LEARNING GEOMETRIC FEATURES WITH DUAL-STREAM CNN FOR 3D ACTION RECOGNITION

Thien Huynh-The, Cam-Hao Hua, Nguyen Anh Tu, Dong-Seong Kim

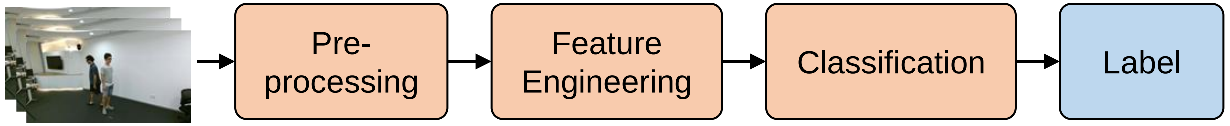
Kumoh National Institute of Technology, S. Korea
Kyung Hee University, S. Korea
Nazarbayev University, Kazakhstan

Introduction

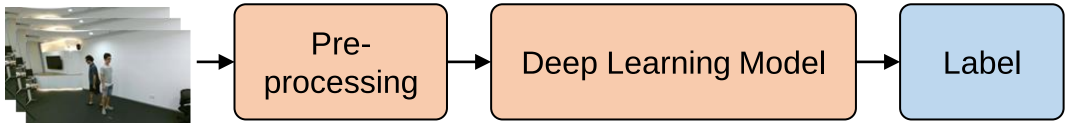
- Human action recognition (HAR) is to identify indoor/outdoor actions that occur in video sequences.
- The key of numerous visual applications
  - Video-based surveillance
  - Daily living assistant
  - Robotic control
  - Healthcare & wellness
  - and other civil and military apps.
- Some critical challenges
  - Viewpoint variation
  - Variant motion velocity
  - Variety of single action and multi-subject interaction in the realistic condition.

Sampe frames of "NTU RGB+D 120" dataset

Background



Conventional machine learning-based approach



Innovative deep learning-based approach

ML-based Human Action Recognition using RGB images

- Pre-processing
  - Object detection
  - Object localization and segmentation
  - And tracking
- Feature engineering
  - Feature extraction: SIFT, HOG, and etc.
  - Feature selection: filter, wrapper, and etc.
- Model learning (classification)
  - Supervised learning: decision tree, support vector machine
  - Unsupervised learning: k-means clustering

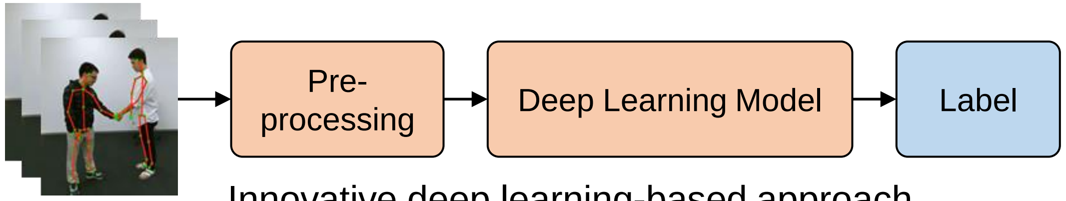
Limitation: Extremely sensitive to illumination, occlusion, and subject appearance.

DL-based Human Action Recognition using RGB images

- Deep learning covers the functionalities of feature engineering and classification
- Advantages of DL (in comparison with ML)
  - Excellent performance on big dataset
  - Without expert knowledge of feature engineering
- Some modern backbone CNNs
  - VGG-16, VGG-19
  - GoogleNet, Inception-v3
  - ResNet, DenseNet

Limitation: Performance is mostly vulnerable by environment and subject's stuffs.

Background



Innovative deep learning-based approach

DL-based Human Action Recognition using 3D Skeleton Data

- Body and limb key-points based skeleton contains higher-level context of subject appearance compared with RGB
- Development and popularity of depth camera
  - More accurate than regular camera with depth information
  - Pose estimation algorithm integrated inside such depth camera like Kinect sensor
- Technical challenges
  - Variable scale (subject-vs-camera distance)
  - Variable viewpoint (camera setup)
  - Intra-class action variation
- Deep learning for 3D action recognition has attracted recently.

