Proceedings of the IEEE International Conference on Image Processing (ICIP 2016), Phoenix, Arizona, USA. September 25-28, 2016.

MULTI-MODALITY AMERICAN SIGN LANGUAGE RECOGNITION

Chenyang Zhang, Yingli Tian

Media Lab
The City College of New York
Department of Electrical Engineering
New York, NY 10031

Matt Huenerfauth

Rochester Institute of Technology
Golisano College of Computing
and Information Sciences
Rochester, NY 14623

ABSTRACT

American Sign Language (ASL) is a visual gestural language which is used by many people who are deaf or hard-of-hearing. In this paper, we design a visual recognition system based on action recognition techniques to recognize individual ASL signs. Specifically, we focus on recognition of words in videos of continuous ASL signing. The proposed framework combines multiple signal modalities because ASL includes gestures of both hands, body movements, and facial expressions. We have collected a corpus of RBG + depth videos of multi-sentence ASL performances, from both fluent signers and ASL students; this corpus has served as a source for training and testing sets for multiple evaluation experiments reported in this paper. Experimental results demonstrate that the proposed framework can automatically recognize ASL.

Index Terms— American Sign Language, Action Recognition, Multi-Modality

1. INTRODUCTION

Within the field of computer accessibility for people with disabilities, many researchers investigate assistive technologies (also called "adaptive technology"), which consists of software and devices that benefit people with disabilities. This technology may enable users to perform tasks with greater efficacy or efficiency, thereby increasing independence and quality of life. People who are deaf or hard-of-hearing (DHH) utilize a variety of methods to communicate, including signed languages, which are natural languages that consist of movements of the hands, arms, torso, head, eyes, and facial expressions. There are more than one hundred sign languages in the world [1]; American Sign Language (ASL) is used throughout the U.S. and Canada, as well as other regions of the world, including West Africa and Southeast Asia. Within the U.S., some researchers estimate that there are approximately 500,000 people who use ASL as a primary language [2]. To promote information access for people who are DHH and to provide more natural methods by which they can interact with computer systems, additional research is needed for automatically recognizing ASL signing from videos of a human.

In addition to enabling new methods by which DHH users could interact with computers, ASL recognition technology could serve as an initial step for future automatic translation technologies from ASL to English. Further, software for automatically recognizing ASL from video could enable new educational tools for students learning the language [3]. Creating a system that could automatically analyze the ASL performance of a student and provide feedback is a key goal of our project, and the sign language recognition work presented in this paper is a necessary component of our future system. ASL is essentially a language communicated by components of human action, therefore it is highly related to human action recognition, gesture recognition and facial expression recognition.

Recognizing human activities from images and videos is a challenging task, one of the challenges is extracting useful visual features from noisy signals. This can be severe when the visual signals vary significantly due to illumination changes, background clutter and scale uncertainties. In recent years, with the success of low-cost depth sensors (such as Microsoft Kinect) many researchers and engineers no longer have to implement tedious signal processing such as human body detection, foreground detection, and illumination normalization by taking advantage of depth sensors [4, 5, 6, 7, 8, 9, 10, 11]. As for automatic ASL recognition, Pugeault *et al.* proposed to learn alphabets from static hand shapes [12]. While much research in this area focuses on the hands, there is also some research work focusing on linguistic information conveyed by the face and head of a human performing sign language, such as [13, 14].

Different from the previous proposed approaches, in this paper, we propose an ASL recognition system which is based on utilization of a Kinect depth sensor and the combination multiple features extracted from different information modalities. Our contributions are: 1) the system will automatically predict lexicon and grammar components for ASL videos in a data-driven manner. 2) Multiple modalities including raw depth sequences, face expressions, and hand gestures signals

are fused to extract compositional visual features. 3) We collaborate with ASL linguistic researchers and record an accurately annotated dataset for ASL sentences, which can benefit the ASL recognition research community. The rest of this paper is organized as follows: Section 2 describes the proposed framework and Section 3 describes our dataset and performance analysis. Finally we conclude this paper in Section 4.

