Developing the Path Signature Methodology and Its Application to Landmark-Based Human Action Recognition

Weixin Yang, Terry Lyons, Hao Ni, Cordelia Schmid, Lianwen Jin

Abstract Landmark-based human action recognition in videos is a challenging task in computer vision. One key step is to design a generic approach that generates discriminative features for the spatial structure and temporal dynamics. To this end, we regard the evolving landmark data as a high-dimensional path and apply path signature techniques to provide an expressive, robust, non-linear, and interpretable representation for the sequential events. We do not extract signature features from the raw path, rather we propose path disintegrations and path transformations as preprocessing steps. Path disintegrations turn a high-dimensional path linearly into a collection of lower-dimensional paths; some of these paths are in pose space while others are defined over a multi-scale collection of temporal intervals. Path transformations decorate the paths with additional coordinates in standard ways to allow the truncated signatures of transformed paths to expose additional features. For spatial representation, we apply the non-linear signature transform to vectorize the paths that arise out of pose disintegration, and for temporal representation, we apply it again to describe this evolving vectorization. Finally, all the features are joined together to constitute the input vector of a linear single-hidden-layer fully-connected network for classification. Experimental results on four diverse datasets demonstrated

Weixin Yang
Mathematical Institute, University of Oxford, UK, e-mail: weixin.yang@maths.ox.ac.uk

Terry Lyons
Mathematical Institute, University of Oxford, UK and Alan Turing Institute, UK, e-mail: tlyons@maths.ox.ac.uk

Hao Ni
Dept. of Mathematics, University College London and Alan Turing Institute, UK, e-mail: ucahhni@ucl.ac.uk

Cordelia Schmid
Inria, France, e-mail: cordelia.schmid@inria.fr

Lianwen Jin
College of Electronic and Information Engineering, South China University of Technology, China, e-mail: lianwen.jin@scut.edu.cn

G. Yin, T. Zariphopoulou (eds.), *Stochastic Analysis, Filtering, and Stochastic Optimization*,
https://doi.org/10.1007/978-3-030-98519-6_18

that the proposed feature set with only a linear shallow network is effective and achieves comparable state-of-the-art results to the advanced deep networks, and meanwhile, is capable of interpretation.

1 Introduction

Human action recognition (HAR) is one of the most challenging tasks in computer vision with a wide range of applications, such as human-computer interaction, video surveillance, behavioral analysis, etc. A vast literature has been devoted to this task in recent years, among which are some informative surveys [1, 2, 3, 4, 5, 6, 7, 8]. An attractive option of HAR is Landmark-based HAR (LHAR) where the object is regarded as a system of correlated labelled landmarks. Johansson's classic moving light-spots experiment [9] demonstrated that people can detect motion patterns and recognize actions from several bright spots distributed on the body, which has stimulated research on pose estimation and LHAR [10, 11, 12]. Different from skeleton-based HAR (SHAR), LHAR, using no knowledge of skeletal structure, is flexible to extend to any landmark data streams with no explicit physical structures, e.g. traffic or people flow.

Although many solutions have been proposed to address the challenge of LHAR, the problem remains unsolved due to two main challenges. First, there is the problem of designing reliable discriminative features for spatial structural representation, and second of modelling the temporal dynamics of motion. In this paper, the path signature feature (PSF) is used and refined as an expressive, robust, non-linear, and interpretable feature set for spatial and temporal representation of LHAR.

The path signature, which was initially introduced in rough paths theory as a branch of stochastic analysis, has been successfully applied to many machine learning tasks. Most existing work can be devided into two categories: sliding-window-based and global-based. In the sliding temporal window approach [13, 14, 15, 16, 17, 18, 19], signatures of small paths are extracted and embedded into multi-channel feature maps as input of a CNN. The signatures herein are merely local descriptors from which the deep models are then trained to learn hierarchical representation. The global-based approaches combine all the cues into a high-dimensional path to compute high-level signatures over the whole time interval [20, 21] or low-level signatures over hierarchical intervals [22, 23]. They are straightforward but not efficient for high dimensional or spatio-temporal data.

To represent spatial pose, most methods [12, 24, 25, 26, 27, 28, 29, 30] used predefined skeletal structures. The connections distributed on a physical body are intuitive spatial constraints but not necessarily the crucial ones to distinguish actions. The connections discarded by imposing a skeletal structure could contain valuable non-local information. To solve this, hand-designed features [31, 32, 33, 34] were employed, but they are limited to encode non-linear dependencies. In this paper, we propose to localize a pose by disintegration into a collection of m-node sub-

paths. The signatures of these paths encode non-local and non-linear geometrical dependencies.

To model temporal dynamics, hand-designed local descriptors [31, 34] were popular, but it is difficult to encode complex spatio-temporal dependences in these. Recently, recurrent neural networks (RNN) [35], especially long short-term memory (LSTM) [36], have gained increasing popularity in handling sequential data, including human actions [37, 24, 38, 39]. In particular, a variation of LSTM [40, 25] succeeded in simultaneously exploring both spatial and temporal information. These deep models play a vital role in feature representation and achieve state-of-the-art performance, but the features learned by them are not as interpretable as hand-designed features. In this paper our temporal disintegration turns the original paths into hierarchical paths, from which the signatures encode multi-scale dynamical dependencies. Moreover, our path transformations decorates the paths with additional coordinates to allow signatures to expose additional valuable features.

To build the spatial and temporal representation, in each frame the spatial PSFs are extracted from the localized paths obtained by pose disintegration. In the clip, the evolution of each spatial feature along the time axis constitutes a spatio-temporal path. After path transformations and temporal disintegration, the temporal PSFs are then extracted from the spatio-temporal paths. Finally, the concatenation of all the features forms the input vector of a linear single-hidden-layer fully-connected network for classification. To extensively evaluate the effectiveness and flexibility of our method, several datasets (i.e., JHMDB [31], SBU [41], Berkeley MHAD [42], and NTURGB+D [39]) collected by different acquisition devices were used for experiments. Using our feature set and only a linear shallow net, we achieve comparable results to the advanced deep learning methods. Moreover, we took a further step toward understanding human actions by analyzing the PSFs and the linear classifier.

Our major contributions lie in four aspects:

1. PSFs are adopted and refined for LHAR with interpretations, proofs, experiments, and discussions of their properties and advantages.

2. Pose disintegration is proposed for non-local spatial dependencies, and temporal disintegration is proposed for multiscale temporal dependencies.

3. Path transformations, decorating the original paths with additional coordinates, are proposed to allow signatures to expose additional features.

4. Using signature-based spatio-temporal representation and only a linear shallow net, we achieve comparable state-of-the-art results to those with deep models. Meanwhile, this interpretable pipeline facilitates the understanding of HAR.

The authors are delighted to dedicate this paper to Mark H. A. Davis for many personal and professional reasons. Mark was wonderfully supportive friend. He was also an adventurous innovator who took mathematical ideas deep into commercial finance. In some sense this paper represents a similar pioneering spirit. It has a long history, and is the first effort to introduce path signature to the central area of action analysis and understanding in computer vision. This stream of research, as we report here, has developed these ideas into a viable methodology for analyzing evolving landmark style data in contexts where the datasets are too small to build effective

deep learning approaches. We hope that by consolidating it here, we will recognize Mark with a paper he would have supported and approved of.

2 Related work

2.1 Path signature feature (PSF)

Rough path theory is concerned with capturing and making precise the interactions between highly oscillatory and non-linear systems [43]. The essential object in rough path theory, called the path signature, was first studied by Chen [44] whose work concentrates on piecewise regular paths. More recently, the path signature has been used by Lyons [45] to make sense of the solution to differential equations driven by very rough signals. It was extended by Lyons' theory from paths of bounded variation [45] to rough paths of finite p-variation for any $p \geq 1$ [46].

Some successful applications of the PSF have been made in the fields of machine learning, pattern recognition and data analysis. First of all, the most notable applications of using PSFs is handwriting understanding. Diehl [21] used iterated integrals of a handwritten curve for recognition and found that some linear functions of the PSF satisfy rotation invariance. Graham [19] used the sliding-window-based PSF as feature maps of a CNN for large-scale online handwritten character recognition, based on which he won the ICDAR2013 competition [47]. Inspired by this, Xie et al. [15, 16] extended the method to handwritten text recognition. Yang et al. [17, 18] explored the higher-level terms of the PSF for text-independent writer identification which requires subtle geometric features. For financial data, useful predictions can be made with only a small number of truncated PSFs [20, 48]. The truncated signature kernel for hand movement classification was presented in [49], and was further extended to an untruncated version [50]. Moreover, PSFs were used on self-reported mood data to distinguish psychiatric disorders [23]. In [51], path signature transform was applied to describe the behaviour of controlled differential equations for modelling temporal dynamics of irregular time series. To model topological data, a novel path signature feature based on the barcodes arising from persistent homology theory was proposed for classification tasks [52]. These applications demonstrate the value of the PSF as an effective and informative feature representation.

The paper has been a long time in development, and the preprints [53] on the ArXiv have already influenced other developments. To name a few, in [54, 55, 56], the extraction of the path signature feature was treated as a flexible intermediate layer in various end-to-end network architectures like CNNs, LSTMs, or Transformer Networks. Also, variants of our proposed feature set were successfully applied to tasks like Arabic handwriting recognition [57], writer identification [58], personal signature verification [59], sketch recognition [60], action/gesture recognition [61, 62], speech emotion recognition [63], etc., showing its generalization ability. The proposed invisibility-reset transformation was further analyzed in [64].

2.2 Landmark-based human action recognition

A human body can be regarded as an articulated system composed of joints that evolve in time [65]. For recent surveys of LHAR, we refer the reader to [8, 66, 67, 68].

Approaches based on hand-designed features for LHAR can be categorized into two classes: joint-based and part-based. The joint-based ones regard the human body as a set of points and attempt to capture the correlation among body joints by using the motion of 3D points [69], measuring the pairwise distances [31, 70, 26, 33, 34], or using the joint orientations [71]. On the other hand, the part-based approaches focus on connected segments of the human skeleton. They group the body into several parts and encode these parts separately [27, 28, 72, 73, 74, 29, 75]. Some methods in this category represent a pose by means of the geometric relations among body parts, for examples, [27, 28] employed quadruples of joints to form a new coordinate system for representation, and [12] considered measurements of the geometric transformation from one body part to another. Some methods assume that certain actions are usually associated with a subset of body parts, so they aim to identify and use the subsets of the most discriminative parts of the joints.

