a-d) Examples of humans interacting with objects

Table 1: Taxonomy of grouping human-object interaction approaches

Variables	Description
Human-centric	Global representation (Santhanam *et al*., 2012)
	Part-based representation (Delaitre *et al*., 2011)
	Skeleton representation, (Xian-Jie *et al*., 2005)
Object-centric	Relative location (Desai *et al*., 2010)
	Relative scale (Yao and Fei-Fei, 2010b)
	Appearance (Prest *et al*., 2012b)
Content	Global coherence (Yao and Fei-Fei, 2012)
	Scene cues (Rabinovich *et al*., 2007)
	Object affordance (Stark *et al*., 2008)
Grasp	Both hands (Kjellstrom *et al*., 2010)
	One hand (Wu *et al*., 2010)
	Precision (Filipovych and Ribeiro, 2008)

Human-Centric approaches start by locating the human in an image, then they look for objects around the

human (Peursum *et al*., 2005; Prest *et al*., 2012a, b; Singh *et al*., 2010; Yao and Fei-Fei, 2010a, 2012; Delaitre *et al*., 2011). The representation of a human can be divided into global (Santhanam *et al*., 2012), part-based and skeleton (Xian-Jie *et al*., 2005). The global representation is one of the simplest models of the human body and can be done by creating a bounding box enclosing the human in the image. The bounding-box model can be helpful, if the human body in the image occupies only a small number of pixels. However, while this representation is simple, it provides less information about the underlying interaction. The part-based representation does not requires locating all parts of the human body. It is only requires locating the parts that can be used to identify the action such as head for drinking. Some parts are challenging to detect due to the size of the part comparing to the rest of the human body, e.g., hand and head. The full-Human Pose Estimation (HPE) (alternative name: skeleton representation) provides a more robust data about the type of human-object-interaction. However, due to the non-rigid (i.e., articulated) nature of the human body, HPE requires an extensive calculations to locate the human parts and angle.

In contrast to the human-centric, object-centric approaches detect the objects first and then analyze the human pose. Detecting object is often a simpler task than detecting high-dimensional human poses. Therefore, such an approach can improve the result by adding constraints when estimating the human pose. For instance when a cigarette is detected in a "smoking-a-cigarette" interaction, the head/hand will likely be very close in space to the cigarette location (Wu *et al*., 2010). However, the presence of multiple subjects could mislead the detector as it is dificult to determine their location with regards to which subject.

Content-based approaches use high-level features to provide cues that can be used to improve recognition such as global coherence (also called, mutual context), scene cue and object affordance. The pose of human and objects can serve as mutual context to each other, so, recognizing one will facilitate the recognition of the other and vice versa (Yao *et al*., 2011). However, at the training stage, global coherence requires a fully supervised training and full human-pose information. Also, the presence of multiple objects could mislead the detector, as it is dificult to determine the object that is related to the subject and to the action within a cluttered scene.

Finally, grasp-based approaches use the visual information from the grasping action to distinguish between different interactions such as when a human grasps a fork, grasps a cup or touches a fork, all these have similar arm motion but different grasping action (Filipovych and Ribeiro, 2008).

MATERIALS AND METHODS

Human-centric: This approach can be categorized in terms of the type of representation used to detect the human, namely, global representation, part-based representation and skeleton representation. In global representation, the human is represented by a rectangle that covers the whole body. This representation is more applicable in low-resolution images and it presents lower computational complexity than the other representations. In part-based representation, all or some of the human body parts (e.g., head, hands and legs) are represented by a rectangle or an ellipse that covers that specificity body part. In skeleton representation, the human is represented by a multi-part (i.e., head, arms, forearms, chest, thighs and legs) hierarchical model such as a tree-structured pictorial structures model (Fig. 2).

Peursum *et al*. (2005) used a human-centric approach by using a Bayesian network such that no shape analysis is used for objects. Also, The human-centric approaches was used by Prest *et al*. (2012b) on still images. As shown on Fig. 3, they introduced an approach for learning HOI automatically from weakly labeled images that works by first localizing the human. Secondly, they localize the

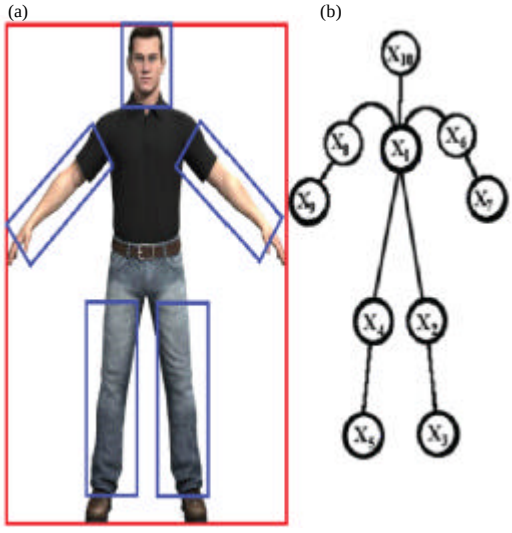

Fig. 2: The global representation is the red rectangle. The part-based representation is the blue rectangles. Right is an illustration of the skeleton using a 10-part tree-structured pictorial structures model (Singh *et al*., 2010)