2. PROPOSED FRAMEWORK

The pipeline of our proposed framework is illustrated in Figure 1. The main features of the system are two-fold: 1) it takes into account multiple signal modalities including depth image sequences, RGB image-based facial expression attributes, hand shapes and keypoint detections for both body joints and facial landmarks. 2) By learning from signing sequences performed by fluent ASL signers and annotations provided by professional linguisticians, our system can recognize different components such as English words and special ASL grammar components, such as facial expressions or head movements that have grammatical meaning during sentences.

2.1. Signal Modalities

There are five signal modalities employed in our system: **1) Depth image sequence.** These are raw depth images with resolution at 512×424. Each depth image is pre-processed by cropping out the smallest bounding box which contains the detected human body. **2) Body skeleton joints.** The skeleton joints are estimated positions for 25 different body parts ranging from *SpineBase*, *SpineMid* to *ThumbLeft* and *ThumbRight*. X and Y coordinates are used for each body joint, thus the feature vector for this modality has 50 dimensions in each video frame. **3) Facial landmarks.** There are five facial landmarks are tracked for each RGB video frame: left and right eyes, left and right mouth corners, and nose. X and Y coordinates are used for each landmark. **4) Hand shapes.** There are five states for each hand: open, close, lasso, unknown and not tracked. **5) Facial expressions/attributes.** There are 8 facial expressions or attributes for each face detected: happy, engaged, wearing glasses, left eye closed, right eye closed, mouth open, mouth moved and looking away.

2.2. Feature Extraction

To extract the information from each modality, the input signals are encoded into fixed dimensional spaces. We employ three feature extraction strategies for the five modalities, as follows:

Depth image sequence. Suppose there are L depth images in the input depth sequence, $\{D_0, ..., D_{L-1}\}$. Depth Motion Maps (DMMs) [15] are computed by accumulating consecutive Motion Energy Images (MEIs) [16] of depth maps:

$$DMM = \Sigma_{i=1}^{L-1} \delta(\|D_i - D_{i-1}\|) \quad (1)$$

where $\delta(\cdot)$ is a point-wise delta function. Since the subtraction in $D_i - D_{i-1}$ is also point-wise, the resulted DMM is a 2D map which has the same dimension as each depth image D_i.

Body joints & Facial landmarks. The body joints and facial landmarks are essentially similar in the aspect that both are 2D tracked locations of a sequence of key-points. Therefore we apply the same strategy to encode these two signal modalities. The information is encoded into two perspectives: structure and motion.

The structure feature encodes the layout information of set of tracked key-points, *i.e.*, their relative positions:

$$S_l = \{V_{i,l} - \hat{V}_l, \forall i\}, l \in \{0, ..., L-1\} \quad (2)$$

where $V_{i,l}$ is the i^{th} keypoint of the l^{th} set of points in the sequence $V = \{V_0, ..., V_{L-1}\}$. V can be either a body joint sequence or a facial landmark sequence. $\hat{V}_l$ denotes the l^{th} anchor point such as *SpineMid* for body joints and $Nose$ for facial landmarks.

The motion feature encodes the temporal offset between two consecutive sets of tracked key-points:

$$M_l = \{V_{i,l} - V_{i,l-1}, \forall i\}, l \in \{1, ..., L-1\} \quad (3)$$

Hand shapes & facial expressions/attributes. These two channels are computed by the Kinect API [1] which provides access to some primary facial and hand gestural characteristics. Hand states include: 1) *Open*, 2) *Closed*, 3) *Lasso*, 4) *Unknown*, and 5) *Not Tracked*. Facial states include: 1) *Happy*, 2) *Engaged*, 3) *Wearing Glasses*, 4-5) *Left/Right Eye Closed*, 6) *Mouth Open*, 7) *Mouth Moved*, and 8) *Looking Away*.

For the hand states, *"Unknown"* and *"Not Tracked"* are combined as one state. Since the hand states are exclusive, *i.e.*, one hand cannot be both "Open" and "Closed" at the same time. For two hands (left and right), there are $4 \times 4 = 16$ combinations. Therefore a 16-dimensional binary vector is used to represent the hand state feature. For facial expressions, since the states are not exclusive, we use "on(1)" or "off(0)" to represent each facial expression state. Therefore a $2^7 = 128$-dimensional binary vector is used to encode the face states (*"Wearing Glasses"* is excluded since it is not related to facial expressions.) We name the encoded information of face and hand as binary facial expression (BFE) features and binary hand shape (BHS) features in the rest of this paper, respectively. Consequently, for a video clip containing L frames, the two types of encoded features (BFE and BHS) are: $BFE \in \{0,1\}^{L \times 128}$ and $BHS \in \{0,1\}^{L \times 16}$, respectively.

