IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 1, 2019, accepted August 19, 2019, date of publication August 26, 2019, date of current version September 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937344

Human Action Recognition in Unconstrained Trimmed Videos Using Residual Attention Network and Joints Path Signature

TASWEER AHMAD[1,2], LIANWEN JIN[1], (Member, IEEE), JIALUO FENG[1], AND GUOZHI TANG[1]
[1]School of Information and Communication Engineering, South China University of Technology, Guangzhou 510000, China
[2]Department of Electrical Engineering, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan

Corresponding author: Tasweer Ahmad (tasveer28@yahoo.com)

This work was supported in part by the NSFC under Grant 61936003, Grant 61673182, and Grant 61771199, in part by the National Key Research and Development Program of China under Grant 2016YFB1001405, in part by the Natural Science Foundation of Guangdong Province (GD-NSF) under Grant 2017A030312006, in part by the Foundation of Guangdong Science and Technology Department (GDSTP) under Grant 2017A010101027, and in part by the Guangzhou Science, Technology and Innovation Commission (GZSTP) under Grant 201704020134. The work of T. Ahmad was supported by the Chinese Scholarship Council (CSC).

ABSTRACT Action recognition has been achieved great progress in recent years because of better feature representation learning and classification technology like convolutional neural networks (CNNs). However, most current deep learning approaches treat the action recognition as a black box, ignoring the specific domain knowledge of action itself. In this paper, by analyzing the characteristics of different actions, we proposed a new framework that involves residual-attention module and joint path-signature feature (JPSF) representation framework. The path signature theory was developed recently in the field of rough path and stochastic analysis, which provides a very efficient way to analyze any temporal sequence data. The proposed n-fold joint path signature features entail the Euclidean distances between joints and respective angles. For our experiment, JPSF for three modalities of joints (spatial location, bi-folds and tri-folds) are computed over the temporal length of action sequence. Then all these PSF are concatenated and fed to a CNN to give the recognition result. Experiments on three benchmark datasets, J-HMDB, HMDB-51 and UCF-101, indicate that our proposed method achieves state-of-the-art performance.

INDEX TERMS Convolutional neural networks, residual-attention, path signature features.

I. INTRODUCTION

Recognizing actions in videos are considered to be a very challenging task in computer vision. A great progress has been embarked in action recognition over the last decade due to convolutional neural networks (CNNs) [1] and Recurrent Neural Networks (RNNs). The task of action recognition becomes challenging due to human articulation, scale and view-point variation, camera motion, occlusion, human-object interaction (riding bicycle, playing guitar etc.) and human-human interaction (dancing, hugging etc.) [2]–[4]. Action Recognition finds numerous pragmatic applications in different areas; as video surveillance, home entertainment, Television and multi-media industry etc. [5]. Interestingly, video-based action recognition has greatly benefited from advancements of image-based CNN-models [6]–[9]. Contrary to 3D-convnet [1], [10], [11], other approaches involve 2D-CNNs over the sequence of frames of input video.

For most actions recognition, some body parts are more important as compared to the rest of the whole human body. Therefore, it seems quite intuitive that emphasis should be made and features should be extracted only from those parts of a frame where the human is present or its nearby. In this regard, attention network seems to be quite appealing to emphasis on the part of the image where human is present. Even for some actions, only part of the human body is important as compared to the whole body, e.g. for example, drinking action only involves hand and mouth, clapping action involves both hands etc.

The signatures of a path are collection of iterated integrals that are used for solving differential equations [31]. Path signature features have been introduced to machine learning and deep learning where they have made significant

The associate editor coordinating the review of this article and approving it for publication was Junchi Yan.

contributions to text recognition, Chinese text recognition and action recognition [33], [36], [37]. Path signatures treated as feature representation technique, are being used as a set of features for convolutional neural networks. A temporal sequence like on-line text recognition is represented as path signature features in order to feed into convolutional neural networks.

Our contribution of this paper is 1) involving residual-attention network for action recognition to emphasize only the most relevant portion of human body part for some action and only extracting corresponding joints. Secondly, we also propose a better n-fold path signature features by involving euclidean distances and corresponding angles between joints. Empirically, we investigate joints bi-folds and tri-folds as they mimic the human limbs which involve 2-3 connected joints. Moreover, our proposed framework can deal with unconstrained videos which may contain one or more subject and may have partial occlusion of the subject.

The rest of paper is organized as, section II entails a literature survey. Section III presents a brief introduction of PSF. The proposed methodology is discussed in section IV. The experimentation and discussion of results have been carried out in section V.

II. LITERATURE REVIEW

A. ACTION RECOGNITION

Recognizing action in images and videos has been quite ubiquitous over the last decade and therefore various image-based action recognition dataset [12], [13] and video based action dataset [14]–[17], have been proposed.

RGB-Flow based Two-stream network is the basic architecture for action recognition in videos [3]. An extension of two-stream framework is the multi-stream network [18], which involve multiple modalities of input videos such as RGB, optical flow, warped flow etc., which are trained by standard CNNs and then predictions are made using all modalities by late-fusion. In CNN+LSTM, at first stage spatial features are modeled by convolutional nets, while temporal features are formulated by recurrent nets at a later stage, [19]. By using 3D-ConvNets [1], [10], videos are represented as spatio-temporal blobs and 3D convolution models are trained for action recognition.

In [20], the authors propose a novel 3D ConvNet, comprising of a temporal transition layer that models variable temporal convolutional kernel depth. The temporal transition layer (TTL) operates on different temporal length, thus contributing the model to capture temporal information at different scales short, medium and longer frame-length. Lattice-LSTM investigates the idea of incorporating two separate LSTM-streams for raw RGB-images and optical flow images, like that of Two-stream architecture [21]. An extension of Two-stream network to inflated 3D ConvNet is proposed in [4] by expanding 3D convolutional networks in order to learn the spatio-temporal features for video classification. Mixed Convolutional Tube (MiCT) unifies 2D CNNs with 3D CNN in order to generate deeper and more informative feature maps, as in [22]. This framework reduces the computational complexity by replacing some of 3D-CNNs with 2D-CNNs. In [23], authors solve the problem of action recognition by introducing Generative Multiple-Instance Learning. Based on the idea of mitigating temporal redundancy in videos, in [24] authors proposed an idea of compressed video action recognition.