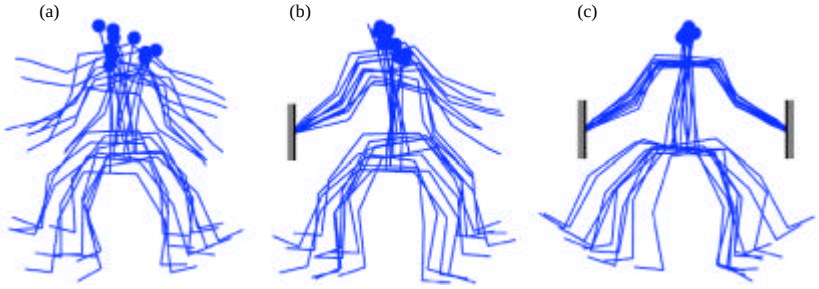

Fig. 3: When the human in contact objects, the freedom of their pose is constrained by the object's location: a) Unconstrained (highly articulated); b) is one hand constrained and c) Both hands constrained (Kjellstrom *et al*., 2010)

action object. Then, they analyze the probability distribution of human object spatial relations (e.g., relative size and relative location).

Yao and Fei-Fei (2010a) used a part-based approach to model a human by using random fields to learn the relation between human body parts and objects. Their model automatically finds the relevant poses for each type of human object interaction as well as the spatial relationships and the connectivity between body parts and objects. However, this approach relies on fully supervised training, for both human body parts and objects. They extend their research by Yao and Fei-Fei (2012) by incorporating a discriminative action classification component that takes the global image information into consideration and learn the overall relationship between difierent actions, human poses and objects rather than modeling each action class separately. Also, they can deal with any number of objects instead of one human and one object. However, the limitations of this method are many. First, they neglect the scene information. Secondly, they rely on fully supervised training for both human body parts and objects. A similar approach was introduced by Delaitre *et al*. (2011) who proposed to replace the standard quantized local HOG/SIFT features, typically used in bag-of-features models with discriminatively trained body part and object detectors. Also, they deploy discriminative selection of interactions using SVM with sparsity-inducing regularizer. A key strength of this method is that they explicitly avoid inferring the complete body configuration.

Object-centric: In a contrasting approach to the human-centric, object-centric approaches to HOI recognition use the appearance of the object to help recognition. For instance, the light that comes from a flashlight, the smoke from the cigarette or the additional constraints used on the human pose while contacting the object. Also, the relative scale size between the human and object can be a helpful cue to improve the accuracy of both object/subject detection (e.g. in case of smoking the size of a cigarette should be smaller than the size of a detected head/hand by a certain amount). However, when a 3-D images are projected to 2-D and if the object is closer to the observer, then the size between subject and object will be inaccurate. The appearance (e.g., texture, color of an object) could also help distinguish between two objects that share the similar sizes such as a tennis ball and a golf ball.

An object-centric approach used by Kjellstrom *et al*. (2010) add more constraints to the degrees of freedom on human pose. When a body part (e.g., hands or feet) is in contact with objects, then more pose restrictions can be added, essentially decreasing the degrees of freedom and improvingthe accuracy of HPE (Fig. 4 and 5). Tracking rigid objects is easier than tracking a highly articulated human body. Kjellstrom *et al*. (2010) state that articulated 3-D tracking of a person can be improved by considering the knowledge about the pose f objects in human's hands. However, they treat objects as "extra body parts" which limits this approach in case of an object not attached to the human (e.g., playing volleyball).

In some event analysis, the uncertainties in the object size, colors and orientation will make the recognition challenging without the use of the object cause (e.g., the use of smoke to detect the object). Wu *et al*. (2010) introduced a mechanism to detect smoking events in video by using anobject-centric approach. In this method, they introduced a color-based ratio histogram analysis that can be used to extract the visual clues from the appearance of an object such as the cigarette light and the smoke. Similarly, Gupta *et al*. (2009) used the lighting of the flashlight to improve the accuracyof their method by encoding such feature.

