LEARNING MOTION DISFLUENCIES FOR AUTOMATIC SIGN LANGUAGE SEGMENTATION

Iva Farag[*]

Saarland University
Saarland Informatics Campus
66123 Saarbrücken, Germany

Heike Brock

Honda Research Institute Japan
8-1 Honcho, Wako-shi
Saitama 351-0188, Japan

ABSTRACT

We introduce a novel technique for the automatic detection of word boundaries within continuous sentence expressions in Japanese Sign Language from three-dimensional body joint positions. First, the flow of signed sentence data within a temporal neighborhood is determined utilizing the spatial correlations between line segments of inter-joint pairs. Next, a frame-wise binary random forest classifier is trained to distinguish word and non-word frame content based on the extracted spatio-temporal features. The output of the classifier is used to propose an automatic word synthesis that achieves reliable and accurate sentence segmentation with an average frame-wise F1 score of 0.89. Evaluation with a baseline data set furthermore shows that the proposed approach can easily be adapted to distinguish between motion transitions and motion primitives for a coarse-action domain.

Index Terms— sign language understanding, temporal segmentation, angular motion features, disfluency detection, binary classification

1. INTRODUCTION

The development of systems that recognize utterances in signed languages constitutes an important step for better inclusion of deaf or hard of hearing (DHH) individuals. However, current sign recognition systems are still far from being applicable in real life scenarios. One of the main reasons for this is the continuous character of sign utterances. Although previous works report recognition accuracies of 90% or higher on isolated signs or finger spelling [1, 2], similar high accuracies could not yet be obtained when using continuous, full sentence data: as in spoken languages, an expression's flow is dependent on speed, content and personal style variations, and two lexical items might merge into one movement without clear separation. Common features of natural language such as a high number of vocabulary and highly imbalanced word occurrence, as well as specialties of the visual movement-based language representation further impede the learning of accurate and reliable classifiers and recognition networks.

To date, best working systems utilize a combination of deep image classifiers such as Convolutional Neural Networks (CNNs) with parallelized sequence modeling [3, 4]. These deep neural networks were trained directly on the continuous sign data, leaving temporal dependencies of full sentence expressions for the network to learn as is. However, accuracy of such networks is largely dependent on the quality and amount of available training data, and most publicly available data bases for sign language recognition cannot be considered sufficiently large and balanced for meaningful end-to-end learning. Therefore, we pursue a step-wise sentence content recognition approach: first, a method should be trained to automatically detect word boundaries and segment the signed expression into its grammatically correct sub-parts, which can then be classified by a neural network.

Given that prior temporal segmentation might reduce overall system accuracy in case of imperfect segmentation [5], we aim to develop a robust segmentation method that is dependent on as little variant parameters as possible. Motivated by previous joint angle based feature descriptors [6], we propose a new angular representation of skeletal correlations that is descriptive for movements with respect to differently-paced sub-motion parts. We then train a binary random forest classifier which simply annotates every frame of a given motion sequence as either word or non-word, and subsequently obtains a valid sentence split proposal. This strategy offers two advantages. First, using motion data on a per-frame basis, we obtain a much larger number of available training data to learn an appropriate level of classifier sensitivity. Second, frequencies of over- or underrepresented corpus words cannot influence the overall accuracy. As a proof of concept, we evaluate our method on two different data set, one data set containing sentence expressions in Japanese Sign Language (JSL) and one popular human activity recognition data set, and demonstrate that the method is able to provide meaningful distinction between unspecified motion transitions and the actual motion actions.

[*]partially supported by the PROMOS scholarship.

2. REPRESENTATION AND SEGMENTATION OF HUMAN ACTIONS

The spatial representation of human motion within its distinct motion primitives is crucial for understanding and processing motion sequences. Addressing the challenges of capturing skeleton structures, researchers have moved from the simplistic raw motion data [7, 8] to more complex skeleton representations. Some of them are based on features computed from joint body positions, either via using angular displacement maps [9], angular joint displacement maps (JADMs) [6] or kinematic-chain induced correlations [10]. Following their success, we adopt a similar strategy and use 3D body positions to extract a compact data representation that captures the characteristics of action primitives. For this we introduce a novel feature computation based on the geometric relations of line segments built from semantically meaningful joint pairs within a defined term of movement.

