Data Augmentation For CNN-Based 3D Action Recognition on Small-Scale Datasets

Thien Huynh-The
ICT Convergence Research Center
Kumoh National Institute of Technology
Gumi, South Korea
thienht@kumoh.ac.kr

Dong-Seong Kim
ICT Convergence Research Center
Kumoh National Institute of Technology
Gumi, South Korea
dskim@kumoh.ac.kr

***Abstract*—Video-based human action recognition recently plays a vital role in many industrial applications thanks to the popularity of depth sensors. A large number of conventional approaches, which have combined handcrafted features and traditional classifiers, cannot deal with various challenges in the field such as the complexity of human actions in the realistic environment. In order to improve recognition performance by exploiting more high-level discriminative features, an efficient skeleton-based action recognition method using deep convolutional neural networks (CNNs) is studied with an image encoder to transform skeleton coordinate data to image-formed data. Since deep learning techniques are fundamentally designed for efficiently working with large datasets, the network overfitting usually occurs if training CNNs on small-scale datasets. To address this issue, a novel data augmentation technique is proposed for both the informative enrichment and overfitting prevention, wherein a skeleton sequence is depicted by manifold action images based on randomly adding some skeleton frames during the data transformation and preparation for the training set. Experimental results on several small-scale challenging datasets demonstrate that the proposed method outperforms state-of-the-art approaches in terms of action recognition accuracy.**

***Index Terms*—Data augmentation, skeleton to image transformation, human action recognition, convolutional neural networks.**

I. Introduction

Nowadays, human action recognition (HAR) have been received considerable attention for multiple areas of industrial applications. For example, human-robot interaction exploits the output context of HAR for personal robotics and industrial robotics for many kinds of purposes as surveillance, assistance, guidance, and so on [1]. Due to several limitations of color image-based approaches (e.g., illumination change, variations of background environment and subject appearance) [2], 3D skeleton-based action recognition becomes a prominent solution thanks to some essential advantages of the depth camera, compared with traditional color cameras (i.e., accurate object detection and skeleton estimation). Based on the skeleton information, an action can be recognized by analyzing spatial joint relations and temporal posture dynamics. Most of existing conventional approaches have adopted handcrafted features with traditional classifiers for recognizing simple actions, but nevertheless, their performance is trivial due to shallow features for weak discrimination [3]. Lately, numerous approaches based on recurrent neural networks (RNNs) and long short-term memory (LSTM) network have been introduced for learning hierarchical features from the raw skeleton data. Despite remarkable efforts of accuracy improvement, RNNs- and LSTM-based learning models have to face with the increment of network complexity (i.e., a vast amount number of input features). By stacking several connected convolutional layers, convolutional neural networks (CNNs) are proficient to extract and learn multiple high-level features at multi-scale representation for classification tasks. Since CNNs are designed for well working with high-dimensional data like image or video, where the deep information of an input is exhaustively captured by sweeping convolution operations through vertical and horizontal dimensions, they cannot process raw skeleton coordinate data directly. Hence, for dealing with this issue, the skeleton data should be transformed to image-form data before training CNNs. Furthermore, CNNs are usually trained on large datasets, which are difficult to collect in realistic environment, to prevent deep networks from overfitting. Therefore, one of the most efficient solutions of overfitting prevention is data augmentation that allows for network generalization.

In this paper, we present an efficient 3D action recognition approach using deep CNNs, in which the information of a skeleton sequence is depicted by an image-based representation. The 3D coordinate data of an acting performance, gathered by many skeleton frames, is transformed into image form, wherein each joint is encoded as a color pixel. For learning multiple high-level features from action images, we adopt end-to-end fine-tuning a pre-trained Inception-v3 network model. By implementing convolution and pooling operations at multi-level feature maps, the spatiotemporal correlations of action appearance are captured entirely. To prevent the deep neural network from overfitting, especially when training/fine-tuning on small-scale datasets, we extra propose a novel data augmentation mechanism, wherein a skeleton sequence or an action sample in the training set is depicted by manifold action images by randomly adding some skeleton frames in the data transformation. Following this strategy, the size of a training set is enlarged, that helps to generalize the

This research was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003.

