Architectures d'apprentissage profond pour la reconnaissance d'actions humaines dans des séquences vidéo RGB-D monoculaires. Application à la surveillance dans les transports publics.

Présentée et soutenue par Huy Hieu PHAM

En vue de l'obtention du grade de Docteur de L'Université de Toulouse

Directeurs de thèse : Denis KOUAMÉ et Louahdi KHOUDOUR

Encadrant : Alain CROUZIL

Membres du comité de suivi : Sergio A Velastin et Pablo Zegers

Toulouse, France, le 19 septembre 2019

Table of contents

1. Introduction to Human Action Recognition in Videos

Human action recognition in RGB-D videos

- A human action can be defined as a spatio-temporal sequence of human body movements that has starting and ending temporal points.
- The main goal of a video-based action recognition system is to automatically analyze ongoing video streams provided by unknown cameras to determine which human actions occur in these videos.
- In computer vision, human action recognition is an automatic labelling process that attempts to label each action with a corresponding name (verb or noun).

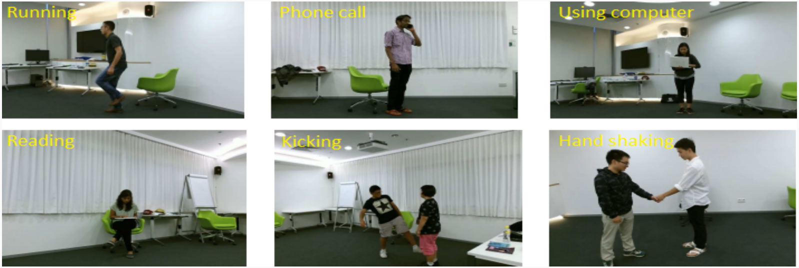


Figure 1: Human action recognition systems usually focus on recognizing daily-life actions.

Motivation

Human action recognition in videos plays a key role in many different intelligent video analysis systems.

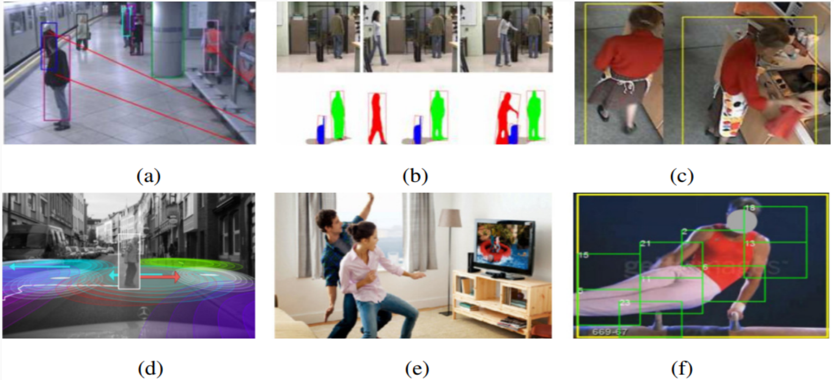


Figure 2: (a) Recognizing actions in intelligent transport systems; (b) stealing detection; (c) remote monitoring service for elderly persons; (d) pedestrian path prediction in self-driving cars; (e) action recognition in the entertainment industry; and (f) action localization in sports videos.

Research challenges

- Large intra-class variations
- Fuzzy boundaries between classes
- Viewpoint variations, camera motion, etc.

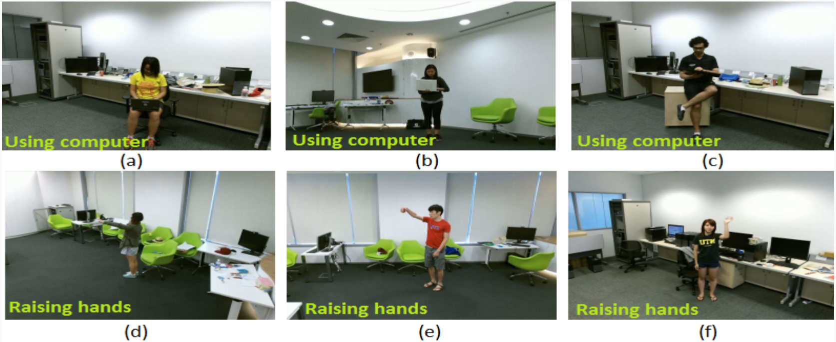


Figure 3: The large intra-class variation and the variety in camera views are two enormous challenges in recognizing human actions.

Research problem and objective

Research problems

- How to recognize correctly what humans do in unknown videos?
- How to learn efficiently spatio-temporal features of human motions by deep convolutional neural networks (D-CNNs)?
- How to build an efficient deep learning framework (*i.e.* higher prediction performance and faster prediction speed) for human action recognition from RGB-D data?

Human action recognition in RGB-D videos

Objective

Developing and validating a deep learning-based approach to analyse human behaviors from RGB-D sequences. Application to public transport monitoring.

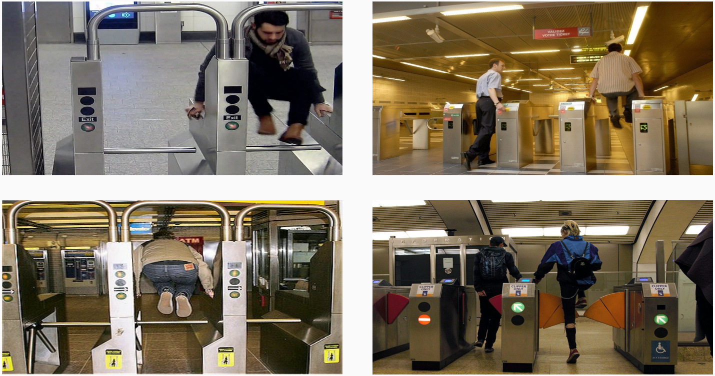


Figure 4: Detecting abnormal behaviors on video surveillance in public transport.

