Precise Motif Sequence Pattern Recognition of American Sign Language (ASL) Letter Signs

Anthony M. Kutscher Sr. and Yanzhen Qu
Colorado Technical University, Colorado Springs, CO, USA
Email: {akutscher, yqu}@coloradotech.edu

***Abstract*—Unsuccessful pattern recognition of the complete set of fingerspelled letter signs were reported by studies using specialized video, RGB video, and Infrared (IR) video cameras, combined with various technologies. One reviewed study scaled (resized) letters to a specific size for ease of pattern recognition. Our study used the five similarly contoured ASL closed hand letter signs (ACLHS) A, M, N, S and T, with differentiation problems due to similar contours to show that motif sequences, formulated from the unique signatures of each of the five ACHLS, are less complex and faster at pattern recognition than scaling each captured letter sign dynamically. Thus, IR photo sensor data can be rotated to specific targets, creating unique signature patterns for each of the five ACHLS, and those unique signatures can be formulated into motif sequences for consistent and accurate pattern recognition of unknown ACHLS, regardless of similar handshape contours.**

***Index Terms*—Kinect, ASL, fingerspelling, similar contours, pattern matching, and motif sequences**

I. Introduction

Imagine utilizing the 3D/IR photo sensors on mobile smart-phones with software that captures the 3D data points of fingerspelled letter signs and accurately translates each fingerspelled letter sign into an audible sound, allowing a hearing person to hear and understand the letters and words being communicated to them by a deaf person. Presently, those in the Deaf and Hearing communities require a human sign-language translator, written text, or that each person communicating use the same sign-language dialect. Therefore, utilizing mobile device 3D/IR photo sensors to capture the 3D data points of fingerspelled letter signs and accurately recognizing and translating each letter sign and word into audible sounds for hearing persons to understand, seamlessly enhances communication from the Deaf-to-Hearing perspective.

A. Deaf to Hearing Communication Issues

Isolation of Deaf communities from many Hearing communities may be the direct result of the amount of time and effort it takes for a hearing person to learn fingerspelling and sign language. There may also be economic hardships that prevent funding for translators in required communication situations. One can see evidence of the communication challenges that exist between deaf and hearing persons when observing them during such routine activities such as doctor visits, fast food orders and grocery shopping; which are activities hearing persons take for granted.

B. Recognition Algorithms: Complexity vs. Speed

One reviewed study [1] used Dynamic Wave Transformation for scaling written alphabet letters, using weighted element matrices for consistent and accurate pattern matching of each letter in the alphabet. Their recognition results were 92.31% using the Euclidean Distance Metric (EDM) and increased to 99.23% when they incorporated their Artificial Neural Network (ANN) recognition scores to their EDM.

Our study has chosen the five similarly shaped ACLHS, representing the letters A, M, N, S and T (Fig. 1), to create five unique signatures and formulate them into motif sequences for pattern matching. Our intent was to show that motif sequences are faster at dynamic pattern recognition than *scaling* captured letter signs.

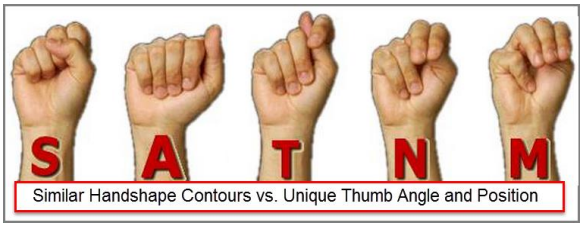

Figure 1. The five ACHLS in the problem set

C. Fingerspelling Benefits and Limitations

Not all words have a sign language gesture to represent them sufficiently. But, all words can be represented with fingerspelled letter signs sufficiently.

Though communication time may be much slower when using fingerspelled letter signs, the time limit is outweighed by the amount of communication that can be accomplished with the fingerspelling approach, since all words can be fingerspelled. Thus, this study's ACHLS interpretation system, using 3D/IR photo sensors, can provide a deaf person with an accurate way to communicate with the Hearing, albeit slower than sign language gestures.