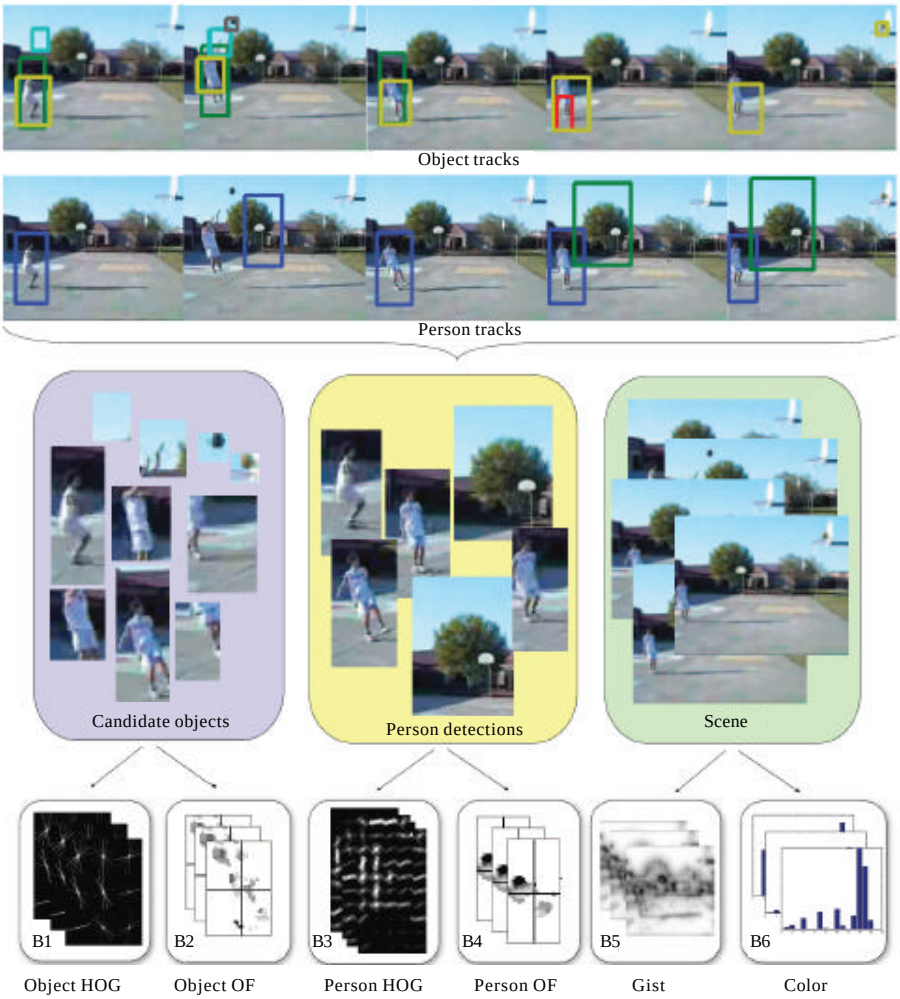

Fig. 4: Ikizler-Cinbis and Sclaroff (2010) use three feature channels, namely scene features, object-centric and the person-centric. After videos are stabilized, they extract candidate object and human tracks. Then, they extract multiple features for all the tracks. A true detections and a noisy detections may be exists

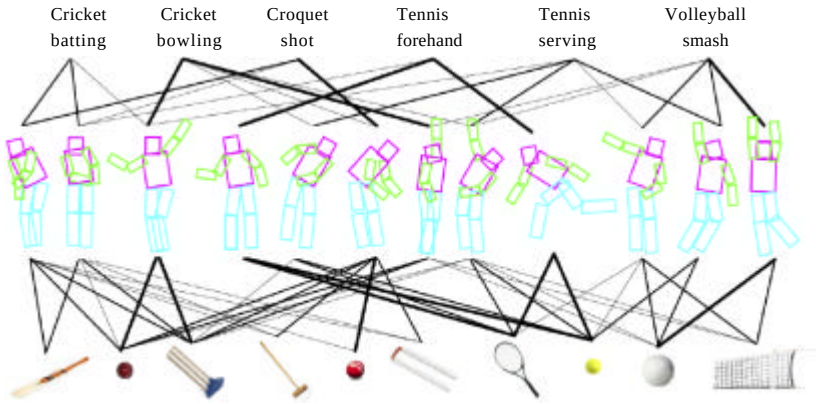

Fig. 5: A graphical illustration of how Mutual Context can improve the recognition by using the relation between object existing and Human Pose Estimation (HPE). Thicker lines indicate stronger connections (Yao and Fei-Fei, 2012)

RESULTS AND DISCUSSION

Content-based: Understanding the surrounding visual information (i.e., context) of an image or video and their different components can be used to improve the recognition rate on many occasions, i.e. when an image has a tennis field that will indicate the type of action that will be preformed. On the other hand, the presence of the ocean will give a different cue of what actions might be taking place. The content orthe high-level feature can be divided into global coherence, scene cue and object affordance. Recently, the use of content in the HOI recognition shows advantages over existing alternatives such as low-level features. Table 2 shows some recent works that deploy the use of content to recognize HOI.

The scene in computer vision refers to the background of an image or video (e.g., sky or tennis court). Scene cues are rich source of information that play an important role in determining and estimating the location of the action whether it has occurred indoors or outdoors. However, using the scene cue requires full image segmentation. The object affordance is an intrinsic property of an object allowing a specific type action to be erformed with the object (e.g., a cup is used to drink not to play tennis), therefore, the accuracy can be improved using the feature. However when more than one object detected in an image, this feature could be misleading.

Gupta *et al*. (2009) refer to the "scene" as node in their Bayesian approach when they refer to the place (e.g., tennis court or a cricket ground) where the action is being performed. However, their approach operates in a fully supervised setting requiring training images with annotation of the human silhouette and the object location. Figure 4 represents the basketball court as an image scene. Marszalek *et al*. (2009) use movie scripts as a means of automatic supervision for training in order to discover the relevant scene classes and their correlation with human actions in movies such as eating, kitchen and running, road. In a similar way, Rabinovich *et al*. (2007) extend the Bag-Of-Features (BOF) Model by incorporating contextual interactions between the scene parts and objects for segmentation and recognition of static images. In real-world applications, scene features can be used to provide a useful information about the possible human object interaction.

Table 2: Content features

Feature	Representative works
Co-occurrence	Wu *et al*. (2017), Al-Akam and Paulus (2018), Sabri *et al*. (2016), Zhu *et al*. (2016), Liu *et al*. (2014), Slimani *et al*. (2014), Yao and Fei-Fei (2012), Delaitre *et al*. (2011), Yao and Fei-Fei (2010b)
Scene cues	Singha *et al*. (2018), Shu *et al*. (2018), Qiu *et al*. (2017), Zhao *et al*. (2017), Ueng and Chen (2016), Gupta *et al*. (2009), Marszalek *et al*. (2009), Rabinovich *et al*. (2007)
Object affordance	Dutta and Zielinska (2019), Dutta and Zielinska (2017), Han *et al*. (2016), Koppula *et al*. (2013, 2012, 2013), Kjellstrom *et al*. (2011), Stark *et al*. (2008)

Recently, there has been increased interest in the use of object affordance for characterizing the HOI and human action. An affordance is an intrinsic property of an object that allows a specific action to be performed on the object. However, the affordance of an object depends on the embodiment of the person performing the action. For instance, a person can use a knife to slice a tomato where as a dog cannot. Hence, the knife affords tomato slicing to human but not to a dog (Kjellstrom *et al*., 2011). However, Koppula *et al*. (2012) encode the information about object affordance as a feature to improve the recognition rate on sub-activity action. They also overcome the problem of detecting small object occluded fully or partially, using the object affordance feature. Object affordance features such as reachable, pourable, movable, drinkable, openable, containable or placeable, can serve as anextra feature that could lead improvement in detecting objects and recognizing actions.

