Skeleton-Based Gesture Recognition Using Several Fully Connected Layers with Path Signature Features and Temporal Transformer Module

Chenyang Li, Xin Zhang,[*] **Lufan Liao, Lianwen Jin, Weixin Yang**

School of Electronic and Information Engineering, South China University of Technology

eexinzhang@scut.edu.cn

Abstract

The skeleton based gesture recognition is gaining more popularity due to its wide possible applications. The key issues are how to extract discriminative features and how to design the classification model. In this paper, we first leverage a robust feature descriptor, path signature (PS), and propose three PS features to explicitly represent the spatial and temporal motion characteristics, *i.e.*, spatial PS (S_PS), temporal PS (T_PS) and temporal spatial PS (T_S_PS). Considering the significance of fine hand movements in the gesture, we propose an "attention on hand" (AOH) principle to define joint pairs for the S_PS and select single joint for the T_PS. In addition, the dyadic method is employed to extract the T_PS and T_S_PS features that encode global and local temporal dynamics in the motion. Secondly, without the recurrent strategy, the classification model still faces challenges on temporal variation among different sequences. We propose a new temporal transformer module (TTM) that can match the sequence key frames by learning the temporal shifting parameter for each input. This is a learning-based module that can be included into standard neural network architecture. Finally, we design a multi-stream fully connected layer based network to treat spatial and temporal features separately and fused them together for the final result. We have tested our method on three benchmark gesture datasets, *i.e.*, ChaLearn 2016, ChaLearn 2013 and MSRC-12. Experimental results demonstrate that we achieve the state-of-the-art performance on skeleton-based gesture recognition with high computational efficiency.

1 Introduction

With the development of intelligent device (*e.g.*, AR, VR and smart-home devices), hand gesture interaction is attracting more attention because of its wide applications for human/computer interaction and communications. Hand gesture recognition is an important and classic problem. Recently, the accurate vision based pose/skeleton estimation gains more popularity due to cost-effective depth sensor (like Microsoft Kinect and Intel RealSense) and reliable real-time body pose estimation development (Wei et al. 2016). Comparing with RGB-D sequence based gesture recognition, the skeleton based methods are robust to illumination changes and view variations, and avoid motion am-
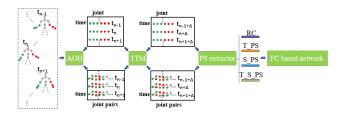
Figure 1: The flowchart of our algorithm.

biguity. In this paper, we focus on the skeleton-based isolated hand gestures recognition, that is, one gesture per one sequence. The key issues are how to extract discriminative spatial temporal features and how to design the classification model.

Hand gestures can be quite different among various users and application scenario. Since human action and hand gesture are similar in terms of motion representation and problem formulation, here we discuss them together. The gesture recognition framework usually involves the feature description and temporal dependency based classification model. Traditional skeleton-based action recognition approaches involve hand-crafted feature extraction. The joint trajectory covariance matrix (Hussein et al. 2013), pairwise relative positions (Wang et al. 2012), 3D joint histogram (Xia, Chen, and Aggarwal 2012) and Lie group embedding (Vemulapalli, Arrate, and Chellappa 2014) are used to represent the skeleton sequences. Human-crafted features are straightforward but with limited representative abilities, which usually require the domain knowledge. Recently, convolutional neural network (CNN) and its extensions are widely used for feature extraction, like 2D-CNN, 3D-CNN, C3D etc. C3D (Tran et al. 2015) is a deep 3D convolutional network model based spatial-temporal feature. It is a generic, compact and implicit representation but requires the large-scale training data. Regarding the action temporal dynamics, Fourier temporal pyramid (FTP) (Veeriah, Zhuang, and Qi 2015) and hidden Markov model (HMM) (Xia, Chen, and Aggarwal 2012) are used with hand-crafted features. For the deep learning method, different structures of Recurrent Neural Networks (RNN), *e.g.*, hierarchical RNN (Du, Wang, and Wang 2015), RNN with regularization (Zhu et al. 2016) differential RNN (Veeriah, Zhuang, and Qi 2015), two-

stream RNN (Wang and Wang 2017) and Long Short-Term Memory (LSTM) (Weng et al. 2018), are popular choice to explore the temporal dependency for recognition. These frameworks have reached state-of-the-art recognition results but the computational complexity may be unacceptable in real-world applications. Hence, we need simple, compact and explicit features to represent global body movements and fine hand motions. Also, the classification model should be simple with temporal dependency.

In this paper, we propose the path-signature feature based hand gesture recognition framework with only few fully connected layers. The flowchart of our algorithm is shown in Fig. 1. The main contributions are as follows:

- We introduce the three different path signature (PS) features , *i.e.*, spatial (S_PS), temporal (T_PS) and temporal spatial PS (T_S_PS) features, to explicitly characterize spatial configuration and temporal dependency of hand gestures. We also propose an AOH principle to define joint pairs for the S_PS and select single joint for the T_PS. In addition, the dyadic method is employed to extract the T_PS and T_S_PS features that encode both global and local temporal dynamics.

- We propose the temporal transformer module (TTM) that can actively produce an appropriate temporal transformation for each input sequence. This is a learning-based module that can be included into standard neural network architecture.