D. Normalization and Recognition Processes

This process (Fig. 2) begins with data captured using a 3D/IR *near mode* photo sensor. The data are then

Manuscript received July 20, 2015; revised December 1, 2015.

doi: 10.18178/ijsps.4.5.405-410

transposed into a 2D square capture matrix for further cleanup of any anomalous contour data

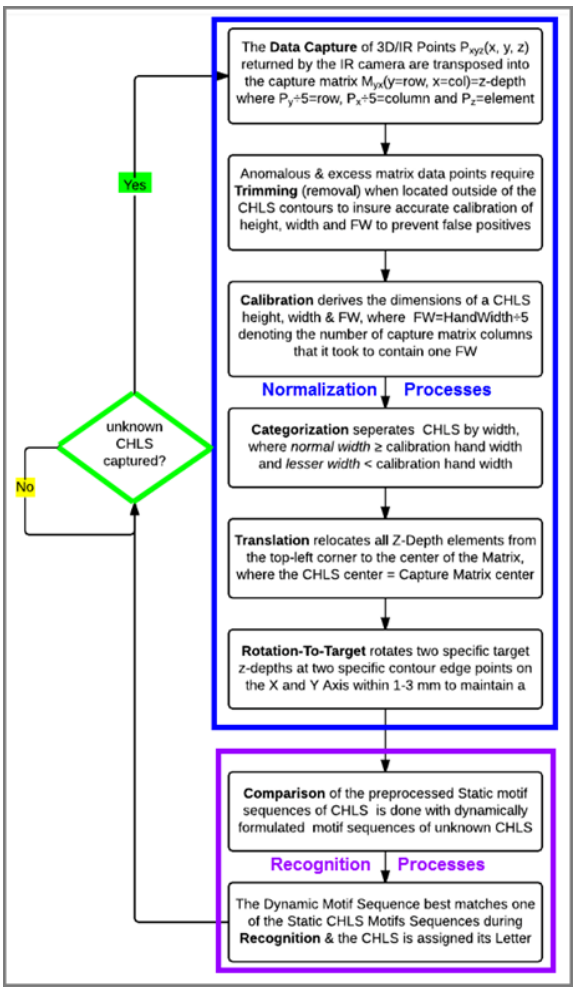

Figure 2. Normalization and recognition processes

Trimming away anomalous data points from the contour edges provides a more precise contour edge representation. Calibrating the height, width and finger width of the handshape allows for easier categorization. Categorization is a divide and conquer method that derives the handshape type, such as a closed- (letter sign A), intermediate- (letter sign I), or open-handshape (letter sign B). Translation is the process of translating the handshape center data point, and all other data points relative to its center point, to matrix center. Rotation-to-target denotes rotating each letter sign's data points to unique target z-depths on its x and y axes, which is required in order to compare with the resulting five unique ACHLS signatures into *Static Motif Sequences* (which are not all in the same areas of the handshape for each of the five CHLS). Each of the five static ACHLS motif sequences are compared with unknown ACHLS and the one of the five ACHLS it matches best is then recognized as the letter sign attached to that motif sequence.

II. Related Work

Most of the work in this field is recent, due to the expense and availability of video technology used to recognize not only the contours of handshapes, but also the location of the thumb and fingers within the handshape. Only since 201, has the Kinect 3D/IR technology been available. Two major categories from our literature review are discussed in the following sections below:

2D/RGB Video Capture and Feature Extraction, specifically used for recognizing fingerspelled letter signs and gestures, was attempted by the following studies, using various *Pattern Matching* algorithms.

Special attachments used for feature extraction

Reference [2] used a *specialized multi-flash camera*, having to position their flashes to cast shadows along the various depth discontinuities within each captured frame scene (frame), using a "shift and scale invariant shape descriptor for fingerspelling recognition."

Reference [3] differentiated right and left hands with colored gloves, from two other segments, and captured full-body gestures, in regions labeled Head, Chest, and Bottom, to separate the concerns of each region.

Edge detection & contours for feature extraction

Reference [4] transformed static frames of colored RGB images to grey scale and used "an edge detection technique … to mark the points in an image at which the intensity changes sharply." Reference [5] converted to gray scale images and applied Polygon Approximation on the static letter sign boundary images, followed by using Difference Freeman Chain Code Direction on the boundary edges, and created feature vectors composed of fingertip counts using difference chain code sequences.