2. State-of-the-Art in Video-based Human Action Recognition

Literature review

Before 2015, traditional approaches for human action recognition in videos are often based on hand-crafted features → usually leads to data dependent methods.

Figure 5: A typical method for video-based human action recognition.

Literature review

Starting from 2015, deep learning-based approaches became a new state-of-the-art in the human action recognition[1].

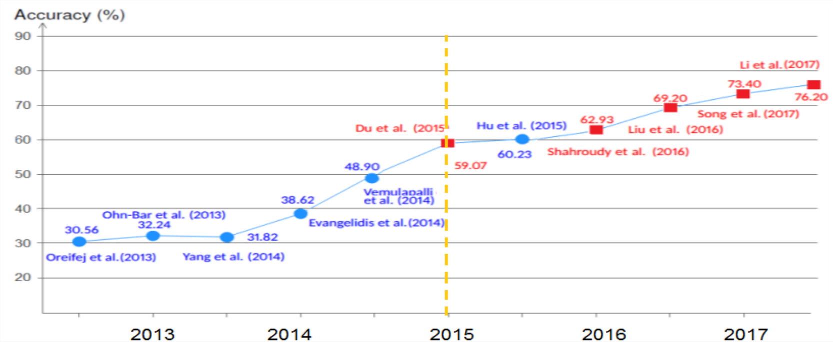


Figure 6: Hand-crafted feature vs. deep learning on the NTU-RGB+D dataset. The traditional approaches are marked with circles, deep learning based approaches are marked with squares.

[1] Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, Sergio A. Velastin, "*Exploiting deep residual networks for human action recognition from skeletal data*" – CVIU 2018.

Literature review

Several important deep learning-based architectures for human action recognition

Architecture 1: 3D Convolutional Neural Network (3D-CNN)[2].

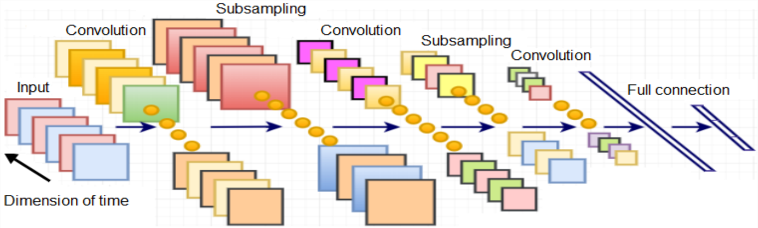


Figure 7: A 3D CNN architecture for human action recognition in which 3D convolutions in the convolution stages of CNNs to compute features from both spatial and temporal dimensions.

[2] Ji, Shuiwang *et al.* "*3D convolutional neural networks for human action recognition*". TPAMI, vol. 35, pp. 221–231, 2015.

Literature review

Several important deep learning-based architectures for human action recognition

Architecture 2: Two-stream CNN[3].

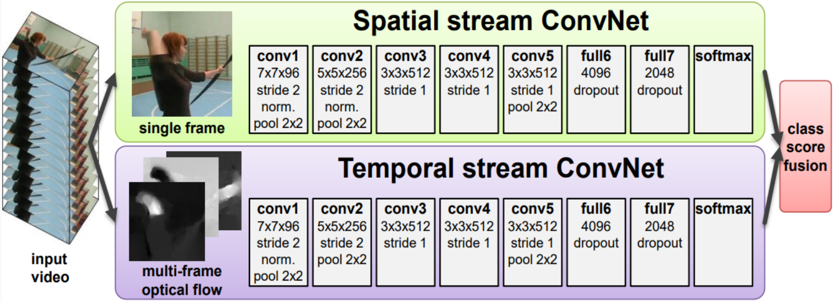


Figure 8: Two-stream CNN framework for human action recognition in videos.

[3] Karen Simonyan and Andrew Zisserman. "*Two-stream convolutional networks for action recognition in videos*". In: NIPS, 2014.

State-of-the-art in Video-based Human Action Recognition

Limitations of previous works and the focus of our study.

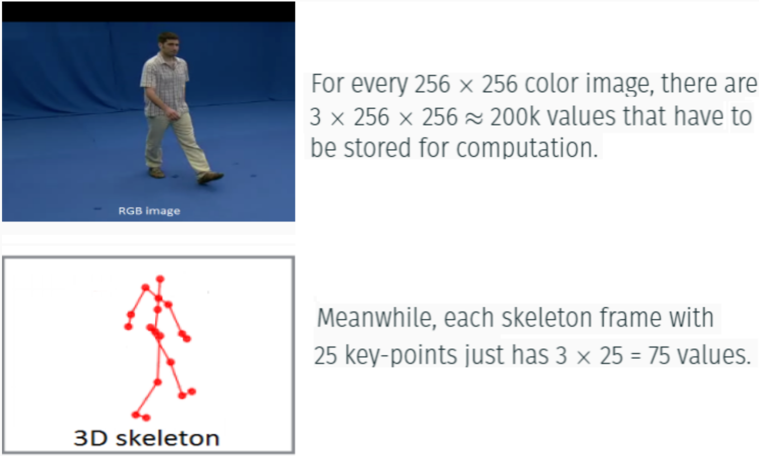


Figure 9: Dimensionality of data: A comparison between RGB data and skeletal data.

3. A New Deep Learning Framework for Action Recognition from Skeleton Sequences

Approach 1: Building a skeleton-based action recognition method using deep neural networks

The proposed method is based on two key ideas:

- Encoding each skeleton sequence into a single color image (called "*action maps*").
- Training state-of-the-art CNN models to learn and classify the action maps.

Proposed method

Motivations

- Human actions can be correctly represented through the skeleton movements.
- The spatio–temporal dynamics of skeleton sequences can be transformed into color images, which can be effectively learned by representation learning models such as D-CNNs.
- Training deep learning models on skeletal data is much faster than training on RGB and depth streams.
- Recent research results indicate that CNNs have achieved outstanding performances in many image recognition tasks.

