2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)

CNN Based Touch Interaction Detection For Infant Speech Development

Qingshuang Chen* Rana Abu-Zhaya †
* *Video and Image Processing Laboratory*
Purdue University
West Lafayette, Indiana USA
chen522@purdue.edu

Amanda Seidl † Fengqing Zhu*
†*Department of Speech, Language, and Hearing Sciences*
Purdue University
West Lafayette, Indiana USA
zhu0@ecn.purdue.edu

***Abstract*—In this paper, we investigate the detection of interaction in videos between two people, namely, a caregiver and an infant. We are interested in a particular type of human interaction known as touch, as touch is a key social and emotional signal used by caregivers when interacting with their children. We propose an automatic touch event recognition method to determine the potential time interval when the caregiver touches the infant. In addition to label the touch events, we also classify them into six touch types based on which body part of infant has been touched. CNN based human pose estimation and person segmentation are used to analyze the spatial relationship between the caregivers hands and the infants. We demonstrate promising results for touch detection and show great potential of reducing human effort in manually generating precise touch annotations.**

***Keywords*-touch event detection; video analysis ;**

I. Introduction

Touch is a key social and emotional signal used by caregivers when interacting with their children [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Touch is presented in an enormous amount of caregiver-infant interactions and its presence has been found to impact infants' attention, arousal levels, behavioral, and emotional states [7], [8], [14], [15], as well as to reduce infants' stress [2]. Touch may be specifically helpful to an infant in language development. Recent work suggests that caregivers do in fact provide their infants with touches that are informative both about the beginnings and ends of words in continuous speech and also about the meanings of words, at least in certain contexts. Specifically, Abu-Zhaya, Seidl, and Cristia[16] recorded caregivers interacting with their infants in a book-reading situation and found that caregiver touch is used in a way that might be helpful to two crucial language learning tasks: segmenting the speech stream into words and mapping words to their referents.

Until recently, the use of touch in mother-infant interactions have employed a micro-genetic approach using frame-by-frame annotation of touch cues yielding a detailed examination of maternal touches during different types of interactions [16]. Not only is annotating these video interactions extremely time consuming, but observers also have to be trained for several hours before they can begin annotating the videos. For example, Abu-Zhaya et al. [16] used ELAN [17] to annotate the touch events. Using their detailed coding, it could take as much as 15 hours to annotate a series of 120 touch events in a 5 minute caregiver-infant interaction. Hence, given the importance of human touch in infant language development, it would be very beneficial to have tools that can easily quantify both the quantity and quality of human touch that infants receive. Having an automatic system that is capable of detecting touch events would greatly reduce the amount of time spent on manually annotating these events. The creation of such an automatic system may also be helpful for medical teams working with special populations and caregivers who have children with special needs.

Touch event is defined as the time when the hands of the caregiver has physical contact with infant in the context of our work. Essentially, a touch will occur when the segmented regions of the hands and the infant overlap. A touch event is further categorized into one of the six touch types (head, arm, hand, torso, leg, foot) based on which body part of infant has been touched. Thus, successfully tracking the hands of the caregiver and clearly detecting the outline of the infant are crucial in our touch event detection. In this paper, we propose an automated method for touch event detection. The contributions of this paper are:

- The proposed method avoids using expensive precision touch event annotations as training data, and takes advantage of training neural network using public available datasets to produce intermediate results, such as human pose information, that are used in the subsequent touch detection step.
- Touch types are also detected based on the position information of caregiver's hands and infant body.
- We experimentally show our method reduces the potential work needed for trained analyst to generates accurate annotations.

II. Related Work

There is a growing need to understand the content of videos, such as human action recognition. Using automated methods to analyze video contents are of great interest due to the high expense and intense labor required to perform these tasks manually. Video action recognition datasets like

UCF101 [18], Hollywood2 [19] have enabled improved performance for these tasks. These datasets [18], [19] consist of small video clips and each video clip has a label for type of actions. A considerable amount of literature has been published on classifying actions in video clips and assign labels to each video clip [20], [21], [22]. However, recognizing pairwise human interaction and making frame level decision is still an open problem.