Reference [6] focused on "recognition based on hand shape alone, not requiring motion cues", hoping to have continuous scaling of recognition percentages as their Lexicon increased in size.

Skin color detection used for feature extraction

Reference [7] focused on Arabic Sign Language (ArSL) recognition used high-resolution video for evaluating skin detection, specifically for the calibration of handshape size and focus.

Reference [8] used RGB video where skin color separated hand-shapes from background noise.

Reference [9] integrated two separate modules, one for detailed face tracking and one for handshape tracking "using color and appearance."

Math formulae used for feature extraction

Reference [10] used start and end video frame pairs of lexical signs in a Bayesian network formulation that would learn and use the positional constraints and normalize "inter-signer variations" for specific handshapes.

Handshape trajectory used for feature extraction

Reference [11] differentiated letters with their "BoostMap embeddings" and recognized handshapes by their differences in trajectory.

Discussion of Feature Extraction with RGB Video

Reference [1] using their Artificial Neural Network (ANN) recognition scores, combined with Dynamic Wave Transformation for scaling written alphabet letters, using weighted element matrices for consistent and accurate pattern matching of each letter in the alphabet is a good example of the majority of methodologies in the

studies reviewed having to maintain a larger number of transformational steps that required applying multiple math formulae to capture and process their 2D/RGB static video frames of data.

The *specialized multi-flash camera* [2] that required extra time to strategically set focus for each set of camera flashes was considered by this study to be unrealistic, with respect to time and added work, being that the flash adjustments were not automated and because data capture in our study used natural scenarios to capture fingerspelled letter signs.

Reference [3] differentiating right and left gesture signing hands with colored gloves was not relevant to our focus on one dominant fingerspelling hand, and neither was [4] transforming 2D/RGB into gray scale frames to use edge detection techniques.

The Reference [5] that used Polygon Approximation was deemed the most significant study reviewed by this study, for our specific set of foci. The most significant difference between our studies was their use of static 2D/RGB images compared vs. our use of dynamic 3D/IR z-depth data usage for pattern matching.

Reference [6] based on scalability for the Lexical database paralleled our ideas with respect to uniqueness of handshape, but our uniqueness was based on z-depth pattern matching, while their uniqueness was based on x and y coordinate contour edge pattern matching.

Neither the Arabic [7] using high-resolution video to calibrate handshape size and focus with skin detection, nor [8] using skin color to separate hand-shapes from background noise, were necessary to use in our study, as we were had "near mode" available on our 3D/IR camera, which only focused on the required right-hand-dominant fingerspelling hand.

References [9]-[11], one integrated separate modules for different body segments, one captured both start and finish positions of each gesture signing handshapes, and one used differences in trajectory to distinguish similar handshapes, respectively. These studies were all considered useful for future research, when capturing gesture signs composed of one or more fingerspelled letter signs simultaneously.

We conclude that the 2D/RGB video studies we reviewed, recognizing gestures or fingerspelled letter signs, experienced increased algorithm complexity, decreases in both recognition time and accuracy.

3D/IR Photo Sensor Capture and Feature Extraction for sign-language recognition, when compared with 2D/RGB, is more accurate and less expensive.

Special attachments used for feature extraction

Reference [12] used Kinect's 3D/IR photo sensor capabilities to configure sensors for their CopyCat system's data collection platform and combed their computer vision with two three-axis accelerometers [on blue and red colored data gloves] in order to recognized signed gestures.

Edge detection & contours for feature extraction

Reference [13] used Kinect's 3D/IR photo sensor capabilities to label handshape pixels for open and closed handshape recognition.

Skin color detection used for feature extraction

Reference [14] used Kinect's 3D/IR photo sensor and RGB color capabilities to identify fingertips and handshape palm centers.

Math formulae used for feature extraction

Reference [15] created a complete general sign language framework for gesture recognition, using photo sensors, Hidden Markov Models and Chaining.

