Research Statement
Supreeth Narasimhaswamy

My research work spans Computer Vision and Machine Learning. My current research focuses on understanding human hands in visual data and combining computer vision methods with augmented and virtual reality (AR/VR) to assist people in performing complex tasks.

1 Research Progress: Visual Reasoning of Human Hands

*Hands* are the central means by which humans interact with their surroundings. Understanding human hands help human behavior analysis and facilitate other visual analysis tasks such as action and gesture recognition. For example, when a person is playing guitar, detecting hands and the interaction object can help us decide that the possible action is *playing guitar*. Hand analysis is vital for augmented and virtual reality devices. For instance, tracking a user’s hands and recognizing their gestures is essential to provide a good user experience in AR & VR applications. A long-standing robotics goal is to build machines that can perform complex tasks like humans. Understanding how humans grasp objects can provide insight into designing robotic arms for manipulating objects. Human hands are exciting objects of study, and a good understanding of hands can help us develop exciting applications.

(1) Hand Detection. The goal of hand detection is to localize all hand instances present in an image. Most of the existing works on hand detection consider images in specific settings such as ego-centric or first-person views [1, 5, 6]. Detecting hands in unconstrained conditions is challenging since hands are highly deformable with many degrees of freedom and can occur in various shapes, sizes, and orientations. Furthermore, occlusions between other objects and skin areas increase the complexity of the task. As such, generic object detection methods do not work well for detecting hands.

One of the issues with using generic object detection methods, such as Mask-RCNN [3], for hand detection is that it mistakes other skin areas for hands and fails to detect many hand instances. This is because the hand locations are estimated based on local Region of Interest (RoI) features which do not have context information. To address this challenge, we proposed a novel attention method to incorporate and learn contextual cues during hand detection. We design the proposed attention module for two types of non-local contextual pooling, feature similarity and spatial relationship between semantically related entities. Intuitively, a region is more likely to be a hand if other nearby pixels have similar skin tones. We can also infer a hand’s location by other semantically related body parts such as the wrist and elbow. The proposed contextual attention encapsulates these two types of non-local contextual pooling operations. We perform these operations efficiently with a few matrix multiplications and additions and learn the attention module’s parameters and other hand detector parameters end-to-end during training. Fig.1 shows some qualitative results of our method and compares it to MaskRCNN. Additionally, we developed a large-scale dataset of images annotated with hand locations. The dataset is one of the first large-scale hand detection benchmarks focusing on third-person views. The proposed dataset will be helpful to train and evaluate hand detectors. We published this work on hand detection as a conference paper [9] at ICCV 2019.

![Figure 1: Hand Detection.](image)

(2) Hand Physical Contact Recognition. Hand contact recognition is a less-explored and relatively new problem. The goal is to classify hand instances into the following categories: (1) No-Contact, (2) Self-Contact, (3) Person-Contact, and (4) Object-Contact. These conditions are not mutually exclusive, and a hand can be in more than one state. Recognizing the physical contact states of hands has many applications, such as hand-object interaction detection and contamination prevention.
Recognizing hand contact states in-the-wild conditions is challenging because the appearance of the hand alone is insufficient to estimate its contact state. This task also requires us to consider the relationships between the hand and other objects in the scene. It can be a complex inference problem for real-world situations, especially where numerous people and objects surround the hand. Furthermore, even for a pair of hands and objects with corresponding segmentation masks, it is not easy to recognize whether the hand is in contact with the object due to the lack of depth information. Fig. 2 shows two images annotated with hand locations and their physical contact states. It would be challenging to detect hand instances in the image and recognize their contact states.

We propose a novel convolutional network that can jointly learn to localize hands and predict their physical contact to tackle hand contact recognition. The network uses outputs from another object detector to obtain the locations of objects present in the scene. It uses these outputs and hand locations to recognize the hand’s contact state using two attention mechanisms. The first attention mechanism learns the hand and a region’s affinity, enclosing the hand and the object, and densely pools features from this region to the hand region. The second attention module adaptively selects salient features from this plausible contact region. To develop and evaluate our method’s performance, we introduce a large-scale dataset called ContactHands, containing images annotated with hand locations and contact states. The proposed network, including the parameters of attention modules, is end-to-end trainable. Fig. 2 shows some qualitative results from our method. We published this work as a conference paper [8] at NeurIPS 2020.

Figure 2: **Hand Contact Recognition.** The first two images show some ground-truth hand contact states. The notation NC, SC, PC, and OC denotes No-Contact, Self-Contact, Other-Person-Contact, and Object-Contact. The last two images are some qualitative results from the proposed method. We visualize the detected hand masks in coded colors as: **No-Contact**, **Self-Contact**, **Other-Person-Contact**, **Object-Contact**

(3) **Hand-Body Association.** The objective of this problem is to detect hands in an image and find the location of the corresponding person for each detected hand. This task is useful for action recognition and scene understanding, especially for multiple-person images and videos. For example, it is helpful to identify people when understanding hand gestures in human-human communication. Another example is to assess the motor and social skills of children with mental disorders by tracking their hands and how hands interact with objects and other people in a tabletop game. Hand-body association helps develop safety applications and assists people working with hand-held tools in manufacturing settings.