Previous work [23] in touch event detection uses hand-crafted features for tracking hand and Grab-Cut [24] segmentation for infant contour generation. Without using any learning based methods, the method proposed in [23] requires manual input from users: draw rectangles to initialize hand positions and provides sure foreground and sure background mask for Grab-Cut segmentation. However, tracker re-initialization was not addressed which became problematic for longer videos if the track cannot be recovered and errors can propagate from frame to frame. Later long-term hand tracking [25] with self-correction capability that can re-initialize the tracker position by integrating the human pose [26] and hand detection [27] information improves hand tracking performance. In this paper, we extend the idea of using human pose information to refine caregiver's hands location. Additionally, we use a robust segmentation method, Mask-RCNN, to predict the infant segmentation instead of using graph-based method in [23].

A. Human Pose Estimation

A key step towards understanding people in images and videos is accurate pose estimation [26]. A good pose estimation system need to be robust to occlusion and invariant to changes in appearance due to clothing or lighting condition. Early research focuses on part based models[28], [29], [30] or pictorial structures [31], [32]. Recent human pose estimations has shifted from classical methods like graphical models to deep neural networks. DeepPose [33] is one of the earliest deep-learning based method to estimate human poses which uses a convolutional architecture to directly regress the coordinates of joints. In [34], [26], a structural heat map is predicted to characterize the probabilities of joints at different locations through multiple resolutions.

Previous work [25] uses stacked hour glasses [26] to estimate caregiver's joints locations and refine hand tracking results. Common top down approaches [26], [33], [34] apply single person detector and then estimate pose for each detection. The top down approaches suffer from early commitment. If the person detector fails in crowed or due to occlusion, it is hard to recovery. A bottom up fashion becomes suitable in our application, since the caregiver and infant are always interacting with each other, and occlusion are commonly seen in recorded sequences.

In this paper, we use a bottom up method proposed in [35] to obtain pose information for caregiver and infant via a two-branch CNN. The first branch learns joints location using a multi-stage CNN [36]. The second branch use the same CNN architecture to learn 2D vector fields, namely part affinity fields, that encode the location and orientation of limbs, where limbs refers to joint pairs for human. Predicted joints are assembled to form full-body pose information for caregiver and infant using a greedy parsing algorithm [35]. Hand joint points are generated using an additional hand model described in [37], that the hand model is trained under multiview.

B. Human Segmentation

Infant segmentation in [23] is obtained using Grab-cut [24] with a manually defined mask in the first frame. The mask contains several strokes indicates sure foreground and sure background, and it is updated every frame using segmentation results from the previous frame. However, the Grab-cut based method is sensitive to errors from the mask and the segmentation error may propagates. In this work, a deep neural network called Mask R-CNN [38] is used for infant segmentation. Mask R-CNN extends the object detector Faster R-CNN [27] by adding a fully convolutional network parallel to the classification and bounding box regression networks, and outputs a binary mask to indicate whether a pixel belongs to a candidate region.

III. Method

The proposed touch detection method performs two tasks: (1) identify a frame is touch or non-touch by checking whether the hand segments of the caregiver overlap with the infant segment, (2) classify a detected touch frame into six different touch type classes based on the spatial relationship between keypoints on caregivers' hand and infant body parts. Hand segment updates are introduced in Section III-A, infant segmentation is described in Section III-B, and touch detection decision is described in Section III-C .