Hand and arm joints used for feature extraction

Reference [16] focused on Kinect's 3D/IR photo sensor data of captured user joints. The matched signs were translated into words or phrases.

Discussion of Photo Sensor Feature Extraction

The use of the Kinect camera in all of these studies was a step in the right direction with respect to accuracy, cost and simplicity. We believe that the majority of the mathematics in all of the studies reviewed was probably unnecessary.

Reference [12] adding two in-house three-axis accelerometers [on blue and red colored data gloves] complicated the process and increased mathematical complexity to recognize letter signs.

References [13] and [14] recognizing handshape fingertips and centers, and whether or not handshapes were closed or open, respectively, had much less complex, contour-only agendas. Our study had also accomplished the same task using the calibration hand dimensions for width, finger width and height of each user's handshape, with the added bonus of capturing handshape volume data points for locating the thumb and fingers. Thus, our heuristics take advantage of the hand's physical properties to decide the Open-, Middle-, Sideways- and Closed-Hand Boundaries, keeping handshape recognition extremely simple.

Reference [15] formulating a general framework for gesture recognition had not considered the specifics of thumb and finger positioning within a letter sign, but was focused on standardizing procedures for recognizing German sign language gestures.

Reference [16] focused on the complete human body skeleton for gestures, which is described by the author as the signing space (the area from top of head to the waist and full arm extension on both sides of the torso) was out of the scope that our study focused on; i.e. the five ACHLS.

III. Targeted Problem and Hypothesis

A. Problem Statement

All existing video recognition software methods will incorrectly recognize the unique depth patterns of ACHLS, due to handshape contour similarities at data capture.

B. Hypothesis Statement

Capturing ACHLS with 3D/IR photo sensors and repositioning them to specific targets, allows for the creation of unique motif sequence invariants for consistent and accurate pattern matching of all five ACHLS, regardless of handshape contour similarities at data capture.

C. Contribution Statement

Accurately recognizing 3D objects that have a unique depth pattern, when using photo sensors and data repositioning, allows for dynamic object pattern recognition in a close proximity.

IV. Methodology

The specific software and lab equipment that was mandatory for 3D/IR photo sensor recognition of letter signs consisted of an ASUS Notebook with a Quad-Core processor, Microsoft's versions of Kinect Software Development Kit (SDK) Version 1.5, Kinect for Windows 3D/IR near mode photo sensor camera, Visual Studio 2012 IDE, Visual C# Programming Language, and Excel 2013.

The details of the normalization and recognition methods are where the importance of this study is focused, rather than the software or instrumentation that was used. As expected, time passes quickly during research and so does the lifetimes of hardware architectures, software applications, and the various ways they are configured become obsolete as well. Thus, it is most likely true that the latest and greatest technologies were probably not used in our study when compared to today's de facto standards. Therefore, the specific brands of the software and lab instruments used in this study will be generalized as much as possible in the discussions that follow.

A. Optimization

Only handshape contour and volume data points were captures with a 3D/IR *near mode* photo sensor, then transformed into unique signatures to consistently and accurately recognize each of the five ACHLS.

B. Identification

The unique signatures transformed from 3D/IR *near mode* photo sensor data captures, were used to formulate static motif sequences (z-depth patterns) for each of the five ACHLS for comparison with each of the dynamically formulated motif sequences created for unknown captured CHLS.

C. Testing

Each of the five ACHLS were replicated 25 times each during the final testing phase, to record results for recognition consistency and accuracy and depict those results in confusion matrices.

