JCSE Journal of Computing Science and Engineering

Invited Paper

Journal of Computing Science and Engineering,
Vol. 10, No. 3, September 2016, pp. 95-101

Enhanced Sign Language Transcription System via Hand Tracking and Pose Estimation

Jung-Ho Kim, Najoung Kim, Hancheol Park, and Jong C. Park*
School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Korea
{jhkim, njkim, hcpark, park}@nlp.kaist.ac.kr

Abstract

In this study, we propose a new system for constructing parallel corpora for sign languages, which are generally under-resourced in comparison to spoken languages. In order to achieve scalability and accessibility regarding data collection and corpus construction, our system utilizes deep learning-based techniques and predicts depth information to perform pose estimation on hand information obtainable from video recordings by a single RGB camera. These estimated poses are then transcribed into expressions in SignWriting. We evaluate the accuracy of hand tracking and hand pose estimation modules of our system quantitatively, using the American Sign Language Image Dataset and the American Sign Language Lexicon Video Dataset. The evaluation results show that our transcription system has a high potential to be successfully employed in constructing a sizable sign language corpus using various types of video resources.

Category: Human computing

Keywords: Sign language recognition; Hand tracking; Depth prediction; Pose estimation; Transcription

I. INTRODUCTION

According to the World Federation of the Deaf [1], there are about 70 million deaf people using sign language as their primary means of communication. Despite this large number of signers, sign languages are still severely under-resourced [2]. This is mainly due to the fact that sign languages, in contrast to spoken languages, contain spatial information which necessitates a different language processing approach. To resolve this issue, researchers have proposed and constructed corpora including information collected via motion recognition. The most prevalent idea in existing research was to use multiple cameras for motion recognition [3-5]. There were also other studies that developed recognition systems utilizing gears such as gloves [6, 7] and infrared sensors [8]. Although such systems could construct high quality resources, data collection and corpus construction using these systems are expensive and do not scale. Some studies have attempted to address the scalability issue through depth-based sign language recognition with lower-cost equipments such as the Kinect Sensor [9, 10] and Leap Motion Sensor [11, 12], but these systems did not yield high accuracy.

The key to resolving the scalability issue is to find an efficient means to collect large amounts of sign language data from various signers. Nowadays, many deaf people are communicating via various channels that support video chats (e.g., PCs, mobile devices, and home appliances with cameras). These video chats would be good sources of data for sign language corpus construction as they contain natural sign language utterances as well as

Open Access http://dx.doi.org/10.5626/JCSE.2016.10.3.95 http://jcse.kiise.org

Received 11 September 2016; **Accepted** 12 September 2016
*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

being scalable due to their accessibility and data format with comparatively lower resolution. Nonetheless, data from video chats would still have to face recognition accuracy issues because most cameras used in video chat environments are not of high-resolution.

To provide a solution that maintains reasonable accuracy under low-resolution environments, this study proposes a method of sign language recognition with a single low-resolution camera that captures RGB color only—the assumed default settings for everyday video chats. As a first step, this study restricts the scope of its sign language recognition to upper body and hands, and the aims within this restricted scope are 1) to successfully track each hand's location and shape and 2) to integrate the individually recognized parts to create sign units that are deterministically mapped to corresponding transcriptions.

As mentioned earlier, the low resolution of the data may hamper the accuracy in the recognition task. This study aims to address the accuracy problem by applying state-of-the-art techniques from computer vision research. For example, the recent work by Tompson et al. [13] predicts joint locations with a convolutional neural network. Such vision-based human body pose estimation approaches inspired Park and Ramanan [14] to propose a deep neural network-based approach to upper body pose estimation. This approach was applied to estimate seven joint body locations—namely left/right hands, left/right elbows, left/right shoulders and head—and this information was used for sign motion recognition. For hand pose estimation, Sun et al. [15] proposed cascaded hand pose regression that can be applied generally without any calibration. More recently, Zhou et al. [16] proposed a model-based deep learning approach that fully exploits articulated hand poses.

After the poses are recognized, our system integrates the individually recognized poses into a single unit that is mapped to a corresponding transcription. Although many options are available for sign language transcription such as the study of Stokoe et al. [17] and HamNoSys [18], we used SignWriting [19], as it is the most widely used writing system for existing relatively large-sized, general-purpose sign language corpora. SignWriting also meets an evaluative goal; we expect to facilitate performance comparison with existing, manually-constructed corpora which also use SignWriting.