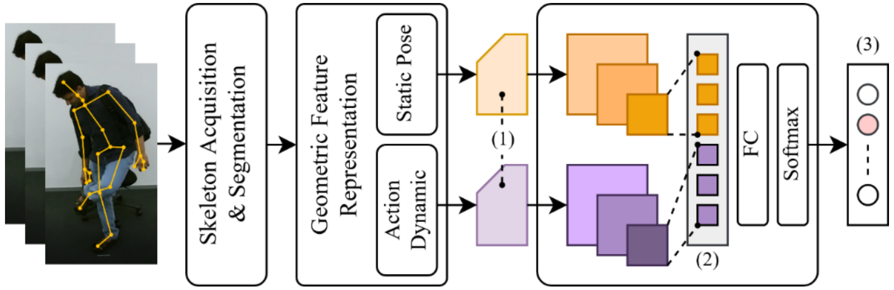
State-of-the-art review

- DL architecture
  - Recurrent Neural Networks/Long Short-Term Memory
  - Convolutional Neural Network
- RNN/LSTM-based HAR approaches
  - Bidirectional RNN [5]
  - Global context-aware attention LSTM [6]
  - Two-Stream Attention LSTM [7]
- CNN-based HAR approaches
  - Skeleton visualization [23]
  - PoF2I + inception-v3 [15]

Limitation

- Incapability of fully covering an entire action sequence
- Lack of concurrently learning spatiotemporal static pose and body transition.

Methodology



The overall action recognition framework with a dual-stream CNN for learning geometric static pose and action dynamic. Annotation: (1)-geometric feature maps, (2)-feature concatenation, and (3)-predicted class scores.

Introduce a CNN with two convolutional streams, namely Deep Geometric Pose-Transition Dual-Stream Network (DGPoT-2^{S}CNN)

- Concurrently learning spatial static pose and temporal action dynamic of an entire sequence.

☞ The joint relation within a skeleton and the body association between two skeletons in consecutive frames are comprehensively encapsulated.

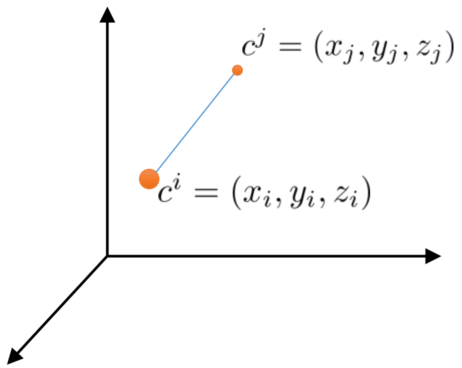
Contribution

- Introduction of an efficient DL-based method for visual-based HAR
- Performance benchmark on NTU RGB+D 120 as the largest and most challenging dataset of action recognition
- Ablation study with various CNN backbones
- Method comparison in terms of recognition accuracy.

Technical highlights

- Static pose + action dynamic ← 3D geometric features of joint-to-joint distance
- Calculation and representation of two geometric feature maps
- Design of a dual-stream CNN architecture with pre-trained inception-v3 for transfer learning
- High compatibility with different pre-trained networks.

Methodology



Measure the distance metric by means of projecting 3D points on three original planes

$$\begin{aligned} \tau^{i,j}_{x=0} &= \left\| c^i_{x=0} - c^j_{x=0} \right\| \\ \tau^{i,j}_{y=0} &= \left\| c^i_{y=0} - c^j_{y=0} \right\| \\ \tau^{i,j}_{z=0} &= \left\| c^i_{z=0} - c^j_{z=0} \right\| \end{aligned} \quad (1)$$

where the general distance in 3D Euclidean space

$$\tau^{i,j} = \left\| c^i - c^j \right\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (2)$$

Note: The triple-value feature is captured for all individual subjects and for interactions

Two categories of geometric feature of

- Human pose representation in the spatial domain

$$\mathbb{F}^{\mathrm{S}} = \left[\boldsymbol{\tau}^{i,j,t=1,\ldots,T}_{k=1,\ldots,K} \right] \quad (3)$$

