To appear in the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK

VIEW-INVARIANT ACTION RECOGNITION FROM RGB DATA VIA 3D POSE ESTIMATION

Renato Baptista, Enjie Ghorbel, Konstantinos Papadopoulos, Girum G. Demisse, Djamila Aouada, Björn Ottersten

Interdisciplinary Center for Security, Reliability and Trust
University of Luxembourg, 29, Avenue JF Kennedy, L-1855 Luxembourg
{renato.baptista, enjie.ghorbel, konstantinos.papadopoulos,
girum.demisse, djamila.aouada, bjorn.ottersten}@uni.lu

ABSTRACT

In this paper, we propose a novel view-invariant action recognition method using a single monocular RGB camera. View-invariance remains a very challenging topic in 2D action recognition due to the lack of 3D information in RGB images. Most successful approaches make use of the concept of knowledge transfer by projecting 3D synthetic data to multiple viewpoints. Instead of relying on knowledge transfer, we propose to augment the RGB data by a third dimension by means of 3D skeleton estimation from 2D images using a CNN-based pose estimator. In order to ensure view-invariance, a pre-processing for alignment is applied followed by data expansion as a way for denoising. Finally, a Long-Short Term Memory (LSTM) architecture is used to model the temporal dependency between skeletons. The proposed network is trained to directly recognize actions from aligned 3D skeletons. The experiments performed on the challenging Northwestern-UCLA dataset show the superiority of our approach as compared to state-of-the-art ones.

Index Terms— Pose Estimation, Skeleton, View- Invariance, LSTM

1. INTRODUCTION

Data acquisition with a particular camera setup depends not only on the observed scene but on the camera configuration as well. The setup leads to pixel-level data variations that are unrelated to the observed scene and subsequently poses a significant challenge in pattern recognition [1]. Particularly, in action recognition systems, tolerance to data variation resulting from differing camera positions (viewpoints) has emerged as one of the main challenges in the field.

This work has been funded by the National Research Fund (FNR), Luxembourg, under the CORE project C15/IS/10415355/3D-ACT/Björn Ottersten. This work was also supported by the European Unions Horizon 2020 research and innovation project STARR under grant agreement No.689947. We thank Oyebade Oyedotun for the fruitful discussions.

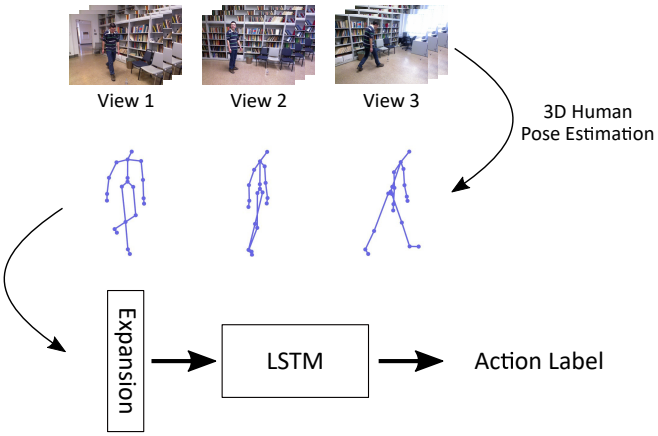

Fig. 1. Overview of the proposed approach.

The introduction of RGB-D cameras played an important role in enhancing view-invariant action recognition. In fact, these low-cost sensors provide in real-time relatively accurate 3D skeletons that have boosted the design of view-invariant approaches [2, 3, 4, 5, 6, 7]. We distinguish two different ways of addressing the issue of viewpoint variability using skeletons provided by RGB-D sensors. The first class of methods carries out a pre-processing of alignment by estimating a transformation matrix between the skeleton and a canonical coordinate system as in [4, 5, 7]. On the other hand, the second class of approaches aims to design motion descriptors which are not affected by viewpoint variability such as: *eigen joints* [2] based on the pairwise distance between joints, *Lie Algebra Representation of body-Parts (LARP)* [6] based on transformation matrices estimated between body-part pairs, etc. These approaches have shown great potential [8].

Unfortunately, they remain hardly applicable to various real-world scenarios due to the two main limitations of RGB-D sensors: (1) skeletons are correctly estimated only within a specific range; and (2) RGB-D cameras are highly affected

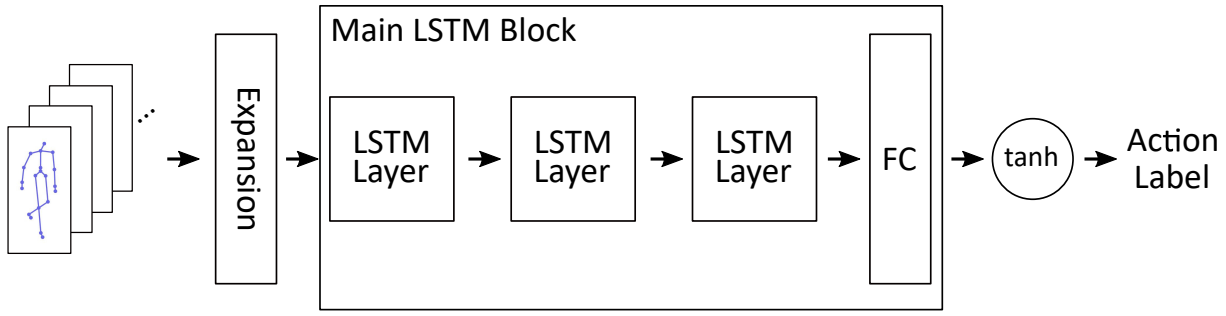

Fig. 2. Proposed network for view-invariant action recognition. FC refers to the fully connected layer at the end of the main LSTM block.

by lightning changes.