B. ATTENTION NETWORKS

In literature, attention networks are originally proposed for video captioning and language processing [25]. Reference [26] presents an attention-based framework for English-to-German and English-to-French translation. In [27], the authors mention a hierarchical attention network for document classification. Recently, attention network has also been proposed for action recognition in videos [28], [29], [30]. In [28], the authors introduce an attention mechanism as low-rank second-order pooling for single image classification. In [29], the authors developed an attention-based neural network in order to model the scene objects interaction for action recognition and video captioning. In [30], authors introduce two separate temporal and spatial attention mechanism to identify the most relevant frame and the most-relevant spatial location in that frame of a particular action.

C. PATH SIGNATURE

In literature, path signatures have revealed significant progress in modeling temporal dependencies. Iterated-integral signatures contain sequences of numbers, derived from a continuous path [31]. These iterated-integrals come from the theory of differential equations driven by rough paths and it finds numerous application in machine learning, statistics and financial data predictions. Log-path signatures contains the same level of information, but with fewer feature dimensions [31]. Lyons et. al. in [32], devised a rough path signature based approach for sound compression, where it is revealed that this method performs better than Fourier and Wavelet methods. In [33], authors reveal that path signatures can be used as a set of features for input to convolutional neural networks (CNNs), which improve the accuracy of on-line character recognition. The authors realize that first, second and third iterated integrals have useful information for recognizing letters, numbers, Assamese and Chinese characters. In [34], authors present an application of iterated integrals for recognizing on-line character recognition with arbitrary rotations. The authors evaluated their approach on pendigits [35] dataset with handwritten digits by 44 writers, 30 for training and 14 for testing. In [36], authors introduce a path-signature based deep CNN architecture for text-independent writer identification. Empirically, authors persuade that path-signature features are viable for recognizing Chinese handwriting. Recently in [37], a PSF-based approach has been employed for action recognition in videos.

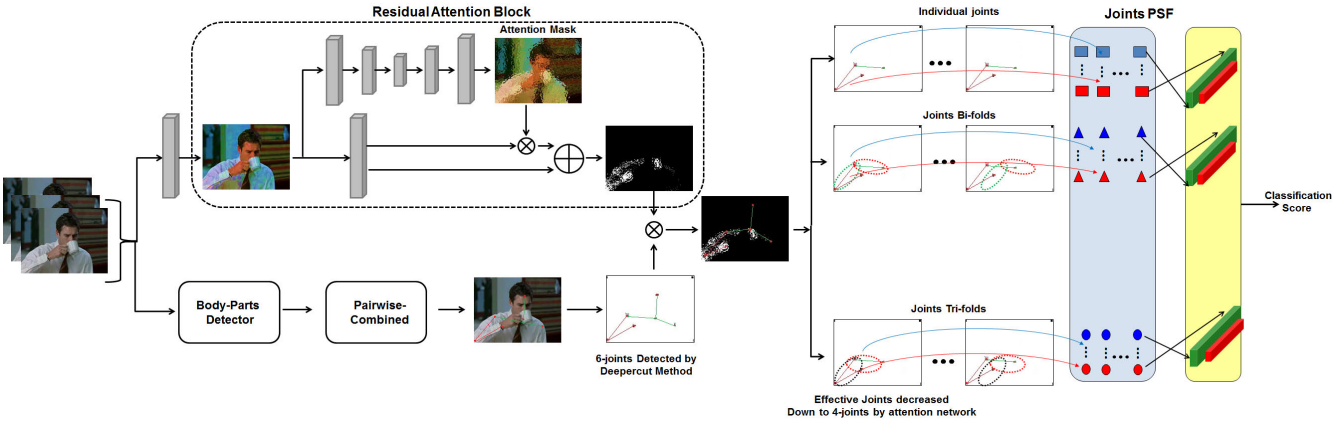


FIGURE 1. Block diagram of proposed framework. First human poses are extracted from input images.Then using attention network proposed architecture figures out the most relevant joints of human body pertaining to a particular action. Finally, path signature features of only those particular joints are extracted and inputted to a CNN to produce the classification results.

III. THE SIGNATURE OF A PATH

The signatures of a path are calculated by using iterated-integrals. For a path $X : [a, b] \mapsto R^d$, coordinate paths are $(X_t^1, \ldots, X_t^d)$ with each real-valued path is defined as $X^i : [a, b] \mapsto R$. A path integral for single index $i \in \{1, \ldots, d\}$ is defined as

$$S(X)_{a,t}^{i} = dX_s^i = X_t^i - X_0^i \tag{1}$$

For k-th fold iterated integral,

$$S(X)_{a,t}^{i_i,\ldots,i_k} = \int_{a<t_k<t} \cdots \int_{a<t_1<t_2} dX_{t_1}^{i_1} \ldots dX_{t_k}^{i_k} \tag{2}$$

For $k = 0$, iterated-integrals corresponds to the original path, $k = 1$ denotes path displacement and $k = 2$ path curvatures. Higher-level path signature features are computed by higher values of k which contain more temporal details, but at the expense of higher feature dimension.

For numerical computation, CoRoPa C++ [45] and iisignature [38] are two open-source libraries, available for computing path signature features.

A. PROPERTIES OF PATH SIGATURES

1) TIME REPARAMETRIZATION

For a multidimensional path $X : [a, b] \mapsto R^d$ and its reparametrized path $\widetilde{X}_1 = \widetilde{X}_\psi$, path signature iterated integrals remain same as

$$S(\widetilde{X})_{a,b}^{i_1,\ldots,i_k} = S(X)_{a,b}^{i_1,\ldots,i_k}, \quad \forall k \geq 0,\ i_1, \ldots, i_k \in 1, \ldots, d.] \tag{3}$$

Intuitively, time reparametrization can be perceived as same action performed by different actors taking different time, however, their path signatures remain same over the length of time.