Given the recent success of deep learning frameworks, some works aim to capture correlation among joint positions using CNNs [76, 77, 78, 79]. In [76], the input feature maps of a CNN were joints colored according to their sequential orders, body parts, or velocity, while in [77] and [78], the CNN's inputs were the concatenation of hand-designed local features. Since human actions are usually recorded as video sequences, it is natural to apply RNNs or LSTMs. HBRNN [24] and Part-aware LSTM [39] contained multiple networks for different groups of joints. Zhu et al. [37] proposed a deep LSTM to learn the co-occurrence of discriminative joints using a mixed-norm regularization term in the cost function. By additional new gating to the LSTM, the Differential LSTM [38] is able to discover the salient motion patterns, and [40, 25] achieved robustness to noise. It is noteworthy that the spatio-temporal RNNs in [40, 25] concurrently encoded both spatial and temporal context of actions within a LSTM. Liu et al. [80] used an attention-based LSTM to iteratively select informative keypoints for recognition. Zhang et al. [81] used a multilayer LSTM to fuse several simple geometric features for recognition. By taking advantage of the graph structure of human skeleton, Graph Convolutional Networks (GCNs) were introduced into the action recognition task. Yan et al. [30] used spatial graph convolutions along with interleaving temporal convolutions. Concurrently, Li et al. [82] proposed a similar approach but introduced a multi-scale module for spatio-temporal modelling. DGNN [83] represented the skeleton as a directed acyclic graph to encode both joint and bone information. MV-IGNET [84] extracted multi-level spatial features and leveraged different skeleton topologies as multi-views to generate complementary action features. MMDGCN [85] proposed a dense graph convolution for local dependencies and used spatial-temporal attention module to reduce the redundancy. These deep learning methods achieved high accuracy on most large-scale action datasets, but they often require a lot of training data and suffer from a lack of interpretability.

3 Path Signature

3.1 Definition and geometric interpretation

The rigorous introduction of the path signature as a faithful description or feature set for unparameterized paths can be found in [43, 86, 87, 88], so in this paper we present it in a practical manner.

A d-dimensional path or stream of timestamped events P over the time interval $[0,T] \subset \mathbb{R}$ can be represented as a continuous map $P : [0,T] \to \mathbb{R}^d$. The coordinates of P at time τ are $P_\tau = \left(P_\tau^1, P_\tau^2, \ldots, P_\tau^d\right)$. To illustrate the idea, we consider the simplest case when $d = 1$. The path is a real-valued path for which the path integral is defined as

$$S(P)_{0,T}^1 = \int_{0<t\leq T} dP_\tau^1 = P_T^1 - P_0^1, \tag{1}$$

which is the increment of this 1-dimensional path over the whole time interval and is called the 1-fold iterated integral. We emphasize that $S(P)_{0,\tau}^1, 0 < \tau \leq T$ is also a real valued path w.r.t τ. The 2-fold iterated integral is

$$S(P)_{0,T}^{11} = \int_{0<\tau_2\leq T} S(P)_{0,\tau_2}^1 dP_{\tau_2}^1 = \frac{1}{2}\left(P_T^1 - P_0^1\right)^2, \tag{2}$$

which is proportional to the square of the increment. Again, $S(P)_{0,\tau}^{11}$ is a real-valued path, so if we continue recursively, the k-fold iterated integral of P is

$$\begin{aligned} S(P)_{0,T}^{11\ldots1} &= \int_{0<\tau_k\leq T} \cdots \int_{0<\tau_2\leq\tau_3} \int_{0<\tau_1\leq t_2} dP_{\tau_1}^1 dP_{\tau_2}^1 \ldots dP_{\tau_k}^1 \\ &= \frac{1}{k!}\left(P_T^1 - P_0^1\right)^k, \end{aligned} \tag{3}$$

which is proportional to the increment to the power of k.

Now, when $d = 2$, the 1-fold iterated integral of the path $\left\{P_\tau^1, P_\tau^2\right\}$ has 2 elements

$$S(P)_{0,T}^1 = \int_{0<t\leq T} dP_\tau^1 = P_T^1 - P_0^1, \tag{4}$$

$$S(P)_{0,T}^2 = \int_{0<t\leq T} dP_\tau^2 = P_T^2 - P_0^2. \tag{5}$$

Each element is the increment of the path on the corresponding axis over the time interval $[0,T]$. They denote the displacement of the given path. The 2-fold iterated integral of this 2D path contains $d^2 = 2^2$ elements

$$S(P)^{11}_{0,T} = \int_{0<\tau_2\leq T}\int_{0<\tau_1\leq t_2} dP^1_{\tau_1} dP^1_{t_2} = \frac{1}{2!}\left(P^1_T - P^1_0\right)^2, \tag{6}$$

$$S(P)^{22}_{0,T} = \int_{0<\tau_2\leq T}\int_{0<\tau_1\leq \tau_2} dP^2_{\tau_1} dP^2_{\tau_2} = \frac{1}{2!}\left(P^2_T - P^2_0\right)^2, \tag{7}$$

$$S(P)^{12}_{0,T} = \int_{0<\tau_2\leq T}\int_{0<\tau_1\leq \tau_2} dP^1_{\tau_1} dP^2_{\tau_2}, \tag{8}$$

$$S(P)^{21}_{0,T} = \int_{0<\tau_2\leq T}\int_{0<t_1\leq t_2} dP^2_{\tau_1} dP^1_{\tau_2}. \tag{9}$$

We note that the first two elements are the same as (2) in the 1-dimensional case. For the other two elements, the geometric intuitions are the areas shown in Fig. 1(a) and Fig. 1(b). Together they represent the Lévy area [86] depicted in Fig. 1(c). The Lévy area, which is a signed area enclosed by the path and the chord connecting the endpoints, can be expressed by

$$A_{0,T} = S(P)^{12}_{0,T} - S(P)^{21}_{0,T}. \tag{10}$$

The sign of the area depends on the sign of the winding number of the path moving around it [89]. The interpretation of the k-fold iterated integral ($k > 2$) of a 2D path is not trivial, so it is not included here. By analogy, for a 3D path, the 1-fold, 2-fold, and 3-fold iterated integrals are units of displacement, area, and volume respectively.

In general, for a path in $\mathbb{R}^d$, the superscript of the k-fold iterated integral, which describes the order of integration, is a multi-index $(i_1, i_2, \ldots, i_k) \in \{1, \ldots, d\}^k$. Therefore, the d^k elements of the k-fold iterated integral of a d-dimensional path can be generally expressed as

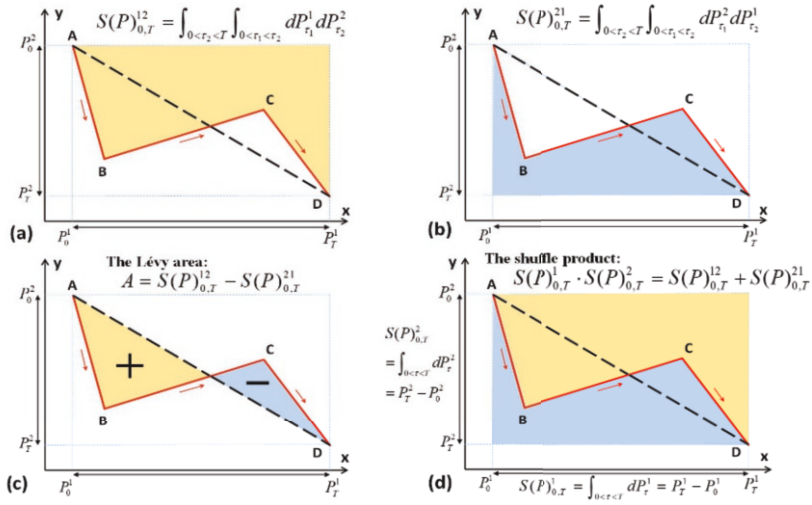

Fig. 1 The geometric intuition of the PSF of a 2D path. The path in red moves from A to D over the time interval [0, T]. The dashed line is the chord connecting the endpoints. Panels (a) and (b) depict two terms of the 2-fold iterated integrals of the path, (c) is the Lévy area enclosed by the path and its chord, and (d) is a demonstration of the shuffle product identity.

$$S(P)_{0,T}^{i_1,i_2,\ldots,i_k} = \int_{0<\tau_k\leq T} \cdots \int_{0<t_2\leq\tau_3} \int_{0<\tau_1\leq\tau_2} dP_{\tau_1}^{i_1} dP_{\tau_2}^{i_2} \ldots dP_{\tau_k}^{i_k}. \tag{11}$$

Then the signature of a path P over the time interval $[0, T]$ is the collection of all the iterated integrals of P:

$$\begin{aligned} S(P)_{0,T} = \Big(&1, S(P)_{0,T}^{1}, \ldots, S(P)_{0,T}^{d}, \\ &S(P)_{0,T}^{1,1}, \ldots, S(P)_{0,T}^{1,d}, S(P)_{0,T}^{2,1}, \ldots, S(P)_{0,T}^{d,1}, \ldots, S(P)_{0,T}^{d,d} \\ &\ldots, S(P)_{0,T}^{1,1,\ldots,1}, \ldots, S(P)_{0,T}^{i_1,i_2,\ldots,i_k}, \ldots, S(P)_{0,T}^{d,d,\ldots,d}, \ldots\Big), \end{aligned} \tag{12}$$

where the 0-th term is conventionally set to 1. Since the signature is defined on top of all the possible indices of finite length, the number of elements in the signature is infinite. In practical use we usually consider the signature truncated at a certain level n written as

$$S_n(P)_{0,T} = \left(1, S(P)_{0,T}^{1}, \ldots, S(P)_{0,T}^{i_1,i_2,\ldots,i_n}, \ldots, S(P)_{0,T}^{d,d,\ldots,d}\right) \tag{13}$$

of which the dimensionality is $\varphi(d, n) = \left(d^{n+1} - 1\right)(d - 1)^{-1}$. The elements of the truncated signature are taken as features (i.e., PSF) encoding the informative geometric properties of sequential data in applications in machine learning. For the feature set, the 0-th term (i.e., a constant value set to 1) is optional, so the dimension can be reduced to

$$\varphi'(d, n) = \left(d^{n+1} - d\right)(d - 1)^{-1}. \tag{14}$$

For the 1-dimensional case ($d = 1$), the feature dimension is exactly equal to n (excluded the 0-th term) according to (1), (2), and (3).