978-1-7281-2927-3/19/$31.00 ©2019 IEEE

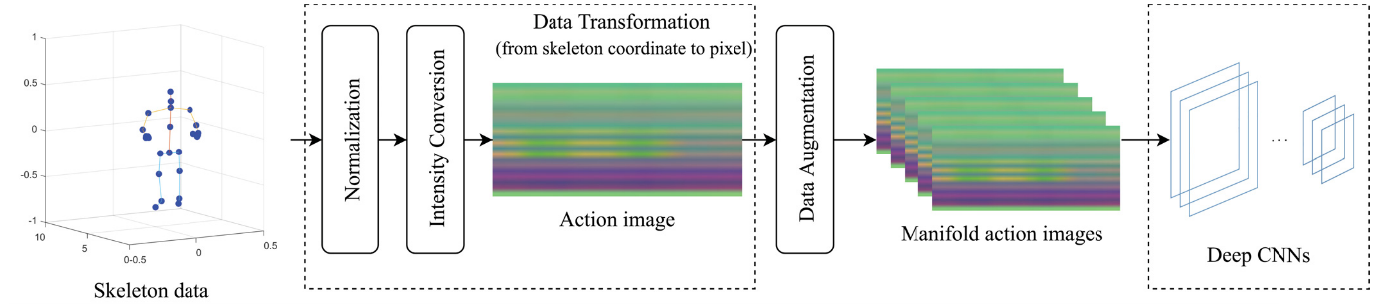


Fig. 1: The schematic overview of our proposed method for 3D action recognition. Data transformation includes two steps of value normalization and intensity conversion to encoding a skeleton sequence as an input to an action image. In the training stage, the data augmentation is adopted for preventing network from overfitting. For learning action recognition model, we fine-tune a pretrained Inception-v3 model.

recognition model which is created by end-to-end fine-tuning a pre-trained Inception-v3 [4] model. Compared with state-of-the-art approaches, including handcrafted feature-based and deep learning-based, our method with the skeleton-to-image transformation and data augmentation achieves significant improvement of recognition accuracy on such small-scale challenging datasets as MSR Action 3D and UTKinect-Action3D.

The main contribution of this paper includes: (i) the development of 3D action recognition method using deep convolutional neural networks, wherein the time series data of joint coordinate is transformed to image-formed data, (ii) the data augmentation mechanism to enlarge the training set for network overfitting prevention, and (iii) the performance analysis and comparison of state-of-the-art skeleton-based action recognition approaches.