Some research has been aimed at applying the use of the spatial co-occurrence (known as Mutual Context) of individual body parts and objects in HOI recognition. The information about the cooccurrence can be useful for coherently modeling the human pose and the used objects within a specific action (Fig. 5). The co-occurrence/mutual context has been introduced by (Yao and Fei-Fei, 2010b) where the human pose estimation and the object detection can benefit from each other. They use a set of training images to model and discover the relevant poses for each type of HOI activity and its spatial relationship with the objects and body parts which significantly improves the performance of both HPE and object detection. However, they rely on a fully supervised training image for both human body parts and objects. Figure 5 shows the learned strength of connectivity between different activities, objects and the human-pose. Therefore, when we have positive result of HPE, it will give a strong indication of what object/objects will be present and their estimate location. Also when, we have positive object detection it will indicate and help facilitate the estimation of the human pose with regard to the detected object. Other than the use of facilitating the estimation of the human pose, Gupta *et al*. (2009) use the co-occurrence clue between the location of the object with regard to human pose by computing the spatial constraints between the human and locations of manipulable objects for different poses (Fig. 6).

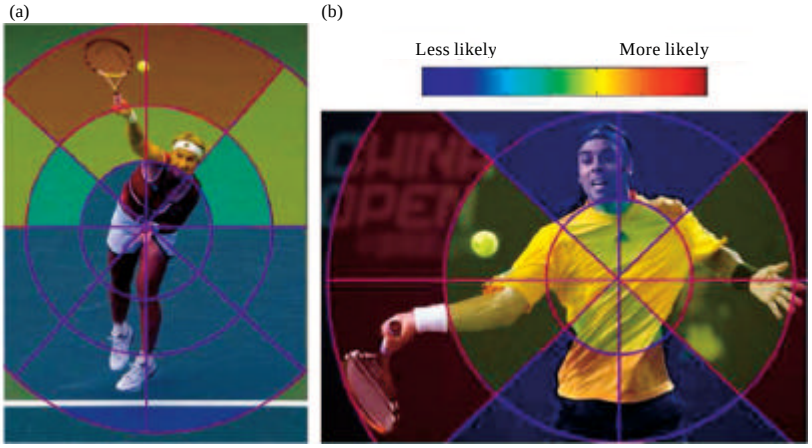

Fig. 6: The co-occurrences spatial constraints between the humans and the locations of manipulable objects and for different body configuration. The left image is when a person preform a tennis-serve, it is more likely to have the ball above the person. Whereas in case of a forehand pose, it is more likely to have the ball on the side. This was done by using two radial bins and eight orientation bins to describe the location of the manipulable object with respect to the person (Gupta *et al*., 2009) promising direction might be to combine

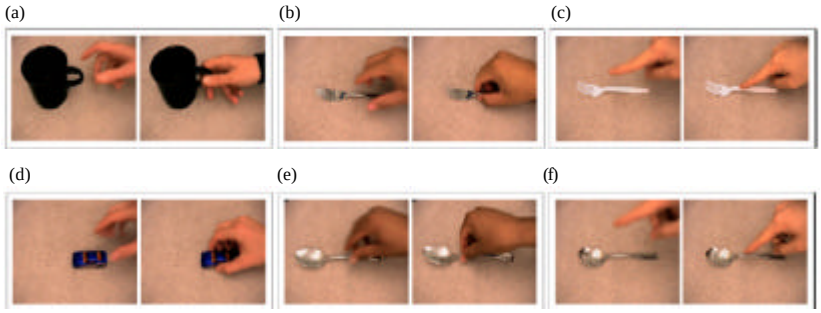

Fig. 7: Different hand poses when performing different actions (Filipovych and Ribeiro, 2008): a) Grasp cup; b) Grasp fork; c) Touch fork; d) Grasp car; e) Grasp spoon and f) Touch spoon

This method shows the power of using the spatial co-occurrence information to relate the spatial location of object and human. However, their method relies on a fully supervised approach for the training part. Similarly, Gupta and Davis (2007) present a graphical Bayesian model that enforces the global coherence between reach motion, object perception, object reaction and manipulation motion for modeling HOI. Also, the co-occurrence can be done not only between the human body as a whole and objects but can also be done between the individual human parts and object as presented by Delaitre *et al*. (2011). A key strength of this method is that they can by passes the difficult problem of estimating the complete human body pose configuration using structural SVMs.

Grasp-based: A gesture is a spatial-temporal pattern which can be either dynamic, static or both. The static shapeof hand is called postures where as the hand movements are called gestures. However, the arm motionand the hand gesture features are promising direction to recognize actions that only involve a moving hand and arm (such as smoking). In this case, the full body model may not be necessary, however, the resolution of the image is more crucial as input. Figure 7 shows the different hand postures on different action. The grasp classes can be useful, if we train the algorithm, for example with actionsthat require the use of two hands such as playing golf or in action that always uses one hand or even a simple action such as holding a cigarette or holding a pen. However, this use of grasp visualinformation requires a high-resolution data.

