DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

Touch Event Recognition For Human Interaction

Qingshuang Chen [1], *He Li* [1], *Rana Abu-Zhaya* [2], *Amanda Seidl* [2], *Fengqing Zhu* [1], *and Edward J. Delp* [1]
[1] *Video and Image Processing Laboratory (VIPER), School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA*
[2] *Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana, USA*

Abstract

This paper investigates the interaction between two people, namely, a caregiver and an infant. A particular type of action in human interaction known as "touch" is described. We propose a method to detect "touch event" that uses color and motion features to track the hand positions of the caregiver. Our approach addresses the problem of hand occlusions during tracking. We propose an event recognition method to determine the time when the caregiver touches the infant and label it as a "touch event" by analyzing the merging contours of the caregiver's hands and the infant's contour. The proposed method shows promising results compared to human annotated data.

Introduction

Touch is a key social and emotional signal used by caregivers when interacting with their children [1–13]. Touch is present 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 for some crucial tasks in development, namely learning language. 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.

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 have to be trained for several hours before they can begin annotating the videos. Hence, given the importance of human touch in human development it would be helpful 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 might be helpful for medical teams working with special populations and caregivers who have children with special needs.

In this paper, we describe a method for automatic touch event detection. The approach we take detects and tracks the caregiver's hands and detects the location of the infant and then defines a "touch" to occur whenever the caregiver's hand contours merge with the infants contour.

Related Work

There has been a great deal of work in human action recognition when humans interact with objects and other humans [17–21]. Many approaches use object detection and analyze the interaction between objects and human body parts to classify actions. Yao and Fei-Fei [22] treat human pose and the object as the context of each other and make improvements in both object detection and pose estimation. Prest [23] regards humans and objects as pairs when localizing and tracking them in space over time. Other work examines group event detection/interaction, for example fighting detection [24] in surveillance video. However, recognizing pairwise human interaction is still an open problem.

The motion and the localization of the human hand is important in characterizing human actions in dynamic scenes. Hand tracking is used in this paper to obtain the spatial information of caregiver's hands with respect to the location of the infant. Hand tracking is very challenging due to large variation in appearance and movement compared to other body parts. Recent advances in articulated model based hand tracking use depth cues [25, 26]. There has been much work in 2D hand tracking based on videos captured from RGB cameras. In [27], a Camshift [28] method is described that reduces tracking failures when the hand moves across other large skin-like areas by classifying velocities. It generally works when the velocity of the hand differs greatly from other skin-like regions and fails to track the hand when there are other skin regions moving along with the hand at similar speed. Hand detection [29] uses hand shape, context, skin colors and deformable part model [30] to detect the hand in static images. Another method [31] combines hand detection with hand tracking and uses an upper body model to refine the hand detection. Our hand tracking method takes advantage of analyzing the motion of the hand inspired from [27], and deals with occlusions occurred while tracking two hands.

In this paper we are interested in detecting a particular action between the caregiver and the infant, namely the touch event, and detecting the moment when the touch occurs. The touch event is defined as the time when the infant is touched by the hands of the caregiver. Thus successfully tracking the hands of the caregiver and clearly detecting the outline of the infant are crucial in our touch event detection. Essentially a touch will occur when the segmented contours of the hands and the infant merge. An example of this is shown in Figure 1. For hand tracking, we use color and motion features, and propose methods to handle the scenario

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

when one hand is fully occluded by the other hand. Foreground detection for determining the infant's location is based on GrabCut [32].

(a) (b)

Figure 1: (a) An example of a frame without a touch event, (b) An example where a touch event has occurred by detecting merging contours.

Touch Event Detection

Our method for touch detection contains three parts, the hand tracker of the caregiver, the infant's contour detector and the touch event detector. Figure 2 shows block diagram of the analysis system.

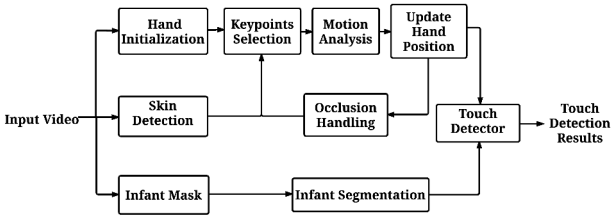

Figure 2: Block diagram of our touch event detection system.

Hand Tracker

The hand tracker uses color and motion features to track the position of both hands in each frame. The hand positions are initialized by manually selecting two bounding boxes containing each hand of the caregiver respectively in the first frame. Each bounding box is considered as a separate independent tracker.