- Body transition description in the temporal domain

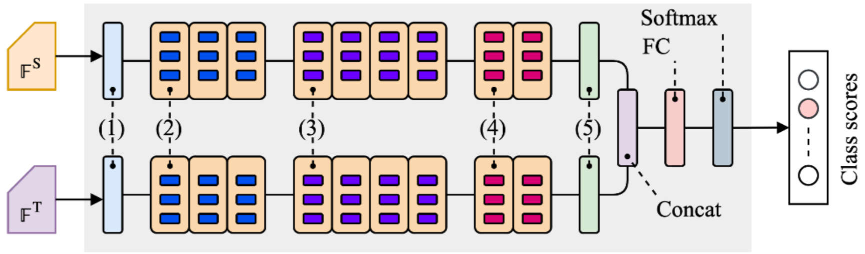
$$\mathbb{F}^{\mathrm{T}} = \left[\boldsymbol{\tau}^{i,j,\Delta t=2,\ldots,T}_{k=1,\ldots,K} \right] \quad (4)$$

Note: Two handcrafted geometric feature maps are written by 3D matrices of volume $K \times T \times 3$

K: No. features extracted from the skeleton data of one frame

T: No. frames

3: Triple-value distance feature



The compact view of the dual-stream CNN architecture with the convolutional flow initialized by Inception-v3. Annotation: (1)-convolutional layers, (2)-inception module A, (3)-inception module B, (4)-inception module C, and (5)-global average pooling layer.

- Deployment of two convolutional streams of pre-trained inception-v3
  - Two inception-v3 models assembled in parallel
  - High-level global pooling features concatenated at the end
- Benefits
  - Learning the intrinsic relationships between multiple joints of intra-subject and inter-subject skeletons
  - Learning the spatial in-frame joint correlations and the temporal frame-wise body associations.
  - Compatibility with different CNN backbone architectures, such as VGG, ResNet, and DenseNet.

Experimental Results

Dataset – NTU RGB+D 120

- 120 of single actions, human-object interactions, and human-human interactions
- 114,480 video sequences of 106 subjects collected via 32 location configurations

Evaluation protocols

- Cross-subject (53/106 subjects for training, remains for testing)
- Cross-setup (16/32 setups for training, remains for testing)

Training parameters

- Stochastic gradient descent with momentum (SGDM) optimizer
- No. fine-tuning epochs: 20
- Mini-batch size: 64
- Learning rate: 0.01 (dropped 90% after 10 epochs)

Performance is measured by recognition rate (%)

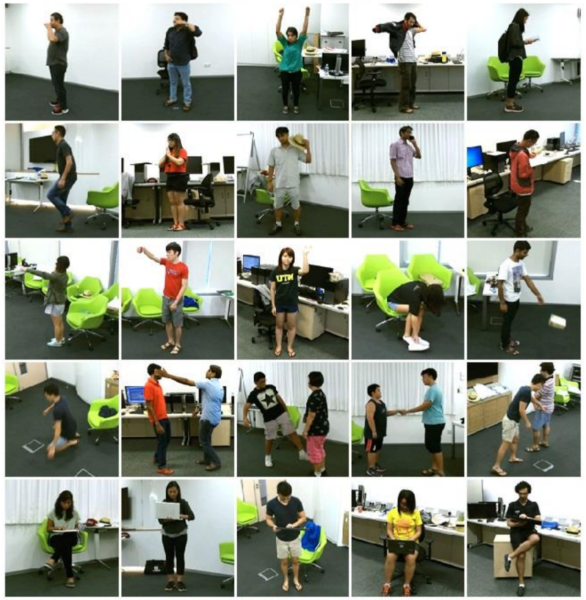
Two experiments are delivered

- Ablation study
- Method comparison

Ablation study

- Single stream with either static pose feature or action dynamic feature
- Different CNN backbones used in dual-stream network