2) SHUFFLE PRODUCT

The Product of two signature terms can be defined as the sum of all possible ways of interleaving signature terms, in the same order. Shuffle product of two terms of signatures can be expressed as a linear combination of higher order terms. For a two-dimensional path $X : [a, b] \mapsto R^2$, shuffle product is expressed as

$$S(X)_{a,b}^{1} S(X)_{a,b}^{2} = S(X)_{a,b}^{1,2} + S(X)_{a,b}^{2,1} \tag{4}$$

3) TIME REVERSAL

Signatures of a path $X : [a, b] \mapsto R^d$ are inverse under tensor product, if the signature are computed backward in time. Under time reversal, for a path $X : [a, b] \mapsto R^d$, it holds that

$$S(X)_{a,b} \otimes S(\overleftarrow{X})_{a,b} = 1. \tag{5}$$

This time reversal identity holds under the power series where $\lambda_0 = 1$ and $\lambda_{i_1 \ldots i_k} = 0$ for all $k \geq 1$.

4) CHEN'S IDENTITY

Chen's identity provides an algebraic relationship between paths and their signatures. Using Chen's identity, we define concatenation as the path $X * Y : [a, c] \mapsto R^d$ and for which $(X * Y)_t = X_t$ for $t \in [a, b]$ and $(X * Y)_t = X_b + (Y_t - Y_b)$ for $t \in [b, c]$.

IV. PROPOSED METHOD

The proposed architecture of our framework is illustrated in figure 1. The first stage of proposed method comprises of attention network that localizes the most salient parts of an image (human body-part) for recognizing some action. Considering only the most relevant joints of human body and then respective path signatures features of only those contributing joints are computed. On one hand, proposed approach significantly reduces down computational complexity, while on the other hand it makes the framework robust to occluded joints,

which may come across if we consider all the joints of human body. It is always computationally exhaustive to compute path signature features of all modalities (spatial locations, bi-folds and tri-folds) of the joints, but if we consider only the most relevant joints of a particular action it can be significantly reduced down computational burden and improve the accuracy by mitigating over-fitting.

A. ATTENTION NETWORK FOR ACTION RECOGNITION

Attention networks are based on the principle of devising attention mask M_i for the i-th features of input, X_i. The attended image features are obtained by element-wise multiplication of attention mask with the input image,

$$X_{i,m} = X_{i,m} * M_{i,m} \tag{6}$$

where i ranges from all spatial locations and m corresponds to all channels. This attention network emphasize strong features in input image while attenuates the less important features, based on mask M_i. This concept is better illustrated in figure 2.

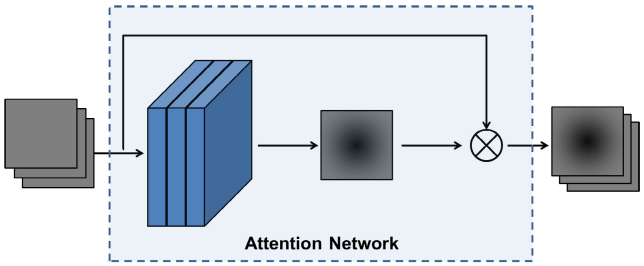


FIGURE 2. Block diagram of general attention network.

1) RESIDUAL-ATTENTION NETWORK

Our proposed framework contains multiple residual attention modules. Each residual-attention module stacks many residual and attention blocks stacked together. Attention modules are responsible for emphasizing the most salient features in images, while residual blocks contain standard residual network for convolution nets. Each attention block contains a mask branch and a trunk branch, where the mask branch is responsible to produce different attention masks in order to figure out different features. Whereas, trunk branch includes standard CNN architecture, e.g VGG or Residual Nets for convolution features. For given input x, mask branch uses top-to-bottom and bottom-to-top (encoder-decoder) architecture in order to learn the same size mask as that for trunk branch. For an attention module f_{att}, the output is formulated as

$$f_{att}(x) = M_{i,m}(x) * I_{i,m}(x) \tag{7}$$

$$M_{i,m}(x) = exp(e_{ti}) / \sum_{k=1}^{L} exp(e_{tk}) \tag{8}$$

where i spans for all spatial locations, whereas m belongs to the index of channel. The attention mask in eq.7 is computed by using emphasizes vectors, $M_{i,m}(x)$ as in eq 8. Each attention module has trunk branch and mask branch, where the mask branch generates feature masks specialized for trunk branch. Moreover, residual attention network is differentiable for both mask branch and trunk branch, where mask branch suppresses noisy labels to update trunk branch.

2) TOP-DOWN AND BOTTOM-UP ATTENTION

For a convolution-net, class prediction is made by

$$Y_{pred} = W * X^T \tag{9}$$

For multiple class prediction, the weight matrix in above equation contains class-specific weights and class-agnostic weights. Top-down attention map corresponds to class-specific weights whereas bottom-up attention map corresponds to the class-agnostic weights, The attention-score for residual-attention encoder-decoder is computed by

$$score_{attenion} = p_k^T * q \tag{10}$$

where attention-score is the dot-product of top-down p_k and bottom-up q attention maps. At each stage, the computed top-down and bottom-up attention map is up-sampled and finally modulated with the trunk branch in order to extract most relevant features, specific to a class k.

3) ATTENTION-BASED RESIDUAL LEARNING

The residual-attention network contains many residual modules and attention modules. The performance of the overall network improves by increasing the number of attention modules. By increasing the number of attention modules improves performance, but it also increases the computational complexity. In residual attention network, the trunk layer depth is determined by $36m + 20$, where $m = 1, 2, 3, 4$ the number of attention modules, [41]. Thus, based on trunk layer depth, residual attention networks are identified as Attention-56, Attention-92, Attention-128, and Attention-164 having one, two, three and four attention modules respectively. An advantage of using a residual attention framework is that multiple attention modules can be stacked together and their performance does not degrade. For residual-attention network, attention modules are devised in such a way to retain useful features by applying dot product while mitigating noisy features. We also replace final softmax-layer in the network with global average pooling (GAP) layer to localize the most salient joints. Then corresponding heat-maps of these localized joints are created for visualization.