The grasping motion has been encoded by Filipovych and Ribeiro (2008) by introducing the concept of actor-object states. Their approach is based on the observation that at the moment of physical contact, both the appearance of actors and the motion are constrained

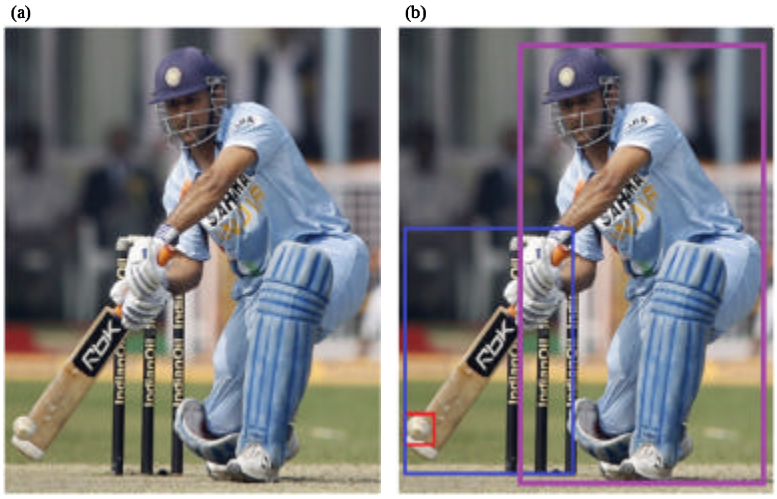

Fig. 8: a) The weakly supervised learning where the image is annotated by the action name, like playing cricket and b) Right is the fully supervised learning where the human, bat and ball are localized in fuchsia, blue and red boxes, respectively, along with the action name (Prest *et al*., 2012b)

by the used object (Fig. 7). They present a probabilistic graphical framework of primitive actor-object interactions that includes information about actor-object static appearances, the interaction's dynamics and the spatial configurations. However, one of the weakness of this approach is the difficulty of discovering the actor-object state in a real-world application. In a different way, Kjellstrom *et al*. (2010) state that with one object, the possible human-object contact (grasping) can be using, right hand-object contact, left hand-object contact, both hands-object contact or no object contact. Then, they use two extra dimensions to encode this information of hand-object contact constraint to improve the accuracy of the arm pose estimation, therefore, creating an accurate global pose (Fig. 8).

Human-object interaction datasets: A number of human-object interaction datasets are publicly available. A frequently used dataset is Gupta dataset (Gupta *et al*., 2009). It consists of ten subjects performing six interactions with four different objects. The objects in the datasets are phone, cup, flashlight and a spray bottle. The human-object interactions with these objects are spraying from a spray bottle, pouring from a cup, making a phone call, answering a phone call, lighting the flashlight and drinking from a cup.

PPMI is a human-object interaction dataset focused on People Playing Musical Instruments (Yao and Fei-Fei, 2010a). It consists of seven different musical instruments, violin, guitar, French horn, flute, saxophone and bassoon. One class includes over 150 images of human playing instruments (called PPMI+). The second class called PPMI and it includes over 150 images of humans holding the instruments without playing. Primitive interactions dataset consists of ten individuals in two different scenarios (clean and cluttered backgrounds) preforming eight different actor-object interaction (Filipovych and Ribeiro, 2008). The interactions are push a toy car, touch a toy car, grasp a toy car, touch a spoon, grasp a spoon, touch a fork, grasp a fork and grasp a cup. Each actor performed interactions with a unique collocation of objects.

Coffee and cigarettes are annotated, comprehensive and realistic interaction datasets created by exploiting the movie "Coffee and Cigarettes" (2003) providing an comprehensive collocation of natural samples for the drinking action (105 samples) and for the smoking action over 140 samples (Laptev and Perez, 2007). However, these actions can appear in different scenes and performed by differentactors while being observed using about two different view points. In addition, they use 32 drinking samples from the movie "Sea of Love" and about 33 drinking examples recorded in their lab.

Recently, a new high quality dataset has been collected by the Center for Research in Computer Visionin University of Central Florida, called UCF101 (Soomro *et al*., 2012). It is an action recognition dataset of realistic action videos containing some human-object interactions collected from YouTube. This dataset consists of interaction with objects including, tennis swing, playing guitar, jump rope, apply lipstick, apply eye makeup, cricket shot and more.

Another interesting but less used datasets are available from many researchers, for example, a grasping

Table 3: Available datasets

Dataset	Description and evaluations
PPMI (Yao and Fei-Fei, 2010a)	People-playing-musical-instruments, includes +150 PPMI+ and +150 PPMI-images
(Gupta *et al*., 2009)	Ten subjects performing six interactions with four different objects.
Primitive interactions (Filipovych and Ribeiro, 2008)	Videos of eight different actor-object interaction types performed by ten individuals in a clean and a cluttered backgrounds
Coffee and cigarettes (Laptev and Perez, 2007)	Pool of natural samples for the action classes smoking and drinking
Grasping object (Romero *et al*., 2010)	Over 1,000,000 images, consisting 33 object grasping actions recorded from 648 different viewpoints
Hammer HOI (Hamer *et al*., 2010)	Hand-object-interaction dataset consists of nine different person male and female with different hand size Human handle camera, can, cup, cup, cup, flute, phone, pliers, sprayer,tennis ball
Hammer depth (Gall *et al*., 2011)	Depth data with registered video sequences for six subjects, 174 object manipulations and 13 action classes
GTEA (Fathi *et al*., 2011)	Seven types of daily activities, each performed by four different subjects. The camera is mounted on a cap worn by the subject
UCF 101 (Soomro *et al*., 2012)	13,320 videos from 101 action categories. It contains realistic action videos collected from YouTub.
Egocentric Dataset (Ren and Philipose, 2009)	Ten video sequences from two human subjects manipulating 42 everyday object instances recorded using a wearable camera
Willow-action (Delaitre *et al*., 2010)	900 images with more than 1000 labeled person from seven different human action classes