Skin color detection

A pixel-based skin detection method is used to obtain the skin mask of the hands [33]. The skin color model is trained using 200 frames from videos with different pairs of caregivers and infants. All skin regions are manually segmented as ground truth skin pixels in every training frame, the remaining pixels are treated as non-skin pixels. The probability of the skin class and non-skin class are defined as follows:

$$P(RGB|skin) = \frac{s(RGB)}{T_s} \tag{1}$$

$$P(RGB|nonskin) = \frac{n(RGB)}{T_n} \tag{2}$$

where $s(RGB)$ represents the number of skin pixels in the histogram and $n(RGB)$ represents the non-skin pixel counts in the histogram. T_s and T_n are the total counts contained in the skin and non-skin histograms, respectively. A color pixel is considered as skin pixel when it satisfies:

$$\frac{P(RGB|skin)}{P(RGB|nonskin)} \geqslant \Theta \tag{3}$$

where

$$\Theta = C \cdot \frac{T_n}{T_s} \tag{4}$$

Θ is a threshold which can be adjusted for trade-off between correct detections and false positives. C is an adaptive parameter. In our experiments described below we empirically choose C (and hence Θ) such that the detected skin regions compared favorably to the training data skin masks. We then used a morphological opening with window size 3 and a constant structuring element on the output of the pixel-based skin classifier to reduce isolated and small skin regions. Figure 3(b) shows the skin detection result of a sample frame in Figure 3(a)

(a) (b)

Figure 3: (a) Original image, (b) Skin color detection, white pixels represent skin region, black pixels represent non-skin region.

Motion Analysis

The motion of the hand is analyzed using optical flow and the Lucas-Kanade registration method on the keypoints that represent the hand [34]. We find the Harris corners [35] inside the hand bounding box and the points on the contour of the hand skin mask and use them as keypoints. By obtaining the optical flow of keypoints using the iterative Lucas-Kanade technique with a pyramid [34, 36], the predicted positions of the keypoints on the next frame can be found. The pyramid structure uses a 3 level pyramid from coarse to fine. The velocity [27] is defined as the Euclidean distance between the keypoints' location in the previous frame and the predicted position in the current frame. Figure 4 illustrates motion analysis for keypoints and the trajectories for 10 frames.

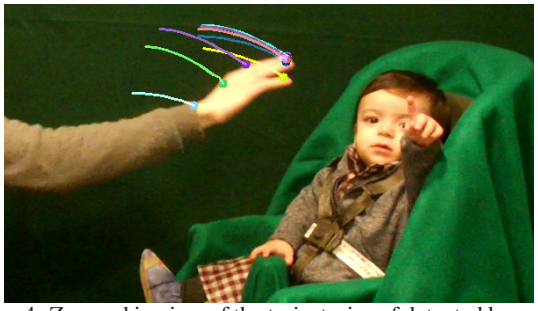

Figure 4: Zoomed in view of the trajectories of detected keypoints in 10 consecutive frames.

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

Keypoints Update

The hand bounding box is updated based on the remaining keypoints after discarding outliers using the velocity and skin color tests described below.

The average velocity v_{avg} is the arithmetic average of the velocity of all keypoints. The velocity test is:

$$|v - v_{avg}| \leqslant 2v_{std} \tag{5}$$

where v is the velocity and v_{std} is the standard deviation of the velocity. Any keypoint that falls outside of the above range are considered an outlier and is discarded. The skin color test checks whether the predicted position of keypoints are on the current skin mask. If not, these keypoints are discarded because we assume the hand is in the skin regions. The new hand bounding box is the smallest rectangle that contains all the remaining keypoints. New keypoints used for motion analysis are the Harris corners and the contour points inside the new bounding box.

Occlusion Handling

Occlusion handling is a challenging problem in object tracking. It is even more difficult in hand tracking, because the hand has a non-rigid shape. In our work, the caregivers are allowed to freely move their hands. Thus, the caregivers often move their two hands together, for example a handclap. The problem then becomes how to keep track of two hands respectively after they separate.

We propose a method to handle occlusion by using the merge and split concept [37]. Figure 5 shows the flow chart of this method. We define a hand flag that indicates whether the two hands are together. In the hand initialization step, we manually select two hands and set the hand flag to 1. The hand tracker described above provides the hand position and the centroid of each hand can easily be obtained. When two hands are approaching, the centroids of hands are also getting closer. A merge happens when the centroids of two hands become one point. When it occurs, the hand flag is set to 0. In our experiments the threshold for the merge and split is set to 50 pixels and was empirically determined.