B. POSE ESTIMATION

The next important part of our proposed architecture is pose estimation, which is carried out by the DeeperCut method [42]. DeeperCut method is a multi-person pose estimation technique, based on integer linear programming (ILP). DeeperCut method locates 14-joints on the human body in an image, using very deep 152-layers ResNet architecture for body part detection that could estimate single or

multi-person pose. ResNet-152 alleviates the problem of vanishing gradient by passing the state through identity layer and devising residual functions. For action recognition, DeeperCut method figures out the spatial location of human joints in each frame, which are subsequently used by path signature feature extraction and classification. The joints extracted by DeeperCut method are enlisted in figure 3. Empirically it is realized that over the length of $N-$frames, joints may appear or disappear due to occlusion/self-occlusion, this problem is mitigated by interpolating the location of joints over the length of frame sequence.

C. JOINTS PATH SIGNATURE FEATURE EXTRACTION

1) BASELINE CASE

For our experiment, baseline case is reported as the spatial locations of individual joints. We define joints n-folds by combining n-different combinations of joints. For a given action video, temporal length of frames is fixed to 10-frames for all cases. The overall feature dimension is determined by $k*(d^{(l+1)}-d)/(d-1)$, where k is the number of active joints in a frame, d is the input dimension and l is the level of path signatures.

2) BI-FOLD JOINT PATH SIGNATURE FEATURES

The bi-fold JPSF are illustrated by considering different pairs of joints and the corresponding euclidean distances between those joints. The total combinations of joints are calculated by C_r^2, where r is the number of contributing joints to some action. The feature dimension for bi-fold JPSF is determined by $C_r^2*(d^{(l+1)}-d)/(d-1)$, where d is the input dimension and l is the level of path signatures. As for 'drinking' action illustrated in fig. 1, it can be perceived that five upper-body joints are related to 'drinking' action, therefore it seems quite appealing that only bi-folds of these joints should be computed. Bi-fold JPSF capture the spatial relationship and dependency among different pairs of joints over a length of $N-$frames. An intuition to bi-fold JPSF is that joints in human body are spatially connected and collectively contribute to some action, as in figure 3.

3) TRI-FOLD JOINT PATH SIGNATURE FEATURE

Subsequently, tri-fold JPSF are defined by considering three different combinations of joints in each frame. The total number of modality combinations for tri-fold JPSF are enlisted by C_r^3, where r is the number of contributing joints to some action. The feature dimension for tri-fold JPSF are determined by $C_r^3*(d^{(l+1)}-d)/(d-1)$, where d is the input dimension and l is the level of path signatures. Tri-fold JPSF involve three joints and models two distances (d_1, d_2) between three joints and the minimum angle theta, θ between distances d_1 and d_2. As the tri-fold JPSF entail three joints and are important to mimic the different body limbs (e.g. arms/legs); cogent for analyzing human articulation and movement. These tri-fold JPSF are vividly illustrated in figure 3, (right side). Similar to bi-fold JPSF, the tri-fold JPSF

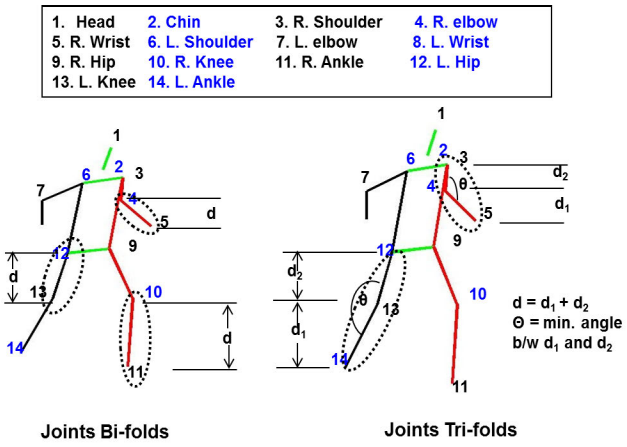


FIGURE 3. The 14-joints with description, as extracted by deeperCut method (Top). Bi-folds of joints, models corresponding distance between joints (Left). Tri-folds of joints, holds corresponding distances and angle between joints (Right).

are also calculated over a length of N-frames. The angles between joints are important in order to figure out different actions. For examples, as referring to figure 4, 'drinking' action and 'shotting' action depend upon same joints, but the angles between these joints are different for 'drinking' and 'shotting' actions

4) HIGHER-FOLD JOINT PATH SIGNATURE FEATURE

The higher-fold JPSF (quad and penta-folds) of joints are investigated by involving more than three combinations of joints. For example, quad-fold JPSF includes four different joints with three corresponding distances d_1, d_2 and d_3, and two angles θ_1 and θ_2. Following the same practice, penta-fold JPSF incorporate four distances d_1, d_2, d_3, and d_4, and three corresponding angles θ_1, θ_2 and θ_3.

5) CLASSIFICATION MODEL

The classification model for our proposed architecture is similar to LeNet, with three stacks of convolution-pooling layers and two fully-connected layers. The network architecture is described like this, Input–> 6C3–> MP2–> 16C3 –> MP2 –> 10C2–> MP2 –> 120FC–> 84FC–> Output. For example for convolution layer 6C3, represents six feature maps with a filter-size of three, the stride size is fixed to one-pixel in all layers. MP2 denotes the two-dimension max-pooling. FC stands for fully connected layers.

V. EXPERIMENTS

A. DATASET

1) J-HMDB DATASET

J-HMDB is joint-annotated subset of HMDB-51 dataset. The dataset contain 21-actions video, having only one main subject performing different tasks [44]. In total, J-HMDB dataset contains 928 videos, where each action class has about 40-50 videos.

(a) Drinking action

(b) Shotting action

FIGURE 4. Significance of joints angles, incorporated by joints tri-folds. Different actions correspond to different angles between joints.

2) HMDB-51 DATASET

HMDB-51 is a trimmed videos dataset which contains 6,766 realistic videos of 51 different actions [14]. For each action, there are at least 100 videos, 70 videos for training and 30 for testing. Three splits of the dataset are provided for cross-validation.