3.2 Calculation of the signature for a discrete path

Although the path signature is initially defined for continuous paths with bounded variation, it is easily extended to discrete paths by linear interpolation [90]. The signature is canonical and does not depend on the choice of timescale used for the interpolation.

Computing the signature of a piecewise linear path does not require integrals. For each line segment of the path, the elements of its signature are given by

$$S(P)_{\tau,t+1}^{i_1,i_2,\ldots,i_k} = \frac{1}{k!} \prod_{j=1}^{k} \left(P_{\tau+1}^{i_j} - P_{\tau}^{i_j}\right), \tag{15}$$

where $P_{\tau}^{i_j}$ is the i_j-th coordinate value of path P at time τ. For the entire path, Chen's identity [44] states that for any time stamps (s, t, u) satisfying that $s < t < u$, we have

$$S(P)_{s,u}^{i,i_2,\ldots,i_k,\ldots,i_n} = \sum_{k=0}^{n} S(P)_{s,t}^{i_1,i_2,\ldots,i_k} S(P)_{t,u}^{i_{k+1},i_{k+2},\ldots,i_n}. \tag{16}$$

This implies that the signature of the entire path can be calculated from the signatures of its pieces.

We recommend the three open-source python software libraries, *esig* (on PyPi), derived from the *CoRoPa* C++ library *libalgebra* [91], *iisignature* [92], and *Signatory* [93] which has a dependency on PyTorch and works well on the CPU as well as the GPU. They all allow fast computation of the path signature.

3.3 Properties of the path signature

3.3.1 Uniqueness

It is proved that the path signature determines a path if and only if the path is not tree-like (this notion is introduced in [45]). A tree-like path is a trajectory containing a section where the path exactly retraces itself. Tree-like paths are common in real-world data streams, for instance, in some human actions, especially periodic ones, like clapping or jumping in place. An effective way to avoid the tree-like situation is adding an extra monotone dimension, such as time, to the original path.

3.3.2 Invariance under translation

The signature computed by (11) or (15) is invariant under translation of the paths, which is a practical advantage and avoids complex recentering normalization.

3.3.3 Invariance under time reparameterization

A time reparameterization of a path is a continuous, nondecreasing substitution for the time variable of a path. It changes the speed of recording of the path. Human actions are largely invariant under changing the speed of the action or viewing speed of the video. The ease with which the signature can completely filter out these changes in the representation is a major advantage for machine learning, substantially reducing the dimensionality of the feature set needed for action classification. The use of the path signature, with its fixed-dimensional feature set, can help the classifer to recognize the same action performed or sampled at different speeds. We refer the reader to [43, 88] for a detailed proof of the invariance of the path signature under time reparameterization.

3.3.4 Nonlinearity of the signature

The shuffle product identity [86] states that the product of two signatures of lower level can be expressed as a linear combination of some higher-level terms. For instance, for the two-dimensional case in section 3.1, we can easily derive the following equation from Fig. 1(d),

$$S(P)^1_{0,T} \cdot S(P)^2_{0,T} = S(P)^{12}_{0,T} + S(P)^{21}_{0,T}. \tag{17}$$

In other words, the nonlinear behavior in terms of lower level terms can be expressed by linear combination of higher-level terms. Therefore, when we incorporate the higher-level terms into the feature representation, we automatically include more nonlinear prior knowledge in our feature set. If the introduced nonlinearity is sufficient, we need only linear classifiers to distinguish the targets.

3.3.5 Fixed dimension under length variations

Another practical property of the path signature is that the dimension of the PSF extracted from the entire path depends on the truncation level of the signature and the intrinsic dimension of the path but is independent of the (sampled) length of the path, as described in 14. For human action recognition, the durations of actions are variable. The use of the path signature allows us to extract a fixed dimension of features and use them with classification methods which require a fixed-length input.

4 Path disintegrations and transformations

The principled and robust representation of unparameterized paths, along with the convenience of reducing polynomial functions on the space of paths to linear ones (which establishes their universality) provide the core motivations for using signatures as features. One can always take the signature of a raw path to remove any dependence on parameterization or translation, but sometimes it is prudent to apply path disintegrations or path transformations as preprocessing to improve the efficiency and effectiveness of PSFs. The disintegrations turn a path into a composition of subpaths while the transformations turn a path into a higher-dimensional path.

4.1 Path disintegrations

4.1.1 Pose Disintegration

In many cases, non-local clues are informative and straightforward, for instance, the non-local displacement between two hand points is a key feature for the action of clapping. To exploit both local and non-local clues in pose, we propose pose disintegration. Landmarks that are labelled with corresponding body parts have no inherent order, so a predefined priority order is randomly chosen and fixed – different random choices of initial order yield comparable results in preliminary experiments. The pose is then regarded as an ordered collection of points in $\mathbb{R}^d$. Our pose disintegration localizes the pose into all possible subposes containing m points. Connecting the m points in each subpose in the inherited order forms a unique m-node sub-path that visits each point once. We end up with a collection of sub-paths which do not need to be parts of physical body and are available for further path transformations or signature extractions.

We consider that functions on a pose can be approximated by functions on the piecewise linear localized paths of its subposes. For convenience, one can view these functions as linear functions on the signature of its localized paths. The terms of the first two levels of signatures cover the displacement and the area information similar to the traditional hand-designed features [31, 34], while the higher-level terms capture more non-linear features. For a pose with N joints, the dimension of the signatures of its localized paths is $C_N^m \cdot \varphi'(d, n)$, where m is the number of points in a subpose, d is the dimension of the path, and n is the truncated signature level. The selections of these parameter values are highly correlated and associated with the uniqueness of the paths. According to [94], any piecewise linear paths in $\mathbb{R}^d$, consisting of at most $m = d + 1$ points, can be uniquely recovered from the signature at the third level. A larger m brings semantically high-level components but requires a larger n for the path uniqueness [95], which exponentially increases the feature dimension according to 14, and means less shareability and more sub-paths. The number of m-node subpaths is in line with Pascal's triangle and increases along with $m(m \geq N/2)$. To avoid feature set of very large dimension, $m \leq 3, n = 3$ for $d = 2$ and $m \leq 4, n = 3$ for $d = 3$ are suggested. Beyond the signature level n required for the unique recovery of a path, the non-linearity (as described in 3.3.4) of the extra high-level terms may still contribute to facilitate the training of the model until the dimensionality of the feature set becomes impractical.

4.1.2 Temporal Disintegration

Temporal disintegration is based on the basic theory of the path signature which suggests that low-level terms of signatures on all intermediate length time intervals can be far more efficient than signatures of high levels over the whole time interval [86]. Therefore, instead of extracting the PSF over the whole time interval, the dyadic path signature features (DPSF) [22] split the entire interval into small intervals with

a dyadic hierarchical structure and then extracts PSF over each small interval. Given a path over the whole time interval $[0, T] \subset \mathbb{R}$ the j-th dyadic level of the hierarchical structure is the collection of subintervals $\left[iT/2^j, (i+1)T/2^j\right], i \in \left[0, 2^j - 1\right], j \in \mathbb{N}$. Note that the 0-th dyadic level contains exactly the whole path. The DPSF over long, medium, and short time intervals describes multi-scale dynamical dependences more efficiently than the PSF over the entire interval, which requires higher-level terms to capture local dependencies.

Slightly different from the hierarchical structure in [22] which may break the events that occur near the conjunctive time stamps $\left\{iT/2^j \mid i \in \left[1, 2^j - 1\right], j \in \mathbb{N}^+\right\}$, we consider an overlapping version over the time intervals $\left[iT/2^{j+1}, (i+2)T/2^{j+1}\right]$, $i \in \left[0, 2 \cdot \left(2^j - 1\right)\right]$. The overlapping DPSF is expected to supplement the original DPSF with additional local details. Its dimension is

$$\hat{\varphi}(h, d, n) = \left(2^{h+1} - h - 2\right) \cdot \varphi'(d, n), \tag{18}$$

where $h \in \mathbb{N}^+$ is the number of the hierarchical level. The selection of h is a tradeoff between improving efficiency and introducing local noises over finer intervals.

4.2 Path transformations

4.2.1 Time-incorporated transformation

The signature is invariant under parameterization, but in many situations, one would like to keep the dependence on time. Adding a monotone increasing time dimension is adopted to encode motion speed. The signature of a time-incorporated path contains two parts: time-independent (TI) and time-dependent (TD). The TI part is exactly the signature of the original path, so its integration order is

$$i_1, i_2, \ldots, i_k \in \{1, \ldots, d\}. \tag{19}$$

The TD part is related with time. Its integration order is

$$i_1, i_2, \ldots, i_k \in \{1, \ldots, d+1\}, \exists m \in [1, k], i_m = d+1, \tag{20}$$

which means each term of the signature in TD is an integral along the time dimension at least once. Given the truncated signature level n, the dimensionality of the TD part is $\varphi'(d+1, n) - \varphi'(d, n)$. The signature of the original path filters out the information about the speed of motion and the sampling rate but the signature of the time-incorporated path allows us to select one and suppress the other according to significance to the classification.