- We propose an extremely simple multi-streams architecture as the classifier with only several fully connected (FC) layers. Different features have their own channels and the FC layer are used for final fusion.

- By only using skeleton data, our method obtains the state-of-the-art results on three major benchmarks, *i.e.*, ChaLearn 2013, ChaLearn 2016 and MSRC 12. Further, our model requires less floating-point multiplication-adds and training memory.

2 Related Work

2.1 Gesture recognition

Skeleton based hand gesture recognition methods are much less than those dealing with the full body skeleton based action recognition (Wang et al. 2018). It is limited by the dataset availability and gesture unique property. Hand gesture mainly involves the finger, palm and hand motion, which only has 1-3 joints in the skeleton obtained from depth data. In (De Smedt, Wannous, and Vandeborre 2016), several skeleton-based features are used together as temporal pyramid, including shape of connected joints, histogram of hand direction and histogram of wrist rotation. In 2D skeleton are superimposed onto original image as dynamic signatures. These features aim to describe the hand motion in detail but the description and representation abilities are limited.

Deep learning have made great process in the area of action recognition. Considering the sequential property, it is natural to apply the RNN, LSTM and their extensions to learn temporal dynamics. In 2017 ChaLearn LAP RGB-D

isolated gesture recognition competition (Wan et al. 2017), the largest hand gesture recognition contest, most participants (including the winner) used C3D and/or LSTM neural networks. C3D architecture has been widely used in the action recognition for appearance and motion modeling because it is more suitable for spatial-temporal feature extraction than 2D CNN. The model inputs are multi-modalities including RGB, depth, optical flow and/or skeleton. Recently, (Weng et al. 2018) proposed a deformable pose traversal convolution method based on 1D convolution and LSTM. These recognition networks usually have multiple LSTMs and temporal streams channels and the final result is multi-stream average fusion. In (Wang and Wang 2017), RNN architecture not only characterizes the temporal dynamics but also considers the spatial configuration in the two-stream architecture. With the proper modeling of skeleton structure and spatial dependency of the action, recognition accuracy increased. In the latest hand gesture recognition research (Narayana, Beveridge, and Draper 2018), by using RGB-D and their flow as inputs, the network has 12 channels representing the large body movement and fine hand motions individually. The fusion channel is a sparsely connected network with one weight per gesture and channel. The RNN/LSTM frameworks deliver the state-of-the-art performance on most action and gesture recognition datasets, indicating the excellent feature and dependency learning capabilities. The only concerns are the architecture complexity, training data requirement and computational efficiency.

2.2 Path signature feature

The path signature (PS) was first proposed in (Chen 1958) in the form of noncommutative formal power series. After that PS was used to solve differential equations driven by rough paths (Lyons 1998; Garrido 2010). Recently, the path signature has been successfully used as a trajectory descriptor and applied to many tasks in the field of machine learning and pattern recognition, such as quantitative finance (Gyurkó et al. 2013; Lyons, Ni, and Oberhauser 2014), handwriting recognition (Lai, Jin, and Yang 2017; Yang et al. 2015), writer identification (Yang, Jin, and Liu 2015; 2016; Liu, Jin, and Xie 2017), human action (Yang et al. 2017) and hand gesture recognition (Li, Zhang, and Jin 2017). (Yang et al. 2017) is the pioneer work employing the path signature feature for skeleton-based action recognition. All joint pairs and temporal joint evolution are considered as *path* and the corresponding path signatures are computed as features. The concatenation of all path signatures are the input vector for classification. In (Li, Zhang, and Jin 2017), the path signature is the firstly used in the gesture recognition by defining the hand trajectory as the path. Path signature can provide the informative representation of sequential data but how to define proper paths and how to deal with their high dimensionality is worthy to be explored.

3 Overview of Path Signature

In this section, we will briefly introduce the mathematical definition, geometric interpretation and some properties of

path signature (PS), which is mainly referred to (Chevyrev and Kormilitzin 2016).

Assume a path $P : [t_1, t_2] \to \mathbb{R}^d$, where $[t_1, t_2]$ is a time interval. The coordinate paths are denoted by $(P_t^1, ..., P_t^d)$, where each $P^i : [t_1, t_2] \to \mathbb{R}$ is a real-value path. For an integer $k \geq 1$ and the collection of indices $i_1, ..., i_k \in \{1, ..., d\}$, the k-fold iterated integral of the path along indices $i_1, ..., i_k$ can be defined as:

$$S(P)_{t_1,t_2}^{i_1,...,i_k} = \int_{t_1 < a_k < t_2} ... \int_{t_1 < a_1 < a_2} dP_{a_1}^{i_1}...dP_{a_k}^{i_k} \quad (1)$$

where $t_1 < a_1 < a_2 < ... < a_k < t_2$.