3) UCF-101 DATASET

UCF-101 dataset contains 101 action classes, with 100 video clips in each class, 70 for training and 30 for testing [15]. In total, UCF-101 dataset contains 13,320 video clips, with each action at-least 70 training and 30-testing videos. UCF-101 dataset also has three splits for cross-validation.

B. TRAINING AND TESTING

For our experiments, We used a mini-batch size of 16-images for training and testing, with a learning rate of 0.001 which was cut down by 0.1 after every 1/3-th of iterations. The path signature features determine input dimension of network, whereas output dimension is the probability distribution of class-labels. Adam optimizer was used as a solver in order to optimize the network. The network was trained for a maximum of 200 epoch, with a momentum of 0.9 and weight-decay of 0.0005. For training, data augmentation was performed by center-cropping images to 224x224, horizontal flipping and by adding some Gaussian noise. For testing purpose, we used a center-crop of 224x224 pixels, with normalization in all image channels. Cross-entropy loss was used as a loss-function to optimize the network. The drop-out was used as a regularizer, with a higher drop-out ratio of 0.90. The network was trained using two NVIDIA TITAN X GPU for 200 epoch. We used Pytorch framework for experimentation, whereas Path Signatures were computed by using iisignature package [38].

C. PIPELINE OF EXPERIMENT

1) POSE ESTIMATION

In the pipeline of experiments, human-pose is estimated in every frame by using DeeperCut method. DeeperCut method takes raw input RGB-images and marks 14-joints on the human body. At the first stage of pose estimation, a semantic segmentation and object classification based human part detector is run on input images in order to find different body parts. Then, body part refinement is carried out by sampling 'D' body part detections and regressing from current location to the relative positions of all other joints, [42]. Non-maximum suppression (NMS) is used for refinement of joint locations. For some actions where the human body is partially occluded or is not fully visible, DeeperCut method even tries to mark all 14-joints which may lead to false action recognition due to over-fitting. This problem is circumvented by introducing residual attention network, which figures out most salient features.

2) ATTENTION NETWORK

The Attention network brings the advantage of emphasizing only those body parts and joints that are most precisely related to some action. The residual-attention network is trained for three different datasets, J-HMDB, HMDB-51 and UCF-101. For J-HMDB dataset, the attention network is trained for 21-action classes, batch size of 32-images and with a learning rate of 0.01. J-HMDB is considered to be simple and less challenging dataset, as compared to HMDB-51 and UCF-101 datasets. The soft-attention mask figures out the location of most relevant joints for action recognition in J-HMDB dataset.

For HMDB-51 and UCF-101 datasets, the batch size for attention network is kept to 16-images, with a learning rate of 0.001. These two datasets contain numerous videos where the human body is partially occluded. Therefore, figuring out most relevant joints is tedious for HMDB-51 and UCF-101 datasets. In attention layers, convolutional kernel size is decreased to three pixels to capture more local features. Attention network encodes the information of only salient joints and subsequent n-fold path signatures of only those joints are computed.

3) PATH SIGNATURE FEATURES

For this experiment, Path signature features are computed by using iisignature package [38]. The iisignature is a python-based package for computing path-signature features. The spatial locations of most salient joints are identified by above-mentioned attention network. For our experiment, the first baseline case is defined by considering only normalized spatial locations of joints. Bi-folds of joints corresponds to the euclidean distance between two joints. Tri-folds and higher n-folds of joints hold all the euclidean distances and angles between joints. The features in each dimension were kept normalized by dividing each value with the maximum value of that dimension.

During experimentation, a detailed study is carried out for baseline, bi-folds and tri-folds of path signature features with the intuition that bi-folds and tri-folds of joints are the most relevant to mimic the motion of human limbs (arms, limbs etc.). Referring to figure 3, it is noticeable that joints

TABLE 1. PSF performance evaluation for different modalities on three datasets. In general, performance exceed as PSF-levels are increased for baseline and bi-folds of joints, thus contributing more feature information.

Signature level	J-HMDB			HMBD-51			UCF-101		
	Baseline Acc. (%)	Bi-fold Acc. (%)	Tri-fold Acc. (%)	Baseline Acc. (%)	Bi-fold Acc. (%)	Tri-fold Acc. (%)	Baseline Acc. (%)	Bi-fold Acc. (%)	Tri-fold Acc. (%)
level-2	44.5	59.5	64.1	35.5	42.0	48.4	47.6	56.4	59.3
level-3	48.6	61.0	65.9	42.7	47.6	53.0	54.1	59.8	65.2
level-4	52.8	63.5	**70.4**	50.1	51.8	**61.6**	56.0	65.5	74.4
level-5	**56.6**	**66.8**	69.0	**52.8**	**56.1**	60.0	58.7	**74.3**	**80.8**
level-6	55.7	65.3	67.8	51.9	54.5	57.2	**61.4**	71.1	78.6

bi-folds have a strong relationship to human limbs, like joint pairs (3, 4), (3, 5) and (3, 6) corresponds to distances between rightShoulder-rightElbow, rightShoulder-rightWrist and rightShoulder-leftShoulder. For joint tri-folds, a similar relationship also exists; for example, tri-fold (3, 4, 5) corresponds to distances between rightShoulder-rightElbow (d_{34}) and rightShoulder-rightWrist (d_{35}) and the angle between d_{34} and d_{35}.

4) TEST EXAMPLE

Considering a test example, having all 14-joints with path signature feature level-2, results in a feature dimension of 168, 1,092 and 15,288 for baseline, bi-fold and tri-fold respectively. However, for 'drinking action' involving attention network points out four most relevant joints (5-joints of arm and neck) are considered then feature dimension substantially reduces down to 100, 60, and 600. This results in a gradual drop in feature dimension about 40%, 94.5% and 96.1%. The given feature dimension is computed by $C_r^n * (d^{(l+1)} - d)/(d - 1)$, where r is the total number of joints, n number of joints considering at a time (bi-folds $n = 2$, tri-folds $n = 3$), d is input dimension and l is signature-level.

D. RESULTS AND ANALYSIS

For our experiment, mean-Average Precision (mAP) is used as an evaluation metric for three splits train-test splits of HMDB-51, UCF-101 and J-HMDB datasets. mAP is calculated by using the following formula, $1/N * \sum_{i=1}^{N} AP_i$, where AP is corresponding average precision for each split of dataset.