II. Related Works

A. Handcrafted Feature-Based Action Recognition

In the last decades, a considerable amount of works have been proposed for 3D human action recognition and analysis using the skeleton information. Compared with depth and color information extracted from images, 3D skeleton shows more advantages of less memory consumption and computational cost for fully describing human posture while conveniently being obtained by depth cameras like Microsoft Kinect sensor [5]–[7]. Different feature extraction techniques and classification methods have been developed to be capable of modeling the representation of human pose for learning actions. One of the most widely used features to explain body movement is skeleton trajectory. Seidenari et al. [8] developed an efficient feature descriptor to capture the overall discriminative volumetric-temporal features from the skeleton data to estimate body movement and handle rotation challenge. By modeling 3D motions as the elements of an exceptional Euclidean group SE(3), Vemulapalli et al. [9] learned each human action as a curve of a Lie group for classification using Support Vector Machine (SVM). To discriminate the actions having a similar joint trajectory in a Riemannian manifold, Devanne et al. [10] formulated 3D points to query trajectories for calculating elastic metrics between trajectory shapes. In another work [11], the skeleton trajectory was comprehensively depicted in a Kendall's shape manifold to effectively address the data corruption issue caused by the variation of execution rates within and across subjects. A parameterization-invariant metric is learned for multiple tasks of aligning, comparing, averaging, and modeling trajectories by uniting the standard Euclidean norm with Transported Square-Root Vector Fields (TSRVFs) of trajectories. Besides the trajectory feature, several existing works exploited other advanced features, such as EigenJoint [12], structured streaming skeleton [13], and pose-transition structures [14]. Yang and Tian [12] developed a novel feature descriptor, called EigenJoint which is the mixture of static posture, motion property, and whole dynamic transition. An EigenJoint feature is established by applying Principal Component Analysis (PCA) to joint differences for noise and computation reduction. For dealing with unsegmented stream and various kinds of a gesture representation, an effective feature descriptor, namely Structured Streaming Skeleton (SSS), was studied by Zhao et al. [13], in which the structure of streaming skeleton is described as a vector of crucial attributes. Huynh-The et al. [14] recommended an effective pose-transition feature, for describing the in-frame human pose and the frame-to-frame human movement, by constructing joint distance, joint angle, and joint-plane distance. Several common geometric features (e.g., joint distance [7], joint angle [15], plane [16], and velocity [17]) extracted from skeleton data are further utilized for portraying human pose and motion. In additions, there are some other ways to delineate human action from 3D joint data, for example, encoding sparse skeleton features to another representative space using covariance matrix [18], Grassmann manifold [19], Markov random field [20] and mapping skeleton data to such multiple geometric viewpoints [21]. Handcrafted features are suitable for several computer vision tasks with quite simple and small-scale datasets thanks to easy implementation and obvious visualization, but nevertheless, their shallowness cannot discriminate complex actions in realistic environments. For example, *sit down* and *stand up* typically share several similar postures together during an action performance, that leads to misclassification of actions.

B. Deep Learning-Based Action Recognition

Recently, deep learning has been studied for many applications of the image processing and computer vision areas due to its impressive power compared with traditional machine learning techniques. Recurrent neural network (RNN) and long-short term memory (LSTM) network (a.k.a., an extension of RNN) have demonstrated their strength for skeleton-based action recognition in numerous existing works. A hierarchical RNN architecture [22] is developed for learning skeleton sequences temporally, wherein five subnets are correspondingly designed for learning the motion of five body parts. Wang et al. [23] captured the spatiotemporal contextual information from skeleton data by a novel two-stream RNNs architecture, wherein a stacked RNN and a hierarchical RNN are built for learning the spatial-dependency of joints. To address the problem of incompetently learning spatiotemporal dynamics, Veerial et al. [24] introduced a differential gating scheme for LSTM networks, in which the informative change of salient motions between two and more successive frames are exposed sufficiently. With a novel gating mechanism, a spatiotemporal LSTM network presented by Liu et al. [25] is able to learn the reliable skeleton data sequentially and also to update the long-term context information selectively. Liu et al. [26] designed Global Context-Aware Attention LSTM (GCA-LSTM) to selectively learn the most discriminative joints from the global contextual information for accuracy improvement. Wang et al. [27] handled the variety of action appearance by incorporating three specific layers (i.e., for beginning joint projection, viewpoint transformation, and spatial dropping out) into a standard RNN. Ensemble Temporal Sliding LSTM (TS-LSTM) [28], a multiple-term LSTM networks architecture, has the ability to apprehend various temporal dependencies among multiple body parts. Some other approaches have considered different kinds of multiple LSTM networks architecture for discriminative feature acquirement, the decision-level fusion of multi-geometric features learning, temporal information maintenance [29]. Although both RNNs and LSTM are able to model short- and long-term actions with performance improvement compared with traditional learning approaches, they directly work with raw skeleton coordinates as an input, that conduct shallow distinctness of action classification. Additionally, the high dimension of input feature rapidly amplifies the network complexity and may cause overfitting.

III. 3D Action Recognition

The overall architecture of our proposed 3D action recognition method is shown in Fig. 1. In this section, we introduce the data transformation for directly encoding raw skeleton data to color image, the data augmentation for enlarging the size of the training set, and then describe the recognition model learning on a pre-trained Inception-v3.