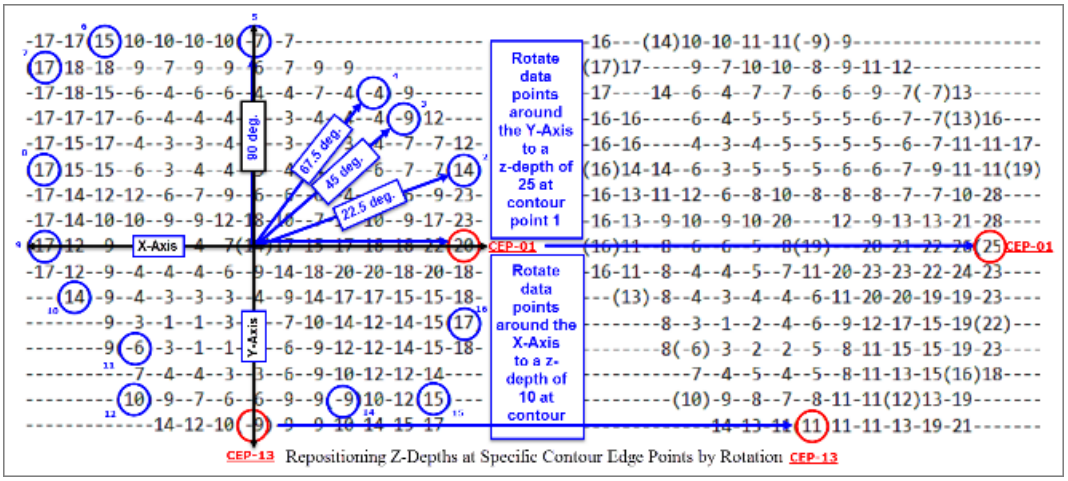

Figure 3. Rotation-To-Target for the letter sign A

V. Evidence and Data

A. Rotate 3D/IR Photo Data into Unique Signatures

This study used a rotate-to-target methodology that rotated the z-depths of two specific CEPs of each of the unknown ACHLS, consistently within a tight range of plus or minus 1 to 3 mm z-depth tolerance. Target z-depths repositioned each of the five ACHLS to their most unique signature position, for recognition consistency and accuracy, during the comparison process between each of the incoming unknown ACHLS dynamic motif sequences, to all five of the ACHLS static motif sequences (Fig. 3).

B. The Five Codified ACHLS Motif Sequences

Each of the *five static ACHLS motif sequences* were formulated and codified to use for pattern matching with incoming captured dynamic motif sequences for each of the unknown ACHLS (Fig. 4).

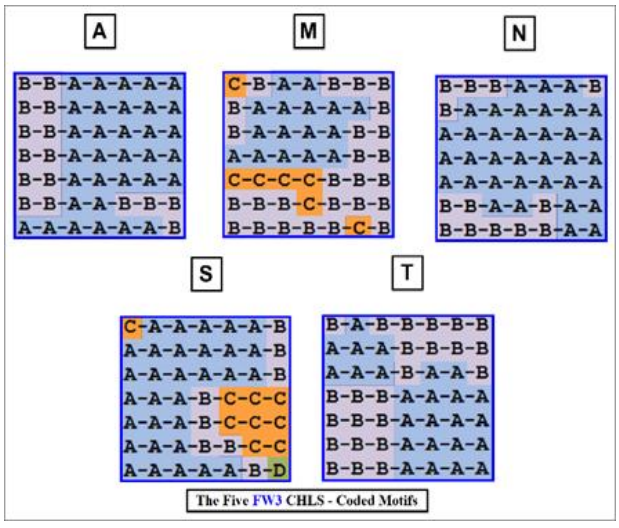

Figure 4. Motif sequences for the five ACHLS

C. Motif Sequence Relaxed and Extra Relaxed Codes

The motif sequence codes (Table I) depict how this study relaxed the z-depth constraints used for pattern matching, allowing for z-depth tolerances of either 1-10 mm or 1-20mm, represented with letters rather than z-depth integers, where A, B, C, and D, represented relaxed Codes and letters X, Y, and Z represented Extra Relaxed Codes.

TABLE I. MOTIF SEQUENCE Z-DEPTH RELAXATION CODE

Z-Depth Ranges for Motif Relaxation Codes	
Strict Ranges	
Z-Depth Range	Motif Code
1-10 mm.	A
11-20 mm.	B
21-30 mm.	C
31-40 mm.	D
Crossover Ranges	
Z-Depth Range	Motif Code
1-20 mm.	X
11-30 mm.	Y
21-40 mm.	Z

VI. SYSTEM TRAINING

Training data was captured to formulate five unique motif sequences for each of the five ACHLS. Three repetitions each, of the five ACHLS were captured from 48 participants, resulting in 720 variables, whose median z-depths were used to formulate the five unique ACHLS motif sequences. These variables were captured from volunteer participants on the campuses of Cal State University Sacramento and NorCal Services for the Deaf and Hard of Hearing in Sacramento.