Proposed method

Approach 1: A two-step learning method for skeleton-based human action recognition with deep convolutional neural networks

Step 1: Encoding skeleton sequences into color images.

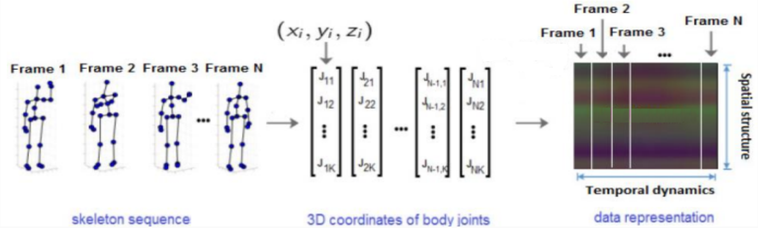


Figure 10: Illustration of the color encoding process.

- Using a transformation function to rescale the joint coordinates into $[0, 255]$.
- Concatenating all transformed skeleton frames over time.

Proposed method

Approach 1: A two-step learning method for skeleton-based human action recognition with deep convolutional neural networks

Step 1: Encoding skeleton sequences into color images.

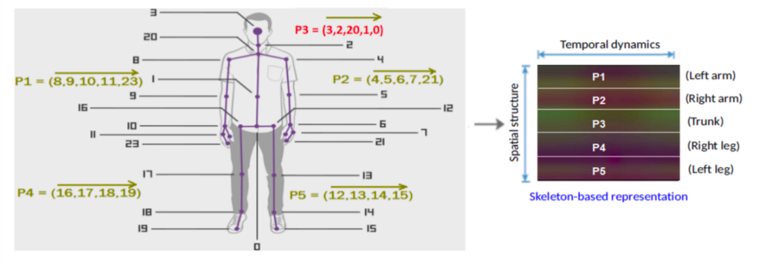


Figure 11: Arranging pixels in color images according to the human body physical structure. This helps to keep the local motion characteristics and to generate more discriminative features in image-based representations.

Proposed method

Approach 1: A two-step learning method for skeleton-based human action recognition with deep convolutional neural networks

Step 1: Encoding skeleton sequences into color images.

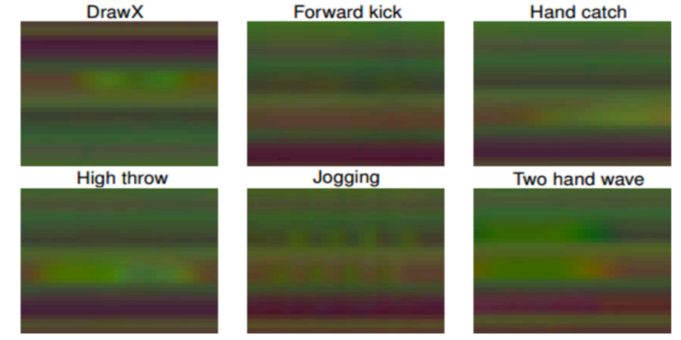


Figure 12: Output of the encoding process obtained from some samples of the MSR Action3D dataset.

Proposed method

Approach 1: A two-step learning method for skeleton-based human action recognition with deep convolutional neural networks

Step 2: Designing and training D-CNNs to learn and classify actions via the color-coded representation.

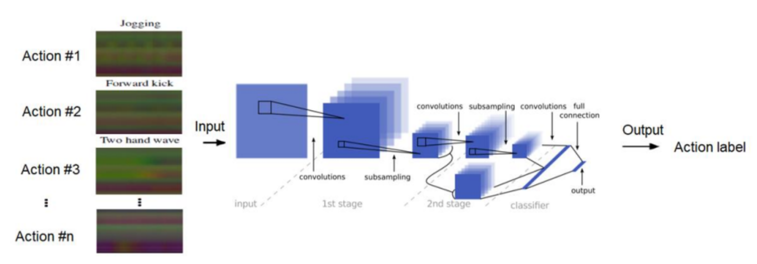


Figure 13: Human action recognition using D-CNNs and the proposed skeleton-based representation.

Network design

ResNet[4] has designed and trained for recognition task. The presence of an identity function $id(x)$ helps ResNet to prevent overfitting and degradation phenomena.

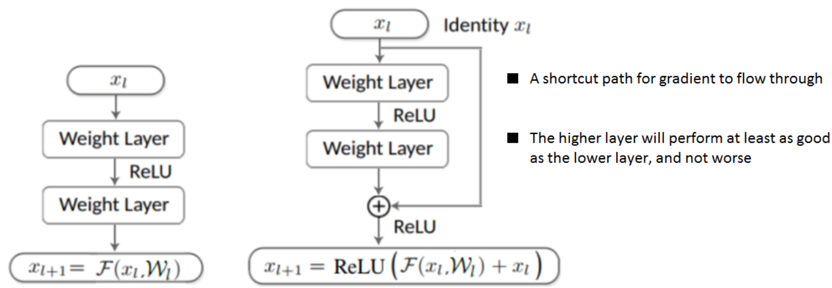


Figure 14: Information flow executed by a traditional CNN (left) and by a ResNet unit (right).

[4] He, Kaiming, *et al.* "*Deep residual learning for image recognition.*" CVPR, 2016.