Detecting hands and linking them to the right people is challenging in scenes containing multiple people. One approach is to detect hands and people separately and subsequently use heuristics based on sizes, distances, or overlapping areas to establish correspondences. This approach, however, performs poorly due to the extreme articulation of both hands and human skeletons, leading to tremendous variation in the relative locations and sizes of a hand and the corresponding human body. Another approach is to run a pose detector and find the hands of each detected pose in the image. However, pose detection by itself is unreliable. For a scene of congregated or interacting people, the hand and arm of a person might be mixed up with the skeleton of another person.

We propose a novel convolutional network that can detect hands and their corresponding person end-to-end. We also introduce a new challenging dataset called BodyHands containing images with hand and corresponding person locations annotations. Finally, we demonstrate the benefits of hand-body association in two important applications, hand tracking and hand contact estimation. Our experiments show that the performance of existing hand tracking and hand contact estimation methods can be improved significantly when reasoning about the hand-body association. We illustrate the hand-body association problem in Fig. 3. This work will appear as a conference paper [7] at CVPR 2022.
(4) Hand Tracking. Hand tracking is a fundamental problem in various application scenarios, from gesture and activity recognition to contact tracing and skill evaluation. One approach for tracking hands is to consider them as parts of a human body and then perform hand tracking based on the tracked human pose. However, pose detection and tracking can be unreliable, especially for partially occluded people or outside the camera’s field of view. Existing multiple-object trackers do not work well for hands even though they have impressive performance tracking pedestrians and vehicles.

Hand tracking is difficult because hands are not ordinary objects, given the extreme articulation of hands and the frequent interaction with other objects. In a short period of a few frames, a hand’s size, shape, location, and visibility can change dramatically and frequently. Many existing multiple-object trackers use the detection and linking paradigm. However, hand detection would fail in motion blur and occlusion, while hand linking across time is complex as a hand’s size, location, pose, and appearance can change drastically. Simultaneously, two different hand instances might look alike, so distinguishing them would be difficult even for a sophisticated re-identification module that has been trained specifically for hands.

We propose HandLer, a novel convolutional architecture that can jointly detect and track hands online in unconstrained videos. We build HandLer upon Cascade-RCNN with additional three novel stages. The first stage is Forward Propagation, where we propagate the features from frame $t-1$ to frame $t$ based on previously detected hands and their estimated motion. The second stage is the Detection and Backward Regression, which uses outputs from the forward propagation to detect hands for frame $t$ and their relative offset in frame $t-1$. The third stage uses an off-the-shelf human pose method to link any fragmented hand tracklets. We train the forward propagation, backward regression, and detection stage end-to-end with the other Cascade-RCNN components. To train and evaluate HandLer, we also contribute YouTube-Hand, the first challenging large-scale dataset of videos annotated with hand locations and trajectories. Experiments on this dataset and other benchmarks show that HandLer outperforms the existing state-of-the-art tracking algorithms by a large margin. Fig. 4 shows some hand tracking examples from our work. This work will appear as a conference paper [4] at CVPR 2022.

2 Ongoing and Future Research: Computer Vision & AR/VR

(1) Perceptual Task Guidance. I am interested in a project that is a part of the Perceptual Task Guidance (PTG) program recently announced by DARPA. This aims to develop AI technologies to help users perform complex physical tasks while making them more versatile by expanding their skill-set and more proficient by reducing their errors. The project seeks to develop visual and audio feedback methods to help with task execution. The goal is to provide users with augmented reality (AR) headsets that allow assistants to provide feedback through speech-aligned graphics. The target assistants will learn about tasks relevant to the user by ingesting knowledge from checklists, illustrated manuals, training videos, and other sources of information. They will then combine this task knowledge with a perceptual environment model to support mixed-
initiative and task-focused user dialogues. The dialogues will assist a user in completing a task, identifying and correcting an error during a task, and instructing them through a new task, considering the user’s level of expertise. A fine-grained understanding of user’s hand interactions is quintessential since people use hands to perform tasks and interact with the environment. This includes detecting and tracking hands, tracking objects the hand is contacting, and interacting during the process.

(2) Writing Guidance for Visually Impaired. Paper documents are ubiquitous in our daily lives. We regularly encounter scenarios that require reading and writing. Examples include paper receipts, bank documents, hospital forms, and legal documents. While any paper-centered activity is straightforward for sighted people, working with standard printed materials is challenging for visually impaired people. While several assistive technologies have enabled visually impaired people to read printed materials, assistive technology to help them write on printed paper is an open-ended and technically challenging problem. Recently, there has been development in technologies that allow impaired people to fill out printed forms independently [2]. Nevertheless, such technologies are limited to specific settings such as filling out forms and do not work well in generic day-to-day scenarios. To address this problem, we are working on a computer vision approach. A user wears a wearable sensor and receives audio feedback about the paper content they are actively interacting. One possible feedback can include the location of the pen/pencil relative to the paper and the current piece of text the user is working on. The approach involves computer vision problems such as text recognition, hand activity understanding, and hand gestures. The problem is challenging since hands often occlude the paper, viewpoint and lighting variations, and domain shifts due to numerous possible writing conditions and environments.
References