Furthermore, current methods study the temporal evolution of the skeleton by exploring statistical relationships of its joints over time by a covariance matrix [11], PCA-based representation [12], alignment kernel computations [13, 14] or self-similarity measures [15, 16]. Most of the approaches however disregard transitional segments and blend them with their neighboring action segments. And while [16] offers a distinction between transition and coherent structures, it relies on the assumption of repetitive or cyclic structures within a motion activity, which is uncharacteristic for sign language motion, in particular. In our method, we additionally apply a kernel transformation over the skeleton features within a given input sequence. Thus, we account for non-linear dependencies over time. This allows us to obtain an accurate temporal segmentation of pure activity sub-sequences that is sensitive to the spatial as well as the velocity components of the motion.

3. SIGN SEGMENTATOR

The main segmentation method is based on the transformation of movement data into a descriptive and robust feature representation. Afterwards, the encoded spatio-temporal information of the skeleton motion is used to train a binary random forest classifier and perform automatic word segmentation.

3.1. Feature Representation

Geometric Descriptors We utilize body joint positions represented in a global 3D coordinate system to compute line segments between pairs of joints and consider their spatial relationship to other line segments.

Assume a skeleton sequence of t frames, where each frame consists of $J_1, \ldots, J_n$ joint positions given by their 3D coordinates, $J_p = (px, py, pz)$. Overall we have $n \times (n-1)/2$ unique joint pairs and their corresponding line segments.

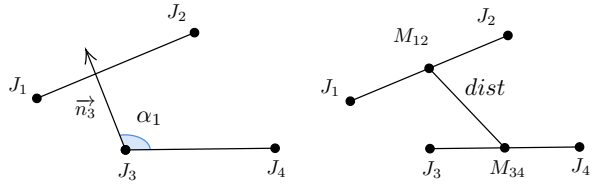

Fig. 1. Geometric angular and distance features between the line segments J_1J_2 and J_3J_4.

Let S be a set of domain-specific pairs of line segments. Each segment pair $s \in S$ is defined as a set of four joints $s = (J_1, J_2, J_3, J_4)$ that uniquely define the respective lines (J_1, J_2) and (J_3, J_4). To capture their spatial relationship, we derive angular and a distance features (Figure 1).

Without loss of generality, we consider (J_3, J_4) to be the reference segment. First, we obtain the normal vector $\overrightarrow{n_3}$ to the line J_1J_2 that passes through the point J_3. Then, we define the corresponding angular feature as

$$\cos\alpha_1 = \frac{\overrightarrow{n_3} \cdot \overrightarrow{J_3J_4}}{||\overrightarrow{n_3}|| \cdot ||\overrightarrow{J_3J_4}||}. \quad (1)$$

Similarly, we compute the second angular feature $\cos\alpha_2$ with respect to J_1J_2's normal vector $\overrightarrow{n_4}$ that passes through the point J_4. Those angles encode the spatial relationship between the line segments with respect to their rotational displacement. To further account for translational motion, we define the distance feature between two line segments to be the Euclidean distance between their respective middle points, $dist = ||M_{12} - M_{34}||$. We concatenate the features derived for each pair of line segments in S and get for each frame i its corresponding geometric feature vector f_i of size $3 \times |S|$. For a given motion sequence of length t the geometric feature matrix is $\mathbf{F} = [f_1 \ \ldots \ f_t] \in R^{(3\times|S|)\times t}$.

Finally, in order to additionally incorporate the temporal information of the skeleton motion sequence, we consider each frame in the context of the features of its neighboring frames within a window of size w_g. This comes down to the following spatio-temporal representation for each frame i: $x_i^{\text{Geo}} = [f_{i-w_g}, \ldots f_i, \ldots f_{i+w_g}]$.

Kernel Descriptors Due to the complexity of the task at hand, we further consider the non-linear relationship between the geometric skeletal features belonging to different frames by applying a kernel transformation.