Table 4: List of available datasets

Dataset	Web address (URL)
YouTube-8M	<https://research.google.com/youtube8m/>
PPMI (Yao and Fei-Fei, 2010a)	<http://www.cs.cmu.edu/~abhinavg>
Primitive interactions (Filipovych and Ribeiro, 2008)	<http://cs.fit.edu/~eribeiro>
Coffee and cigarettes (Laptev and Perez, 2007)	<http://www.di.ens.fr/~laptev/download.html>
Grasping object (Romero *et al*., 2010)	<http://www.cas.kth.se/~jrgn>
Hammer HOI (Hamer *et al*., 2010)	<http://www.vision.ee.ethz.ch/~gallju>
Hammer depth (Gall *et al*., 2011)	<http://www.vision.ee.ethz.ch/~gallju>
GTEA (Fathi *et al*., 2011)	<http://www.cc.gatech.edu/~afathi3>
UCF 101 (Soomro *et al*., 2012)	<http://crcv.ucf.edu/data/UCF101.php>
Egocentric dataset (Ren and Philipose, 2009)	<http://seattle.intel-research.net/~xren/egovision09>
Willow-action (Delaitre *et al*., 2010)	<http://www.di.ens.fr/willow/research/stillactions/>

Table 5: Comparison of HOI recognition methods

Variables	Advantages	Disadvantages
Global rep	Easy to detect in low-resolution	Less information and cue about the human lower computational complexity
Part-based rep	Easy to detect some parts (head) No need to locate all body parts	Hard to detect other parts (legs, hand)
Skeleton rep	Provides a very strong indication about HOI Helpful for object detection	Expensive computational complexity not helpful in all H_{0I} (e.g., smoking)
Relative location	Provide a cue about the human parts location	Mislead if more than one subject/object within a frame
Relative scale	Improve the accuracy of both object/subject detection	When a 3-D image projected in 2-D, if the object is closer to the observer, then the size scale between subject and object will be inaccurate
Appearance	Helps distinguishing between objects that has same shape (e.g., tennis ball and golf ball)	Requires colored data
Global coherence	Increase the accuracy of HOI where as more logical constrains are added to enforce the global coherence	At training stage, it requires fully supervised training
	Easy to calculate	Misleading, if more than one subject within a frame
Scene cues	Determine indoor or outdoor	HOI Requires full frame segmentation
Object affordance	Provide a useful information about HOI	Could misled if there is more than one object in a frame
Both hands	Does not required fingers pose analysis	Self-occlude and misleading, if more than one human involve
One hand	Cue to find the manipulated object location	Fingers pose estimation is required require high-resolution data
Precision	Useful to distinguish between tiny HOI	Require high-resolution data

object dataset (Romero *et al*., 2010). It contains more than 1,000,000 images, consisting of five different time steps of 33 object grasping actions recorded from 648 different viewpoints. The Hammer HOI dataset consists of nine different persons males and females with different hand sizes (Hamer *et al*., 2010). The human is handling camera, can, cup, flute, phone, pliers, sprayer and tennis ball. The hammer depth dataset is a dataset that comprises depth data with registered video sequences for 13 action classes, six subjects and 174 object manipulations (Gall *et al*., 2011). Willow-action contains about 900 images with more than 1000 labeled persons from seven different human actionclasses (Delaitre *et al*., 2010). The classes are walking, running, riding horse, riding bike, playing instrument, photographing and interacting with computer.

Finally, there are view datasets for HOI called egocentric where a wearable camera is placed on thehead of the subject to present the action from the view of the human. For example, Georgia Tech Egocentric Activities (GTEA) dataset contains seven types of daily activities, each performed by four different subjects (Fathi *et al*.,

2011). The camera is mounted on a cap worn by the subject. Also, Ren and Philipose (2009) collected a comprehensive egocentric dataset using a high-quality wearable video camera. This dataset includes ten video sequences from two human subjects manipulating 42 every day object. For a complete list of datasets, please refer to Table 3 and for URL to these datasets, please refer to Table 4.

Future directions: In the previous sections, we provided an overview of the HOI recognition methods. Despite these challenges, researchers have made continuous and substantial progress in all aspects of the HOI recognition. However, the lack of accurate HPE, grasping details and contextual information cause some failures to many methods. Table 5 describes each method and identifies its weaknesses and advantages.

CONCLUSION

However, there are still many open problems and other avenues available for future study and investigation. These directions include: the encapsulating of the physical grasping motion. The grasping of object can be different based on the shape of the object. This could be used to improve the classification rate. In this case, we expect that the hand gesture is constrained by the target object (Filipovych and Ribeiro, 2008), therefore, the recognition of one facilitates the recognition of the other.

The use of contextual information and the logical relation between human, object and thescene can be used to improve the classification score. These features can be used to serve in the as final stage as contextual facilitation to improve the recognition rate (Yao and Fei-Fei, 2010b). The development of a strong human pose estimation and object detection. However, when human make contact with an object, the object's location can be used to reduce the degree of the pose freedom which will improve the accuracy of HPE (Kjellstrom *et al*., 2010).