Network design

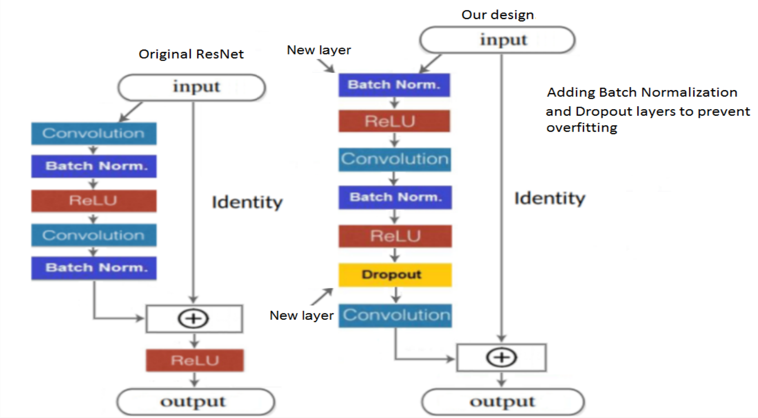


Figure 15: A ResNet building unit that was proposed in the original paper (**left**). Our proposed ResNet building (**right**). The symbol $\oplus$ denotes element-wise addition.

Experiments

Datasets and settings: The proposed method was evaluated on three public datasets: MSR Action3D[5], KARD[6], and NTU-RGB+D[7].

• **MSR Action3D dataset**: 20 actions, 557 skeleton sequences. Three subsets: AS1, AS2, and AS3.

• **KARD dataset**: 18 actions, 540 skeleton sequences. Three subsets: Action Set 1, Action Set 2, and Action Set 3.

• **NTU-RGB+D dataset**: the largest RGB-D dataset currently available with 56,000+ videos, 60 action classes. Two evaluation settings: Cross-Subject and Cross-View.

---

[5] Li *et al.*. "*Action recognition based on a bag of 3D points*". In CVPR, 2010.
[6] Gaglio *et al.* "*Human activity recognition process using 3D posture data*". IEEE Trans. Hum.-Mach. Syst. 2015.
[7] Shahroudy *et al.* "*NTU-RGB+D: A large scale dataset for 3D human activity analysis*" in CVPR, 2016.

Experiments

Training methodology

- All networks are designed for the acceptable images with the size of 32×32 pixels as input features and classifying them into n categories corresponding to n action classes in each dataset.
- Using a mini-batch of 128 samples.
- The learning rate starts from 0.01 for the first 75 epochs, 0.001 for the next 75 epochs and 0.0001 for the remaining 50 epochs.
- Data augmentation techniques (*i.e.* random cropping, flipping) were used to reduce overfitting.

Experimental results

Model	Cross-Subject	Cross-View
Original-ResNet-20	73.90%	80.80%
Original-ResNet-32	75.40%	81.60%
Original-ResNet-44	75.20%	81.50%
Original-ResNet-56	75.00%	81.50%
Original-ResNet-110	73.80%	80.00%
Proposed-ResNet-20	76.80%	83.80%
Proposed-ResNet-32	76.70%	84.70%
Proposed-ResNet-44	77.20%	84.80%
Proposed-ResNet-56	**78.20%**	**85.60%**
Proposed-ResNet-110	78.00%	84.60%

Our best model

Table 1: Results on the NTU-RGB+D dataset for Cross-Subject and Cross-View evaluations.

Proposed method

Experimental results

Method (protocol of Shahroudy et al., 2016)	Cross-Subject	Cross-View
HON4D (Oreifej and Liu, 2013)	30.56%	7.26%
Super Normal Vector (Yang and Tian, 2014)	31.82%	13.61%
HOG^2 (Ohn-Bar and Trivedi, 2013)	32.24%	22.27%
Skeletal Quads (Evangelidis, Singh, and Horaud, 2014)	38.62%	41.36%
Shuffle and Learn (Misra, Zitnick, and Hebert, 2016)	47.50%	N/A
Key poses + SVM (Cippitelli et al., 2016a)	48.90%	N/A
Lie Group (Vemulapalli, Arrate, and Chellappa, 2014)	50.08%	52.76%
HBRNN-L (Du, Wang, and Wang, 2015)	59.07%	63.97%
FTP Dynamic Skeletons (Hu et al., 2015b)	60.23%	65.22%
P-LSTM (Shahroudy et al., 2016)	62.93%	70.27%
RNN Encoder-Decoder (Luo et al., 2017)	66.20%	N/A
ST-LSTM (Liu et al., 2016b)	69.20%	77.7%
STA-LSTM (Song et al., 2017)	73.40%	81.2%
Res-TCN (Kim and Reiter, 2017)	74.30%	83.1%
DSSCA - SSLM (Shahroudy et al., 2017)	74.86%	N/A
Joint Distance Maps + CNN (Li et al., 2017a)	76.20%	N/A%
Our best model (Proposed-ResNet-56)	**78.20%**	**85.60%**

Table 2: Performance comparison of our proposed ResNet model with the state-of-the-art methods on the NTU-RGB+D dataset.

Proposed method

Experimental results[8]

	MSR 3D (overall)	KARD (overall)	NTU-RGB+D Cross-Subject	NTU-RGB+D Cross-View
Prior works	96.50%	99.31%	76.20%	83.10%
Our results	99.90%	99.98%	78.20%	85.60%
Improvements	**3.40%**	**0.67%**	**2.00%**	**2.50%**

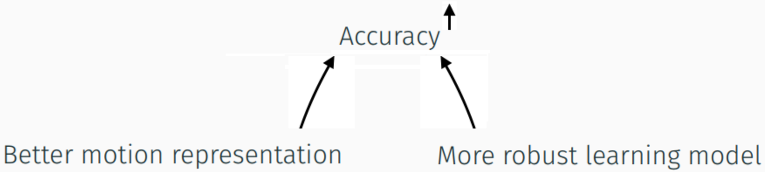
Previous state-of-the-art recognition performance that have been reported in the Literature.

Table 3: The best of our results compared to the best prior results on MSR Action3D, KARD, and NTU-RGB+D datasets.

[8]This comparison was conducted at the end of 2017 and may not be complete at the time being.

Proposed method

There is still a lot of room for improvement.



Approach 2: A new 3D motion representation for skeleton-based human action recognition with deep convolutional neural networks.