For a motion sequence of t frames and its corresponding geometric feature matrix $\mathbf{F}$, the frame-wise kernel matrix is defined as,

$$\mathbf{K} = \phi(\mathbf{F})^\top \phi(\mathbf{F}) \in R^{t\times t}, \quad (2)$$

where $k_{i,j}$ characterizes the similarity between the spatial feature vectors f_i and f_j of frames i and j in terms of

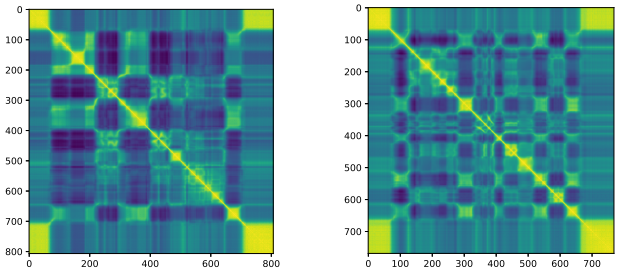

Fig. 2. Laplacian kernel matrix between the geometric feature vectors with normalized distance measure for two example sequences of JSL motion.

the kernel function $\phi(f_i)^\top \phi(f_j)$ (Figure 2). We observe a distinct relationship between the geometric descriptors of neighboring frames. Therefore, we additionally construct the kernel feature vector for a given frame i and window size w_k as the flattened upper triangular sub-kernel $x_i^{\text{Ker}} = \text{triang}[\dots K_{uv} \dots]$, $\forall\, u, v \in [i - w_k, i + w_k]$. From this definition we can derive further high-level understanding of skeleton movement dependencies over time.

Finally, the complete feature descriptor for each frame i of a motion data is given as $x_i = [x_i^{\text{Geo}}, x_i^{\text{Ker}}]$. Those per-frame feature vectors simultaneously capture spatial information about the relative position of body joints with respect to each other as well as their temporal evolution over the whole motion sequence.

3.2. Random Forest Based Split Proposal

We exploit the proposed features for learning a random forest model. The training set is defined as $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ with $(x_1, \dots, x_n)$ the uniform-length feature vectors for each frame of the motion sequence data as described above and $(y_1, \dots, y_n)$ the corresponding binary labels identifying frames belonging to an action primitive (sign word) and transitional frames. For the given classification task, we train a balanced binary random forest with bootstrapping. The learned model is then used to classify frames from unseen motion sequences. Consecutive frames belonging to the same class are interpreted as an action segment (class 1) or an action transition (class 0). As a final post-processing step, we remove the singleton artifacts for a refined segmentation.

4. EXPERIMENTAL SETTINGS

We train and evaluate our method on two diverse data sets built using optical motion capture systems, our own JSL sentence data set (DJSLC) for implementation of a future JSL translation system [17] and parts of the CMU motion capture data set as a common baseline comparison [18]. For both data sets, we train a random forest with 300 decision trees and window sizes $w_g = 2$ and $w_k = 10$. We apply a Laplacian kernel transformation with $\gamma = 1/\#features$ for its better performance on high dimensional data.

For the DJSLC, we use a feature vector of size 815 built from geometric and kernel values of 41 domain-specific segment pairs between joint positions of the upper body (hip to head) and finger joints. In particular, we focus on inter- and intra-hand segment pairs as well as hand-body segments. DJSLC contains 1432 sequences of 3 to 12 daily domain words captured with a temporal resolution of 60Hz. We perform an 8-fold cross validation, meaning that in every fold we use 1253 sentences for training and the remaining 179 sentences for testing so that in total every sentence is part of a test set once. Within every test fold the ratio of class 0 and class 1 frames was approximately 65:35.

For the CMU data set, we adapt our choice of line segment pairs to the characteristics of the given motion. We focus mainly on segments connecting high activity joints such as wrists, knees, feet, chest and their dependency over time. In total, we obtain 50 pairs of segments and a final feature vector of size 960 for each input frame. From the CMU data set, we consider sequences 1-8 performed by subject 86 as containing distinct movements.Since our objective is to distinguish between high activity action motion and in-between transitions, we further label actions such as walking and standing as transitional motion. We perform leave-one-out cross-validation over all input samples with sampling rate of 30 Hz.