The rest of this paper is organized as follows. Section II presents an overview of our system and its features. Section III explains how we evaluate our system. Section IV presents the evaluation results and analysis for each evaluation and discusses limitations of our system. Finally, Section V summarizes our work and provides remarks on potential improvements.

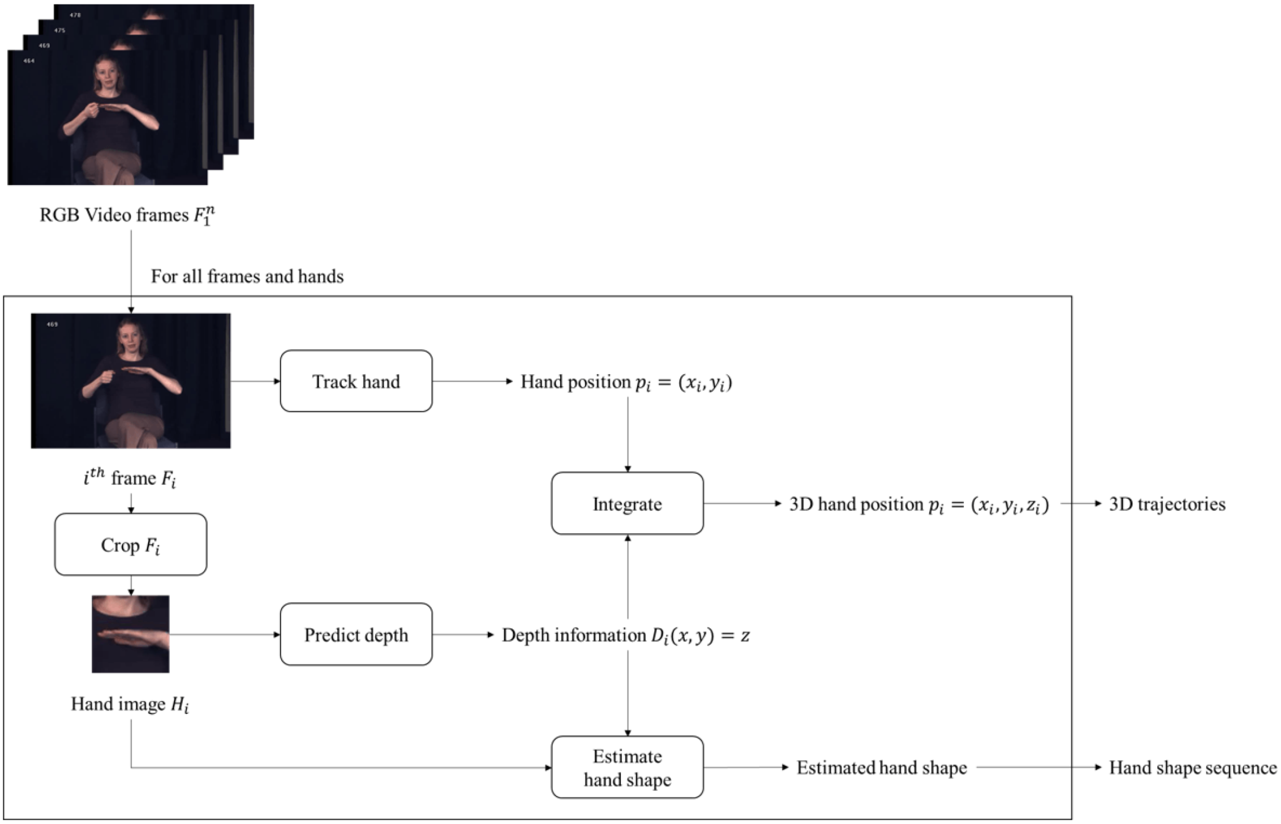


Fig. 1. System overview.

II. SYSTEM OVERVIEW AND FEATURES

A. System Overview

Fig. 1 illustrates the structure of our proposed system. The system takes a video recording of continuous sign motions as input, tracks the locations of hands which can be represented as two dimensional trajectories by connecting all tracked locations in a sequence, and predicts the depth information for each video frame. Then it generates three dimensional trajectories using these individually calculated trajectories and a sequence of depth information. Finally, it estimates the pose of both hands using depth information and integrates poses and trajectories as a preparatory step of transcribing sign language or generating sign language animation. The following sections explain the key features of our system in detail.