4.2.2 Invisibility-reset transformation

The signature capturing relative position information is invariant under translation. Given that the absolute position may be essential for some scenarios (e.g., HAR under static CCTVs), we propose the invisibility-reset transformation of a path to retain the absolute position information in signatures. For a path P in $\mathbb{R}^d$ within the interval $[0, T]$, we add two time steps T+1 and T+2 with value P_T and 0 respectively at the end of P and add a visibility dimension v with values 1 in $[0, T]$ and 0 in $(T, T + 2]$. In other words, the invisibility-reset path P_{IR} in $\mathbb{R}^{d+1}$ first becomes invisible at time T+1 and then is reset to the origin at T+2. According to (15) and (16), we have

$$S(P_{IR})^{i_1,i,\cdots,v,v}_{0,T+2} = -S(P)^{i,i_2\cdots i_k}_{0,T}, i_1, i_2, \ldots, i_k \in \{1, \ldots, d\} \tag{21}$$

which means certain terms in $S(P_{IR})$ encode the relative positions as in $S(P)$. Moreover, the terms of the first level of $S(P_{IR})$, given by

$$S(P_{IR})^{i_1}_{0,T+2} = -P^{i_1}_0, i_1 \in \{1, \ldots, d\}, \tag{22}$$

are the absolute position of the initial point. This simple transformation retains different position information in signatures and thus allows the network to select one and suppress the other according to significance to the task.

4.2.3 Multi-delayed lead-lag transformation

The lead-lag transformation proposed in [20, 87, 90] is designed to exploit the quadratic cross-variation between the original path and its delayed path. To extend its capability to describe long-term dependencies of sequential events, our modified lead-lag transformation, as shown in Fig. 2, is constructed by the original path and its multiple delayed paths (instead of one delayed path in [20]). We denote the dimension of a lead-lag transformed path as d_{LLT}. The signatures of lead-lag paths with smaller d_{LLT} encode short-term dependencies, while those with larger d_{LLT} explore more long-term dependencies.

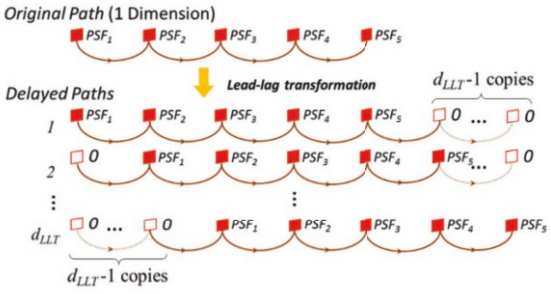

Fig. 2 The illustration of multi-delayed lead-lag transformation. The dimen-sion of lead-lag paths is d_{LLT}. The delayed paths are padded with zeros to ensure a fix length for each dimension.

Table 1 Proposed features for LHAR

# of joints	Spatial structural features (for each frame)	Temporal dynamical features (along the time axis)
1 (a single joint)	S-J: The d-dimensional coordinates of each of the predefined N joints are used.	T-J-PSF: The temporal PSF over the evolution of each joint up to signature level n_{TJ} is extracted.
2 (joint pair)	S-P-PSF: The PSF over each pair of joint up to signature level n_{SP} is extracted.	T-S-PSF: The evolution of each dimension of spatial PSF is treated as a path, over which the temporal PSF up to signature level n_{TS} is extracted.
3 (joint triple)	S-T-PSF: The PSF over each triple of joint up to signature level n_{ST} is extracted.	

5 Feature extraction for human action recognition

Our proposed feature set for LHAR, which we describe in this section, is outlined in Table 1. We note that an unofficial Python implementation of the feature set is available on GitHub [96].

5.1 Spatial structural features

First of all, the basic information describing the spatial structure is the d-dimensional coordinates of each of the N joints of the body. The keyword **S-J** denotes the spatial coordinate values of the joints. The dimension of this part is $D_{SJ} = N \cdot d$ for each sampled frame.

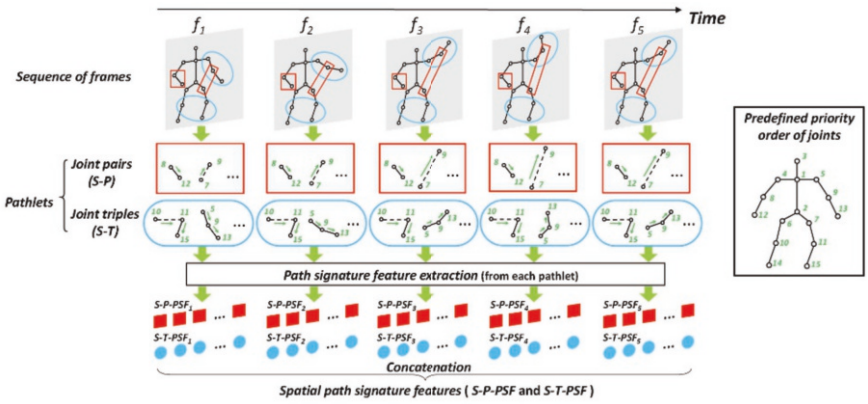

Fig. 3 The illustration of spatial feature (**S-P-PSF** and **S-T-PSF**) extraction. Note that we predefine the priority order of all the N joints (N = 15 in this figure). The red quadrangles denote the feature extraction of joint pairs, while the blue ellipses denote that of joint triples. All possible pairs and triples of joints are considered.

To encode the spatial context we use pose disintegration with $m = 2$ and $m = 3$, which means joint pairs and joint triples are used as illustrated in Fig. 3. The signatures of each of these subpaths are computed to model the spatial constraints in each frame. The spatial PSF of joint pairs and joint triples are denoted by **S-P-PSF** and **ST-PSF** respectively. If the truncation level of the signature of pairs and triples are n_{SP} and n_{ST} respectively, then the feature dimensions per frame are $D_{SP} = C_N^2 \cdot \varphi'(d, n_{SP})$ and $D_{ST} = C_N^3 \cdot \varphi'(d, n_{ST})$ respectively. Finally, the spatial features from all sampled frames are extracted and concatenated. The dimension of **S-P-PSF** and **S-T-PSF** per frame is denoted by $D_S = D_{SP} + D_{ST}$.

5.2 Temporal dynamical features

Inspired by the works in [40, 25] which jointly learned the spatial and temporal contexts in a variant of LSTM, we suggest that the dynamics of landmark-based human action can be described by the evolution of spatial context. The spatial context herein are the features we extracted in section 5.1, although other spatial features can be used alternatively. From these, we are going to extract two kinds of temporal features **T-J-PSF** and **T-S-PSF**.

The **T-J-PSF**, illustrated in Fig. 4, is the temporal PSF of the evolution of each joint along the time. The evolution of each joint is naturally a time-sequence, so we consider its time-incorporated transformation. For N-joint bodies in $\mathbb{R}^d$, the dimension of **T-J-PSF** is $D_{TJ} = N \cdot \varphi'(d+1, n_{TJ})$, where n_{TJ} is the truncation level of the signature.

Since each dimension of the spatial contextual features (**S-P-PSF** and **S-T-PSF**) characterizes one particular spatial constraint of a pose, the evolution of this spatial constraint along the time forms a spatio-temporal path which delivers temporal constraints of a motion. The temporal PSF of these spatio-temporal paths is denoted by **T-S-PSF** and illustrated in Fig. 5. Since the signature of a spatiotemporal path (i.e., a 1D path) is just the increments to a certain power, the multi-delayed lead-lag transformation is applied to each path to enrich the PSF with cross variations among

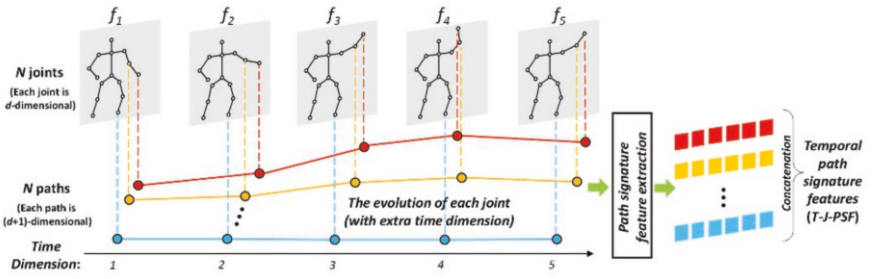

Fig. 4 Illustration of temporal features extracted from the evolution of each corresponding joint (**T-J-PSF**).

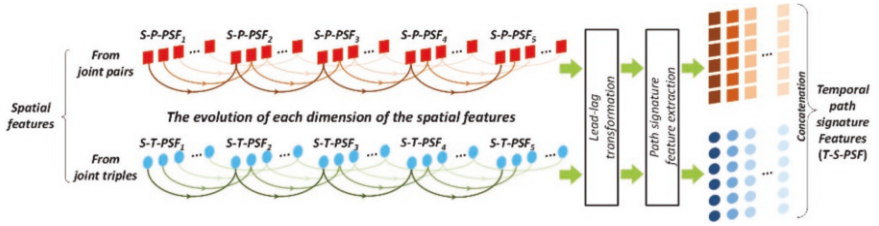

Fig. 5 Illustration of temporal features extracted from the evolution of spatial context (**T-S-PSF**). Each dimension of the spatial features is treated equally and individually.

events of the path. If the truncation level of the signature is n_{TS} and the dimension of the lead-lag paths is d_{LLT}, the dimension of **T-S-PSF** from all spatio-temporal paths is $D_{TS} = D_S \cdot \varphi'(d_{LLT}, n_{TS})$. Considering there might exist complicated or long-range actions, the temporal disintegration in section 4.1.2 can be applied. If so, the dimensions are $D_{TJ} = N \cdot \hat{\varphi}(h_{TJ}, d+1, n_{TJ})$ and $D_{TS} = D_S \cdot \hat{\varphi}(h_{TS}, d_{LLT}, n_{TS})$ where h_{TJ} and h_{TS} are the corresponding hierarchical levels.

The dimension of all temporal PSFs is $D_T = D_{TJ} + D_{TS}$. Finally, the total dimension of spatial and temporal features per clip is $D = M \cdot (D_{SJ} + D_S) + D_T$, where M is the number of sampled frames. Moreover, the spatial components can be covered by the temporal PSFs extracted from invisibility-reset paths which require no sampling step.