5. RESULTS AND ANALYSIS

We evaluate the quality of our temporal segmentation in terms of correct binary detection of all class 1 frames and compare it to current state-of-the art algorithms.

5.1. JSL Data

To evaluate the predictive abilities of our skeleton feature vectors, we consider a random forest model trained on JADM features as defined in [6] and a model trained only over our proposed geometric features. We compare them to our full classifier based on both geometric and kernel relationships. We report the average performance over an 8-fold cross-validation in Table 1. We observed that our proposed geometric skeleton representation is more descriptive than the JADM features in the context of this segmentation task. Furthermore, the best F1 score is achieved by combining the spatial features and their kernel transformation with strong statistical significance (average p=10.41 in a chi-square based McNemar test between $\text{RF}_{\text{Geo+Ker}}$ and RF_{Geo}), confirming the advantage of incorporating non-linear relationships.

We investigate the practical application of our method by examining the exact word proposals in comparison to the ground truth segmentation (Figure 3). In most cases, our algorithm was able to correctly identify the position of words within the sentences. In some cases however it splits com-

Table 1. Analysis of the binary classification performance of random forest classifiers trained with different skeleton features on the DJSLC.

	precision	*recall*	$F1$	*accuracy*
RF_{JADM}	0.87	0.82	0.84	0.89
RF_{Geo}	**0.89**	0.85	0.87	0.91
$RF_{Geo+Ker}$	**0.89**	**0.90**	**0.89**	**0.92**

plex sign words into sub-movements or fails to detect fast and short sign segments (bottom two sequences in Figure 3). Such discrepancies could potentially be further refined based on the confidence values provided by the classifier. Their post-processing as well as impact on subsequent full sentence classification should be investigated as a next step.

5.2. CMU Data

We examine the performance of our method on the general full body motion of the CMU data set. We compare our strategy to a baseline random forest model trained only over the absolute positions of the skeleton joints as well as the Region Growing action partitioning proposed by [16]. We report the statistical results of a strict frame-wise evaluation in Figure 4. Our segmentation technique consistently outperforms the baseline and achieves F1 score rates similar to a current state-of-the-art temporal segmentation algorithm.

Figure 5 shows the action primitives suggested by our model and compares them to the Region Growing and the ground truth annotation. We were able to detect sensitive transitions between separate motion segments, while sometimes directly identifying their underlying substructures. However, if motions flow into each other without visible transition, the actions were merged together. The results obtained from the CMU data set are a proof-of-concept and demonstrate the practical application of the proposed method.

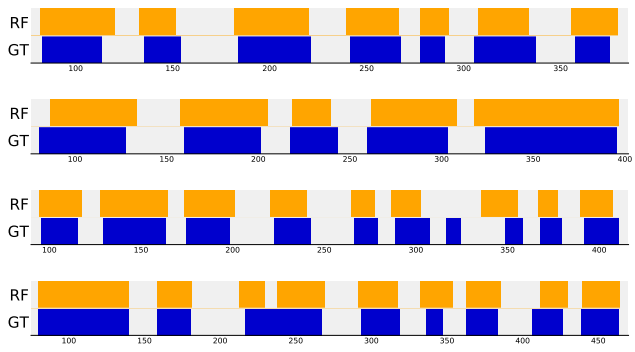

Fig. 3. Segmentation of Japanese sign language motion sequences returned by our algorithm in comparison to the ground truth.

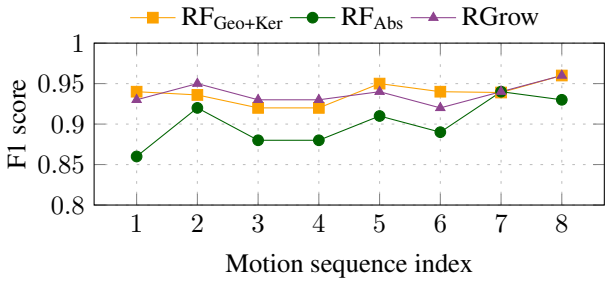

Fig. 4. F1 score of strict frame-wise classification of sequences 1-8 for subject 86 obtained by our algorithm ($RF_{Geo+Ker}$), a simpler baseline version of it (RF_{Abs}) and the Region Growing (RGrow) segmentation by [16]).