The signature of path P, denoted by $S(P)_{t_1,t_2}$, is the collection (infinite series) of all the iterated integrals of P:

$$\begin{aligned} S(P)_{t_1,t_2} = &(1, S(P)_{t_1,t_2}^1, S(P)_{t_1,t_2}^2, ..., S(P)_{t_1,t_2}^d, \\ &S(P)_{t_1,t_2}^{1,1}, ..., S(P)_{t_1,t_2}^{1,d}, ..., S(P)_{t_1,t_2}^{d,d}, \\ &..., \\ &S(P)_{t_1,t_2}^{1,...,1}, ..., S(P)_{t_1,t_2}^{i_1,...,i_k}, ..., S(P)_{t_1,t_2}^{d,...,d}, \\ &...) \end{aligned}$$
$$(2)$$

The k-th level PS is the collection (finite series) of all the k-fold iterated integral of path P. The 1-st and 2-nd level represents path displacement and path curvature respectively. By increasing k, higher levels of path information can be extracted, but the dimensionality of iterated integrals enlarge rapidly as well. Note that the 0-th level PS of path P is equal to 1 by convention.

In practice, we often truncate the $S(P)_{t_1,t_2}$ at level m to ensure the dimensionality of the PS feature in a reasonable range. The dimensionality of $S(P)_{t_1,t_2}$ truncated at level m is calculated through $M = d + \cdots + d^m$ (without zeroth term).

The path is considered as the piecewise linear path after sampling. The PS of a discrete path with finite length can be easily calculate based on linear interpolation and Chen's identity (Chen 1958). For each straight line of a path, the element of its PS can be calculates by:

$$S(P)_{t,t+1}^{i_1,...,i_k} = \frac{1}{k!} \prod_{j=1}^{k} S(P)_{t,t+1}^{i_j} \quad (3)$$

For the entire path, Chen's identity states that for any (t_s, t_m, t_u) satisfying: $t_s < t_m < t_u$, then,

$$S(P)_{t_s,t_u}^{i_1,...,i_k,...,i_n} = \sum_{k=0}^{n} S(P)_{t_s,t_m}^{i_1,...,i_k} S(P)_{t_m,t_u}^{i_{k+1},i_{k+2},...,i_n} \quad (4)$$

PS has two excellent properties for path expression. First, PS is the unique representation of a non tree-like path (Hambly and Lyons 2010). A tree-like path is a trajectory that retraces itself (such as clapping). For time-sequential data, it's natural and effective to add an extra time dimension into the original path to avoid the tree-like situation. Second, shuffle product identity (Lyons et al. 2004) indicates that the product of two signature of lower level can be expressed as a linear combination of some higher level terms. Hence, adoption of higher level terms of PS actually brings more nonlinear prior knowledge, which reduces the need for the classifier of high complexity. More properties and related details can be found in (Chevyrev and Kormilitzin 2016).

We recommend an open-source python library named *iisignature*, which can be easily installed through *pip*.

4 Approach

In this section, we first introduce an "attention on hand" (AOH) principle for PS extraction, which deals with global body movements and fine hand motions. Then we propose a novel temporal transformer module (TTM) to alleviate the sequence temporal variation. Finally a multi-stream architecture is presented to fuse different types of features.

4.1 AOH principle and PS extraction

AOH principle Before calculating the PS feature, we need to consider about what path to be used and how to design paths efficiently for the recognition. In (Yang et al. 2017), the first work leveraging PS features in the human action recognition problem, single joint, joint pair and joint triple are utilized to define paths. They use all N joints and exhaust all the possible pairs and triples (*i.e.*, C_N^2 and C_N^3), which brings performance improvement but increases dimensionality dramatically. In the context of gesture recognition, we propose the AOH principle to select single joints and joint pairs.

For the *single joint*, only the joint belongs to the hand part (including elbow, wrist and hand 3 joints, *i.e.*, $N_J = 3 \cdot 2 = 6$) are selected, as Fig. 2 (a) shows. For the *joint pair*, three kinds of pairs are considered (as depicted in Fig. 2 (b) 1)-3)). The first kind of *joint pair* is inside the same hand part, describing the geometric characteristics of hand explicitly. The second kind of *joint pair* is from two hand parts, indicating the relative state of two hands. The third kind of *joint pair* is across hand part and body part (upper body joints except hand part), characterizing the related location of hand and body. The selected single joint and joint pairs defined by AOH principle can not only model global hand relative body movements and fine hand motions but also make the PS features more compact.

Path definition and PS feature extraction Based on the selected single joint and joint pairs obtained by AOH principle, we further define one spatial path and two kinds of temporal paths for PS feature extraction. We regard each joint pair as a spatial path for the PS feature extraction. The first type of temporal path is the evolution of each selected single joint along the time, as shown in Fig. 2 (c). Another type of temporal path is the evolution of spatial correlations among joints, as shown in Fig. 2 (d). The summary of three PS features are shown in Table 1.