A. From Skeleton Data to Image Form

Given a skeleton sequence S consisting of N frames, where $S = \{F^t\}$ with $t \in [1, N]$, and a skeleton including m joints. Let p_i^t be the 3D coordinate of the i-th joint of a skeleton in frame F^t. Each joint is defined as $p = (x, y, z)$ in space $\Re^3$. In order to convert skeleton data to image-formed data, each joint is encoded to be a color pixel, where the x, y, and z values are normalized and then mapped to the intensity values g^R, g^G, and g^B of three color channels of Red, Green, and Blue, respectively, by a general intensity conversion function as follows

$$g = \frac{p - \min(p)}{\max(p) - \min(p)} \times (g_{\max} - g_{\min}), \tag{1}$$

where g refers to a color pixel which is formed by $g = \begin{bmatrix} g^R & g^G & g^B \end{bmatrix}$ in the deep dimension, and $[g_{\min}, g_{\max}]$ is the range of gray-scale value, usually $g_{\min} = 0$ and $g_{\max} = 255$ for a full-scale conversion to 24-bits color image. Corresponding to a sequence S constructed by N skeleton frames as follows

$$S = \begin{bmatrix} F^1 & \cdots & F^N \end{bmatrix} = \begin{bmatrix} p_1^1 & \cdots & p_1^N \\ \vdots & \ddots & \vdots \\ p_m^1 & \cdots & p_m^N \end{bmatrix}, \tag{2}$$

an action image I is generated with a defined form as follows

$$I = \begin{bmatrix} g_1^1 & \cdots & g_1^N \\ \vdots & \ddots & \vdots \\ g_m^1 & \cdots & g_m^N \end{bmatrix}. \tag{3}$$

The action image I has the size of $(N \times m)$ (a.k.a., the image resolution), where N is the width corresponding to the number of skeleton frames and m is the height corresponding to the number of body joints. It can be seen that each skeleton frame is encoded to be a pixel column of I in (3).

B. Randomly Generative Data Augmentation

Fundamentally, the strength of CNNs is the ability of learning more contextual information by multiple convolutional layer stacks and consequently improves the overall performance of the image classification task. Nevertheless, to achieve an outstanding accuracy in a comparison with traditional classification methods, they need to be trained on a large dataset as the prerequisite to avoid overfitting. Unlike such datasets for image classification task as ImageNet [30], most available 3D action recognition datasets are quite small due to some difficulties of data recording and synchronization. In this work, we propose a productive data augmentation mechanism for the overfitting prevention and informative enrichment. Ordinarily, only one action image I, which has the resolution of $(N \times m)$, is generated from one skeleton sequence, correspondingly. However, our mechanism can produce manifold action images while almost maintaining the whole action appearance. By adding some skeleton frames F, randomly selected from a sequence S, we can generate several secondary versions J from the principal action image I. This processing can be done directly on the image I by adding some pixel columns corresponding to skeleton frames (i.e., after converting joint coordinate to pixel, some pixel columns are selected for duplication). Particularly, for each

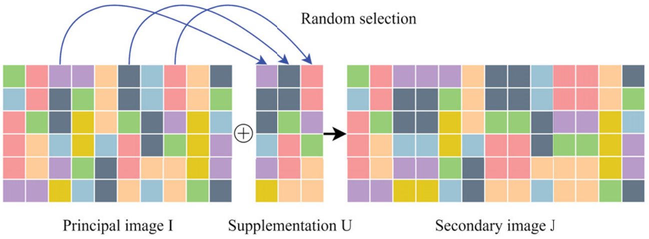


Fig. 2: An illustration of randomly generative data augmentation. Three frames corresponding to three pixel columns (called as the supplementation part U), randomly selected from the principal image I, are added for producing the secondary image J in ordering.