[1] https://msdn.microsoft.com/en-us/library/dn758675.aspx

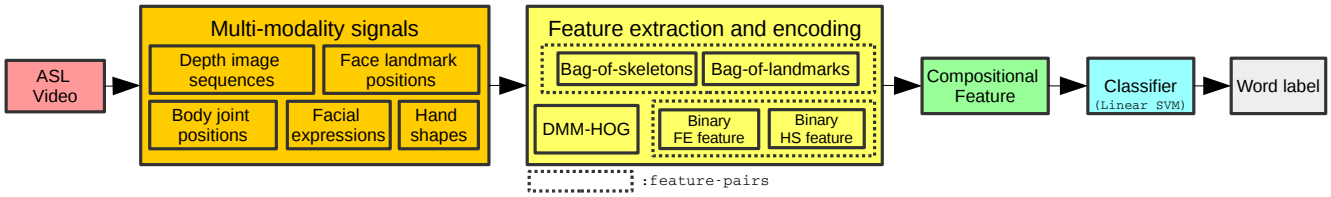


Fig. 1. Pipeline of our system. The input ASL video clip extracts five modalities of signals including different aspects of ASL. Features are extracted from each of the five signals and then combined together to represent that ASL video clip. Then the compositional features are used to train an ASL word classifier.

2.3. Representation and Recognition of ASL Lexical Items

For each feature modality, we need to encode varied length video clips into the same embedding space (fixed dimension) except for the DMM feature, because the resulted energy map is already summed over the temporal dimension. For the feature channels of body joints and facial landmarks, we apply a bag-of-words model which maps the input feature set into a fixed-length vector. K-means and soft assignment are then applied to encode the feature vector. For the latter two feature channels (BFE feature and BHS feature), a histogram over all possible states is used to generate the representation of ASL words and grammars. Linear SVM is employed for ASL word and grammar recognition.

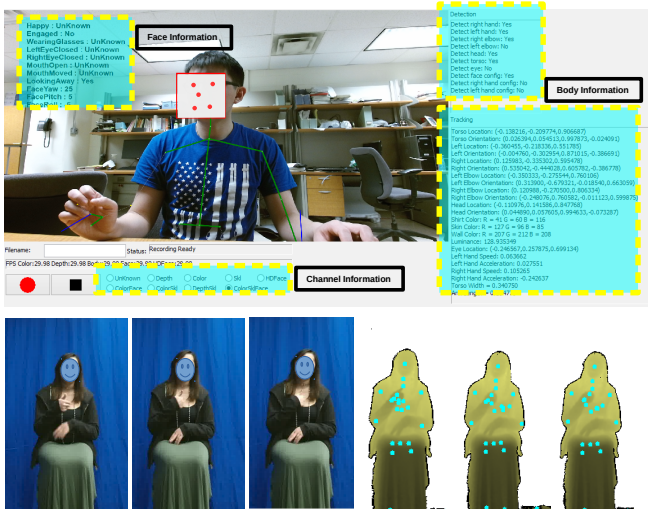


Fig. 2. Top: our interactive UI which shows all extracted information includes RGB, depth, body joints and other facial/body information. Bottom: sample frames from our collected dataset showing the ASL word "I" in both RGB and depth channels (blue dots are body joints.) Faces are blocked for privacy. Please zoom in for more details.

3. EXPERIMENTS

3.1. Dataset

The dataset in this paper is collected by ourselves with a Kinect-based recorder shown in Fig 2. It contains 61 video sequences recorded from five fluent ASL signers, each of which is a multi-sentence performance. All the video sequences are annotated by a team of ASL linguists, who produced a timeline of the words in the video (and a set of facial expressions or head movements with grammatical significance). Using the timeline annotation, we segmented the video collection along word boundaries into a set of 673 video clips, each of which is either a single word (*e.g.*, "I" as in Fig. 2) or an ASL grammar component (*e.g.* "FACE_WHQ"– facial expression for wh-word question). There are total of 99 unique lexical items and 27 of them with sample number larger than 5 are selected in our experiment. [2] For each lexical item, 50% of the samples are used for training the rest half are used for testing.