Spatial PS features The fundamental description of spatial structure is the d-dimensional raw coordinates. We concatenate the coordinates of single joints in each frame as a RC vector obeying chronological order as Fig. 2 (c) shows. Further, due to the poor characterization ability of RC and noise interference, we extract PS features over selected joint
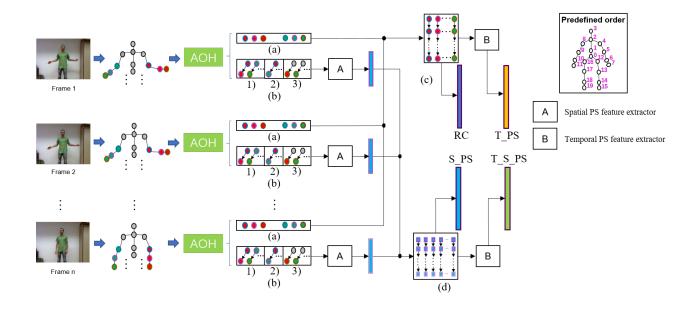
Figure 2: The illustration of PS features extraction (T_PS, S_PS and T_S_PS). (a) and (b) are single joints and joint pairs selected following AOH. (c) and (d) are the temporal paths of single joint and each dimension of S_PS. Note that TTM (proposed in Section 4.2 and illustrated in Fig. 4) should be implemented between AOH and PS extractors, but we omit it here for clarity.

pairs to explore the spatial relation between joints. The implementations of spatial PS feature extractor (Box A in Fig. 2) are as follows: i) Select elements that need to be calculated in Eq.1 according to the truncated level m_S. ii) Calculate the truncated spatial PS of a joint pair (a straight line in Eq.3) by Eq.1 and Eq.3 (The start and end points are defined according to the predefined order in Fig. 2). iii) Finally concatenate the truncated spatial PS of all input joint pairs as the spatial PS (S_PS) feature.

Dyadic temporal PS features The dyadic method with PS was firstly used in (Yang et al. 2016) for the writer ID identification. Since the gesture always contain global and local variation, we employ the dyadic method for temporal PS feature extraction.

The dyadic method divides the entire path into dyadic pieces and set up a hierarchical representation of the path. It can extract both the global and local feature of entire path, and reduce the feature dimensionality as well. If the dyadic level is L_D, then an entire path can be divided into $2^{(L_D+1)} - 1$ subpaths.

To characterize the temporal dynamic of single joint, the evolution of each single joint is treated as an entire temporal path as shown in Fig. 2 (c). An extra monotone time dimension is added to ensure the path uniqueness (*i.e.*, to avoid tree-like path as discussed in Section 3).

To further explore kinematic constraints of the joint pairs, the evolution of each dimension in the S_PS of every frame also can be regarded as another kind of entire temporal path, as Fig. 2 (d) shows. As a result, we acquire a series of 1D paths. However, the signature of a 1D path is just the increments to a certain power, which can be easily get from Eq. 3. To alleviate this problem, we use the lead-lag trans-

formation as (Yang et al. 2017) does over the 1D path to enrich the temporal contextual information.

The implementations of temporal PS feature extractor (Box B in Fig. 2) are as follows: i) Select elements that need to be calculated in Eq.1 according to the truncated level m_T or m_{T_S}. ii) Every subpath generated by dyadic method is an entire path (consist of several straight lines) in Eq. 4. According to Eq. 1, Eq. 3 and Eq. 4, the truncated temporal PS of a subpath can be calculated. iii) Concatenate the truncated temporal PS of all subpaths as temporal PS (T_PS) or temporal spatial PS (T_S_PS) feature.

Table 1: The proposed feature for skeletal hand gesture recognition.

Feature types	Feature description
Raw coordinates (**RC**)	The d-dimensional coordinates of N_J single joints.
Spatial PS (**S_PS**)	The PS over each predefined joint pair truncated at level m_S.
Temporal PS (**T_PS**)	The PS over the temporal evolution of each single joint truncated at level m_T.
Temporal Spatial PS (**T_S_PS**)	The PS over the temporal evolution of each dimension of S_PS truncated at level m_{T_S}.

4.2 Temporal Transformer Module (TTM)

Motivation Although deep neural network achieves break through in the sequential recognition task, it's still limited

by the lack of ability to be temporally invariant to the input sequence in a computationally and parameter efficient manner. In the context of action recognition, the time-stamps of keyframes are variant among different clips, which makes the model difficult to catch the key information along time dimension.

There are mainly two existing types of methods to solve this problem: structure driven method and data driven method. The structure driven method mostly use LSTM to model the temporal contextual dependence of sequence data. The data driven method is to provide more diverse samples by temporal shift data enhancement. However, LSTM model requires large training data and unnegligible training cost. If we used a simple network like FC layer as the classifier, the temporal consistency is also learned as part of features, which is the unwanted result. For example, we visualize the weight matrix of the first FC layer of a trained one-stream network (will be introduced in the following), which takes RC as input, as shown in Fig. 3. The x-axis denotes the input dimensionality (obey chronological order), and the y-axis denotes the neuron number (64 in the first FC layer). The brighter position means the corresponding weight is larger, that is, this connection is more important. The FC layer pay more attention on several time stamp, indicating the position of key frames. If there is the temporal variation between the model and testing sequence, the recognition result is worse. Even if the training data is augmented by temporal shift, the model capacity is too small to fit any arbitrary temporal situation.

Recent work (Cao et al. 2017) proposed a spatiotemporal transform method to deal with the spatiotemporal variation, but their method is for RGB-D video. As mentioned in Section 4.1, we have employed the temporal PS features to represent temporal dynamics within the sequence. This is what exactly (Yang et al. 2017) has done. The inter-sequence temporal difference might be alleviated by the temporal transformation. To this end, we design a differentiable module called temporal transformer module (TTM). This module can actively transform the input data temporally and finally adjust the key frame to the best time stamp for the network.