The use of object affordance (Koppula *et al*., 2012) can be used to reduce the search space by excluding the non-logical relations between the object's functionality (e.g., drinkable) and an action (e.g., tennis-serve). Kjellstrom *et al*. (2011) used the afford ance to classify objects grasping in a fully labeled training data. The use of semi-supervised training to retrieve the affordance of an object can be a promising direction.

Finally, a possible direction of investigation is to further study the hand gesture when manipulatingan object (Hamer *et al*., 2009). However, most works on HOI recognition are focused on estimating the human-pose and little attention has been given to the study of hand gesture when interacting with an object. A promising direction might be to combine the analysis of the hand configuration with the existing features.

REFERENCES

Al-Akam, R. and D. Paulus, 2018. Dense 3D optical flow co-occurrence matrices for human activity recognition. Proceedings of the 5th International Workshop on Sensor-based Activity Recognition and Interaction (iWOAR'18), September 20-21, 2018, ACM, Berlin, Germany, pp: 1-8.

Delaitre, V., I. Laptev and J. Sivic, 2010. Recognizing human actions in still images: A study of bag-of-features and part-based representations. Proceedings of the 21st British Conference on Machine Vision (BMVC 2010), August 31-September 3, 2010, Aberystwyth, UK., pp: 1-11.

Delaitre, V., J. Sivic and I. Laptev, 2011. Learning Person-Object Interactions for Action Recognition in Still Images. In: Advances in Neural Information Processing Systems 24, Shawe-Taylor, J., R.S. Zemel, P.L. Bartlett, F. Pereira and K.Q. Weinberger (Eds.). Curran Associates Inc., New York, USA., pp: 1503-1511.

Desai, C., D. Ramanan and C. Fowlkes, 2010. Discriminative models for static human-object interactions. Proceedings of the 2010 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 9-16.

Dutta, V. and T. Zielinska, 2017. Action prediction based on physically grounded object affordances in human-object interactions. Proceedings of the 2017 11th International Workshop on Robot Motion and Control (RoMoCo), July 3-5, 2017, IEEE, Wasowo, Poland, ISBN:978-1-5386-3927-6, pp: 47-52.

Dutta, V. and T. Zielinska, 2019. Predicting human actions taking into account objecta ordances. J. Intell. Rob. Syst., 93: 745-761.

Fathi, A., X. Ren and J.M. Rehg, 2011. Learning to recognize objects in egocentric activities. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2011), June 20-25, 2011, IEEE, Colorado, USA., pp: 3281-3288.

Filipovych, R. and E. Ribeiro, 2008. Recognizing primitive interactions by exploring actor-object states. Proceedings of the 2008 IEEE International Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, IEEE, Anchorage, Alaska, USA., pp: 1-7.

Filipovych, R. and E. Ribeiro, 2011. Robust sequence alignment for actor-object interaction recognition: Discovering actor-object states. Comput. Vision Image Understanding, 115: 177-193.

Gall, J., A. Fossati and L. Van Gool, 2011. Functional categorization of objects using real-time markerless motion capture. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2011), June 20-25, 2011, IEEE, Colorado, USA., pp: 1969-1976.

Gupta, A. and L.S. Davis, 2007. Objects in action: An approach for combining action understanding and object perception. Proceedings of the 2007 IEEE International Conference on Computer Vision and Pattern Recognition, June 17-22, 2007, Minneapolis, Minnesota, USA., pp: 1-8.

Gupta, A., A. Kembhavi and L.S. Davis, 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE. Trans. Pattern Anal. Mach. Intell., 31: 1775-1789.

Hamer, H., J. Gall, T. Weise and L. Van Gool, 2010. An object-dependent hand pose prior from sparse training data. Proceedings of the 2010 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 671-678.

Hamer, H., K. Schindler, E. Koller-Meier and L. Van Gool, 2009. Tracking a hand manipulating an object. Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, September 29-October 2, 2009, IEEE, Kyoto, Japan, pp: 1475-1482.

Han, J.H., S.J. Lee and J.H. Kim, 2016. Behavior hierarchy-based affordance map for recognition of human intention and its application to human-robot interaction. IEEE. Trans. Hum. Mach. Syst., 46: 708-722.

Ikizler-Cinbis, N. and S. Sclaroff, 2010. Object, scene and actions: Combining multiple features for human action recognition. Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), September 5-11, 2010, Greece, pp: 494-507.

Kjellstrom, H., D. Kragic and M.J. Black, 2010. Tracking people interacting with objects. Proceedings of the 2010 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 747-754.

Kjellstrom, H., J. Romero and D. Kragic, 2011. Visual object-action recognition: Inferring object affordances from human demonstration. Comput. Vision Image Understanding, 115: 81-90.

Koppula, H.S., R. Gupta and A. Saxena, 2012. Human activity learning using object affordances from RGB-D videos. Comput. Vision Pattern Recogn., 1: 1-10.

Koppula, H.S., R. Gupta and A. Saxena, 2013. Learning human activities and object affordances from RGB-D videos. Intl. J. Rob. Res., 32: 951-970.

Laptev, I. and P. Perez, 2007. Retrieving actions in movies. Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, October 14-21, 2007, IEEE, Rio de Janeiro, Brazil, pp: 1-8.

Liu, H., M. Liu and Q. Sun, 2014. Learning directional co-occurrence for human action classification. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-9, 2014, IEEE, Florence, Italy, pp: 1235-1239.