A. Hand Location

The pose estimator [35] we used is trained on Microsoft COCO [39] dataset and a foot dataset [35] with a total of 25 joints. After obtaining wrist and elbow position from the pose estimator, the hand joints detector [37] is applied by assuming hand is located at an extend region of forearm in the same direction, we denote as $I_{crop} \in \mathbb{R}^{w \times h \times 3}$. The hand joints detector $h(\cdot)$ maps cropped hand region I_{crop} to N joints locations $\mathbf{x_n}$ associated with a score c_n, where N is 21 in this model. An example of human pose estimation and hand joints estimation are shown in Figure 1, where joints information for infant will be used in Section III-B and Section III-C. We say a hand joint is detected if Equation 1 is equal to one, where $\mathbb{1}(\cdot)$ is an indicator function, α and β are empirically set to be 0.5 and 10, respectively.

$$Confidence = \mathbb{1}((\sum_{n \in [1 \ldots N]} \mathbb{1}(c_n > \alpha)) > \beta) \qquad (1)$$

We extend the tightest bounding box enclosing all hand joints ten pixels in horizontal and vertical direction, and use the extended rectangular as the hand bounding box. If the detected hand is not confident or the pose estimation does not provide wrist or elbow position, then a feature based tracking method is enabled to continue updating hand position. The hand tracker uses color and motion features as keypoints to track in each frame. Due to the lack of feature points in small hand bounding boxes, instead of using SIFT features [40] to detect keypoints, our tracker uses contour points and Harris corners [41] to preserve temporal information. Contour points are generated using a pixel-based skin detection method [42]. Similarly, the final hand bounding box is the extended rectangular enclosing all keypoints. The hand tracker is disabled when a confident hand joints prediction is available. The final hand segment is obtained by applying skin detection inside the hand bounding box.

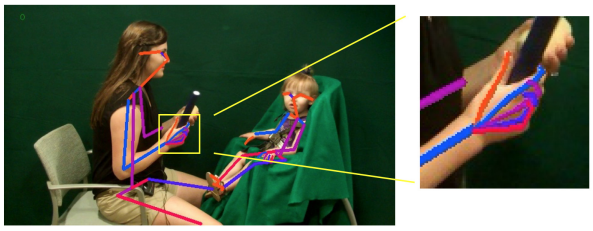


Figure 1. Human pose estimation and hand joints estimation

B. Infant Segmentation

Mask-RCNN [38] is trained on Microsoft COCO [39] and is used to generate infant segmentation. Infant body parts may be occluded by caregiver or other objects during their interaction, and Mask-RCNN tends to exclude the occluded region, an example is shown in Figure 2(b). However, excluding occluded region is not consistence with the way we detect touch event. Thus, we proposed a temporal refinement to recovery the occluded region. We use the confidence score of infant joints to assess the confidence of an infant segmentation by assuming when parts are missing in the infant segmentation, the confident score of occluded joints is also low. Equation 1 is used to determine whether a infant segmentation is confident, where α and β are empirically set to be 0.3 and 20 respectively, and c_n is the confidence score for a infant joint here. An invalid infant segmentation at t_1 is recovered by using a confident segmentation from previous frame at t_0 as shown in Figure 2(c).

C. Touch Detection

The proposed method makes decision to label a frame as "touch" or "non-touch" first, and then assigns a touch type label L_i to detected "touch" frame based which

(a) (b) (c)

Figure 2. Infant segmentation temporal refinement, (a) valid infant segmentation at time t_0 (b) occluded segmentation at time t_1 (c) infant segmentation recovered by temporal refinement at time t_1.

part of infant body has been touched, where $L_i \in \{\text{“}head\text{”}, \text{“}hand\text{”}, \text{“}torso\text{”}, \text{“}arm\text{”}, \text{“}leg\text{”}, \text{“}foot\text{”}\}$ and i is the index. Whether touch occurs in a given frame is determined by checking if caregiver's hand segments, obtained from Section III-A, overlap with the infant segmentation from Section III-B.