In Table 1 for both J-HMDB and HMDB-51 datasets for baseline and bi-fold of joints, performance increases as the PSF-levels are increased up to level-5, however, it slightly drops down from level-5 to level-6. For the tri-fold of joints, performance tweaks up to PSF level-4 and then it comes down. For UCF-101 dataset for baseline case, the performance increases as the signature levels are increased. Likewise for bi-folds and tri-folds of joints, mAP also surpass as the PSF-levels are increased up to level-5, after that performance begins to drop down.

An intuitive justification for a rise in accuracy by increasing signature levels is that as path levels are increased, more and more supplementary details are incorporated by path signature features for final classification. However, the performance from path level-5 to level-6 slightly decreases due to over-fitting as the feature dimension tremendously increases. From Table 1, it is also conceived that performance for baseline case is marginal because baseline modality only comprises of spatial location of joints whereas bi-folds, tri-folds and higher folds of joints entail normalized euclidean distances (and angles) between joints.

1) MODALITY FUSION

After computing PSF of individual modalities, we concatenated different modalities (Baseline, Bi-folds and Tri-folds) and excite them to CNN for final classification, as referred in Table 2. The feature dimension of different modalities are made same by padding and then encoded as three different channels of the input. For above-mentioned three datasets, it is observed that performance gradually increases as different modalities are fused together. For J-HMDB dataset and HMDB-51, the best performance for baseline and bi-folds of joints are reported for PSF level-5, and for tri-fold of joints the best performance is revealed for level-4. Referring to Table 2 for J-HMDB dataset, it is conceived that baseline case furnish a performance of 56.6%, adding bi-folds surpass the performance to 72.0%, then further by adding tri-folds result in an improved performance of 81.3%. For HMDB-51 dataset, the baseline case generates a performance of 52.8%, adding bi-folds surpass the performance to 69.5%, then further adding tri-folds results in an improved performance of 79.2%. Likewise for UCF-101 dataset, the best performance for baseline of joints is reported for PSF level-6, for tri-fold and bi-folds of joints the best performance is disclosed for level-5. Referring to Table 2 for UCF-101 dataset, it is conceived that baseline case furnish a performance of 61.4 %, adding bi-folds surpass the performance to 81.6 %, then further adding tri-folds result in an improved performance of 97.3 %. For all three datasets, best performance is claimed when considering all three modalities due to the reason that all three modalities (baseline, bi-fold, tri-folds) contain complementary details when unified results in best performance.

Using path signature feature, the performance for action recognition increases as we use bi-folds and tri-folds of joints. But as we involve higher (quad or penta) folds of joints, the performance gradually drops down. This may be due to the reason that higher folds of joints contain more euclidean distances and angles, resulting in higher feature dimension for

TABLE 2. Illustration of performance improvement by unifying Baseline of joints with Bi-fold and Tri-fold modalities.

Datasets	Baseline Acc. (%)	Base+Bi Acc. (%)	Base+Bi+Tri Acc. (%)
J-HMBD	56.6	72.0	81.3
HMBD-51	52.8	69.5	79.2
UCF-101	61.4	81.6	97.3

TABLE 3. Performance evaluation for joints bi-folds, tri-folds, quad-folds and penta-folds. It is observed that network performance for quad and penta-folds degrades due to over-fitting.

Dataset		J-HMBD Acc. (%)	HMDB-51 Acc. (%)	UCF-101 Acc. (%)
Bi-folds	level-3	61.0	47.6 %	59.8 %
Tri-folds		65.9	53.0	65.2
Quad-folds		57.3	40.6	55.6
Penta-folds		53.1	36.6	50.4

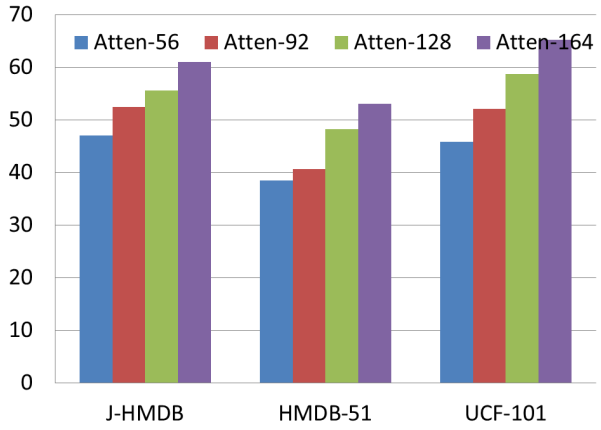


FIGURE 5. Residual attention network-56, 92, 128 and 164 having one, two, three and four attention modules. It is noticeable that as attention modules are increased, performance also tweaks. (Best viewed in color).

path signature features. This trend is empirically explained in Table 3.

In figure 5 comparisons have been made among different attention networks for path signature level-3. In this table, it is realized that for all three datasets as we increase the attention blocks in residual attention network, it results in improved final classification accuracy. The best performance is reported by using four attention blocks which results in Attention-164 architecture.

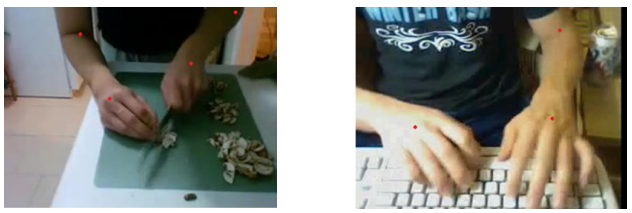
E. COMPARISON WITH STATE-OF-THE-ART METHODS

In Table 4, a comparison of our proposed method with contemporary methods has been presented for J-HMDB, HMDB-51 and UCF-101 datasets. For J-HMDB dataset, [44] encompass high-level features and results in an accuracy of 76.0%. [46] investigates pose-based convolutional neural networks and results in an improved performance. In [37], path signature based features has been included for action

TABLE 4. Comparison of proposed method with previous methods.