B. System Features

1) Hand Tracking

Tracking hands in still images or video frames is still a challenging problem because the shape of hands is hard to describe computationally. To overcome this problem, most proposed hand tracking systems assume the following: The color of the hands is fixed, the color of background can be removed easily, and hands are moving faster than any other objects in the video. The Kalman filter [20] and particle filters [21] are introduced to track moving hands based on the assumptions above. Yuan et al. [22] propose temporal filtering to detect candidates of hand locations based on the hand color and motion residue. In this study, we adopt a pictorial structure model [23], which is designed to operate on uncontrolled images with difficult illumination conditions and cluttered backgrounds. When a still image is given as input, it detects the human in the image, generates a bounding box, and then finally predicts the location of hands.

2) Depth Map Prediction

In order to predict the depth map from a single image, we used a state-of-the-art depth map prediction tool by Eigen et al. [24]. This tool provides the prediction model that consists of two *deep network stacks* [24] in which one makes a coarse global prediction from the input image, and the other locally refines the prediction. They trained the model using a convolutional neural network on NYU Depth v2 and KITTI datasets [25, 26], in which each data contains an outdoor or indoor scene. In this study, we used the pre-trained model that this tool provides, because most of the state-of-the art tools including this tool have rarely provided the training code.

3) Hand Pose Estimation

Hand pose estimation requires more detailed recognition compared to upper body pose estimation. Even if we use additional depth information, recognition is difficult when there is occlusion, and the larger base number of joints presents additional difficulty for accurate recognition. To resolve this issue, we adopt Zhou et al.'s model-based deep learning approach [16] that fully exploits hand model geometry. This method enables us to obtain all joint positions in both hands. They used NYU and ICVL datasets [13, 27] that are widely used in depth-based hand pose estimation to train their model. NYU dataset consists of 72,757 images for training and 8,252 images for testing. ICVL dataset has over 300k images for training but is less accurate than NYU.

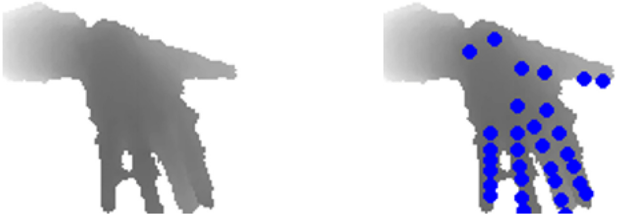


Fig. 2. Input (raw) and output (estimated) hand poses.

As the model requires an image of size 128×128 pixel in order to recognize a hand, we crop out a 128×128-pixel image that places the wrist position in the center from each whole image frame. Then we extract only the depth value from the cropped image and use this as an input for the hand pose estimation model. Fig. 2 compares the input and the resulting output, which is the estimated pose for a hand.

We evaluate the results by examining how accurately the hand shapes were transcribed according to the agreement of symbols. The metric above has Boolean true/false values obtained by comparing the values of the output and reference (from the American Sign Language Image Dataset; see Section III-A). In order to compare these, we mapped the hand shape information of the ref-

Table 1. The number of hand poses for each SignWriting category

Symbol	Category name	# of poses
	Index	14
	Index middle	16
	Index middle thumb	38
	Four fingers	8
	Five fingers	56
	Baby finger	30
	Ring finger	22
	Middle finger	19
	Index thumb	40
	Thumb	16

Table 2. Mapping between SignWriting and ASLLVD categories

SignWriting category	ASLLVD category	# of categories
Index	1, D, X, bent-1, flat-F/flat-G, sml-C/3, tight-C/2, flat-O/2, alt-P, I	11
Index middle	alt-G/bent-L, U/H, crvd-U, bent-U, cocked_U, V, crvd-V, bent-V,N, alt-N,	20
Index middle thumb	bent-N, P/K, L, L-X, crvd-L, R, Y, bent-I-L-Y, Horns, O/2-Horns	6
Four fingers	U-L, bent-U-L, 3, crvd-3, R-L, I-L-Y	5
Five fingers	4, B-xd, bent-B-xd, E, full-M	20
Baby finger	5, crvd-5, 5-C, 5-C-L, 5-C-tt, B, flat-B, B-L, Vulcan, crvd-B,	4
Ring finger	crvd-flat-B, crvd-sprd-B, bent-B, bent-B-L, C, tight-C, flat-O,	2
Middle finger	fanned-flat-O, loose-E, Rlxd	5
Index thumb	6, W, crvd-W, bent-W	5
Thumb	7, open-7	8

erence into SignWriting categories. We use SignWriting with 259 hand pose expressions to transcribe the hand poses. Table 1 shows the number of hand poses that each category covers.