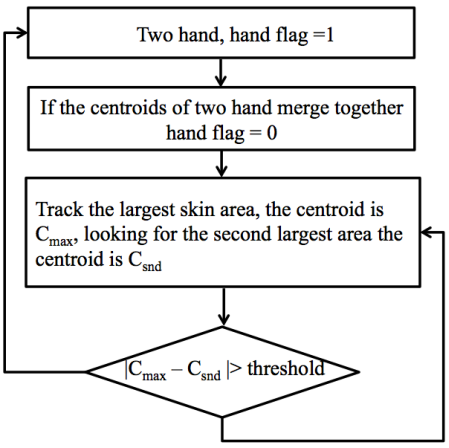

Figure 5: Flowchart of the merge and split method.

After the two hands merge together, the independent trackers are tracking the same region. The hand trackers not only track the contour points for largest skin region, but also search for the second largest skin region. Once the Euclidean distance between the centroid of the largest skin region and the centroid of the second largest skin region is greater than the threshold, a split occurs. The one tracker tracks the largest skin region and the other tracks the second largest skin region. The hand flag is then set back to 1.

The two hands merge and split method proposed above works for most of occlusion cases we have observed because hands represent large skin regions in the scene. However, the hand tracker fails in some special cases. For example, during the splitting of two merged hands, one hand may be fully occluded by objects other than the hand.

Infant Detector

We also need to find the location of the infant. We use Grab-Cut [32] to detect the contour of the infant in every frame of the video sequence. Grab-cut segmentation is an iterative method based on Graph-Cut [38], which is described by the Gibbs energy:

$$E(x) = \sum_{i \in I} D(x_i) + \lambda \sum_{i \in I, j \in N_i} V(x_i, x_j) \tag{6}$$

where i is a pixel that belongs to image I, N_i is the neighboring pixels of i, x_i takes on the value of 0 for sure background, 1 for sure foreground, 2 for probably background, and 3 for probably foreground. $D(x_i)$ is the data term, and $V(x_i, x_j)$ is the smoothing term. The data term $D(x_i)$ is modeled by a Gaussian Mixture Model (GMM), where we estimate the probability distribution of the background and the foreground. A mask is generated by the user that marks the foreground as RGB color white $(255, 255, 255)$, background as RGB color black $(0, 0, 0)$, and the unknown region as a RGB color different than black and white. Based on this initial graph, the Grab-Cut method finds a minimum cost to the energy function. A zoomed in view of the original image and its mask image is shown in Figure 6(a) and 6(b). The infant segmentation can be seen in Figure 6(c).

(a) (b) (c)

Figure 6: (a) Original image (b) Infant mask (c) Infant Contour.

Touch Event Detector

After obtaining the caregiver's hands position and infant's contour from the hand tracker and baby detector in each frame, a touch event can be defined as the time period during which contours merge. We examine each hand of the caregiver with the infant's contour separately. Once a hand touches the infant, the contour of the hand and the contour of the baby will merge into one contour. When this situation occurs, we can declare that a touch event has occurred. The same method is used to detect whether a touch event has occurred for the other hand. Either hand touching the baby will result in a touch event detection and labeling that frame as a touch event. Figure 7(a) shows a frame

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

without a touch event, and Figure 7(b) shows an example of a potential touch event by detecting the merging contours of the hands and the infant.

(a) (b)

Figure 7: (a) An example of a frame without a touch event, (b) An example where a potential touch event has occurred by detecting merging contours.

Experimental Results

To test the performance of our method, we recorded the interactions between a caregiver and an infant in a lab setting. 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 video sequences were acquired from different pairs of caregivers and infants at different times and dates while under the same recording settings. The videos were recorded at a resolution of 1280×720 and with a frame rate of 30 fps.

The video were processed using our automatic touch detector. In our experiments the threshold C in the skin detector was set to 0.0001 and the threshold for the merge and split for the hand occlusion is set to 50 pixels. We compared results with the ground-truth data annotated by a trained analyst. Figure 8 shows a sample result of the performance comparison. The green bars indicate potential touch events detected and the blue bars are touch events noted by a trained analyst. The automatic touch detector successfully captured all touch events, but included one false alarm. 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 view. The current automatic touch detector cannot differentiate different types of touch, resulting in one touch event detected instead of three consecutive touches (resting, grabbing, moving) as indicated by the trained analyst between 11.2 - 14.2 seconds.

Conclusion

We described a touch event detection system that combines hand tracking and infant segmentation. The recognition results are sensitive to the accuracy of the hand tracking and infant segmentation. Either losing track of the caregiver's hands or inaccurate infant segmentation may trigger a false alarm. For future work, we will focus on re-identifying hands when the hand tracking fails in order to avoid the touch recognition error propagating to subsequent frames. The development of precise hand tracking combined with learning based approaches will also improve the accuracy and incorporate the identification of touch event types. We will also investigate approaches that have robust, non-empirical, way of selecting the threshold C in the skin detection.

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.