Marszalek, M., I. Laptev and C. Schmid, 2009. Actions in context. Proceedings of the CVPR 2009-IEEE International Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, IEEE, Miami, Florida, USA., pp: 2929-2936.

Peursum, P., G. West and S. Venkatesh, 2005. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05) Vol. 1, October 17-21, 2005, IEEE, Beijing, China, pp: 82-89.

Prest, A., C. Schmid and V. Ferrari, 2011. Weakly supervised learning of interactions between humans and objects. IEEE. Trans. Pattern Anal. Mach. Intell., 34: 601-614.

Prest, A., V. Ferrari and C. Schmid, 2012. Explicit modeling of human-object interactions in realistic videos. IEEE. Trans. Pattern Anal. Mach. Intell., 35: 835-848.

Qiu, Z., T. Yao and T. Mei, 2017. Learning spatio-temporal representation with pseudo-3D residual networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, IEEE, Venice, Italy, pp: 5533-5541.

Rabinovich, A., A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie, 2007. Objects in context. Proceedings of the International Conference on (ICCV) Vol. 1, October 14-21, 2007, IEEE, Rio de Janeiro, Brazil, pp: 1-8.

Ren, X. and M. Philipose, 2009. Egocentric recognition of handled objects: Benchmark and analysis. Proceedings of the 2009 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 20-25, 2009, IEEE, Miami, Florida, USA., pp: 1-8.

Romero, J., H. Kjellstrom and D. Kragic, 2010. Hands in action: Real-time 3D reconstruction of hands in interaction with objects. Proceedings of the 2010 IEEE International Conference on Robotics and Automation, May 3-7, 2010, IEEE, Anchorage, Alaska, USA., pp: 458-463.

Sabri, A.Q.M., J. Boonaert, E.R.M.F. Abdullah and A.M. Mansoor, 2016. Spatio-temporal co-occurrence characterizations for human action classification. Malaysian J. Comput. Sci., 30: 154-173.

Santhanam, T., C.P. Sumathi and S. Gomathi, 2012. A survey of techniques for human detection in static images. Proceedings of the 2nd International Conference on Computational Science, Engineering and Information Technology (CCSEIT '12), October 26-28, 2012, ACM, India, pp: 328-336.

Shu, T., Y. Peng, L. Fan, H. Lu and S.C. Zhu, 2018. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. Top. Cognit. Sci., 10: 225-241.

Singh, V.K., F.M. Khan and R. Nevatia, 2010. Multiple pose context trees for estimating human pose in object context. Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 17-24.

Singha, J., A. Roy and R.H. Laskar, 2018. Dynamic hand gesture recognition using vision-based approach for human-computer interaction. Neural Comput. Appl., 29: 1129-1141.

Slimani, K.N.H., Y. Benezeth and F. Souami, 2014. Human interaction recognition based on the co-occurence of visual words. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, June 23-28, 2014, IEEE, Columbus, Ohio, USA., pp: 461-466.

Soomro, K., A.R. Zamir and M. Shah, 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. Comput. Vision Pattern Recogn., 1: 1-7.

Stark, M., P. Lies, M. Zillich, J. Wyatt and B. Schiele, 2008. Functional object class detection based on learned affordance cues. Proceedings of the International Conference on Computer Vision Systems (ICVS2008), May 12-15, 2008, Springer, Berlin, Germany, pp: 435-444.

Ueng, S.K. and G.Z. Chen, 2016. Vision based multi-user human computer interaction. Multimedia Tools Appl., 75: 10059-10076.

Wu, C., J. Zhang, O. Sener, B. Selman and S. Savarese *et al*., 2017. Watch-n-patch: Unsupervised learning of actions and relations. IEEE. Trans. Pattern Anal. Mach. Intell., 40: 467-481.

Wu, P., J.W. Hsieh, J.C. Cheng, S.C. Cheng and S.Y. Tseng, 2010. Human smoking event detection using visual interaction clues. Proceedings of the 2010 20th International Conference on Pattern Recognition, August 23-26, 2010, IEEE, Istanbul, Turkey, pp: 4344-4347.

Xian-Jie, Q., W. Zhao-Qi, X. Shi-Hong and L. Jin-Tao, 2005. Estimating articulated human pose from video using shape context. Proceedings of the 5th IEEE International Symposium on Signal Processing and Information Technology, December 21, 2005, IEEE, Athens, Greece, pp: 583-588.

Yao, B. and L. Fei-Fei, 2010a. Grouplet: A structured image representation for recognizing human and object interactions. Proceedings of the 2010 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 9-16.

Yao, B. and L. Fei-Fei, 2010b. Modeling mutual context of object and human pose in human-object interaction activities. Proceedings of the 2010 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, IEEE, San Francisco, California, USA., pp: 17-24.

Yao, B. and L. Fei-Fei, 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE. Trans. Pattern Anal. Mach. Intell., 34: 1691-1703.

Yao, B., A. Khosla and L. Fei-Fei, 2011. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. Proceedings of the 28th International Conference on Machine Learning (ICML), June 28- July 02, 2011, Bellevue, Washington, USA., pp: 1-8.

Zhao, C., J. Wang and H. Lu, 2017. Learning discriminative context models for concurrent collective activity recognition. Multimedia Tools Appl., 76: 7401-7420.

Zhu, W., J. Hu, G. Sun, X. Cao and Y. Qiao, 2016. A key volume mining deep framework for action recognition. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, IEEE, Las Vegas, Nevada, USA., pp: 1991-1999.