VII. INTERPRETATION

The confusion matrix (Table II) depicts results of 100% recognition consistency and accuracy for letter signs A, N, S and T, and results of 96% consistency and accuracy for letter sign M. The "*FW3*" in Table II's stems from the training data resulting in three separate handshape sizes, labeled FW2 (small), FW3 (medium) and FW4 (large), where the integers (2, 3 and 4) denoted the number of matrix columns used to contain one finger. Thus, the *FW3 ACHLS Confusion Matrix* (Table II) represents 25 replicates each of the five ACHLS from one FW3 hand sized participant.

TABLE II. CONFUSION MATRIX FOR FW3 ACHLS

Percentage of Recognition - Confusion Matrix - 25 Reps					
	A	M	N	S	T
A	100%	0%	0%	0%	0%
M	0%	96%	0%	0%	4%
N	0%	0%	100%	0%	0%
S	0%	0%	0%	100%	0%
T	0%	0%	0%	0%	100%

VIII. CONCLUSION

This study proposes that inherent problems caused by *similarly contoured handshapes*, are remedied when capturing ACHLS with 3D/IR *photo sensors*, repositioning them to specific z-depth targets to create unique signatures, and formulating those unique signatures into unique *motif sequence invariants* for consistent and accurate pattern matching of all five ACHLS, regardless of *handshape contour similarities.*

IX. FUTURE WORK

Moving forward in our research, we intend to continue using a 3D/IR photo sensor for consistent and accurate data capture, rotation-to-target to create unique z-depth signatures, and formulating unique motif sequences from those unique signatures for recognition via pattern matching. These methods will be enhanced to include both static and non-static (moving) fingerspelled letter signs. This will require new methodologies for capturing, repositioning and recognizing the unique patterns of sign language in motion.

We intend to continue enhancing methods until they are able to recognize motion-based signs, gestures and facial expressions, which would most certainly revolutionize ASL recognition, using *unique motif sequences* for *pattern* matching. Thus, the use of mobile technologies, such as smart-phones, tablets, and laptops would replace the inconvenience of having to transport a 3D/IR photo sensor video camera.

With video sign language recognition, it will be unnecessary for the Hearing to learn sign language in order to communicate with the Deaf. A deaf person would be able to fingerspell into a Smart-Phone, with photo sensor capture capabilities, and our system would translate sign language into voice. Therefore, the Deaf-To-Hearing problem of communication would be solved.

ACKNOWLEGEMENTS

Authors would like to thank Mr. Stefan Stegmueller for letting this study to modify his handshape contour software.

Authors also would like to thank California State University at Sacramento and NorCal Services for the Deaf and Hard of Hearing for helping to set up data capture sessions and for recruiting the 48 volunteers.

REFERENCES

[1] D. K. Patel, T. Som, and M. K. Singh, "Multiresolution technique to handwritten English character recognition using learning rule and Euclidean distance metric," in *Proc. International Conference on Signal Processing and Communication*, Noida, 2013, pp. 207-212.

[2] R. Feris, M. Turk, R. Raskar, K. Tan, and G. Ohashi, "Exploiting depth discontinuities for vision-based fingerspelling recognition," in *Proc. Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 155.

[3] B. Tsai and C. Huang, "A vision-based Taiwanese sign language recognition system," in *Proc. International Conference on Pattern Recognition*, Istanbul, 2010, pp. 3683-3686.

[4] V. S. Kulkarni and S. D. Lokhande, "Appearance based recognition of American sign language using gesture segmentation," *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 560-565, 2010.