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

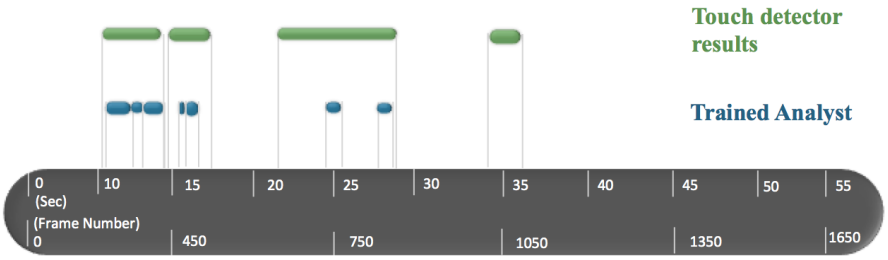

Figure 8: The comparison of touch event detection with a trained analyst.

[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] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Journal of Computing Surveys*, vol. 43, pp. 1–43, April 2011.

[18] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, October 2009.

[19] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 1473–1488, November 2008.

[20] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9–16, June 2010, San Francisco, CA.

[21] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, pp. 428–440, March 1999.

[22] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17–24, June 2010, San Francisco, CA.

[23] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 835–848, April 2013.

[24] Y. Chen, L. Zhang, B. Lin, Y. Xu, and X. Ren, "Fighting detection based on optical flow context histogram," *Proceedings of the Second International Conference on Innovations in Bio-inspired Computing and Applications*, pp. 95–98, December 2011, Shenzhen, China.

[25] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," *Proceedings of the British Machine Vision Conference*, pp. 101.1–101.11, September 2011, Nethergate, UK.

[26] B. Stenger, P. R. Mendonça, and R. Cipolla, "Model-based 3d tracking of an articulated hand," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 310–315, December 2001, Kauai, HI.

[27] C. Chen, M. Zhang, K. Qiu, and Z. Pan, "Real-time robust hand tracking based on camshift and motion velocity," *Proceedings of the IEEE International Conference on Digital Home*, pp. 20–24, November 2014, Guangzhou, China.

[28] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision*, pp. 214–219, October 1998, Princeton, NJ.

[29] A. Mittal, A. Zisserman, and P. H. Torr, "Hand detection using multiple proposals," *Proceedings of the British Machine Vision Conference*, pp. 75.1–75.11, September 2011, Nethergate, UK.

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, September 2010.

[31] N. H. Do and K. Yanai, "Hand detection and tracking in videos for fine-grained action recognition," *Proceedings of the Asian Conference on Computer Vision Workshops*, pp. 19–34, November 2014, Singapore, Singapore.

[32] Y. Li, J. Sun, C. Tang, and H. Shum, "Lazy snapping," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 303–308, August 2004.

[33] 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.

[34] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 674–679, August 1981, Vancouver, British Columbia.

[35] 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.

[36] J. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Technical Report*, 1999, Intel Corporation, Santa Clara, CA.

[37] T. Yang, Q. Pan, J. Li, and S. Z. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 970–975, June 2005, San Diego, CA.

[38] 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.

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-465

Author Biography

Qingshuang Chen received her BS in Electrical Engineering from the Purdue University (2014) and began her PhD studies in the VIPER lab (Video and Image Processing Laboratory) at Purdue. She works on people interaction recognition and object tracking.

He Li received his BS in Computer Engineering from Purdue University in 2014. He is currently pursuing a PhD degree at Purdue University. His work has focused on object recognition, image classification and retrieval.

Rana Abu-Zhaya is a Ph.D. student in the Speech, Language, and Hearing Sciences Department at Purdue University. She has worked as a speech-language pathologist with different clinical populations in Israel. In her PhD work, she is exploring the role of dyadic interactions in the course of language acquisition, and she is focusing on effects of caregiver touch on infant word segmentation and learning.

Amanda Seidl received her Ph.D. in Linguistics from the University of Pennsylvania in 2000. This was followed by a post-doctoral fellowship in Cognitive Science at the Johns Hopkins University. She is currently a Professor in the Speech, Language, and Hearing Sciences department at Purdue University. The overarching goal of her research is to discover how language comes to the child.

Fengqing Zhu received her Ph.D. in Electrical and Computer Engineering from Purdue University in 2011. Prior to joining Purdue in 2015, she was a Staff Researcher at Huawei Technologies (USA). She is currently an Assistant Professor of Electrical and Computer Engineering at Purdue University. Her research interests include image processing and analysis, augmented reality, and video compression.

Edward J. Delp was born in Cincinnati, Ohio. He is currently the Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering at Purdue University. His research interests include image and video compression, multimedia security, medical imaging, multimedia systems, communication and information theory.