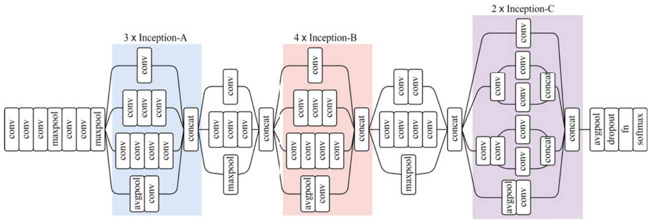


Fig. 3: The overall architecture of Inception-v3. To guarantee the network compatibility, the size of input images must be rescaled to 299 × 299, known as the pre-defined size for Inception-v3 and the last fully connected layer (a.k.a., the output layer) is modified to fit the number of action classes.

secondary image, k pixel columns corresponding to a set of frames $\{F^{n_1}, F^{n_2}, \ldots, F^{n_k}\}$ are randomly selected from I in (3) for supplementation in ordering, denoted U as follows

$$U = \begin{bmatrix} g_1^{n_1} & g_1^{n_2} & \cdots & g_1^{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ g_m^{n_1} & g_m^{n_2} & \cdots & g_m^{n_k} \end{bmatrix} \tag{4}$$

The secondary image J is constructed by a frame-wise ordering inclusion of the principal image I and the supplementation part U as follows

$$J = I \oplus U \tag{5}$$

where $\oplus$ refers to as the inclusion operation. An example of this mechanism is illustrated in Fig. 2. Following this mechanism, M images can be conducted from only one principal action image, that helps to enlarge the size of training set. It should be noticed that the image size is changed from N to $N+k$ with $(k \leq N)$. At this point, the training set is ready to use transfer learning for the action classification task.

C. Deep CNNs for Action Learning

It is known that fine-tuning a pre-trained network with transfer learning technique is more advantageous if compared with training a network from scratch with randomly initialized weights (e.g., saving training time while inheriting valuable feature set learned from a huge dataset in advance). Due to the power of pre-trained networks and also the usage convenience, most of the current CNNs-based action recognition approaches either exploit a pre-trained model to learn new patterns or develop a network with weights of convolution layers initialized from a pre-trained model. In this research, Inception-v3 [4], a state-of-the-art deep CNNs model, is fine-tuned for learning human actions on the augmented training set. Inception-v3 , which is fundamentally developed for the image classification task in computer vision, has achieved a very impressive performance in terms of classification accuracy. Inspired by GoogleNet (a.k.a., Inception-v1), Inception-v3 is improved with the idea of convolution factorization (the use of smaller and asymmetric convolutions), regularization of auxiliary classifier, and efficient grid size reduction. In the view of overall architecture, Inception-v3 is designed with three major different types of inception module as shown in Fig. 3. In details, Inception-v3 has 3 Inception-A modules (where a sequence of 3 × 3 convolutional layers is adopted to reduce the number of parameters), 4 Inception-B modules (where several asymmetric convolutional layers 1 × 7 and 7 × 1 are connected to significantly save computational cost), and 2 Inception-C modules (where filter bank outputs are expanded to generate high dimensional sparse representation). By passing the asymmetric filters along the input (known as the output of previous layer) vertically, the network learns the temporal information of skeleton dynamics thoroughly.