Proposed TTM Inspired by STN (Jaderberg et al. 2015), TTM is a differentiable module that applies a temporal transform to RC. The TTM contains two steps: Localization network (LN) and temporal shifting.

Firstly, we use LN to learn the temporal transform factor delta (Δ), as shown in Fig.4 (b). It takes the input vector $I \in \mathbb{R}^{D_{RC}}$ (D_{RC} denotes the dimensionality of RC) and output Δ as shown in Fig. 4 (b), *i.e.*, $\Delta = f_{LN}(I)$. Note that the network function $f_{LN}()$ can take any form, such as FC layer or 1D convolution layer, but should finally regress to one neuron.

Secondly, the input vector I is reshaped as a matrix $V^i \in \mathbb{R}^{d \cdot N_J \times F}$, where N_J is the number of single joints (*i.e.*, 6 in Fig. 2) and F denotes the frame number. Each column of V^i is a vector v_x^i which consists of the coordinates of single joints in the same frame, $x \in [1, F]$. And if we denote the matrix after shifting as V^o, then each column of it can be

Figure 3: The visualization of the weight matrix of first fully connected layer in a trained one-stream network.

calculate by:

$$v_x^o = f_I(x - \Delta, V^i) = (1 - \alpha) \cdot v_{\lfloor x-\Delta \rfloor}^i + \alpha \cdot v_{\lceil x-\Delta \rceil}^i \quad (5)$$

Here, because Δ is a decimal, we use linear interpolation function $f_I()$ to generate the V^o. α is calculated by $(x - \Delta) - \lfloor x - \Delta \rfloor$. Note that $x - \Delta$ is clip by value [1, F]. Eventually, V^o is reshaped back as a output vector O. The code of TTM is made publicly available[1].

4.3 Multi-stream Architecture

As discussed in Section 4.1, we define spatial and temporal path based on AOH principle. Corresponding PS features encode the spatial and temporal information of the action. Previous work (Yang et al. 2017; Li, Zhang, and Jin 2017) concatenates these features together as whole and one classification model is employed, *e.g*, FC layers. We believe temporal and spatial features should be treated separately because one represents the joint evolution and the other describes the body configuration. The final result is fusion of these multiple channels.

So we utilize the multi-stream architecture to process different kinds of information separately. As a result, we design three kinds of network architectures, one-stream network (1s_net), two-stream network (2s_net) and three-stream network (3s_net), as shown in Fig. 4(c)-(e) . The 1s_net directly concatenates all the features as one input vector and feed it to a 2-fc-layer network, similar to (Yang et al. 2017). The 2s_net has two inputs, RC and PS, representing basic information and extracted compact information. As defined in Table 1, the T_PS and T_S_PS represent temporal information and S_PS is spatial feature. Hence, the 3s_net has three streams with two FC layers separately. The final fusion result is obtained through a FC layer as the weighted summation.

5 Experimental Results and Discussion

5.1 Datasets

ChaLearn 2013 dataset: It is the ChaLearn 2013 Multi-model gesture dataset (Escalera et al. 2013), which contains 23 hours of Kinect data with 27 persons performing 20 Italian gestures. This dataset provides RGB, depth, foreground segmentation and Kinect skeletons. Here, we only use skeletons for gesture recognition as done in the literature (Wang and Wang 2017).

ChaLearn 2016 dataset: ChaLearn Isolated dataset (Wan et al. 2016), the largest gesture recognition dataset consisting of RGB and depth videos, includes 35,878 training, 5,784 validation and 6,271 test videos for totally 249 gestures.

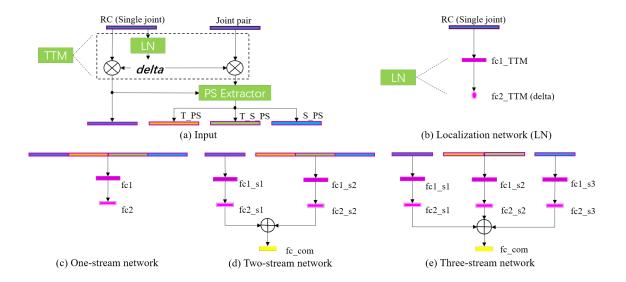[1]https://github.com/LiChenyang-Github/Temporal-Transformer-Module

Figure 4: Three different types of network architectures: one-stream network (1s_net), two-stream network (2s_net) and three-stream network (3s_net). TTM is the temporal transformer module. LN is the localization network that generates the transformation parameters Δ. $\otimes$ and $\oplus$ denote temporal shift and weighted sum. The fc2 in each stream denotes a FC layer with a softmax activation function.

We use Openpose (Wei et al. 2016) to estimate the skeleton joints in all videos as (Lin et al. 2018) did. It can be downloaded from our homepage[2].

MSRC-12 dataset: MSRC-12 gesture dataset (Fothergill et al. 2012) includes 6 iconic and 6 metaphoric gestures performed by 30 people. We use 6 iconic gestures from the dataset that amounts to 3,034 instances and employ 5-fold leave-person-out cross-validation as in (Jung and Hong 2014).