3.2. Recognition Results

Firstly, two concepts about experiment settings will be discussed:

Feature Pairs. ASL is a language that conveys information through simultaneous movements of the hands, arms, torso, head, face, and eyes. We previously discussed how our system uses five feature channels: 1) DMM-HoG [15], 2) Bag-of-Skeleton, 3) Bag-of-Landmarks, 4) Binary Facial Expression feature, and 5) Binary Hand Shape feature. Since features 2) 3) and also 4) 5) are similar but contain complementary information (gestural and facial), we group them into two feature pairs.

Vocabulary Groups. The 27 lexical items are divided into 4 groups according to their frequencies in the corpus because 1) mean accuracies on class categories with similar frequencies are more informative than on unbalanced corpus. The full listings of lexical items can be seen in Fig. 3. We also compare the accuracies with different feature combinations among all groups and without grouping (putting all lexical items into the same big group.)

[2] The 27 lexical items are: where, father, mother, there, animal, favorite and etc. Please zoom Fig 3 for the full list.

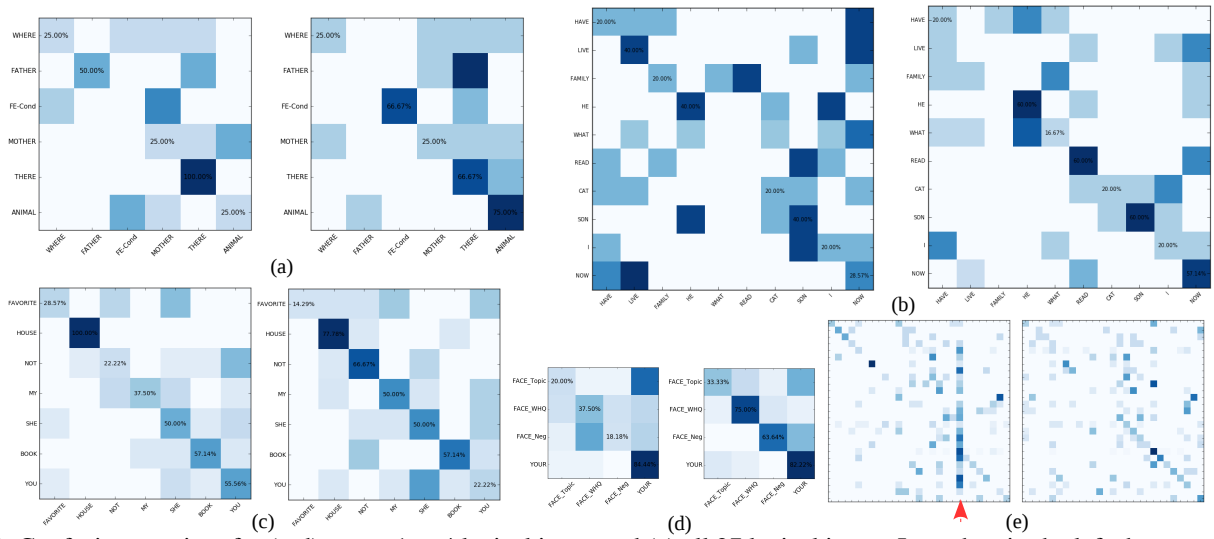


Fig. 3. Confusion matrices for (a-d) group 1 to 4 lexical items and (e) all 27 lexical items. In each pair, the left shows results of DMM only and the right shows using all features. Due to the space limitation, the text for (e) is omitted. Red arrow shows an undesired classification result due to unbalanced data.

Recognition results across all feature combinations and vocabulary groupings are shown in Table 1. By comparing with DMM, feature pairs 1 and 2, we observe that for low-frequency group (Group-1), feature pair 2 reaches the best result (45.45%). The reason is that the cardinality of the training set is small (due to low vocabulary frequency) and feature pair 2 contains the lowest dimensional feature while other feature channels are of higher dimensions, which make the data scarcity more severe. But with the cardinality of the training samples per class increases (Group-2, 3, 4 and "All"), this phenomenon is not severe any more. In general, combining all features together performs the best, especially in Group-4 and "All". As expected, in Group-4 test, features without sufficient information from facial expression (only DMM) performs inferior than all other groups with facial information.