6 Experimental results and analysis

6.1 Datasets

Monocular videos recorded by 2D cameras are capable of collecting spontaneous actions, but their sensitivity to viewpoint variations and occlusions makes recognition a difficult task [1]. Intuitively, human body is general in 3D space, so marker-based motion capture systems [97] were designed to collect highly accurate locations of human joints. However, they are often expensive and impractical for recording realistic action videos. Fortunately, costeffective depth cameras (e.g. Kinect sensor [98]) were designed to provide reliable joint locations via real-time pose estimation algorithms (e.g., [99]). Our method is general enough to be applied to various kinds of data. To extensively evaluate the proposed methods, we conducted experiments on four datasets containing examples of all three types of data: JHMDB [31], SBU [41], Berkeley MHAD [42], and NTURGB+D [39]. The information we used herein for action recognition is the locations of landmarks in all the frames. However, it is worth noting that our method is flexible and additional information such as visibility state or confidence score can be included.

The JHMDB dataset [31] is a 2D human action dataset. There are 928 clips, each clip containing between 15 and 40 frames. A clip captures only one person doing one of 21 actions. The 2D joint positions are manually annotated. There are 3 splits separating the whole dataset into training and testing set. The final result is the average of them. The sub-JHMDB is a subset of JHMDB with the full body inside the frame. The challenges are the spontaneity of the actions captured by the clips from YouTube and the loss of information due to 2D projection.

The SBU Interaction [41] is a 3D Kinect-based dataset. It has 282 clips categorized into 8 classes of two-actor interactions, and has 30 joints per frame. Its depth information suffers from self-occlusion, causing measurement errors in the estimated joint locations.

The third dataset is Berkeley MHAD dataset [42] captured by a marker-based motion capture system. It consists of 659 clips, of which 384 clips, performed by 7 actors, are used for training and 275 clips by 5 different actors are used for testing. The 3D locations of 43 joints captured by LED markers are very accurate.

The Kinect-based NTURGB+D [39] is one of the largest 3D action recognition datasets and contains 56 thousand clips of 60 classes. The large viewpoint variations and unconstrained number of actors pose considerable challenges for analysis of this dataset.

Note that the quantitative analysis was conducted on JHMDB, and all the parameters were determined by 5- fold cross validation on the training set of the first split.

6.2 Network configurations

Since PSFs are rich non-linear features, we adopted a single-hidden-layer linear neural network as our classifier (1- layer net also works well in preliminary experiments). The network is fully-connected and the activation of the hidden neurons is the identity function. The input dimension is decided by the PSF (i.e., **S-P-PSF**, **S-T-PSF**, **T-J-PSF**, **T-S-PSF**, or some combinations of them) and the output is a probability distribution given by a softmax layer over all the class labels in a dataset. The single hidden layer has 64 neurons. The networks are trained by stochastic gradient descent on the cross-entropy with momentum 0.7 and mini-batch size 30. The learning rate updates in accordance to $\alpha(t) = \alpha(0) \cdot \exp(-\lambda t)$ where $\alpha(0) = 0.005, \quad \lambda = 0.005$. The maximum epoch is 300 for all experiments.

Dropconnect [100], a generalization of Dropout [101], randomly omits a proportion of connections at each training iteration. It is applied to the connections between the input and the single hidden layer for regularization. A high ratio of Dropconnect is essential to mitigate overfitting because the features herein are of very high dimension. Additionally, since the actions of some joints are highly correlated with each other, a small proportion of joints or features may already be sufficient to distinguish some actions. Based on our preliminary experiments, the Dropconnect rate is set to 0.95.

6.3 Data preprocessing and benchmark

We used two kinds of data augmentation. One is horizontal flipping, and the other one is adding Gaussian noise (inspired by [31]) over joint coordinates to simulate the noisy positions caused by estimation or annotation.

To cope with translation variation, we normalized the joints from world coordinate system to person-centric coordinate system by placing the center point of the body at the origin. To compensate for the biometric differences, we normalized the coordinate values to the range of [-1,1] over the entire clip. For feature normalization, each feature was divided by the maximum absolute value of the corresponding dimension and normalized to [-1, 1].

The spatial components (**S-J**, **S-P-PSF**, and **S-T-PSF**) are calculated for each frame. To obtain a fixed-length input to the network, we uniformly sampled M (in this paper, $M = 10$) frames from each clip. As the signature has a fixed dimension under length variation, the temporal features (**T-J-PSF** and **T-S-PSF**) are extracted from all the frames without subsampling. Our baseline method is using **S-J**, the d-dimensional coordinate values of all N joints. This leads to MNd-dimensional feature set, for which we obtained a validation error rate of 57.54 ± 3.26%.

6.4 Investigation of the spatial features

As described in section 4.1.1 and 5.1, by pose disintegration with $m = 2$ and $m = 3$, we constructed all the joint pairs and triples as localized paths for **S-P-PSF** and **S-TPSF** respectively. The error rates on the validation set obtained by these feature sets are shown in Table 2 and Table 3. The performance improves when higher terms of the signature are considered, but the improvements tend to be negligible and the variance increases when the dimension of the feature grows exponentially with the signature level n. For the joint pairs, a suitable truncation level n_{SP} is 2 or 3, while for the joint triples, the level n_{ST} needs to be as high as 3 or 4, which suggests the choice of n should depend on m. We refer the reader to [95] which discusses the relationship among m, n, and the path dimension d from the view of path recovery. For the following experiments, we chose to fix $n_{SP} = 2$, $n_{ST} = 3$.

Table 2 Effect of different signature levels on the performance of S-P-PSF

Type of subpaths	Signature level n_{SP}	Feature dim.	Error rate (%)
Joint Pairs	1	2100	32.79 ± 4.49
	2	6300	25.41 ± 4.55
	3	14700	**24.10 ± 5.65**
	4	31500	24.10 ± 5.72

Table 3 Effect of different signature levels on the performance of S-T-PSF

Type of subpaths	Signature level n_{ST}	Feature dim.	Error rate (%)
Joint Triples	1	9100	43.93 ± 2.87
	2	27300	32.46 ± 3.26
	3	63700	26.39 ± 3.99
	4	136500	**24.75 ± 4.79**
	5	282100	23.77 ± 6.41
	6	573300	25.24 ± 6.44

6.5 Investigation of the temporal features

6.5.1 Investigation of T-J-PSF

First, we investigated the effect of the time-incorporated transformation and the truncation level n_{TJ} of the **T-J-PSF**. As shown in Fig. 6, if the truncation level n_{TJ} (the horizontal axis) is 1, adding a time dimension (the green plot) only improves the performance a little. This is because the first level term related to the time dimension is only the duration of the action. When n_{TJ} increases, the performance improvements of using time-incorporated PSF are more significant, showing the effectiveness of the time-incorporated path transformation. As to the truncation level, when n_{TJ} increases, the results have lower bias together with gradually higher variance, so a trade-off is required. Here, $n_{TJ} = 5$ is a good choice.

In addition, we investigated the effect of the signature of the time-incorporated path at different frame rates. We artificially increased the frame rate by interpolating additional frames at random time stamps of the original clips. Bodies of the additional frames were copied from those of their adjacent frames. On the other hand, we decreased the frame rate by random subsampling. The networks were trained using the training clips at original frame rate (30fps) and tested 10 times using artificial validation clips at each of the frame rates ranging from 6fps to 90fps in 6fps steps. As shown in Fig. 7, when the frame rate increases from 30fps to 90fps, the error rates of using the time-independent part (TI) stay the same, while those of using

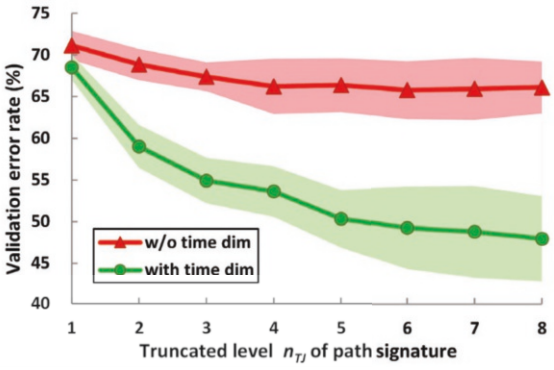

Fig. 6 Comparison of **T-J-PSF** w/ and w/o using time-incorporated paths. The colored areas are the error bands.

the time-dependent part (TD) raise rapidly. It demonstrates the TI (i.e. the signature of original path) is invariant under time reparameterization while the TD is very sensitive to speed variation. The larger the signature level n, the more sensitive the TD is to speed variation. Similarly, in the other direction, when the frame rate decreases from 30fps to 6fps, the influence to TD is far more significant than that to TI, showing the tolerance of TI under missing frames.

If we replace the PSF with the overlapping DPSF, then an appropriate hierarchical level h_{TJ} needs to be chosen. As shown in 8, in terms of performance, the low-level (e.g., $n_{TJ} = 2$) overlapping DPSFs over the hierarchical intervals (e.g., $h_{TJ} = 3$) often outperform the high-level (e.g., $n_{TJ} = 5$) PSFs over the whole interval ($h_{TJ} = 1$), which shows the efficiency of using temporal disintegration. However, when the disintegrated paths are too fragmented to avoid being dominated by local noises (e.g., when $h_{TJ} > 3$), the additional features are harmful. We thus fixed $h_{TJ} = 3$. Another observation is that the improvements from $h_{TJ} = 1$ to $h_{TJ} = 3$ become less significant along with the increasing n_{TJ}, demonstrating a trend that the high-level PSF and lowlevel DPSF yield similar information eventually.

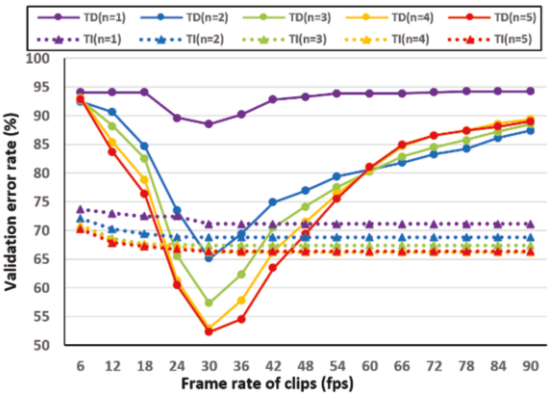

Fig. 7 Sensitivity of the time-dependent and time-independent part of the time-incorporated PSF to different frame rates.