5.2 Data Preprocessing and Network Setting

We first normalize skeletons by subtracting the central joint, which is the average position of all joints in a video clip. Then all coordinates are further normalized to the range of [-1, 1] over the entire video clip. Finally, we sample all videos clips to 39 frames by linear interpolation or uniform sampling. The data enhancement methods we use are three-fold. The first one is temporal augmentation by randomly temporal shift the frame in range of [-5, 5]. The second one is adding Gaussian noise with a standard deviation of 0.001 to joints coordinates. The last one is rotating coordinates along x, y, z three axes in range of $[-\pi/36, \pi/36]$, $[-\pi/18, \pi/18]$ and $[-\pi/36, \pi/36]$.

For ChaLearn 2013 and MSRC-12 two datasets, we set the neuron number of each 2-fc-layer stream to 64 and C (64_fc-C_fc), where C is the gesture class number. For the largest dataset ChaLearn 2016, we use 256_fc-C_fc. We adopt 64_fc-1_fc architecture for $f_{LN}()$. DropOut (Hinton et al. 2012) layer is added after the first FC layer of each stream to avoid over fitting. The mini-batch size and dropout rate are set to 56 and 0.5. We use the method of stochastic gradient descent with a momentum value equal to 0.7. The learn-

ing rate updates in accordance to $\alpha(n) = \alpha(0) \cdot exp(-\lambda n)$, where n is a positive integer starting from 1 and increasing by 1 for each 1/2 epoch. $\alpha(0)$ and λ are set to 0.01 and 0.001. The L_D for T_PS and T_S_PS calculation are set to 3 and 2. The lead-lag dimensionality is set to 2.

5.3 Ablation Study

We do some ablation experiments on the ChaLearn 2013 dataset to explore the truncated level of PS and examine the effectiveness of PS, TTM and multi-stream architecture.

Investigation of the truncated level of PS We utilize the one-stream network without TTM (1s_net_w/o TTM) to explore the contributions of different truncated level PS. The validation accuracy rate is 77.84% with only RC. The accuracy rate after adding different PS are shown in Table 2. It is noted that the T_S_PS is calculated from S_PS truncated at level 2 (The abbreviation can be referred to Table 1). The performance improves after adding any type of PS truncated at any level, which indicates the effectiveness of PS. The last column of each row illustrates that all types of PS feature are complementary.

It is worthy to note that the contributions trend to be negligible and even vanish when truncated level is greater than a certain value. There is a trade-off between validation performance and feature dimensionality. As a result, we choose to set m_T, m_S, m_{T_S} as 4, 2, 3.

Investigation of the TTM We use the 1s_net with RC as input to examine the effectiveness of TTM. Results are presented in Table 3. 1s_net_w/o TTM can be roughly regarded as the method proposed by (Yang et al. 2017).

Firstly, we use temporal augmentation to test whether the data driven methods can make the improvement. We shift

---
[2]http://www.hcii-lab.net/data/

Table 2: The ablation study of PS features on ChaLearn 2013. The truncated level of PS can be referred to Section 3.

PS truncated level	+T_PS	+S_PS	+T_S_PS	+AllPS
1	77.98	82.30	85.35	**85.50**
2	81.28	87.62	87.80	**88.76**
3	84.42	88.12	88.81	**89.17**
4	85.27	88.06	88.09	**89.20**

the samples in range of [-5, 5] to provide more diverse samples. However, it doesn't improve the performance. Hence, directly data driven methods cannot work well.

Then we directly add TTM to the 1s_net (1s_net_TTM), and the result improves from 77.84% to 80.55%. The temporal transformation parameter learned by TTM can fit the key moment of an action and the active part of the FC layer.

At last we add the same temporal augmentation for 1s_net_TTM, and the accuracy rate increases from 80.55% to 81.33%. This attractive observation indicates that TTM makes good use of the diverse samples provided by temporal enhancement. In other words, the network is more adaptive after adding TTM.

Table 3: The ablation study of TTM on ChaLearn 2013. Temporal Enhancement is abbreviated to Temp. Enh.

Components	Accuracy rate (%)
1s_net_w/o TTM	77.84
1s_net_w/o TTM + Temp. Enh.	77.46
1s_net_TTM	80.55
1s_net_TTM + Temp. Enh.	81.33

Investigation of different network architectures For the estimation of different architecture, we use network without TTM, with RC and all PS selected above as input, which can be roughly regarded as the architecture utilized by (Yang et al. 2017). As shown in Table 4, the performance improves clearly, which indicates that the multi-stream architectures allow each stream to dig deeply into one type of feature and finally provide more discriminative information.

Table 4: The ablation study of network architectures.

Components	1s_net	2s_net	3s_net
Accuracy rate (%)	89.43	89.63	90.19

5.4 Comparison with the State-of-the-arts

In this subsection, we use the best parameter setting and network structure getting from our ablation study. We also do all the data augmentation methods mentioned above for our network.

ChaLearn 2013 The results on the ChaLearn 2013 dataset are shown in Table 5. Currently, methods achieving the

best performance are mainly benefited from powerful characterization ability of CNN and LSTM models. (Du, Fu, and Wang 2015) organizes the raw coordinates as the spatial temporal feature maps then feeds it to the hierarchical spatial-temporal adaptive filter banks CNN architecture. (Wang and Wang 2017) propose a two-stream LSTM to model temporal dynamics and spatial configurations separately. Compared with these method, our FC based network achieves the best results with less multiplication-adds operation as Table 6 shows. Note that the AOH principle dramatically reduces the Multiplication-Adds compared with the method without AOH (Yang et al. 2017).