Method	Accuracy in (%) J-HMDB	HMDB-51	UCF-101
HLPF, [45]	76.0	-	-
P-CNN, [46]	79.5	-	-
Path Sig., [36]	80.4	-	-
Two Stream CNN, [3]	-	59.4	88.0
TSN (3 modalities) , [18]	-	69.4	94.2
Lattice LSTM, [21]	-	68.5	94.0
ShuttleNet, [39]	-	71.7	95.4
I3D-CNN, [4]	-	66.4	93.4
Compressed Video, [24]	-	70.2	94.9
Mixed 2D-3D Conv., [22]	-	70.5	94.7
OFF with ResNet-20, [40]	-	74.2	96.0
Proposed Method	**81.3**	**79.2**	**97.3**

(a) In order to find main actor, absolute motion of each subject is computed by using normalized motion of individual joints a length of frames, N.

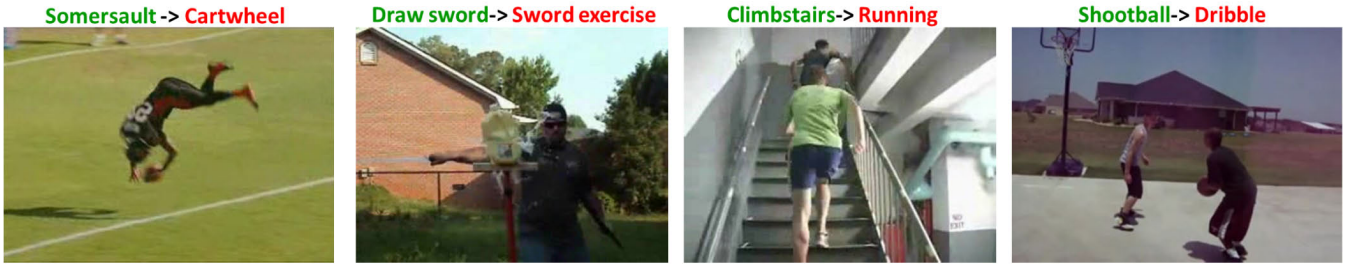


(b) For above cutting and typing actions, Deepercut method is unable to find human joints. In such cases, joints are manually located.

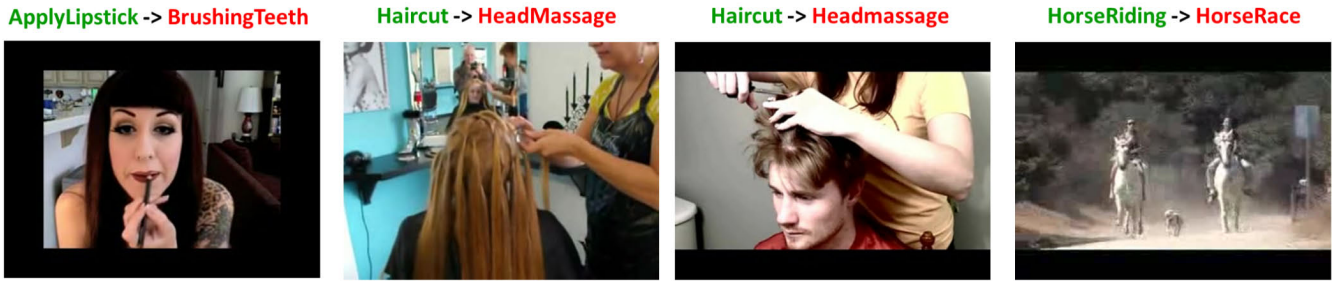
FIGURE 6. Illustration of some extreme cases.

recognition in J-HMDB dataset. This path signature approach in [37], resulted in an improved performance as compared to previous high-level features or pose-based features methods. Then, we propose attention-based framework using path signature features, which results in state-of-the-art performance over existing methods for J-HMDB dataset.

Reference [3] includes two stream convolutional network and serves as a baseline for HMDB-51 and UCF-101 datasets. As our proposed method involves RGB and pose modalities, final version of TSN [18] takes 3-modalities (RGB+optical flow+iDT) as input. Lattice LSTM involves complex ResNet and temporal segment network [21]. Reference [39] includes CNN+shuttlenet, where shuttle-net is a multi-layer complex RNN architecture. Inflated 3-D convolutional network [4], have exhibited unprecedented performance for Kinetics dataset. Reference [22] involves complex architecture of mixed 2D and 3D-CNN in two streams architecture. Reference [40] entails optical flow guided features (OFF) along with RGB and optical flow images, where OFF features are fed to ResNet-20. Our proposed architecture is quite intuitive as it first extracts most relevant parts of an image using the attention network and then the path-signature features of only those joints are extracted for classification.



(a) Examples of some Failure cases for HMDB-51 dataset.



(b) Example of some Failure cases for UCF-101 dataset

FIGURE 7. Some failure cases have been reported where within-class variance is high, while between class variance is low. Therefore, such cases are misclassified.

From Table 4, it can be seen that our method achieves the best performance for all the three datasets. It is also worth to note that on the challenging HMDB-51 dataset, our method outperforms previous methods with a large margin.

F. DISCUSSION

During experimentation, some extreme cases are reported where our approach is able to deal with such cases, as illustrated in figure 6. The first extreme case is reported as when there is more than one subject for the same action, for example actions such as sword and shake-hands involve more than one subject, refer fig 6a. In such scenario, Deepercut method estimates multiple human poses, then for each frame one main subject is identified by computing the normalized motion of joints for each subject. The subject with higher normalized motion over a length of frames is identified as main actor and the path signatures of only this subject are further computed.

In figure 6b, another extreme case is figured out when only a part of the human body is exposed to camera. For example, for actions cutting, mixing and typing in UCF-101 dataset, only hands are exposed to camera, Deepercut method cannot estimate human pose in such extreme cases [42]. For such actions, human poses are manually estimated in each frame, that seems to be a tedious task. For the sake of generality, such extreme actions are not included for J-HMDB dataset, on which previous methods have been exercised [37]. During the sequence of frames, it is quite possible that a subject may disappear or partially/fully occluded. For such cases, the coordinates of subject are interpolated by using bi-cubic interpolation.