6. CONCLUSION

We introduced a novel method for the segmentation of continuous motion data for application with continuous sentence expressions in Japanese Sign Language. The method is based on the composition of spatio-temporal angular and distance features between domain-specific pairs of joint segments. A binary random forest model was trained on the extracted features for automatic word synthesis from motion sequences. Strict frame-wise evaluation of the classifier reaches an average F1 score of 0.89 for a 8-fold cross-validation cycle. This suggests that the proposed combination of statistical signal processing and machine learning is able to reveal hidden characteristics in the sign motion that can be retrieved as indicators for data segmentation. The universal properties of our segmentation strategy were also tested on a full-body human motion data set popular for general-purpose activity recognition. Results show that the algorithm is able to distinguish between dynamic and static motion phases and reaches similar segmentation accuracy as previous state-of-the-art methods.

Next, our method should be employed in a two-stage classifier for the given continuous sign motion data, promising to provide opportunities for higher recognition accuracy in this challenging setting. Additionally, it should be further applied to other data sets in order to examine the quality of its performance on different sign languages and data qualities.

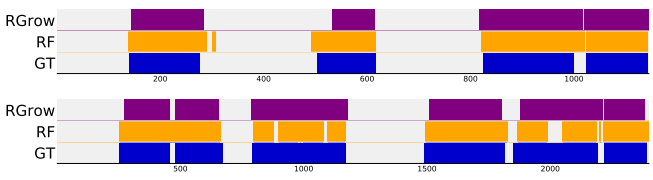

Fig. 5. Ground truth segmentation for sequences 1 (top) and 2 (bottom) for subject 86 of the CMU data and the segmentation by Region Growing (RGrow) proposed in [16].

7. REFERENCES

[1] Chao Sun, Tianzhu Zhang, Bing-Kun Bao, Changsheng Xu, and Tao Mei, "Discriminative exemplar coding for sign language recognition with Kinect," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1418–1428, 2013.

[2] Ao Tang, Ke Lu, Yufei Wang, Jie Huang, and Houqiang Li, "A real-time hand posture recognition system using deep neural networks," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 2, pp. 21, 2015.

[3] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden, "Deep Sign: Hybrid CNN-HMM for continuous sign language recognition," in *Proceedings of the British Machine Vision Conference 2016*, 2016.

[4] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *IEEE International Conference on Computer Vision*, 2017, vol. 1.

[5] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li, "Video-based sign language recognition without temporal segmentation," *32nd AAAI Conference on Artificial Intelligence*, 2018.

[6] E Kiran Kumar, PVV Kishore, ASCS Sastry, M Teja Kiran Kumar, and D Anil Kumar, "Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 645–649, 2018.

[7] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 3995–4001.

[8] Ajjen Joshi, Camille Monnier, Margrit Betke, and Stan Sclaroff, "Comparing random forest approaches to segmenting and classifying gestures," *Image and Vision Computing*, vol. 58, pp. 86–95, 2017.

[9] Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[10] Qiuhong Ke, Senjian An, Mohammed Bennamoun, Ferdous Sohel, and Farid Boussaid, "SkeletonNet: Mining deep part features for 3-D action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 731–735, 2017.

[11] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 2466–2472.

[12] Jernej Barbič, Alla Safonova, Jia-Yu Pan, Christos Faloutsos, Jessica K Hodgins, and Nancy S Pollard, "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface 2004*, 2004, pp. 185–194.

[13] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," in *8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[14] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 582–596, 2013.

[15] Anna Vögele, Björn Krüger, and Reinhard Klein, "Efficient unsupervised temporal segmentation of human motion," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2014, pp. 167–176.

[16] Björn Krüger, Anna Vögele, Tobias Willig, Angela Yao, Reinhard Klein, and Andreas Weber, "Efficient unsupervised temporal segmentation of motion data," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 797–812, 2017.

[17] Heike Brock and Kazuhiro Nakadai, "Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese sign language expressions," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*, May 2018.

[18] CMU, "CMU graphics lab motion capture database," Available: `http://mocap.cs.cmu.edu`, 2013, [Online].