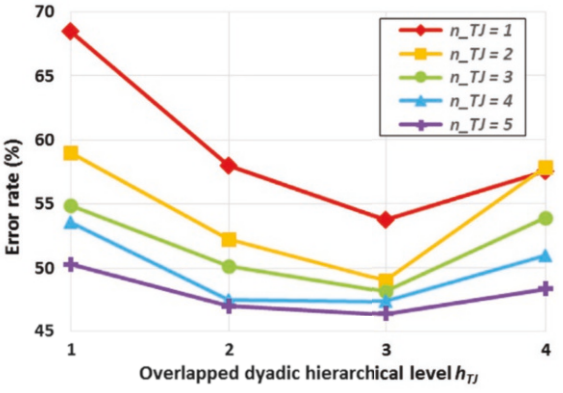

Fig. 8 Comparison of **T-J-PSF** with different dyadic hierarchical level h_{TJ} and different truncation level n_{TJ} of signature.

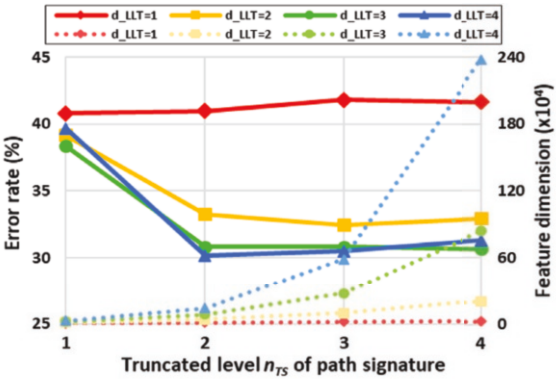

Fig. 9 Comparison of the error rates (solid) and the feature dimensions (dashed) of using **T-S-PSF** with different dimensions d_{LLT} of lead-lag paths and different truncation level n_{TS} of signature.

6.5.2 Investigation of T-S-PSF

Regarding the PSF derived from the evolution of the spatial context (**T-S-PSF**), two factors were evaluated: the dimension d_{LLT} of the lead-lag path and the truncation level n_{TS} of the signature. As shown in Fig. 9, the results improve when a higher dimension d_{LLT} of the lead-lag path is adopted, but the marginal improvement is less obvious when $d_{LLT} \geq 3$. For the truncation level n_{TS}, the improvements are significant from $n_{TS} = 1$ to $n_{TS} = 2$, but they are negligible when $n_{TS} > 2$. The dashed lines in Fig.9 show the trends of feature dimension under different parameters. By making a trade-off between model complexity and performance, we fixed $d_{LLT} = 3, n_{TS} = 2$.

By using the overlapping DPSF instead of PSF, the validation error rates are $30.82 \pm 7.00\%$, $26.07 \pm 6.12\%$, $26.39 \pm 5.51\%$, and $26.07 \pm 5.23\%$, when the hierarchical level h_{TS} is 1, 2, 3, and 4 respectively. Thus, we fixed h_{TS} to 3.

6.6 Ablation study

For the ablation study of our features on the JHMDB [31], we used the parameter setting for each feature based on the foregoing analysis. We retrained the network using the whole training set (including the validation set) and took the final result as the average of the three splits. The results are shown in Table 4. We can see that adding the spatial PSF (Ex. 4) to the baseline (Ex. 1) gives an improvement of about 20%, and further adding the temporal PSF (Ex. 9) contributes an additional 10%. The spatial context may be alternative between joint pairs and joint triples, for example Ex. 2 vs. Ex. 3, or Ex. 7 vs. Ex. 8, but they are complementary as shown in Ex. 4 and Ex. 9.

Applying the invisibility-reset transformation to all the paths before taking the temporal signatures allow us to remove all the spatial components **S-J**, **S-P-PSF**, and **S-TPSF**, while obtain the same accuracy as that of Ex. 9.

Also, we evaluated the method which directly takes all the evolving N landmarks in $\mathbb{R}^d$ as a Nd-dimensional path for signature extraction. Together with **S-J**, it achieves 55.0% in accuracy. The dimension of this PSF is $\varphi'(Nd, n) = 838,230$ when $n = 4$ and will be impractical when $n > 4$. This shows the cost-effectiveness of using pose and temporal disintegration.

Table 4 Effect of different signature levels on the performance of S-T-PSF

Ex.#	S-J	S-P-PSF	S-T-PSF	T-J-PSF	T-S-PSF (S-P)*	T-S-PSF (S-T)*	Accuracy (%)
1	o						48.9
2	o	o					68.4
3	o		o				68.7
4	o	o	o				69.2
5	o			o			62.0
6	o	o	o	o			73.5
7	o	o		o	o		79.1
8	o		o	o		o	78.3
9	o	o	o	o	o	o	80.4

∗ S-P (S-T) means the temporal features are only on the base of spatial joint pairs (joint triples).

6.7 Comparison with the state-of-the-art methods

To achieve our best results, we adopted the best settings of parameters from the foregoing analysis. For the JHMDB dataset [31], the results were given in the previous subsections. For the other three datasets, we followed the evaluation criteria in [40].

6.7.1 Comparison over small datasets

For the JHMDB dataset, previous state-of-the-art methods are high-level pose feature (HLPF) [31] and its modified version (i.e. Novel HLPF [34]), dense trajectory features [102] encoded by Fisher vectors [103], and the pose-based CNN features (P-CNN) [79]. As shown in Table 5, our method, which uses only the joint locations, achieve better performance than the P-CNN which requires additional RGB information. Further, our method manages the high degree of nonlinearity, and outperforms other methods using hand-designed features like HLPF. Also, the computation of our feature extraction is very fast. The average speed using *esig* [91] on a single thread of an Intel 2.4GHz Xeon Gold 6240R CPU is 85 fps on the JHMDB dataset.

Moreover, we used the off-the-shelf pose estimation called Alphapose (with Poseflow) [104] to get a set of 17 estimated joints from the RGB videos of the sub-JHMDB dataset, and then trained and tested the network using the estimated poses. By using only location information, our test accuracy is 68.2%, which outperforms that

of PCNN [79] (66.8%), PA-AP [105] (61.5%), JointAP [106] (61.2%), or HLPF [31] (54.1%). As an example of the flexibility of our method on additional clues, taking the confidence scores from the pose estimation as an additional dimension of landmarks raises the accuracy rate to 75.7%. However, a gap of accuracy still exists between using estimated poses and ground truth poses (84.23% by ours).

Table 5 Comparison of methods on JHMDB using ground-truth landmarks

Methods	Accuracy (%)
DT-FV [102]	65.9
P-CNN [79]	74.6
HLPF [31]	76.0
Novel HLPF [34]	79.6
Path Signature (Ours)	80.4

For the SBU Interaction dataset, the two human bodies are regarded as one united articulated system with a total of 30 joints in 3D. As shown in Table 6, the proposed method using PSF significantly outperforms the other skeleton-based methods including many RNN-based or LSTM-based ones. Aside from the accuracy, the interpretable PSF could facilitate further understanding of interactions.

Table 6 Comparison of methods on SBU dataset

Method	Accuracy (%)
Yun et al. [41]	80.3
Ji et al. [107]	86.9
CHARM [108]	83.9
HBRNN [24] (reported by [37])	80.4
Deep LSTM (reported by [37])	86.0
Co-occurrence LSTM [37]	90.4
STA-LSTM [109]	91.5
ST-LSTM-Trust Gate [40, 25]	93.3
SkeletonNet [110]	93.5
GC-Attention-LSTM [80]	94.1
Path Signature (Ours)	96.8

Table 7 Comparison of methods on MHAD dataset

Method	Accuracy (%)
Vantigodi et al. [111]	96.1
Ofli et al. [73]	95.4
Vantigodi et al. [112]	97.6
Kapsouras et al. [113]	98.2
HBRNN [24]	100
ST-LSTM-Trust Gate [40, 25]	100
Path Signature (Ours)	100

For the Berkeley MHAD dataset, we achieve the same accuracy (100%) as the state-of-the-art methods shown in Table 7, showing the effectiveness of PSF for recognizing actions with accurate joint locations.

These results show that the proposed hand-designed feature set with single-layer linear network can outperform most deep learning methods on small datasets.

6.7.2 Comparison over large-scale datasets

We also conducted experiments on the large-scale NTURGB+D data.

For normalization, we applied the same 3D rotation and scaling as those in [39], so the body in the first frame faces the camera directly and those in the following frames are compensated accordingly. Since in this dataset different actions contain different number of detected actors, we applied a two-stage classification. The first stage is a binary classifier separating the actions into two types: 1-body or 2-body actions, then the second stage is the corresponding classifier (1-body or 2-body classifier) for each type. The supervised label of the binary classification at the first stage can be found by going through all the training samples and calculating the average number of actors in each action class. The range of the numbers is [1.02, 1.06] for the first 49 classes which are annotated as 1-body actions, while the range is [1.87, 2.04] for the remaining 11 classes which are annotated as 2-body actions.

Before feature extraction, we ranked all the detected actors in each clip based on the magnitudes of their movements. Then, for the 1-body classifier, features were extracted from the most active actor. For the first-stage binary classifier and the 2-body classifier, the two most active actors were regarded as one evolving object; this means we ended up having twice the number of joints per frame (i.e., 50 joints per frame). If a body is missing in the entire clip, the coordinates of this body are set to 0; if a body is missing in some medial frames, its coordinates are filled in using cubic spline interpolation [114].

The final results were given by two-stage classification as shown in Table 8. Table 9 shows that our method outperforms many deep learning based methods. The GCN [30] and its variants [83, 84, 85] achieve the current state-of-the-art accuracy on NTURGB+D dataset by taking advantage of the human skeleton structure. To utilize this skeleton structure as a prior knowledge to reduce complexity in our feature set is worth further studying.