[5] M. Geetha, R. Menon, S. Jayan, R. James, and G. V. V. Janardhan, "Gesture recognition for American sign language with polygon

approximation," in *Proc. IEEE International Conference on Technology for Education*, Chennai, Tamil Nadu, 2011, pp. 241-245.
[6] S. Liwicki and M. Everingham, "Automatic recognition of fingerspelled words in British sign language," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Miami, FL, 2009, pp. 50-57.
[7] N. Albelwi and Y. M. Alginahi, "Real-Time Arabic Sign Language (ArSL) recognition," in *Proc. International Conference on Communications and Information Technology*, Saudi Arabia, 2012.
[8] M. P. Paulraj, S. Yaacob, M. S. B. Z. Azalan, and R. Palaniappan, "A phoneme based sign language recognition system using skin color segmentation," in *Proc. 6th International Colloquium on Signal Processing & Its Applications*, Mallaca City, 2010, pp. 1-5.
[9] J. Piater, T. Hoyoux, and W. Du, "Video analysis for continuous sign language recognition," presented at the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, (Workshop at 7th International Conference on Language Resources and Evaluation), Malta, May 2010.
[10] A. Thangali, J. P. Nash, S. Sclaroff, and C. Neidle, "Exploiting phonological constraints for handshape inference in ASL video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2011, pp. 521-528.
[11] H. Yang and S. Lee, "Robust sign language recognition with hierarchical conditional random fields," in *Proc. International Conference on Pattern Recognition*, Istanbul, 2010, pp. 2202-2205.
[12] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the Kinect," in *Proc. 13th International Conference on Multimodal Interfaces*, Alicante, Spain, 2011, pp. 279-286.
[13] M. Tang. Recognizing hand gestures with Microsoft's Kinect. [Online]. Available: http://www.stanford.edu/class/ee368/Project_11/Reports/Tang_Hand_Gesture_Recognition.pdf
[14] J. L. Raheja, A. Chaudhary, and K. Singal, "Tracking of fingertips and centres of palm using KINECT," in *Proc. 3rd International Conference on Computational Intelligence, Modeling & Simulation*, Langkawi, 2011, pp. 248-252.
[15] S. Lang, M. Block-Berlitz, and R. Rojas, "Sign language recognition with Kinect," in *Proc. 11th International Conference on Artificial Intelligence and Soft Computing*, Poland, 2012.
[16] K. F. Li, K. Lothrop, E. Gill, and S. Lau, "A web-based sign language translator using 3D video processing," in *Proc. 14th International Conference on Network-Based Information Systems*, Tirana, 2011, pp. 356-361.

Anthony M. Kutscher Sr. received his Bachelor in Information and Computer Science in 1994 and Master in Information and Computer Science, software engineering in 1998, both from the University of California in Irvine California, USA; and his Doctor of Computer Science, from Colorado Technical University in Colorado Springs, Colorado, USA in 2015. He is currently a Lead Faculty of Computer Science and IT at CTU's Denver Campus. Dr. Kutscher's recent research interests include artificial intelligence, pattern recognition and image processing as well as mobile computing. He has previously worked for Brown College, MN, USA as the Chair of the Software and Networking Departments, and at Rockwell Semiconductors Inc. as a Sr. Software Engineer, and for Disney Worldwide Inc. as an SAP analyst. Dr. Kutscher has been a student member of both IEEE and ACM since 2013.

Yanzhen Qu currently is the university dean of College of Computer Science and Technology, and professor in Computer Science at Colorado Technical University, USA. He received his B.Eng. in Electronic Engineering from Anhui University, China, M. Eng. in Electrical Engineering from The Chinese Academy of Sciences, and Ph.D. in Computer Science from Concordia University, Canada. Over his industrial career characterized by many the world first innovations, he has served at various senior or executive level Product R&D and IT management positions at several multinational corporations. He was also the chief system architect and the development director of several world first very large real-time commercial software systems. At Colorado Technical University, Dr. Qu is the dissertation supervisor of many computer science doctoral students, and his recent research interests include cloud computing, cyber security, data engineering, software engineering process and methods, data mining over non-structured data, affective computing, artificial intelligence, scalable enterprise information management system, big data analytics as well as embedded and mobile computing. He served as general/program/session chair or keynote speaker in various professional conferences or workshops. He is also a visiting professor of over twenty universities. He has published many research papers in the peer reviewed conferences and professional journals, and is currently serving as a member of editorial board of several professional journals. He is a senior member of IEEE and IACSIT.