III. EVALUATION SETTINGS

To account for the performance of our proposed system, we evaluate each feature individually. This is to identify which feature is responsible for a high error rate and make individual reinforcements in the future to improve the system performance as a whole. However, as pose estimation itself is not our novel proposal, we do not measure the distance errors between estimated pose and reference. Instead, we present accuracy evaluations for hand tracking and hand pose estimation. We use the following two datasets as references for our evaluation.

A. Dataset

We use the American Sign Language Image Dataset (ASLID) [28] to evaluate the accuracy of tracking hands. ASLID contains 479 images captured from the American Sign Language Lexicon Video Dataset (ASLLVD) [29, 30] videos by two ASL native signers. Annotations on these images contain position information and two dimensional points of the upper body such as head, shoulders, elbows, and wrists on the image. We use 1,804 video clips from ASLLVD, each of them mapped to a corresponding word in sign language, to evaluate the accuracy of hand pose estimation.

B. Evaluation Metrics

1) Hand Tracking

We apply a quantitative evaluation metric [28] for measuring the accuracy of hand detection. The tracking is considered to be successful if the distance between the estimated position (p_e) and reference (p_r) is less than a threshold (θ).

$$Accuracy(p_e, p_r, \theta) = \begin{cases} 1, \ if \ \| p_e - p_r \| \leq \theta \\ 0, \ otherwise \end{cases} \quad (1)$$

2) Hand Pose Estimation

In order to measure the accuracy of our hand pose estimation, we transcribed it in SignWriting and mapped all hand shapes in ASLLVD dataset to SignWriting hand shape categories. Table 2 shows how the hand shape information in ASLLVD [31] is mapped, where each ASLLVD hand shape only maps to exactly one SignWriting category.

IV. EVALUATION RESULTS AND DISCUSSION

In this section, we evaluate the accuracy of hand tracking and hand pose estimation and then interpret the result.

A. Hand Tracking

Fig. 3 shows the accuracies of our tracking system and Gattupalli et al.'s system [28] when the distance threshold increases. Although we cannot directly compare our tracking system to Gattupalli et al.'s system, it showed competitive results considering the complexity of our tracking system.

Also, there are no differences between the accuracies of left hand and right hand even though signers' dominant hand is the right hand; the dominant hand is more frequently used than the non-dominant hand to express sign language.

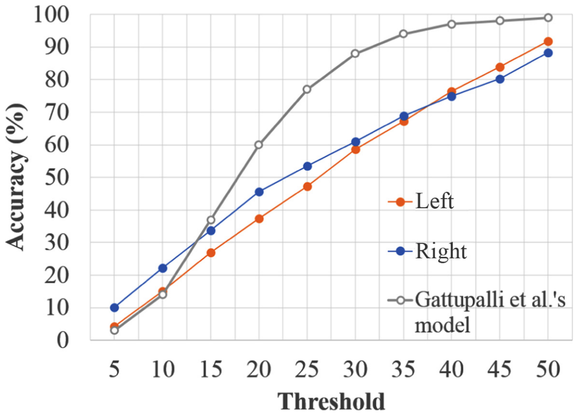


Fig. 3. Accuracy of tracking hands.

B. Hand Pose Estimation

Table 3 shows the average accuracy of estimating hand poses for each category and as a whole. The total count of symbols is 1,804.

As for this result, we believe that our system just showed its possibility because the average accuracies are not well balanced. A probable reason is that some complicated categories such as ring and middle fingers make it difficult to extract depth information since the model is not fine-tuned (we used the model that is trained on indoor and outdoor scene images). Therefore, there is a further possibility of improvement in our system if we employ a depth prediction model specifically fitted to our purpose.