Building a better skeleton-based representation called **SPMF** for human action recognition in videos. Each action map contains two key components: **Pose Features** (**PF**) and **Motion Features** (**MF**).

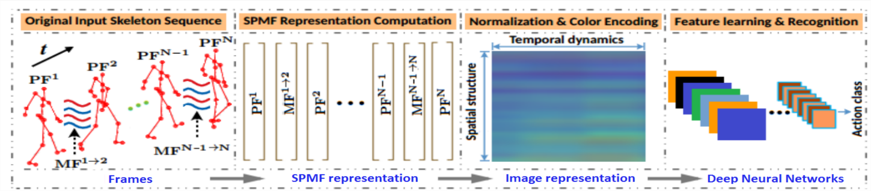


Figure 16: Encoding a skeleton sequence into a single action map.

Pose Features (PF)

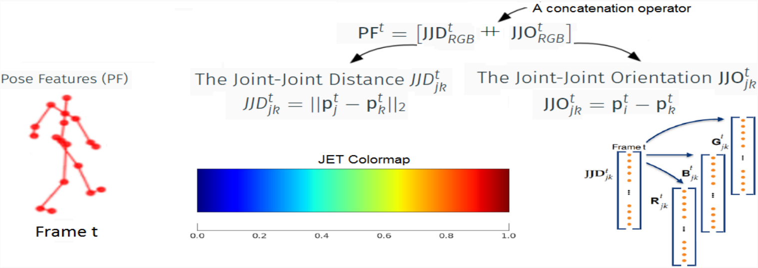


Figure 17: Computing Pose Features (PF) from skeletons.

- The pose vector (**PF**) was computed from joint-joint distances and concatenated with joint-joint orientations.
- The JET colormap was used to convert joint-joint distances to color points.

Action recognition using deep networks

Motion Features (MF)

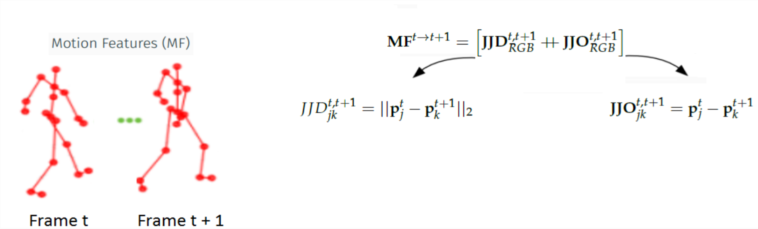


Figure 18: Computing Motion Features (MF) from skeletons.

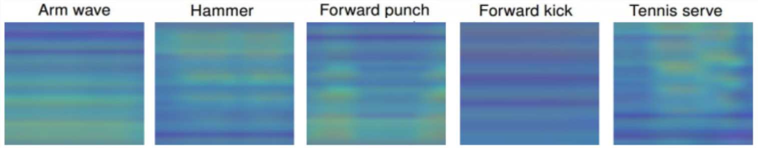


Figure 19: The SPMFs obtained from some samples of the MSR Action3D dataset.

Color enhancement

The Adaptive Histogram Equalization (AHE) algorithm was then used to highlight the motion map and form the Enhanced-SPMF.

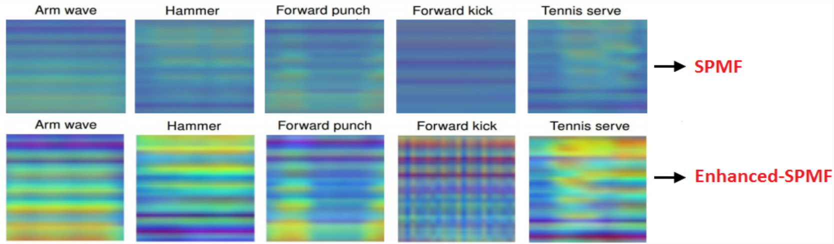


Figure 20: The proposed **Enhanced-SPMF** representation for human action recognition from skeleton sequences.

Learning model based on DenseNet[9]

DenseNet-16, DenseNet-28, DenseNet-40 were used for learning and recognition task on the proposed Enhanced-SPMFs.

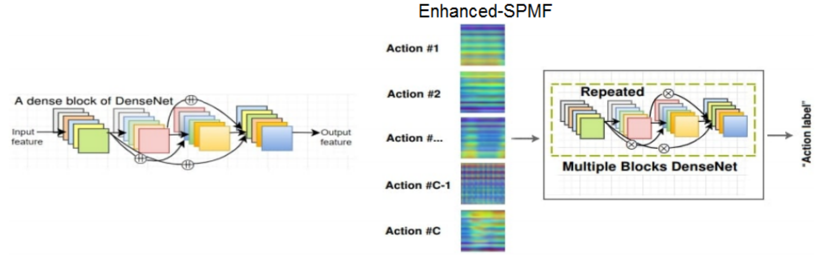


Figure 21: The proposed Enhanced-SPMFs are fed into a DenseNet for classifying action maps.

[9] Huang, Gao, et al. "Densely Connected Convolutional Networks." IEEE CVPR, 2017.

Comparison with state-of-the-art

Method (protocol of [33])	Cross-Subject	Cross-View
Lie Group [39]	50.10%	52.80%
Hierarchical RNN [6]	59.07%	63.97%
Dynamic Skeletons [13]	60.20%	65.20%
Two-Layer P-LSTM [33]	62.93%	70.27%
ST-LSTM Trust Gates [21]	69.20%	77.70%
Geometric Features [50]	70.26%	82.39%
Two-Stream RNN [40]	71.30%	79.50%
Enhanced Skeleton [24]	75.97%	82.56%
GCA-LSTM [22]	76.10%	84.00%
SPMF [27]	78.89%	86.15%
Enhanced-SPMF DenseNet-16 (**ours**)	77.89%	**86.55%**
Enhanced-SPMF DenseNet-28 (**ours**)	**79.07%**	**86.82%**
Enhanced-SPMF DenseNet-40 (**ours**)	**79.95%**	**87.52%**

Ours →

Table 4: Recognition accuracy on the large-scale NTU-RGB+D dataset.

