Literature Review on Human-Object Interaction

Na Zhang

List

- [2020] A paper list of HOI (Human-Object Interaction)
- [2021] A detailed paper list of HOI (image/video)

Review

- [2021] A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images
- [2020] Human-Object Interaction Detection: A Quick Survey and Examination of Methods

- **[2021] A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images**

- concentrates on human-centric interactions, which can be categorized as human-to-human and human-to-objects.
- 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.
- 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 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

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

- Earlier work on vision-based recognition mainly focused on action recognition.
- 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
  - interaction with humans.
- 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.

[1] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: a review,” ACM Computing Surveys (CSUR), vol. 43, no. 3, pp. 1-43, 2011.

Challenges

- While dealing with HOI recognition and detection many challenges come in the way which makes this
- Following are some of the while dealing with HOI.
- **A. Occlusion**
  - To carefully detect the HOIs **bject**.
  - But sometimes **the part of** e actor itself while performing the interaction.
  - This issue affects the perfo
  - For example, when a perso e, 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.

- **B. Inter-class similarity**
- If the model


mage or video, it does not i ecause there may be mul inter-class similarity.
- Fig. 2 demo object.
- In Fig. 2(a),
- in Fig. 2(b),
- 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.

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

- **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**
- **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**
- **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.

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

Dataset name	Images	Interactions	Example activities	Yea
PPMI [16]	-	7	Playing violin, playing guitar, playing a flute	201
TBH [50]	341	3	Playing trumpet, wearing hat	201
Stanford40 Action [51]	9,532	40	Brushing the teeth, cleaning the floor	201
89 Action [52]	2,038	89	'Sitting on a chair, 'drinking from a bottle'	201
TUHOI [42]	10,805	2,974	Playing ping-pong, using a laptop, holding a computer mouse	201
MPII [43]	40,522	823	Playing violin, riding a bus, horse grooming	201
HICO [45]	47,774	600	Feeding a giraffe, sailing a boat, talking on a cellphone	201
V-COCO [1]	10,346	26	Laying on the bed, reading a book, working on a laptop, kicking sports ball	201
VRD [2]	5,000	70/37,993 (predicates/relationships)	Person kicking ball, a person on top of the ramp	201
HCVRD [47]	52,855	9,852	Man holding surfboard, a man wearing kneepad	201
HICO-DET [45]	47,776	600	Tying a boat, feeding a bird, riding an airplane	201
HOI-A [24]	38,668	10	Reading document, talking on a mobile phone, kicking sports ball	202

hand-crafted techniques

- SIFT/HOG
- Generative probabilistic models (fully supervised model)
- Discriminative Model (HOG descriptor+Latent SVM)
- Replaced local SIFT/HOG features with trained object and body part detectors
- Exemplar Based Modeling
- Discriminative grouplets (low-level SIFT features)

Deep learning-based

- RCNN
- Fast RCNN + VGG16
- Faster RCNN and deep metric learning module
- Graph RCNN (Relation Proposal network (RePN) + attentional Graph VG convolutional Network(aGCN))
- ResNet
- (Pose aware Multi-level Feature Network

- 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.

- **[2020] Human-Object Interaction Detection: A Quick Survey and Examination of Methods**

- We have classified the methods of solving human-object interaction detection problems into the two classes:
  - multi-stream architectures
  - graph networks.
- **Multi-stream architectures** produce promising results and are easily augmented with supplemental information detection methods such as pose and gaze.
- **Graph neural networks** intuitively connect objects in the image in a graphical form of nodes and connected images, that represent the relationships between objects in the image.

- Human-object interaction detection is a relatively new task in the world of computer vision and visual semantic information extraction.
- With the goal of machines identifying interactions that humans perform on objects, there are many real-world use cases for the research in this field.
- We provide a basic survey of the developments in the field of human-object interaction detection.
- Many works in this field use multi-stream convolutional neural network architectures, which combine features from multiple sources in the input image.
- Most commonly these are the humans and objects in question, as well as a spatial quality of the two.
- As far as we are aware, there have not been in-depth studies performed that look into the performance of each component individually.
- In order to provide insight to future researchers, we perform an individualized study that examines the performance of each component of a multi-stream convolutional neural network architectures for human-object interaction detection.
- Specifically we examine the HORCNN architecture as it is a foundational work in the field. In addition, we provide an in-depth look at the HICO-DET dataset, a popular benchmark in the field of human-object interaction detection.
- Code and papers can be found at https://github.com/SHI-Labs/Human-Object-Interaction-Detection.