Further observation can be obtained from Figure 3 by comparing DMM-HOG and all feature combination. In general, feature combination is superior to DMM-HOG in all cases except for (c) Group-3. The performance gap is more obvious in (d) Group-4 and (e) All lexical items. In (e), the red arrow indicates DMM-HOG tends to generate many false-positives by assigning many incorrect samples to the most frequent term (the corresponding vocabulary class is "YOU", which is the most frequent word in our vocabulary) and feature combination has no such issue.

Table 1. Recognition results of the proposed framework with different combinations of feature modalities.

	Group-1	Group-2	Group-3	Group-4	All
DMM[15]	36.36%	22.64%	**50.88%**	56.32%	28.77%
Pair-1	36.36%	28.30%	**50.88%**	62.07%	15.98%
Pair-2	**45.45%**	24.53%	28.07%	63.22%	24.66%
All	40.91%	**32.08%**	49.12%	**70.11%**	**36.07%**

3.3. Discussion

Since our data collection is on-going and we are expecting more data, the complexity and the size of vocabulary of this problem will be also growing. In the future, we will scale this framework so that the system's accuracy is maintained as it is expanded to include a larger vocabulary of lexical items. We note that the feature combination in the current system is performed through simple concatenation. As the complexity of the system increases in the future (as we attempt to recognize larger vocabulary sizes) there will be more feature channels included; thus, more effective fusion techniques and automatic ASL sentence segmentation will be explored.

4. CONCLUSION

In this paper, we have described how to identify specific lexical items in a video recording (RGB+depth) of ASL signing. This research is part of a project to develop educational tools to provide feedback for ASL students [3]. The proposed system learn from labeled ASL videos captured by a Kinect depth sensor and predict ASL components in new input videos. In addition to the system we are developing, the corpus of ASL videos that we are collecting (using a Kinect sensor, along with linguistic annotations of lexical items and other grammatical features) will serve as a valuable dataset for research on sign recognition technologies.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and IIS-1400802.

6. REFERENCES

[1] M Paul Lewis, Gary F Simons, and Charles D Fennig, "Ethnologue: Languages of the world, 17 edn. dallas: Sil international," 2013.

[2] Ross E Mitchell, Travas A Young, Bellamie Bachleda, and Michael A Karchmer, "How many people use asl in the united states? why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[3] Matt Huenerfauth, Elaine Gale, Brian Penly, Mackenzie Willard, and Dhananjai Hariharan, "Comparing methods of displaying language feedback for student videos of american sign language," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 139–146.

[4] Chenyang Zhang, Xiaodong Yang, and YingLi Tian, "Histogram of 3d facets: A characteristic descriptor for hand gesture recognition," in *Automatic Face and Gesture Recognition (FG), 10th IEEE International Conference on*. IEEE, 2013, pp. 1–8.

[5] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[6] Antonio W Vieira, Erickson R Nascimento, Gabriel L Oliveira, Zicheng Liu, and Mario FM Campos, "Stop: Space-time occupancy mail patterns for 3d action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 252–259. Springer, 2012.

[7] Yu Zhu, Wenbin Chen, and Guodong Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *CVPR Workshops*. IEEE, 2013, pp. 486–491.

[8] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *CVPR*. IEEE, 2014, pp. 588–595.

[9] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu, "Robust 3d action recognition with random occupancy patterns," in *ECCV*. 2012, pp. 872–885, Springer.

[10] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*. IEEE, 2012, pp. 1290–1297.

[11] Lu Xia and JK Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *CVPR*. IEEE, 2013, pp. 2834–2841.

[12] Nicolas Pugeault and Richard Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *ICCV Workshops)*. IEEE, 2011, pp. 1114–1119.

[13] Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N Metaxas, and Carol Neidle, "Recognizing eyebrow and periodic head gestures using crfs for non-manual grammatical marker detection in asl," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[14] D Metaxas, B Liu, F Yang, P Yang, N Michael, and C Neidle, "Recognition of nonmanual markers in asl using non-parametric adaptive 2d-3d face tracking," in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association*, 2012.

[15] Xiaodong Yang, Chenyang Zhang, and YingLi Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *ACM Multimedia*. ACM, 2012, pp. 1057–1060.

[16] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.