IV. Experiments

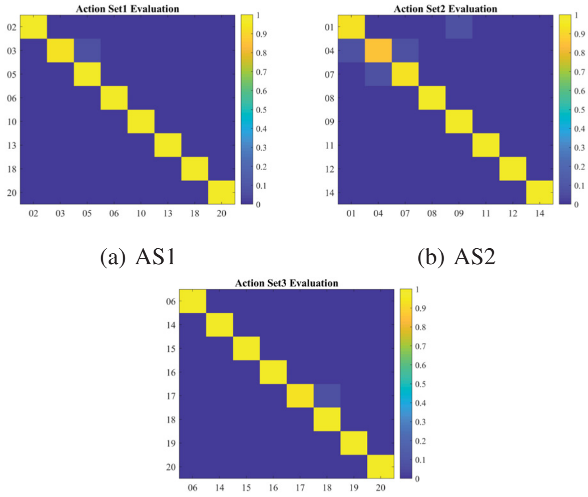
A. Datasets

MSR Action3D [5]: There are totally 557 sequences of 20 different single actions, in which all actors perform each action 2-3 times. The skeleton data is recorded at 15 fps (frames per second) by a depth camera that provides 20 joints of each skeleton. Wide-ranging body movements of several sport-oriented actions and unstable motion velocity of some body parts are the challenges of MSR Action3D The dataset is evaluated with the cross-subject protocol, where half of subjects are for training and remaining subjects are for testing.

UTKinect-Action3D [6]: This dataset, collected by using a single stationary Kinect sensor, has 200 sequences of 10 human-object interactive action classes performed by 10 participants. Each action is captured twice for every subject and each skeleton detected in a frame has 20 joints. As a challenge, some body parts are failed in tracking their coordinates due to occlusion. We follow the standard leave-one-out cross validation configuration, where one sequence is used for testing and reaming samples are for training.

B. Experimental Setup

In the data augmentation, we empirically set $M = 30$ for the number of secondary images with $k = 5$ frames randomly selected for supplementation. Regarding the hyper-parameter configuration for end-to-end fine-tuning Inception-v3, the stochastic gradient descent with momentum (SGDM) optimizer is tuned. With the mini-batch size 16, the network is fine-tuned in 30 epochs with the learning rate initialized at



(a) AS1

(b) AS2

(c) AS3

Fig. 4: Confusion matrices on three MSR Action3D subsets (the class IDs from 01 to 20 stand for high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup & throw.)

0.001 and reduced 90% after every 15 epochs. All experiments are implemented on a system using one NVIDIA GeForce GTX 1080Ti card.

C. Results and Comparisons

MSR Action3D dataset: Following the standard evaluation protocol, the dataset is split into three actions subsets, including Action Set1 (AS1), Action Set2 (AS2), Action Set3 (AS3). The recognition results are correspondingly reported by three confusion matrices in Fig.4. It is observed that some samples of *hammer* are misclassified to *forward punch* in AS1, two actions *hand catch* and *draw x* are confused together in AS2, and there are some errors of *tennis* actions recognition in AS3. Furthermore, we compare the proposed method with other state-of-the-art approaches in terms of recognition accuracy, where the comparison results are summarized in Table I. Our method achieves the highest average accuracy with 98.2%, that is much better than such handcrafted feature-based approaches as Histogram of 3D joints [6], Structured Streaming Skeleton [13], EigenJoints [12] and such deep learning-based approaches as dRNN [24], HBRNN-L [22].

TABLE I: Comparison on MSR Action3D Dataset

Method	Accuracy (%)
Histogram of 3D joints [6]	79.0
Structured Streaming Skeleton [13]	81.7
EigenJoints [12]	83.3
Joint angle similarities [15]	83.5
TSRVFs [11]	89.0
Grassmann Manifold [19]	90.2
Covariance descriptors [18]	90.5
Lie Group [9]	90.9
Pattern-based M^{4}IL [20]	91.0
dRNN [24]	92.0
Riemannian Manifold [10]	92.1
Combined features with RFs [16]	94.3
HBRNN-L [22]	94.5
TriViews + STIPs [21]	94.9
PAM + Pose-Transition Feature [14]	97.1
Ensemble TS-LSTM v2 [28]	97.2
Proposed	**98.2**

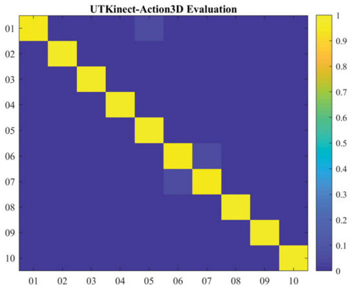


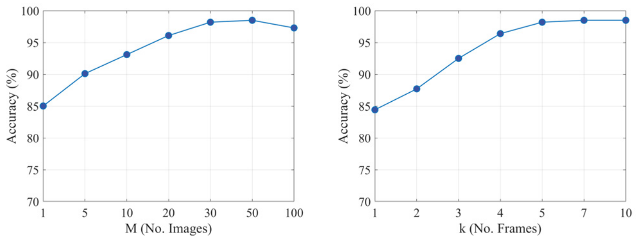
Fig. 5: Confusion matrices of the proposed method on the UTKinect-Action3D dataset (the class IDs from 01 to 10 stand for walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands).