- State-of-the-art accuracy on four challenging datasets: MSR Action3D, KARD, SBU Interaction and NTU-RGB+D.
- Less computation for training and inference.

Result of the proposed combinations

The proposed method is able to obtain a high-level of performance due to:

- New action representations that are suitable for the problem of human action recognition.
- Using state-of-the-art deep learning models for the classification task.
- A good training procedure and optimization.

Model	Input	MSR Action3D (overall)	KARD (overall)	SBU Kinect (overall)	NTU-RGB+D (cross-subject)	NTU-RGB+D (cross-view)
ResNet-44	Image-coded	99.90%	99.98%	N/A	77.20%	84.80%
Inception-ResNet-222	SPMF	98.56%	N/A	N/A	78.89%	86.15%
DenseNet	Enhanced-SPMF	99.10%	N/A	96.67%	80.11%	86.82%

Table 5: Summary of the proposed models (architecture + representation) and their experimental results on all datasets.

SPMF vs Enhanced-SPMF: A comparison

Setting

Training DenseNet on the SPMFs and Enhanced-SPMFs provided by the SBU dataset using the same training methodology (*e.g.* learning rate, batch size, optimizer.).

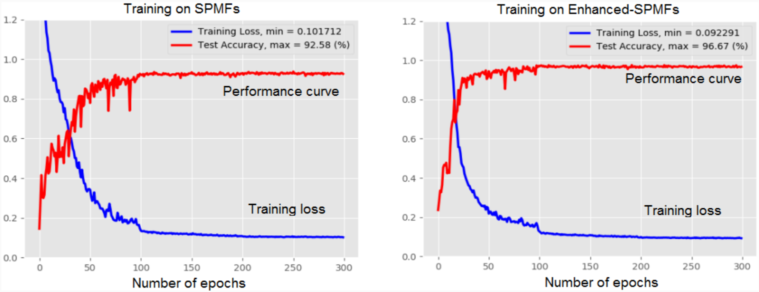
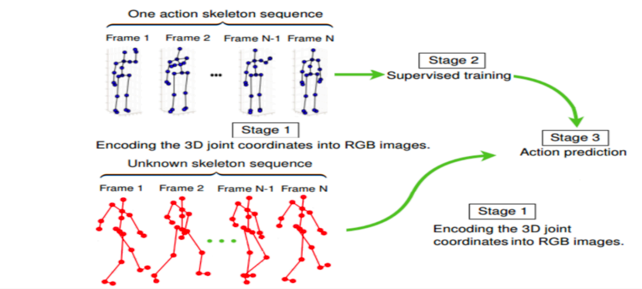


Figure 22: Test accuracy of the proposed DenseNet on SPMFs (left – **92.58%**) and on Enhanced-SPMFs (right – **96.67%**).

Computational efficiency evaluation



Component	Average processing time
Stage 1	7.83×10^{-3}s per sequence (Intel Core i7 3.2GHz CPU)
Stage 2	1.27×10^{-3}s per sequence (GTX 1080 Ti GPU)
Stage 3	8.31×10^{-3}s per sequence (GTX 1080 Ti GPU)

Figure 23: Three main stages of the proposed deep learning framework for recognizing human actions from skeleton sequences. The inference stage, including the stage (1) that is executed on a CPU and the stage (3), takes an average of 8.31×10^{-3}s per sequence without parallel processing.

CEMEST-Tisséo dataset

- A new real-wold surveillance dataset containing both normal and anomalous events for studying human behaviors in public transport.
- 203 video samples containing RGB videos, depth map sequences, and 3D skeletal data.
- Three action classes: *crossing (franchir) normally over the barriers*, *jumping (sauter) over the ticket barriers*, and *sneaking (se faufiler) under the ticket barriers*.

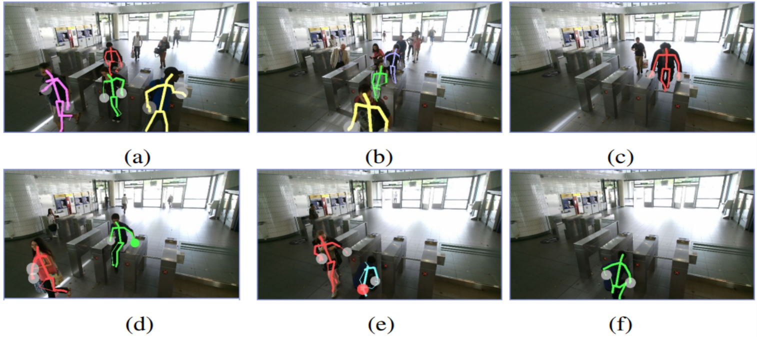


Figure 24: Some samples from the CEMEST-Tisséo dataset.

Experimental results

- We achieved an accuracy of **91%** with the DenseNet-40 when training from scratch.
- We reached an accuracy of **95%** with transfer learning, increasing the performance by more than **4%** compared to the first setting.

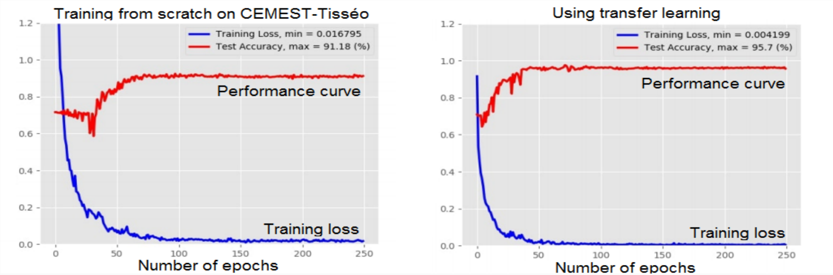


Figure 25: Learning curves of DenseNet-40 trained on the CEMEST-Tisséo dataset.