To classify touch type, we analyze the spatial relationship between caregiver's hands and infant body parts. We define six infant body parts corresponding to six touch type labels, and each part contains a set of limbs, where limbs are pairs of adjacent joint points belongs to that part. For example, left elbow and left wrist are a pair of adjacent joint points, they form the limb left forearm and belongs to the part "arm". We use a straight line connecting from one joint to another to fit the body limbs, and using a set of points to represent the fitted line, they are linearly spaced in 0.1 pixel in horizontal direction. Then for a given frame I, there are sets S_i for $i \in [1 \dots 6]$ contains fitted points for each part respectively to represent infant body parts. We evaluate the Euclidean distance of joint points of caregiver's hands to infant body parts. For each hand point $\mathbf{x_n}$, we get a label z_n using Equation 2.

$$z_n = \underset{i}{\operatorname{argmin}} \|\mathbf{x_n} - \mathbf{x_p}\|_2^2, \forall \mathbf{x_p} \in S_i \tag{2}$$

The final touch type is determined as the majority vote of z_n, where $n \in [1 \dots N]$, and N is the total number of caregiver's hand points.

IV. Experiments

A. Dataset

We evaluate the performance of our method on a testing set which records the interactions between a caregiver and an infant in a lab setting. Our testing dataset contains five 2500 frames video sequences and two long video sequences (more than 9000 frames each). The video sequences were acquired from different pairs of caregivers and infants at different times and dates while under the same recording settings. In these experiments, the caregivers were asked to interact with the infant as they would normally do during playtime. The infant was secured in a high chair and the caregiver sat on a chair facing the infant. The lab where the experiments were conducted had a green wall as background and the high chair was also covered by a green blanket. A

RGB camera and a clip-on wireless microphone were used to record video and audio data. The videos were recorded at a resolution of 1280×720 at 30 fps.

The precise labels for each video in the testing set are annotated by trained analyst, and this information is used as groundtruth for our evaluation. The testing set contains $25,043$ out of $31,374$ non-touch frames. The type of touch occured are listed in Table I.

Table I
TOUCH TYPES OCCURRENCES

Labels	Head	Hand	Torso	Arm	Leg	Foot
frames	403	472	367	168	3346	1575

B. Evaluation Metrics

Precision and recall are used to assess the performance of the automated touch detection method. Our method aims to narrow down the potential touch time intervals and provide less frames for trained analyst to annotate compared to original video sequences. Thus, recall of this method is more important than precision. Another metric reduced amount in numbe of frames is used to show how much work is reduced for trained analyst when annotating only the potential touch frames after using automated touch detection compared to annotating the entire sequence. In other words, trained analyst could skip annotating predicted non-touch frames ($FN+TN$), because they were less likely to contain touch frames.

$$ReducedAmount = \frac{FN+TN}{TP+FP+FN+TN} \quad (3)$$

For touch type detection, we use a confusion matrix to evaluate the quality of our detection results.

C. Experiment Results

The touch/non-touch detection results are showed in Table II. The proposed automatic touch detector successfully captured 99.19% of touch frames with a precision score of 48.13%. Reduced amount in table II shows the trained analyst could skip 58.41% number of frames, which is a great reduction compared to annotating every frame. The proposed method outperforms previous work [23] by having less missed touches, higher precision in predicted touches and less frames needed for trained analyst to annotate.

The confusion matrix in Figure 3 illustrates the performance of touch type detection. We observed that "Head" class and "Foot" class have higher scores compare to other classes. Considering those two classes are located in the top and bottom part of an infant segmentation respectively, they are less likely to be confused with other body parts. Taking "Torso" class for an example, 77% of "Torso" class are predicted correctly with 12% are classified as neighbor class "Arm" and 9% are labeled as "Leg". Because infant body part torso are spatially close to legs and arms.