TABLE II: Comparison on UTKinect-Action3D Dataset

Methods	Accuracy (%)
Grassmann Manifold [19]	88.5
Histogram of 3D joints [6]	90.9
Riemannian Manifold [10]	91.5
TriViews + STIPs [21]	92.9
MLSTM + Weight Fusion [29]	96.0
Ensemble TS-LSTM v2 [28]	97.0
ST-LSTM [25]	97.0
PAM + Pose-Transition Feture [14]	97.0
Lie Group [9]	97.1
Proposed	**98.5**
GCA-LSTM (stepwise) [26]	99.0

UTKinect-Action3D dataset: The recognition result of our method is detailed reported by the confusion matrix in Fig. 5. Most of actions are recognized precisely, except *throw* and *push* are confused together. The performance comparison is additionally delivered in Table II, wherein the proposed method presents the second best result (less than GCA-LSTM (stepwise) [26] 0.5%). Besides augmenting training sets, some much informative frames in a sequence are duplicated for increasing action discrimination, that helps to improve accuracy.

Additionally, the influence of parameters (i.e., the number of secondary images M and the number of randomly selected frames k) in the data augmentation process on the overall recognition performance is benchmarked, where the results are plotted in Fig. 6. Compared with non-augmentation (i.e., $M = 1$), the recognition accuracy is significantly improved at $M = 30$. However, a great value of M may reduce accuracy due to more duplicated action images. Similarly, a larger value of k conducts a higher performance thanks to the diversity of



(a) (b)

Fig. 6: Overall recognition rate of MSR Action3D on different parameter configurations: (a) Number of generated secondary images and (b) Number of randomly selected frames.

augmented training set.

V. Conclusion

In this paper, we have presented an efficient 3D action recognition method using deep convolutional neural networks, in which a novel encoder to convert body joint coordinate data to image-formed data is developed. To prevent the network from overfitting when training CNNs on small-scale datasets, we also introduced an efficient data augmentation mechanism by adding some skeleton frames randomly selected from a skeleton sequence. For learning action recognition model, a pre-trained Inception-v3 model is fine-tuned end-to-end on the augmented training set. Experimental results on the MSR Action 3D and UTKinect-Action3D datasets demonstrate that our method mostly outperforms other state-of-the-art action recognition approaches, including handcrafted feature-based and deep learning-based, in terms of accuracy.

References

[1] T. Huynh-The, C.-H Hua and D. Kim, "Encoding Pose Features To Images With Data Augmentation For 3D Action Recognition," *IEEE Trans. Ind. Informat* [Early Access].

[2] N. A. Tu, T. Huynh-The, K. U. Khan and Y. Lee, "ML-HDP: A Hierarchical Bayesian Nonparametric Model for Recognizing Human Actions in Video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 800-814, March 2019.

[3] T. Huynh-The, B.-V. Le, S. Lee, Y. Yoon, "Interactive activity recognition using pose-based spatio–temporal relation features and four-level Pachinko Allocation Model," *Inf. Sci.*, vol. 369, pp. 317-333, 2016.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, Las Vegas, NV, 2016, pp. 2818-2826.

[5] W. Li, Z. Zhang and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comp. Vis. Pattern Recogn. Workshops*, San Francisco, CA, 2010, pp. 9-14.

[6] L. Xia, C. Chen and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comp. Vis. Pattern Recogn. Workshops*, Providence, RI, 2012, pp. 20-27.