Table 5: Comparison of methods on the ChaLearn 2013.

Method	Accuracy rate (%)
HiVideoDarwin Wang, Wang, and Wang	74.90
VideoDarwin Fernando et al.	75.30
D-LSDA Su et al.	76.80
CNN for Skeleton Du, Fu, and Wang	91.20
Two-stream LSTM Wang and Wang	91.70
3s_net_TTM	**92.08**

Table 6: Comparison of the Multiplication-Adds. [1]: Du, Fu, and Wang [2]: Wang and Wang [1]: PS calculation.

Method	CNN[1]	2s_LSTM[2]	3s_net w/o AOH	3s_net
Mult-adds (Million)	11.54	358.34	14.89^1+15.06	**2.69^1+2.00**

ChaLearn 2016 The results on ChaLearn 2016 are summarized in Table 7. Our model outperforms the skeleton based method (Lin et al. 2018) by around 4.5%. We also notice that the accuracies of skeleton based methods are inferior to video frame based models. The reasons are mainly two-fold. Firstly, the precision of OpenPose is affected by the drastic background and illumination changes. Secondly, a lot of gesture classes requires recognizing the static hand gesture instead of dynamic hand motion. The recognition performance on these classes mainly depends on the estimation precision of hand joints. It is worth noting that our model is the simplest one. We argue that the performance will improve if more accurate joints locations are provided.

Table 7: Comparison on ChaLearn 2016 dataset. [1]: Wang et al. [2]: Miao et al. [3]: Narayana, Beveridge, and Draper [4]: Lin et al. RGB, depth, optical flow and skeleton are abbreviated as R, D, O and S.

Method	Test acc. (%)	Modality				Model
		R	**D**	**O**	**S**	
AMRL[1]	65.59	√	√			8*CNN+4ConvLSTM
ASU[2]	67.71	√	√	√		4*C3D+2*TSN+1*SVM
FOANet[3]	**82.07**	√	√	√		12*CNN
SkeLSTM[4]	35.39				√	1*LSTM
3s_net_TTM	**39.95**				√	3*FC

MSRC-12 (Wang et al. 2016) proposed Joint Trajectory Maps (JIM), which are a set of 2D images that encode spatiotemporal information carried by 3D skeleton sequences in three orthogonal planes. In (Jung and Hong 2014), a novel framework called Enhanced Sequence Matching (ESM) is leveraged to align and compare action sequences based on a set of elementary Moving Poses (eMP). (Garcia-Hernando and Kim 2017) proposed "transition forests", an ensemble of randomized tree classifiers that learnt both static pose information and temporal transitions. All these methods show the importance of spatio-temporal information modelling. Our method extracts spatial, temporal and joint spatial-temporal features and achieves the state-of-the-art accuracy of 99.01%, as shown in Table 8.

Table 8: Comparison of methods on the MSRC-12 dataset.

Method	Accuracy rate (%)
JTM Wang et al.	93.12
DFM Lehrmann, Gehler, and Nowozin	94.04
ESM Jung and Hong	96.76
RJP Garcia-Hernando and Kim	97.54
MP Garcia-Hernando and Kim	98.25
3s_net_TTM	**99.01**

6 Conclusion

In this paper, we first leverage S_PS, T_PS and T_S_PS three PS features to explicitly represent the spatial and temporal motion characteristics. In the path definition, we propose the AOH principle to select single joint and joint pairs, which ensures the feature robust and compact. Furthermore, the dyadic method employed to extract the T_PS and T_S_PS features that encode global and local temporal dynamics with less dimensionality. Secondly, we propose a differentiable module TTM to match the sequence key frames by learning the temporal shifting parameter for each input. Finally, we design a multi-stream FC layer based network to treat spatial and temporal features separately. The ablation study has shown the effective of every contribution. We have achieved the best result on skeleton-based gesture recognition with high computational efficiency on three benchmarks. We will explore the possible combination of the attention scheme and PS features.

7 Acknowledgments

References

Cao, C.; Zhang, Y.; Wu, Y.; Lu, H.; and Cheng, J. 2017. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3763–3771.

Chen, K.-T. 1958. Integration of paths–a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society* 89(2):395–407.

Chevyrev, I., and Kormilitzin, A. 2016. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*.

De Smedt, Q.; Wannous, H.; and Vandeborre, J.-P. 2016. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–9.

Du, Y.; Fu, Y.; and Wang, L. 2015. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, 579–583. IEEE.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Escalera, S.; Gonzàlez, J.; Baró, X.; Reyes, M.; Lopes, O.; Guyon, I.; Athitsos, V.; and Escalante, H. 2013. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 445–452. ACM.

Fernando, B.; Gavves, E.; Oramas, J. M.; Ghodrati, A.; and Tuytelaars, T. 2015. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5378–5387.

Fothergill, S.; Mentis, H.; Kohli, P.; and Nowozin, S. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1737–1746. ACM.

Garcia-Hernando, G., and Kim, T.-K. 2017. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 432–440.

Garrido, A. 2010. System control and rough paths (oxford mathematical monographs).