Table 8 Accuracy (%) of the two-stage classification on NTURGB+D dataset

Task	The 1st stage	The 2nd stage		Final
		1-body	2-body	
Cross-subject	99.2	75.7	91.9	78.3
Cross-view	99.3	82.5	94.4	86.1

Table 9 Comparison of methods on NTURGB+D dataset

Method	Deep networks?	Cross-subject	Cross-view
Dynamic Skeletons [115]	X(SVM)	60.2	65.2
HBRNN [24]	✓(RNN)	59.1	64.0
Part-aware LSTM [39]	✓(LSTM)	62.9	70.3
ST-LSTM-Trust Gate [40, 25]	✓(LSTM)	69.2	77.7
STA-LSTM [109]	✓(LSTM)	73.4	81.2
SkeletonNet [110]	✓(CNN)	75.9	81.2
Joint Distance Maps [116]	✓(CNN)	76.2	82.3
GC-Attention-LSTM [80]	✓(LSTM)	74.4	82.8
Deep STGC [82]	✓(GCN)	74.8	86.3
ST-GCN [30]	✓(GCN)	81.5	88.3
Path Signature (Ours)	X(Single-layer NN)	78.3	86.1

6.8 Toward understanding of human actions

The interpretable geometric properties of PSF facilitate the understanding of human actions. By using a linear classifier the importance of each feature to each action class can be evaluated by the product of the two-layer weight matrices. For each class of sub-JHMDB, we ranked the joint pairs/triples according to the average over the weights connecting the features of joint groups and the corresponding class label. The top-3 joint pairs/triples for spatial and temporal features are shown in Fig. 10. The spatial ones often emphasize static constraints while the temporal ones highlight dynamic variations. Notice that many top pairs/triples are physically non-local, which demonstrates the effectiveness of the pose disintegration method.

Moreover, by using temporal disintegration ($h = 3$), we can evaluate the importance of different timescales and time intervals. As shown in Fig. 11, discriminative motions often appear in various intervals of finer timescales, e.g., the start of "catch" or "pick", the middle of "kick ball" or "swing ball", and the end of "golf" or "jump".

7 Conclusions

In this paper, we refined the path signature as a robust, nonlinear, and interpretable feature for landmark-based data. Path disintegrations and transformations are proposed to improve the effectiveness and efficiency of signature features. Based on these, we designed and built the signature-based spatio-temporal representation of action sequences. Experimental results show that using our feature set, a linear shallow fully-connected neural network achieves comparable results to advanced methods including CNN-based and RNN-based ones, especially on small datasets.

For future work, one could reduce the size of the representation of the body or feature set based on our analysis and understanding of human actions. It would also be interesting to integrate our landmark-based representation with other informative cues (e.g., appearance) to improve the performance of HAR. Moreover, our method

Fig. 10 Top-3 most important joint pairs/triples for (a) spatial features and (b) temporal features based on our linear network.

is general enough for other landmark-based objects where the given information in each landmark can be diverse.

Acknowledgements T. L. and H. N. are supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was supported by ERC advanced grant ESig (no. 291244). C. S. was supported in part by ERC advanced grant ALLEGRO. This work was supported in part by the Alexander von Humbolt Foundation. W. Y. and L. J. were supported in part by NSFC (Grant no.: 61472144, 61673182), the National Key Research & Development Plan of China (no.2016YFB1001405), GD-NSF (no.2017A030312006), GDSTP (Grant no.:2015B010101004, 2015B010130003), GZSTP (no.201607010227). W. Y. was supported by Royal Society Newton International Fellowship (NIF/R1/181466).

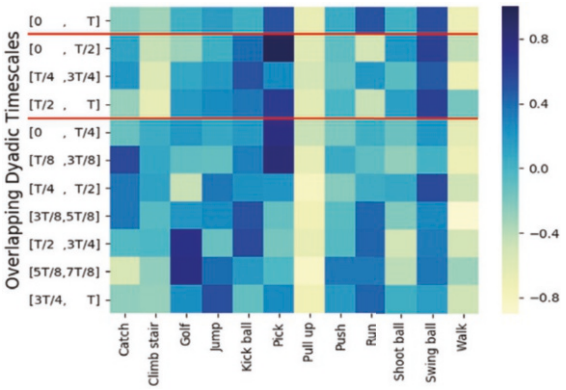

Fig. 11 Visualization of the important timescales and time periods for the actions in sub-JHMDB dataset. The darker in color, the more important it is.

References

[1] T. B. Moeslund and E. Granum. "A survey of computer vision-based human motion capture". In: *Computer vision and image understanding* 81.3 (2001), pp. 231–268.

[2] S. Mitra and T. Acharya. "Gesture recognition: A survey". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.3 (2007), pp. 311–324.

[3] R. Poppe. "A survey on vision-based human action recognition". In: *Image and vision computing* 28.6 (2010), pp. 976–990.

[4] D. Weinland, R. Ronfard, and E. Boyer. "A survey of vision-based methods for action representation, segmentation and recognition". In: *Computer vision and image understanding* 115.2 (2011), pp. 224–241.

[5] M. Ziaeefard and R. Bergevin. "Semantic human activity recognition: A literature review". In: *Pattern Recognition* 48.8 (2015), pp. 2329–2345.

[6] G. Guo and A. Lai. "A survey on still image based human action recognition". In: *Pattern Recognition* 47.10 (2014), pp. 3343–3361.

[7] C. H. Lim, E. Vats, and C. S. Chan. "Fuzzy human motion analysis: A review". In: *Pattern Recognition* 48.5 (2015), pp. 1773–1796.

[8] L. L. Presti and M. La Cascia. "3D skeleton-based human action classification: A survey". In: *Pattern Recognition* 53 (2016), pp. 130–147.

[9] G. Johansson. "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* 14.2 (1973), pp. 201–211.

[10] S. Sadanand and J. J. Corso. "Action bank: A high-level representation of activity in video". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 1234–1241.

[11] A. Ciptadi, M. S. Goodwin, and J. M. Rehg. "Movement pattern histogram for action recognition and retrieval". In: *European conference on computer vision*. Springer. 2014, pp. 695–710.

[12] R. Vemulapalli, F. Arrate, and R. Chellappa. "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. 2014, pp. 588–595.

[13] W. Yang et al. "Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge". In: *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE. 2015, pp. 551–555.

[14] W. Yang et al. "DropSample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition". In: *Pattern Recognition* 58 (2016), pp. 190–203.

[15] Z. Xie et al. "Fully convolutional recurrent network for handwritten chinese text recognition". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 4011–4016.

[16] Z. Xie et al. "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.8 (2017), pp. 1903–1917.

[17] W. Yang, L. Jin, and M. Liu. "Chinese character-level writer identification using path signature feature, DropStroke and deep CNN". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 546–550.

[18] W. Yang, L. Jin, and M. Liu. "Deepwriterid: An end-to-end online text-independent writer identification system". In: *IEEE Intelligent Systems* 31.2 (2016), pp. 45–53.

[19] B. Graham. "Sparse arrays of signatures for online character recognition". In: *arXiv preprint arXiv:1308.0371* (2013).

[20] L. G. Gyurkó et al. "Extracting information from the signature of a financial data stream". In: *arXiv preprint arXiv:1307.7244* (2013).

[21] J. Diehl. "Rotation invariants of two dimensional curves based on iterated integrals". In: *arXiv preprint arXiv:1305.6883* (2013).

[22] W. Yang et al. "Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 4083–4088.

[23] I. P. Arribas et al. "A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder". In: *Translational psychiatry* 8.1 (2018), pp. 1–7.

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

[25] J. Liu et al. "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates". In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017), pp. 3007–3021.

[26] T. Kerola, N. Inoue, and K. Shinoda. "Spectral graph skeletons for 3D action recognition". In: *Asian conference on computer vision*. Springer. 2014, pp. 417–432.

[27] M. Müller, T. Röder, and M. Clausen. "Efficient content-based retrieval of motion capture data". In: *ACM SIGGRAPH 2005 Papers*. 2005, pp. 677–685.

[28] G. Evangelidis, G. Singh, and R. Horaud. "Skeletal quads: Human action recognition using joint quadruples". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 4513–4518.

[29] M. Li and H. Leung. "Multiview skeletal interaction recognition using active joint interaction graph". In: *IEEE Transactions on Multimedia* 18.11 (2016), pp. 2293–2302.

[30] S. Yan, Y. Xiong, and D. Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[31] H. Jhuang et al. "Towards understanding action recognition". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 3192–3199.

[32] J. Wang et al. "Mining actionlet ensemble for action recognition with depth cameras". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 1290–1297.

[33] X. Yang and Y. L. Tian. "Effective 3D action recognition using EigenJoints". In: *Journal of Visual Communication & Image Representation* 25.1 (2014), pp. 2–11.

[34] J. Fan, Z. Zha, and X. Tian. "Action recognition with novel high-level pose features". In: *IEEE International Conference on Multimedia & Expo Workshops*. 2016.

[35] R. J. Williams and D. Zipser. "A learning algorithm for continually running fully recurrent neural networks". In: *Neural computation* 1.2 (1989), pp. 270–280.

[36] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[37] W. Zhu et al. "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.

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

[39] A. Shahroudy et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1010–1019.

[40] J. Liu et al. "Spatio-temporal lstm with trust gates for 3d human action recognition". In: *European conference on computer vision*. Springer. 2016, pp. 816–833.

[41] K. Yun et al. "Two-person interaction detection using body-pose features and multiple instance learning". In: *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, pp. 28–35.

[42] F. Ofli et al. “Berkeley mhad: A comprehensive multimodal human action database”. In: *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE. 2013, pp. 53–60.

[43] T. Lyons. “Rough paths, signatures and the modelling of functions on streams”. In: *arXiv preprint arXiv:1405.4537* (2014).

[44] K.-T. Chen. “Integration of paths–A faithful representation of paths by noncommutative formal power series”. In: *Transactions of the American Mathematical Society* 89.2 (1958), pp. 395–407.

[45] B. Hambly and T. Lyons. “Uniqueness for the signature of a path of bounded variation and the reduced path group”. In: *Annals of Mathematics* (2010), pp. 109–167.