In figure 7, some failure cases have been studied for both HMDB-51 and UCF-101 datasets. In fig.7a, the action somersault is misclassified as cartwheel, because of this particular video it closely resembles with cartwheel. Likewise for the other three cases (Drawsword, Climbstairs and Shootball) from HMDB-51, these closely resembles with false class. Likewise in fig. 7, the action ApplyLipstick is misclassified as BrushingTeeth as in this particular video this action looks like BrushingTeeth. Similarly other false classification results for UCF-101 have been disclosed. The strong reason for misclassification of these actions is their high within class variance and low between class variance to their targeted false class.

VI. CONCLUSION

In this paper, we propose an intuitive idea of paying attention only to the most relevant part of human body corresponding to some action using attention network. Then body-joints of only those portions are extracted and corresponding path signature features of those joints are computed. It is justified because for most actions only some portion of human body is the most relevant. Then, we propose n-folds (bi-folds and tri-folds) of joint path signature features. These n-folds of JPSF better model the orientation and euclidean distances between joints over the temporal length of sequence. Above-mentioned PSF of individual joints, bi-folds and tri-folds are concatenated for final classification, where they perform state-of-the-art. We also investigated higher folds of joints (quad-folds and penta-folds), and realize that computational complexity get increased by using higher folds of joints, but performance surpass meagerly. Experiments have

shown that the proposed architecture is viable to recognize actions on three benchmark datasets, J-HMDB, HMDB-51, and UCF-101. In future, log-PSF and Laplace-PSF which are variants of PSF and will be investigated.

REFERENCES

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[2] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," 2015, *arXiv:1507.02159*. [Online]. Available: https://arxiv.org/abs/1507.02159

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 568–576.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[5] M. A. Bagheri, Q. Gaom, and S. Escalera, "Support vector machines with time series distance Kernels for action classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–7.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[10] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[11] A. J. Piergiovanni, C. Fan, and M. S. Ryoo, "Title learning latent subevents in activity videos using temporal attention filters," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4247–4254.

[12] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1017–1025.

[13] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1331–1338.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[15] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[16] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3398–3405.

[17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: https://arxiv.org/abs/1705.06950

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[19] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[20] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, L. Van Gool, "Temporal 3D ConvNets: New architecture and transfer learning for video classification," 2017, *arXiv:1711.08200*. [Online]. Available: https://arxiv.org/abs/1711.08200

[21] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2166–2175.

[22] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 449–458.

[23] Y. Zhu, Y. Long, Y. Guan, S. Newsam, and L. Shao, "Towards universal representation for unseen action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9436–9445.

[24] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Compressed video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6026–6035.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2048–2057.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[27] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[28] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 34–45.

[29] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6790–6800.

[30] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, "Interpretable spatio-temporal attention for video action recognition," 2018, *arXiv:1810.04511*. [Online]. Available: https://arxiv.org/abs/1810.04511

[31] J. Reizenstein, "Calculation of iterated-integral signatures and log signatures," 2017, *arXiv:1712.02757*. [Online]. Available: https://arxiv.org/abs/1712.02757

[32] T. J. Lyons and T. J. Lyons, "Sound compression: A rough path approach," in *Proc. 4th Int. Symp. Inf. Commun. Technol.*, 2005, pp. 223–228.

[33] B. Graham, "Sparse arrays of signatures for online character recognition," 2013, *arXiv:1308.0371*. [Online]. Available: https://arxiv.org/abs/1308.0371

[34] J. Diehl, "Rotation invariants of two dimensional curves based on iterated integrals," 2013, *arXiv:1305.6883*. [Online]. Available: https://arxiv.org/abs/1305.6883

[35] F. Alimoglu and E. Alpaydin, "Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition," in *Proc. 15th Turkish Artif. Intell. Artif. Neural Netw. Symp. (TAINN)*, 1996.

[36] W. Yang, L. Jin, and M. Liu, "Chinese character-level writer identification using path signature feature, DropStroke and deep CNN," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 546–550.

[37] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, and J. Chang, "Leveraging the path signature for skeleton-based human action recognition," 2017, *arXiv:1707.03993*. [Online]. Available: https://arxiv.org/abs/1707.03993

[38] G. Benjamin and J. Reizenstein. *The Iisignature Package*. Accessed: Jan. 8, 2018. [Online]. Available: https://github.com/bottler/iisignature

[39] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 716–725.

[40] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1390–1399.

[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[42] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 34–50.

[43] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 4898–4906.
[44] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[45] CoRoPa. *Computational Rough Paths Software library*. Accessed: Feb. 28, 2019. [Online]. Available: http://coropa.sourceforge.net/
[46] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.

TASWEER AHMAD received the bachelor's degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2007, and the master's degree in electronic and communication engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2009. He is currently pursuing the Ph.D. degree with the South China University of Technology, China. He was an Instructor with the Government College University, Lahore, from 2010 to 2015, and with the COMSATS Institute of Information Technology, Sahiwal Campus, Pakistan, from 2015 to 2016. He is on study leave from the COMSATS Institute of Information Technology to complete his Ph.D. degree. His current research interests include image processing, computer vision, and machine learning.

LIANWEN JIN (M'98) received the B.S. degree from the University of Science and Technology of China, Anhui, China, in 1991, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1996, where he is currently a Professor with the College of Electronic and Information Engineering. He has authored over 100 scientific articles. His research interests include handwriting analysis and recognition, image processing, machine learning, and intelligent systems. He is a member of the IEEE Computational Intelligence Society, the IEEE Signal Processing Society, and the IEEE Computer Society. He has received the New Century Excellent Talent Program of MOE Award and the Guangdong Pearl River Distinguished Professor Award.

JIALUO FENG received the bachelor's degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently pursuing the master's degree in communication and information system with the South China University of Technology, Guangzhou, China. His current research interests include handwriting analysis, document layout analysis, and machine learning.

GUOZHI TANG received the bachelor's degree in information technology from Yunnan University, Yunnan, China, in 2019. He is currently pursuing the master's degree with the South China University of Technology, Guangzhou. His research interests include deep learning and car license plate detection.