4. A Unified Deep Learning Framework for 3D Pose Estimation and Action Recognition from RGB Videos

A unified deep learning framework for 3D pose estimation and action recognition from RGB videos

Objective: Learning for 3D human pose estimation from a single RGB image using deep neural networks.

• Using a state-of-the-art 2D pose estimator (*e.g.* OpenPose) to obtain 2D human poses from RGB image sequences.

• Building a deep learning network for learning and estimating 3D human poses from 2D poses.

Given an input RGB image $\mathsf{I} \in \mathbb{R}^{W \times H \times 3}$. Denoting 2D keypoints as $\mathsf{p}_{2D} \in \mathbb{R}^{2 \times N}$ and the estimated 3D pose as $\hat{\mathsf{p}}_{3D} \in \mathbb{R}^{3 \times M}$. A neural network can be trained to produce

$$\hat{\mathsf{p}}_{3D} = f(\mathsf{p}_{2D}, \theta), \tag{1}$$

in a supervised manner, where θ is a set of trainable parameters of the function f.

A unified deep learning framework for 3D pose estimation and action recognition from RGB videos

Objective: Learning for 3D human pose estimation from a single RGB image using deep neural networks.

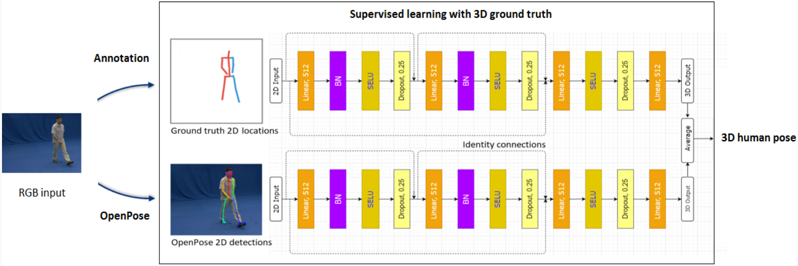


Figure 26: Diagram of the proposed two-stream network for training our 3D pose estimator.

A unified deep learning framework for 3D pose estimation and action recognition from RGB videos

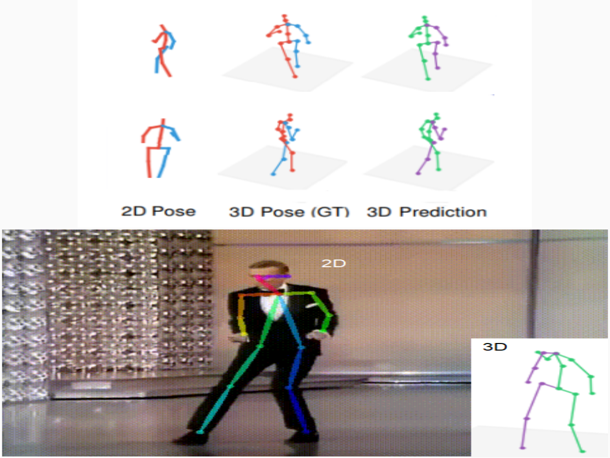


Figure 27: Visualization of 3D output of the proposed estimation algorithm.

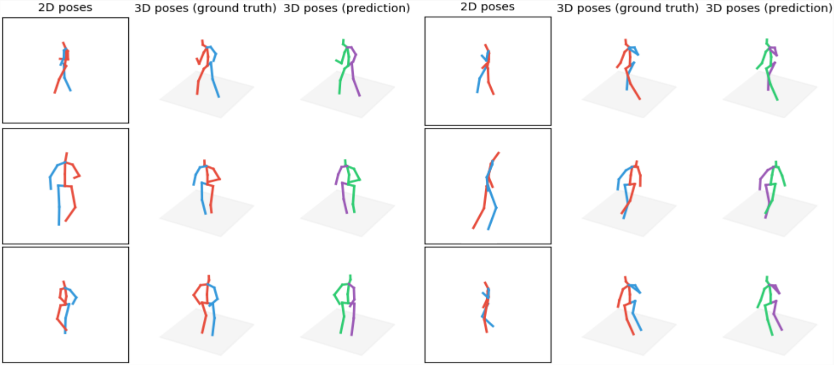


Figure 28: Visualization of 3D output of the estimation algorithm with many different human poses from the test set of Human3.6M.

A unified deep learning framework for 3D pose estimation and action recognition from RGB videos

Experimental result on Human3.6M dataset

Method	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Ionescu et al., 2014†	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Du et al., 2016*	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Tekin et al., 2016	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Park, Hwang, and Kwak, 2016*	100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou et al., 2016*	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Xingyi et al., 2016*	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Pavlakos et al., 2017	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Mehta et al., 2017a*	67.4	71.9	66.7	69.1	71.9	65.0	68.3	83.7	120.0	66.0	79.8	63.9	48.9	76.8	53.7	68.6
Martinez et al., 2017*	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Shuang, Xiao, and Yichen, 2018	52.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	53.4	67.2	54.8	53.4	47.1	61.6	59.1
Luvizon, Picard, and Tabia, 2018	49.2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	44.4	57.9	48.9	53.2
Martinez et al., 2017†	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Ours	**36.6**	**43.2**	**38.1**	**40.8**	**44.4**	**51.8**	**43.7**	**38.4**	**50.8**	**52.0**	**42.1**	**42.2**	**44.0**	**32.3**	**35.9**	$\boxed{\textbf{42.4}}$

Figure 29: Experimental results and comparison with previous state-of-the-art 3D pose estimation approaches on the Human3.6M dataset. Results are reported by the average error in millimeters between the ground truth and the corresponding predictions over all joints.