[7] T. Huynh-The, B.-V. Le and S. Lee, "Describing body-pose feature - poselet - activity relationship using Pachinko Allocation Model," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, Budapest, 2016, pp. 40-45.

[8] L. Seidenari, V. Varano, S. Berretti, A.D. Bimbo, P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. Workshops*, Portland, OR, 2013, pp. 479–485.

[9] R. Vemulapalli, F. Arrate and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, Columbus, OH, 2014, pp. 588-595.

[10] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi and A. Del Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340-1352, July 2015.

[11] B. B. Amor, J. Su and A. Srivastava, "Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 1-13, 1 Jan. 2016.

[12] X. Yang and Y. Tian, "Effective 3D action recognition using Eigen-Joints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2-11, 2014.

[13] X. Zhao, X. Li, C. Pang, Q. Z. Sheng, S. Wang and M. Ye, "Structured Streaming Skeleton - A New Feature for Online Human Gesture Recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1s, pp. 22:1–22:18, Oct. 2014.

[14] T. Huynh-The, C.-H. Hua, N. A. Tu, T. Hur, J. Bang, D. Kim, M. B. Amin, B. H. Kang, H. Seung, S.-Y. Shin, E.-S. Kim, S. Lee, "Hierarchical topic modeling with pose-transition feature for action recognition using 3D skeleton data," *Inf. Sci.*, vol. 444, pp. 20-35, 2018.

[15] E. Ohn-Bar and M. M. Trivedi, "Joint Angles Similarities and HOG^2 for Action Recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. Workshops*, Portland, OR, 2013, pp. 465-470.

[16] Y. Zhu, W. Chen and G. Guo, "Fusing Spatiotemporal Features and Joints for 3D Action Recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. Workshops*, Portland, OR, 2013, pp. 486-491.

[17] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Comp. Soc. Conf. Comp. Vis. Pattern Recogn. Workshops*, Providence, RI, pp. 28-35, 2012.

[18] M. E. Hussein, M. Torki, M. A. Gowayyed and Motaz El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. Twenty-Third Int. Joint Conf. Artif. Intell. (IJCAI '13)*, Beijing, 2013, pp. 2466-2472.

[19] R. Slama, H. Wannous, M. Daoudi and A. Srivastava, "Accurate 3D action recognition using learning on the Grassmann manifold," *Pattern Recogn.*, vol. 48, no. 2, pp. 556-567, 2015.

[20] X. Cai, W. Zhou, L. Wu, J. Luo and H. Li, "Effective Active Skeleton Representation for Low Latency Human Action Recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141-154, Feb. 2016.

[21] W. Chen, G. Guo, "TriViews: A general framework to use 3D depth data effectively for action recognition," *J. Vis. Commun. Image Represent.*, vol. 26,pp. 182-191, 2016.

[22] Yong Du, W. Wang and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, Boston, MA, 2015, pp. 1110-1118.

[23] H. Wang and L. Wang, "Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, Honolulu, HI, 2017, pp. 3633-3642.

[24] V. Veeriah, N. Zhuang and G. Qi, "Differential Recurrent Neural Networks for Action Recognition," in *Proc. IEEE Conf. Comp. Vis.*, Santiago, 2015, pp. 4041-4049.

[25] J. Liu, A. Shahroudy, D. Xu and G. Wang, "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition," in *Proc. European Conf. Comp. Vis.*, Amsterdam, 2016, pp. 816-833.

[26] J. Liu, G. Wang, L. Duan, K. Abdiyeva and A. C. Kot, "Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586-1599, April 2018.

[27] H. Wang and L. Wang, "Learning robust representations using recurrent neural networks for skeleton based action classification and detection," in *Proc. IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, Hong Kong, 2017, pp. 591-596.

[28] I. Lee, D. Kim, S. Kang and S. Lee, "Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks," in *Proc. IEEE Int. Conf. Comp. Vis.*, Venice, 2017, pp. 1012-1020.

[29] S. Zhang et al., "Fusing Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330-2343, Sept. 2018.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.