Also, the accuracy of the hand pose estimation module could be improved further by a categorization algorithm reflecting a better understanding of the joints. Although the current algorithm uses all combinations of angles between adjacent joints to categorize the symbols, incorporating knowledge about possible and impossible combinations of joints into the algorithm is expected to make the module more accurate and efficient.

Table 3. Accuracy of estimating hand poses

SignWriting category	# of words	# of correct words	Accuracy
Index	268	192	0.716
Index middle	492	282	0.573
Index middle thumb	117	82	0.701
Four fingers	99	42	0.424
Five fingers	422	388	0.919
Baby finger	67	39	0.582
Ring finger	47	9	0.191
Middle finger	63	23	0.365
Index thumb	72	61	0.847
Thumb	157	100	0.637
Total	1804	1218	0.675

V. CONCLUSION AND FUTURE WORK

This study proposed a system that accurately transcribes sign language video into SignWriting by applying hand tracking and pose estimation. We quantitatively evaluated the modules of our system via two evaluation metrics, and confirmed that pose estimation improves the recognition accuracy of a single camera.

It is to be noted that we could not find an existing system to which we could compare the transcription performance, to the extent of our knowledge. However, as the calibrated results obtained by applying pose estimation are more accurate compared to transcriptions from raw data, we believe that our system provides a reasonable starting point for future efforts for more advanced sign language transcription systems.

Without a doubt, it would be useful to include evaluations in terms of intelligibility as well as accuracy as presented in this study, as it would be important that the transcriptions generated by this system are indeed intelligible to signers who use SignWriting. For example, we could ask signers to give a score of how intelligible the output is, or ask them to judge whether the original meaning is well preserved in the output.

For our future work, we aim to present an improved system that incorporates non-manual expressions by adding a regression model for facial expressions. In addition, we will reinforce the accuracy of the transcription module to make the system more stable, as well as carrying out a user study to evaluate the intelligibility of the result transcriptions.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (NSIP) (No. NRF-2014R1A2A1A11052310).

REFERENCES

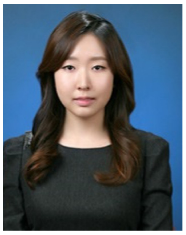
1. World Federation of the Deaf, “Sign language,” http://wfdeaf.org/human-rights/crpd/sign-language.
2. L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: a survey,” *Speech Communication*, vol. 56, pp. 85-100, 2014.
3. H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, “The recognition algorithm with non-contact for

Japanese sign language using morphological analysis," in *Gesture and Sign Language in Human-Computer Interaction*, Heidelberg: Springer, 1997, pp. 273-284.

4. C. Vogler and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis," in *Proceedings of 6th International Conference on Computer Vision*, Bombay, India, 1998, pp. 363-369.
5. A. Utsumi and J. Ohya, "Multiple-hand-gesture tracking using multiple cameras," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, 1999, pp. 473-478.
6. P. Lu and M. Huenerfauth, "Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects," in *Proceedings of the 11th international ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh, PA, 2009, pp. 83-90.
7. C. Oz and M. C. Leu, "American sign language word recognition with a sensory glove using artificial neural networks," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 7, pp. 1204-1213, 2011.
8. C. H. Morimoto, D. Koons, A. Amir, and M. Flickner, "Pupil detection and tracking using multiple light sources," *Image and Vision Computing*, vol. 18, no. 4, pp. 331-335, 2000.
9. Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, Alicante, Spain, 2011, pp. 279-286.
10. S. Lang, M. Block, and R. Rojas, "Sign language recognition using kinect," in *Proceedings of the International Conference on Artificial Intelligence and Soft Computing*, Zakopane, Poland, 2012, pp. 394-402.
11. L. E. Potter, J. Araullo, and L. Carter, "The leap motion controller: a view on sign language," in *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, Adelaide, Australia, 2013, pp. 175-178.
12. G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, Paris, 2014, pp. 1565-1569.
13. J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics*, vol. 33, no. 5, article no. 169, 2014.
14. D. Park and D. Ramanan, "Articulated pose estimation with tiny synthetic videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 58-66.
15. X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 824-832.
16. X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, "Model-based deep hand pose estimation," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, New York, NY, 2016.
17. W. C. Stokoe, D. C. Casterline, and C. G. Croneberg, *A Dictionary of American Sign Language on Linguistic Principles*, Silver Spring, MD: Linstok Press, 1976.
18. S. Prillwitz, *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide*, Hamburg: Signum, 1989.
19. V. Sutton, *Lessons in Sign Writing*, La Jolla, CA: SignWriting, 1995.
20. R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.
21. A. Doucet, N. De Freitas, and N. Gordon, "An introduction to sequential Monte Carlo methods," in *Sequential Monte Carlo Methods in Practice*, Heidelberg: Springer, 2001, pp. 3-14.
22. Q. Yuan, S. Sclaroff, and V. Athitsos, "Automatic 2D hand tracking in video sequences," in *Proceedings of 7th IEEE Workshops on Application of Computer Vision*, Breckenridge, CO, 2005, pp. 250-256.
23. M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190-214, 2012.
24. D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2366-2374, 2014.
25. N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision*, Springer, Florence, Italy, 2012, pp. 746-760.
26. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231-1237, 2013.
27. D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: structured estimation of 3D articulated hand posture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 3786-3793.
28. S. Gattupalli, A. Ghaderi, and V. Athitsos, "Evaluation of deep learning based pose estimation for sign language," 2016; http://arxiv.org/pdf/1602.09065v3.pdf.
29. C. Neidle and C. Vogler, "A new web interface to facilitate access to corpora: development of the ASLLRP data access interface (DAI)," in *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC)*, Istanbul, Turkey, 2012, pp. 137-142.
30. C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the American sign language lexicon video dataset (ASLLVD) corpus," in *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC)*, Istanbul, Turkey, 2012, pp. 143-150.
31. Handshapes data from the National Center for Sing Language and gesture resources, http://www.bu.edu/asllrp/cslgr/pages/ncslgr-handshapes.html.