A unified deep learning framework for 3D pose estimation and action recognition from RGB videos

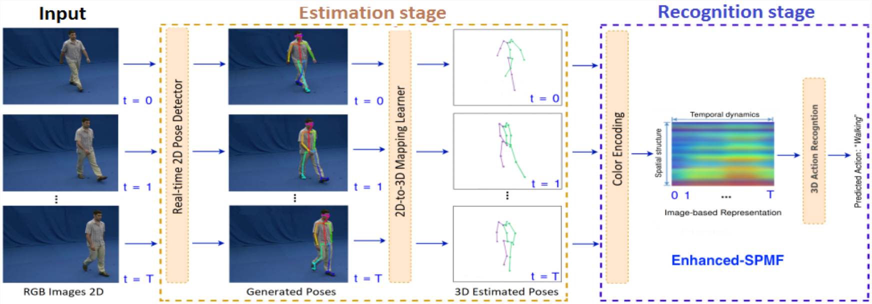


Figure 30: Overview of our method for 3D pose estimation and action recognition from RGB videos. In the recognition stage, the 3D estimated poses were encoded via Enhanced-SPMF and finally fed into a CNN for supervised classification, which is automatically searched by the Efficient Neural Architecture Search (ENAS) algorithm.

Experimental results

MSR Action3D

Method	AS1	AS2	AS3	Aver.
Li, Zhang, and Liu, 2010	72.90	71.90	71.90	74.70
Chen, Liu, and Kehtarnavaz, 2013	96.20	83.20	92.00	90.47
Vemulapalli, Arrate, and Chellappa, 2014	95.29	83.87	98.22	92.46
Du, Wang, and Wang, 2015	99.33	94.64	95.50	94.49
Liu et al., 2016b	N/A	N/A	N/A	94.80
Wang et al., 2016b	93.60	95.50	95.10	94.80
Weng, Weng, and Yuan, 2017	91.50	95.60	97.30	94.80
Xu et al., 2015a	99.10	92.90	96.40	96.10
Lee et al., 2017	95.24	96.43	100.0	97.22
Enhanced-SPMF DenseNet (L=250, k=24)	**98.83**	**99.06**	**99.40**	**99.10**
Proposed method	**97.87**	**96.81**	**99.27**	**97.98**

SBU Kinect Interaction

Method	Accuracy (%)
Song et al., 2017	91.51
Liu et al., 2016b	93.30
Weng et al., 2018	93.30
Ke et al., 2017	93.57
Tas and Koniusz, 2018	94.36
Wang and Wang, 2017	94.80
Liu et al., 2018	94.90
Zhang et al., 2019 (using VA-RNN)	95.70
Zhang et al., 2019 (using VA-CNN)	97.50
Enhanced-SPMF DenseNet (L=250,k=24)	**97.86**
Proposed method	**96.30**

Ours

Table 6: Test accuracies (%) on the MSR Action3D et SBU Kinect Interaction datasets.

5. Conclusion and Perspectives

Contributions

In general, the main contributions of this thesis include:

- Propose, develop and validate different deep learning-based approaches for determining which human actions occur from monocular RGB-D video sequences.
- Review the most prominent state-of-the-art deep learning algorithms applied to the recognition of human actions in videos.
- A new deep learning approach for human action recognition by encoding skeleton sequences into color images.
- Two new 3D skeleton-based representations, namely SPMF and Enhanced-SPMF.
- A new deep learning architecture for estimating 3D human poses from RGB images/videos.
- Collect a new RGB-D dataset called CEMEST-Tisséo for analysing passenger behaviors in public transport. The dataset was opened for research purposes.
- Contribute to 6 publications in international journals (CVIU 2018, IET Computer Vision 2019, Intelligent Sensors 2019) and conferences (ICPRS 2017, IEEE ICIP 2018, ICIAR 2019) and two preprints.

Limitations

- Lack of evaluation of the proposed 3D pose estimation method on the CEMEST-Tisséo dataset.
- Invalid or missing data of local fragments in the input sequences may lead to drop in the recognition rate.
- Recognizing human actions on continuous video streams.

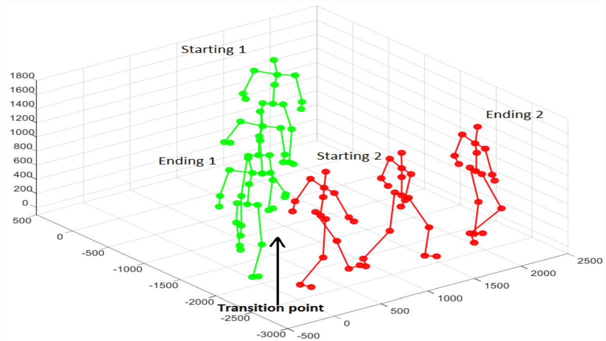


Figure 31: How to determine the starting point and the ending point of an action?

Perspectives

- Recurrent Neural Networks with Long Short-Term Memory units
- Graph Convolutional Networks
- Temporal Convolutional Networks
- Attention Temporal Networks
- Multi-Stream Deep Neural Networks

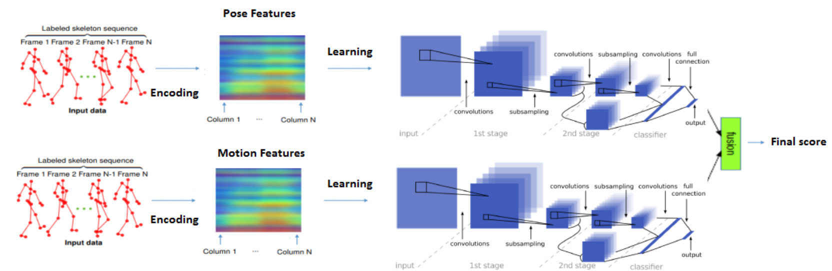


Figure 32: A two-stream deep neural network for parallel learning pose and motion features.

Thank you for your attention!