Sampe frames of "NTU RGB+D" dataset



Sampe frames of "NTU RGB+D 120" dataset

Experimental Results

Accuracy (%) with Different Backbone Networks

Configurations	GoogleNet		VGG-19		ResNet-101		DenseNet-201		Inception-v3	
	C-Sub	C-Set	C-Sub	C-Set	C-Sub	C-Set	C-Sub	C-Set	C-Sub	C-Set
Single stream $\mathbb{F}^S$	69.63	71.69	72.39	74.05	73.79	76.22	73.58	75.79	76.06	77.95
Single stream $\mathbb{F}^T$	69.14	71.79	72.32	73.99	73.64	75.91	73.44	76.00	74.63	77.55
DGPoT-2^{S}CNN	71.77	73.85	73.36	74.92	74.20	76.48	74.15	76.39	76.33	78.91

Method Comparison

Methods	C-Sub	C-Set
Part-Aware LSTM [22]	25.5	26.3
Dynamic Skeleton [1]	50.8	54.7
Internal Feature Fusion [8]	58.2	60.9
GCA-LSTM [6]	58.3	59.2
Skeleton Visualization [23]	60.3	63.2
Two-Stream Attention LSTM [7]	61.2	63.3
Multi-Task CNN with RotClips [14]	62.2	61.8
Body Pose Evolution Map [10]	64.6	66.9
PoF2I + Inception-v3 [15]	67.2	68.8
DGPoT-2^{S}CNN (GoogleNet)	71.8	73.9
DGPoT-2^{S}CNN (VGG-19)	73.4	74.9
DGPoT-2^{S}CNN (ResNet-101)	74.2	76.5
DGPoT-2^{S}CNN (DenseNet-201)	74.2	76.4
DGPoT-2^{S}CNN (Inception-v3)	76.3	78.9

Analysis

- Static pose is more important than action dynamic
- Fusing deep features via dual-stream strategy improve recognition rate properly
- Cross-subject is more challenging than cross-setup
- GoogleNet reports the worst accuracy, while Inception-v3 yields the best score.
- Outperformance of both state-of-the-art CNNs-based and LSTM-based approaches*

* The comparison is made with 3D skeleton-based HAR methods

Conclusion

DGPoT-2^{S}CNN

- Calculation of joint-to-joint distance metric to explain the static pose and action dynamic
- Development of a dual-stream CNN with pre-trained Inception-v3
- Mining intrinsic intra-subject joint relationships and inter-subject skeleton associations in the spatiotemporal dimension.
- Compatibility of various CNN backbones
- Outperformance of existing LSTM- and CNN-based approaches

Limitation

- Sensitive to diverse subject appearance

Future

- Enhancement with more robustly geometric metrics for pose description and action transition explanation

Reference

[1] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov 2017.

[2] N. A. Tu, T. Huynh-The, K. U. Khan, and Y. Lee, "Mlhdp: 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, and 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] T. Huynh-The et al., "Hierarchical topic modeling with pose-transition feature for action recognition using 3d skeleton data," *Inf. Sci.*, vol. 444, pp. 20 – 35, 2018.

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

[6] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *2017 IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, July 2017, pp. 3671–3680.

[7] 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.

[8] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec 2018.

[9] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, June 2017.

[10] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *2018 IEEE/CVF Conf. Comp. Vis. Pattern Recogn.*, June 2018, pp. 1159–1168.

[11] C. Li, S. Sun, X. Min, W. Lin, B. Nie, and X. Zhang, "End-to-end learning of deep convolutional neural network for 3d human action recognition," in *2017 IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, July 2017, pp. 609–612.

[12] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2016 IEEE Comp. Vis. Pattern Recogn. Workshops (CVPRW)*, July 2017, pp. 1623–1631.

[13] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Syst.*, vol. 158, pp. 43 – 53, 2018.

[14] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2842–2855, June 2018.

[15] T. Huynh-The, H. Hua-Cam, and D. Kim, "Encoding pose features to images with data augmentation for 3d action recognition," *IEEE Trans. Ind. Informat.*, pp. 1–1, 2019.

Thank you for your attention

Dr. Thien Huynh-The
Kumoh National Institute of Technology
Email: thienht@kumoh.ac.kr