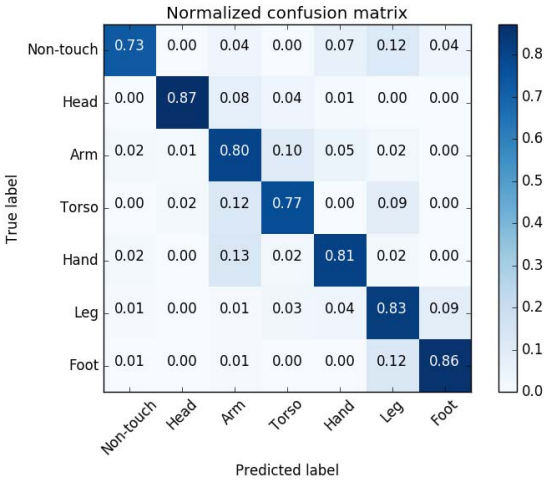
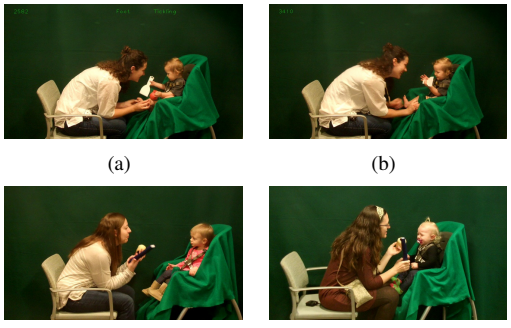


Figure 3. Confusion matrix of touch type labels

With 48.13% of precision for touch detection results, the proposed method still detected more false alarm than true touches. This was mainly due to the lack of precise hand contour detection for some frames in the video and difficulty in dealing with occlusions due to the camera viewing angle. In addition, without the third dimension information, it is difficult to distinguish from a true touch to fake touch illustrate in Figure 4(a) and Figure 4(b). Furthermore, we feel these challenging potential touch frames require a second look from trained analyst. From our results, the reduced amount is larger for videos sequences where the caregiver is well separated from the infant when they are not interacting (Figure 4(c)) than those caregiver and infant pairs that are in close-proximity (Figure 4(d)) for entire sequences.



(a) (b) (c) (d)

Figure 4. Examples frames from testing set, (a) true touch (b) false touch (c) well-separated (d) close-proximity

V. CONCLUSION

We proposed an automatic touch event detection system that detects and tracks the caregiver's hands, detects the location of the infant and then defines a "touch" to occur whenever the caregiver's hand contours overlap with the infants contour. The touch type label is assigned to predicted

Table II
TOUCH INTERACTION DETECTION RESULTS FROM OUR DATASET

	Total Frames	Method	Recall	Precision	Reduced Amount
Testing sequences	31,374	[23]	72.03%	24.66%	41.07%
		Proposed	99.19%	48.13%	58.41%

touch frames based on the spatial relationship between caregiver's hands and infant body parts. The proposed method avoids using expensive precise touch annotations for training. Instead it takes advantage of CNN models that are trained on large public datasets to produce intermediate results needed to identify touches. The proposed method allows trained analyst skip annotating 58.41% of frames and still be able to capture more than 99% true touch frames.

REFERENCES

[1] E. Anisfeld, V. Casper, M. Nozyce, and N. Cunningham, "Does infant carrying promote attachment? An experimental study of the effects of increased physical contact on the development of attachment," *Child Development*, vol. 61, no. 5, pp. 1617–1627, October 1990.

[2] R. Feldman, M. Singer, and O. Zagoory, "Touch attenuates infants physiological reactivity to stress," *Developmental Science*, vol. 13, no. 2, pp. 271–278, March 2010.

[3] S. G. Ferber, "The nature of touch in mothers experiencing maternity blues: The contribution of parity," *Early Human Development*, vol. 79, no. 1, pp. 65–75, August 2004.

[4] S. G. Ferber, R. Feldman, and I. R. Makhoul, "The development of maternal touch across the first year of life," *Early Human Development*, vol. 84, no. 6, pp. 363–370, June 2008.

[5] F. Franco, A. Fogel, D. S. Messinger, and C. A. Frazier, "Cultural differences in physical contact between hispanic and anglo mother–infant dyads living in the united states," *Early Development and Parenting*, vol. 5, no. 3, pp. 119–127, May 1996.