Jung-Ho Kim

Jung-Ho Kim has received his B.S. degree in Computer Science and Engineering from Hanyang University in 2014 and M.S. degree in School of Computing from Korea Advanced Institute of Science and Technology (KAIST) in 2016. He is currently a Ph.D. student in School of Computing at KAIST. His research interests include recognition, corpus construction, and translation of sign language.

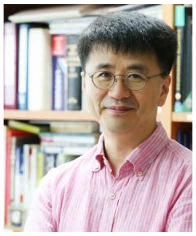


Najoung Kim

Najoung Kim is a PhD student in the Department of Cognitive Science at Johns Hopkins University. She earned her B.A. degree in English Literature and Linguistics from Seoul National University and her M.St. degree in General Linguistics & Comparative Philology from the University of Oxford. She has recently worked as a visiting researcher in the School of Computing at KAIST. Her main research interests are in computational models and the cognitive basis of language.

Hancheol Park

Hancheol Park received his B.S. degree in industrial and information system engineering from Ajou University in 2012. He is now in the integrated Master's and PhD program at the School of Computing, Korea Advanced Institute of Science and Technology (KAIST). His research interests are in paraphrase extraction, generation, statistical machine translation.



Jong C. Park

Jong C. Park is Professor in School of Computing and Associate Vice President of Information Service for Knowledge and Culture at KAIST, in charge of its libraries and IT teams. He earned his B.E. and M.S.E. degrees in Computer Engineering from Seoul National University and Ph.D. in Computer and Information Science from the University of Pennsylvania in Philadelphia, Pennsylvania, USA. At KAIST, he has been working on computational linguistics (e.g., syntax-semantics interface), natural language processing (e.g., biomedical information extraction), cognitive science (e.g., modeling language loss), bioinformatics (e.g., BioNLP in general), and quality of life technologies (e.g., for patients with dementia and mild cognitive impairment, for the Deaf, and for older people). He serves as founding and managing co-Editor-in-Chief of Journal of Computing Science and Engineering (JCSE) since 2007. He also serves as founding General co-Chair of the International Symposium on Languages in Biology and Medicine (LBM, since 2005) and PC Chair of PACLIC (2016), and served as PC co-Chair of IJCNLP (2013) and co-Chair of the Local Organizing Committee of ACL (2012). He participates in about 8 TPCs of international conferences each year. He is on the editorial boards of Journal of Natural Language Engineering (JNLE), ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), BMC Journal of Biomedical Semantics, and Journal of Language and Information. He is a permanent member of Sigma Xi, KIISE, and KSLI.