Gyurkó, L. G.; Lyons, T.; Kontkowski, M.; and Field, J. 2013. Extracting information from the signature of a financial data stream. *arXiv preprint arXiv:1307.7244*.

Hambly, B., and Lyons, T. 2010. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics* 109–167.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hussein, M. E.; Torki, M.; Gowayyed, M. A.; and El-Saban, M. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, volume 13, 2466–2472.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Jung, H.-J., and Hong, K.-S. 2014. Enhanced sequence matching for action recognition from 3d skeletal data. In *Asian Conference on Computer Vision*, 226–240. Springer.

Lai, S.; Jin, L.; and Yang, W. 2017. Toward high-performance online hccr: A cnn approach with dropdistortion, path signature and spatial stochastic max-pooling. *Pattern Recognition Letters* 89:60–66.

Lehrmann, A. M.; Gehler, P. V.; and Nowozin, S. 2014. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1314–1321.

Li, C.; Zhang, X.; and Jin, L. 2017. Lpsnet: A novel log path signature feature based hand gesture recognition framework. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, 631–639. IEEE.

Lin, C.; Wan, J.; Liang, Y.; and Li, S. Z. 2018. Large-scale isolated gesture recognition using a refined fused model based on masked res-c3d network and skeleton lstm. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, 52–58. IEEE.

Liu, M.; Jin, L.; and Xie, Z. 2017. Ps-lstm: Capturing essential sequential online information with path signature and lstm for writer identification. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, 664–669. IEEE.

Lyons, T.; Caruana, M.; Lévy, T.; and Picard, J. 2004. Differential equations driven by rough paths. *Ecole d'été de Probabilités de Saint-Flour XXXIV* 1–93.

Lyons, T.; Ni, H.; and Oberhauser, H. 2014. A feature set for streams and an application to high-frequency financial tick data. In *Proceedings of the 2014 International Conference on Big Data Science and Computing*, 5. ACM.

Lyons, T. J. 1998. Differential equations driven by rough signals. *Revista Matemática Iberoamericana* 14(2):215–310.

Miao, Q.; Li, Y.; Ouyang, W.; Ma, Z.; Xu, X.; Shi, W.; Cao, X.; Liu, Z.; Chai, X.; Liu, Z.; et al. 2017. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3047–3055.

Narayana, P.; Beveridge, J. R.; and Draper, B. A. 2018. Gesture recognition: Focus on the hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5235–5244.

Su, B.; Ding, X.; Wang, H.; and Wu, Y. 2018. Discriminative dimensionality reduction for multi-dimensional sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(1):77–91.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Veeriah, V.; Zhuang, N.; and Qi, G.-J. 2015. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, 4041–4049.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 588–595.

Wan, J.; Zhao, Y.; Zhou, S.; Guyon, I.; Escalera, S.; and Li, S. Z. 2016. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 56–64.

Wan, J.; Escalera, S.; Anbarjafari, G.; Escalante, H. J.; Baró, X.; Guyon, I.; Madadi, M.; Allik, J.; Gorbova, J.; Lin, C.; et al. 2017. Results and analysis of chalearn lap multi-modal isolated and con-tinuous gesture recognition, and real versus fake expressed emotions challenges. In *ICCV Workshops*, 3189–3197.

Wang, H., and Wang, L. 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *e Conference on Computer Vision and Pa ern Recognition (CVPR)*.

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 1290–1297. IEEE.

Wang, P.; Li, Z.; Hou, Y.; and Li, W. 2016. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 2016 ACM on Multimedia Conference*, 102–106. ACM.

Wang, H.; Wang, P.; Song, Z.; and Li, W. 2017. Large-scale multimodal gesture recognition using heterogeneous networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3129–3137.

Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; and Escalera, S. 2018. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*.

Wang, H.; Wang, W.; and Wang, L. 2015. Hierarchical motion evolution for action recognition. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, 574–578. IEEE.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4724–4732.

Weng, J.; Liu, M.; Jiang, X.; and Yuan, J. 2018. Deformable pose traversal convolution for 3d action and gesture recognition. In *Europen Conference on Computer Vision (ECCV)*, volume 2, 6.

Xia, L.; Chen, C.-C.; and Aggarwal, J. K. 2012. View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, 20–27. IEEE.

Yang, W.; Jin, L.; Xie, Z.; and Feng, Z. 2015. Improved deep convolutional neural network for online handwritten chinese character recognition using domain-specific knowledge. *arXiv preprint arXiv:1505.07675*.

Yang, W.; Jin, L.; Ni, H.; and Lyons, T. 2016. Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 4083–4088. IEEE.

Yang, W.; Lyons, T.; Ni, H.; Schmid, C.; Jin, L.; and Chang, J. 2017. Leveraging the path signature for skeleton-based human action recognition. *arXiv preprint arXiv:1707.03993*.

Yang, W.; Jin, L.; and Liu, M. 2015. Chinese character-level writer identification using path signature feature, dropstroke and deep cnn. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 546–550. IEEE.

Yang, W.; Jin, L.; and Liu, M. 2016. Deepwriterid: An end-to-end online text-independent writer identification system. *IEEE Intelligent Systems* 31(2):45–53.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X.; et al. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, volume 2, 6.