[6] E. Herrera, N. Reissland, and J. Shepherd, "Maternal touch and maternal child-directed speech: Effects of depressed mood in the postnatal period," *Journal of Affective Disorders*, vol. 81, no. 1, pp. 29–39, July 2004.

[7] M. J. Hertenstein, "Touch: Its communicative functions in infancy," *Human Development*, vol. 45, no. 2, pp. 70–94, March 2002.

[8] A. D. Jean and D. M. Stack, "Functions of maternal touch and infants affect during face-to-face interactions: New directions for the still-face," *Infant Behavior and Development*, vol. 32, no. 1, pp. 123–128, January 2009.

[9] A. D. Jean, D. M. Stack, and A. Fogel, "A longitudinal investigation of maternal touching across the first 6 months of life: Age and context effects," *Infant Behavior and Development*, vol. 32, no. 3, pp. 344–349, June 2009.

[10] R. Moszkowski and D. M. Stack, "Infant touching behavior during mother-infant face-to-face interactions," *Infant and Child Development*, vol. 16, no. 3, pp. 307–319, June 2007.

[11] D. W. Muir, "Adult communications with infants through touch: The forgotten sense," *Human Development*, vol. 45, no. 2, pp. 95–99, March 2002.

[12] I. Nomikou and K. J. Rohlfing, "Language does something: Body action and language in maternal input to three-month-olds," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 2, pp. 113–128, June 2011.

[13] D. M. Stack and S. L. Arnold, "Changes in mothers' touch and hand gestures influence infant behavior during face-to-face interchanges," *Infant Behavior and Development*, vol. 21, no. 3, pp. 451–468, June 1998.

[14] A. D. Jean and D. M. Stack, "Full-term and very-low-birth-weight preterm infants self-regulating behaviors during a still-face interaction: Influences of maternal touch," *Infant Behavior and Development*, vol. 35, no. 4, pp. 779–791, December 2012.

[15] D. M. Stack and D. W. Muir, "Tactile stimulation as a component of social interchange: New interpretations for the still-face effect," *British Journal of Developmental Psychology*, vol. 8, no. 2, pp. 131–145, June 1990.

[16] R. Abu-Zhaya, A. Seidl, and A. Cristia, "Multimodal infant-directed communication: How caregivers combine tactile and linguistic cues," *Journal of Child Language*, To appear.

[17] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media/multi-modal resources with elan." *LREC*, 2004.

[18] "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *CRCV-TR-12-01*, 2012, University of Central Florida, Orlando, FL.

[19] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, 2009.

[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169–3176, 2011.

[21] 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, 2013.

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[23] Q. Chen, H. Li, R. Abu-Zhaya, A. Seidl, F. Zhu, and E. J. Delp, "Touch event recognition for human interaction," *Proceedings of IS&T International Symposium on Electronic Imaging*, vol. 2016, no. 11, pp. 1–6, February 2016, San Francisco, CA.

[24] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, August 2004.

[25] Q. Chen and F. Zhu, "Long term hand tracking with proposal selection," *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6, 2018.

[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Proceedings of the European Conference on Computer Vision*, pp. 483–499, September 2016, Amsterdam, The Netherlands.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 91–99, December 2015, Montreal, Canada.

[28] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[29] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3681, 2013.

[30] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3487–3494, 2013.

[31] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 67–92, 1973.

[32] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1014–1021, 2009.

[33] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, June 2014, Columbus, Ohio.

[34] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1799–1807, December 2014, Montreal,Canada.

[35] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," *arXiv preprint arXiv:1812.08008*, 2018.

[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[37] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *Proceedings of the European conference on computer vision*, pp. 740–755, 2014.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[41] C. Harris and M. Stephens, "A combined corner and edge detector," *Proceedings of the Fourth Alvey Vision Conference*, vol. 15, p. 50, August 1988, Manchester, UK.

[42] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, January 2002.