[46] H. Boedihardjo et al. “The signature of a rough path: uniqueness”. In: *Advances in Mathematics* 293 (2016), pp. 720–737.

[47] F. Yin et al. “ICDAR 2013 Chinese handwriting recognition competition”. In: *2013 12th international conference on document analysis and recognition*. IEEE. 2013, pp. 1464–1470.

[48] I. P. Arribas. “Derivatives pricing using signature payoffs”. In: *arXiv preprint arXiv:1809.09466* (2018).

[49] F. J. Király and H. Oberhauser. “Kernels for sequentially ordered data”. In: *Journal of Machine Learning Research* 20.31 (2019), pp. 1–45.

[50] C. Salvi et al. “The Signature Kernel is the solution of a Goursat PDE”. In: *arXiv e-prints* (2020), arXiv–2006.

[51] P. Kidger et al. “Neural controlled differential equations for irregular time series”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020.

[52] I. Chevyrev, V. Nanda, and H. Oberhauser. “Persistence paths and signature features in topological data analysis”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.1 (2018), pp. 192–202.

[53] W. Yang et al. “Leveraging the path signature for skeleton-based human action recognition”. In: *arXiv preprint arXiv:1707.03993* 1 (2017).

[54] P. Bonnier et al. “Deep Signature Transforms”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2019.

[55] S. Liao et al. “Learning stochastic differential equations using RNN with log signature features”. In: *arXiv preprint arXiv:1908.08286* (2019).

[56] J. P. Biyong et al. “Information Extraction from Swedish Medical Prescriptions with Sig-Transformer Encoder”. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 2020, pp. 41–54.

[57] D. Wilson-Nunn et al. “A path signature approach to online arabic handwriting recognition”. In: *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*. IEEE. 2018, pp. 135–139.

[58] S. Lai, Y. Zhu, and L. Jin. “Encoding Pathlet and SIFT Features With Bagged VLAD for Historical Writer Identification”. In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 3553–3566.

[59] S. Lai and L. Jin. "Recurrent adaptation networks for online signature verification". In: *IEEE Transactions on Information Forensics and Security* 14.6 (2018), pp. 1624–1637.

[60] G. Biau and A. Fermanian. "Learning with Signatures". In: *International Workshop on Functional and Operatorial Statistics*. Springer. 2020, pp. 19–26.

[61] T. Ahmad et al. "Human Action Recognition in Unconstrained Trimmed Videos Using Residual Attention Network and Joints Path Signature". In: *IEEE Access* 7 (2019), pp. 121212–121222.

[62] C. Li et al. "Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 8585–8593.

[63] B. Wang et al. "A path signature approach for speech emotion recognition". In: *Interspeech 2019*. ISCA. 2019, pp. 1661–1665.

[64] Y. Wu et al. "Signature features with the visibility transformation". In: *arXiv preprint arXiv:2004.04006* (2020).

[65] V. M. Zatsiorsky and V. M. Zaciorskij. *Kinetics of human motion*. Human kinetics, 2002.

[66] J. K. Aggarwal and M. S. Ryoo. "Human activity analysis: A review". In: *ACM Computing Surveys (CSUR)* 43.3 (2011), pp. 1–43.

[67] M. Ye et al. "A survey on human motion analysis from depth data". In: *Time-of-flight and depth imaging. sensors, algorithms, and applications*. Springer, 2013, pp. 149–187.

[68] L. Wang, D. Q. Huynh, and P. Koniusz. "A comparative review of recent kinect-based action recognition algorithms". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 15–28.

[69] M. E. Hussein et al. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations". In: *Twenty-third international joint conference on artificial intelligence*. 2013.

[70] C. Ellis et al. "Exploring the trade-off between accuracy and observational latency in action recognition". In: *International Journal of Computer Vision* 101.3 (2013), pp. 420–436.

[71] L. Xia, C.-C. Chen, and J. K. Aggarwal. "View invariant human action recognition using histograms of 3d joints". In: *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, pp. 20–27.

[72] F. Ofli et al. "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition". In: *Journal of Visual Communication and Image Representation* 25.1 (2014), pp. 24–38.

[73] C. Wang, Y. Wang, and A. L. Yuille. "An approach to pose-based action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 915–922.

[74] P. Wei et al. “Concurrent action detection with structural prediction”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 3136–3143.

[75] A. Shahroudy et al. “Multimodal multipart learning for action recognition in depth videos”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.10 (2015), pp. 2123–2129.

[76] Y. Hou et al. “Skeleton optical spectra-based action recognition using convolutional neural networks”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.3 (2016), pp. 807–811.

[77] C. Li et al. “Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation”. In: *arXiv preprint arXiv:1804.06055* (2018).

[78] M. Rhif, H. Wannous, and I. R. Farah. “Action recognition from 3D skeleton sequences using deep networks on lie group features”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 3427–3432.

[79] G. Chéron, I. Laptev, and C. Schmid. “P-cnn: Pose-based cnn features for action recognition”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3218–3226.

[80] J. Liu et al. “Global context-aware attention lstm networks for 3d action recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1647–1656.

[81] W. Yang et al. “Developing the path signature methodology and its application to landmark-based human action recognition”. In: *arXiv preprint arXiv:1707.03993* (2017).

[82] C. Li et al. “Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition”. In: *Proceedings of the AAAI conference on artificial intelligence*. 2018.

[83] L. Shi et al. “Skeleton-based action recognition with directed graph neural networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7912–7921.

[84] M. Wang, B. Ni, and X. Yang. “Learning Multi-View Interactional Skeleton Graph for Action Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[85] H. Xia and X. Gao. “Multi-Scale Mixed Dense Graph Convolution Network for Skeleton-Based Action Recognition”. In: *IEEE Access* 9 (2021), pp. 36475–36484.

[86] T. J. Lyons. “Differential equations driven by rough signals”. In: *Revista Matemática Iberoamericana* 14.2 (1998), pp. 215–310.

[87] H. Ni. “A multi-dimensional stream and its signature representation”. In: *arXiv preprint arXiv:1509.03346* (2015).

[88] I. Chevyrev and A. Kormilitzin. “A primer on the signature method in machine learning”. In: *arXiv preprint arXiv:1603.03788* (2016).

[89] H. Boedihardjo, H. Ni, and Z. Qian. “Uniqueness of signature for simple curves”. In: *Journal of Functional Analysis* 267.6 (2014), pp. 1778–1806.

[90] G. Flint, B. Hambly, and T. Lyons. "Discretely sampled signals and the rough Hoff process". In: *Stochastic Processes and their Applications* 126.9 (2016), pp. 2593–2614.

[91] T. Lyons et al. *Esig on PyPi derived from CoRoPa: Computational Rough Paths software library.* https://github.com/datasig-ac-uk/esig,https://github.com/terrylyons/libalgebra, http://coropa.sourceforge.net/. 2007–2021.

[92] J. Reizenstein and B. Graham. "The iisignature library: efficient calculation of iterated-integral signatures and log signatures". In: *arXiv preprint arXiv:1802.08252* (2018).

[93] P. Kidger and T. Lyons. "Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU". In: *International Conference on Learning Representations*. 2020.

[94] M. Pfeffer, A. Seigal, and B. Sturmfels. "Learning paths from signature tensors". In: *SIAM Journal on Matrix Analysis and Applications* 40.2 (2019), pp. 394–416.

[95] C. Améndola, P. Friz, and B. Sturmfels. "Varieties of signature tensors". In: *Forum of Mathematics, Sigma*. Vol. 7. Cambridge University Press. 2019.

[96] *Implementation of the Path-signature-feature methodology for creating datasets for human action recognition from landmark data.* https://github.com/kschlegel/PSFDataset/. 2020.

[97] *CMU graphics lab motion capture database.* http://mocap.cs.cmu.edu/. 2003.

[98] *Microsoft Kinect.* http://www.xbox.com/en-US/Kinect/.

[99] J. Shotton et al. "Real-time human pose recognition in parts from single depth images". In: *Communications of the ACM* 56.1 (2013), pp. 116–124.

[100] L. Wan et al. "Regularization of neural networks using dropconnect". In: *International conference on machine learning*. PMLR. 2013, pp. 1058–1066.

[101] G. E. Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[102] H. Wang and C. Schmid. "Action recognition with improved trajectories". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 3551–3558.

[103] D. Oneata, J. Verbeek, and C. Schmid. "Action and event recognition with fisher vectors on a compact feature set". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 1817–1824.

[104] H.-S. Fang et al. "Rmpe: Regional multi-person pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2334–2343.

[105] U. Iqbal, M. Garbade, and J. Gall. "Pose for action-action for pose". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 438–445.

[106] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. "Joint action recognition and pose estimation from video". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1293–1301.

[107] Y. Ji, G. Ye, and H. Cheng. "Interactive body part contrast mining for human interaction recognition". In: *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. 2014, pp. 1–6.

[108] W. Li et al. "Category-blind human action recognition: A practical recognition system". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4444–4452.

[109] S. Song et al. "An end-to-end spatio-temporal attention model for human action recognition from skeleton data". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[110] Q. Ke et al. "Skeletonnet: Mining deep part features for 3-d action recognition". In: *IEEE signal processing letters* 24.6 (2017), pp. 731–735.

[111] S. Vantigodi and R. V. Babu. "Real-time human action recognition from motion capture data". In: *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*. IEEE. 2013, pp. 1–4.

[112] S. Vantigodi and V. B. Radhakrishnan. "Action recognition from motion capture data using meta-cognitive rbf network classifier". In: *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*. IEEE. 2014, pp. 1–6.

[113] I. Kapsouras and N. Nikolaidis. "Action recognition on motion capture data using a dynemes and forward differences representation". In: *Journal of Visual Communication and Image Representation* 25.6 (2014), pp. 1432–1445.

[114] C. De Boor and C. De Boor. *A practical guide to splines*. Vol. 27. springer-verlag New York, 1978.

[115] J.-F. Hu et al. "Jointly learning heterogeneous features for RGB-D activity recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5344–5352.

[116] C. Li et al. "Joint distance maps based action recognition with convolutional neural networks". In: *IEEE Signal Processing Letters* 24.5 (2017), pp. 624–628.