As an alternative, the topic of view-invariant action recognition using 2D image sequences, i.e., RGB data without the depth modality, has been explored by researchers. Among the most successful RGB-based approaches, one can mention the methods based on the concept of *knowledge transfer* [9, 10]. The latter aim to estimate a view-independent latent space where the features are mapped and compared. This is done by generating 3D synthetic data using a Motion Capture (MoCap) system. The acquired 3D data are then projected in various viewpoints. These approaches usually use 2D features which are by definition not view-invariant such as *Trajectory Shape Descriptor (TSD)* [9, 10, 11, 12]. As a result, these descriptors do not include radial motion information and do not take into account the human body structure.

In this paper, we propose a skeleton-based approach for view-invariant action recognition using a monocular camera. Our work builds on the recent effective Convolutional Neural Network (CNN) based methods for the estimation of 3D skeletons from a single RGB image [13, 14, 15]. We propose a full system, as illustrated in Fig. 1, where 3D skeletons are firstly extracted from RGB images using a CNN-based estimator. To achieve view-invariance, an effective pose-based data alignment is carried out. Then, the temporal dependency of the estimated pose sequences is modelled using a *Long Short-Term Memory* (LSTM) network. To address potentially noisy pose estimates, we train a feed-forward network together with the LSTM in an end-to-end training scheme. The feed-forward network is designed to expand the pose estimates to a higher dimensional space and potentially decouple several explanatory factors before modelling the temporal-dependency. In summary, the contributions of this paper are:

1. A novel approach for view-invariant action recognition using only RGB data.
2. An LSTM–based temporal model that is effective for estimating the temporal dependency between noisy skeletal pose estimates.

To evaluate and validate the proposed system, experiments on the Northwestern-UCLA dataset are realized. The obtained results show that our method outperforms RGB-based state-of-the art approaches on the same dataset.

This paper is structured as follows: Section 2 describes the proposed approach. Then, Section 3 depicts the conducted experiments. Finally, Section 4 concludes this work.

2. PROPOSED APPROACH

In this section, we describe the two main components of the proposed approach; *3D pose estimation from RGB and data alignment* which is described in Section 2.1, and *pose sequence modelling* described in Section 2.2.

2.1. 3D pose estimation and data alignment

Recent development in deep learning has enabled hierarchical systems to learn powerful filters from data [16]. Furthermore, filters that are learned from large datasets can effectively be used for problems with insufficient data, using what is called transfer learning. The *VNect* approach proposed in [15] is one of the systems that use transfer learning for effective 3D pose estimation directly from RGB images. VNect is based on a CNN pose regression that allows the real-time estimation of 2D and 3D skeletons using a single RGB image. For each estimated human joint, the network is trained to estimate a 2D confidence heatmap along with locations maps (for each of the three dimensions).

One of the main advantages of estimating a 3D pose is the ability to estimate the positions of corresponding 3D points in different viewpoints. In which case, 3D pose alignment can be estimated with a closed-form solution. To further explain, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the estimates of the same subject's 3D pose from two different viewpoints. Assuming the mean of the estimated pose is centered, the alignment of the estimated pose is performed by estimating the rotation $\mathbf{R}$ through the following optimization:

$$\arg\min_{\mathbf{R}} ||\mathbf{x}_1 - \mathbf{R}\mathbf{x}_2||_2^2. \quad (1)$$

The formulation (1) has a closed-form solution given as

$$\widetilde{\mathbf{R}} = \mathbf{V}\mathbf{U}^T, \tag{2}$$

where $\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{x}_1\mathbf{x}_2^T$, with $\mathbf{U}$ and $\mathbf{V}$ being unitary matrices and Σ a diagonal matrix corresponding to the singular value decomposition (SVD) of $\mathbf{x}_1\mathbf{x}_2^T$. The matrix $\widetilde{\mathbf{R}}$ denotes the estimated rotation matrix. Given two sequences of n poses estimated from two different viewpoints $\mathbf{X}_1 = \{\mathbf{x}_1^1, \cdots, \mathbf{x}_1^n\}$ and $\mathbf{X}_2 = \{\mathbf{x}_2^1, \cdots, \mathbf{x}_2^n\}$, we estimate the alignment between the first corresponding poses $\mathbf{x}_1^1$ and $\mathbf{x}_2^1$ using (2). Afterwards, the estimated rotation matrix $\widetilde{\mathbf{R}}$ is used to align the rest of the subsequent poses of the sequence.

2.2. Pose sequence modelling

In general, 3D pose estimation from RGB data can be noisy depending on the estimation model and the available training dataset. In this subsection, we propose an LSTM–based temporal model that is suitable for estimating the temporal dependency between noisy skeletal pose estimates. Our approach has two main components: (1) a feed-forward network for expanding the data to a high-dimensional space, and (2) multi-layer LSTM units for modelling the temporal dependency, see Fig. 2.

Data expansion: An estimated 3D skeleton with J number of joints is a vector in $\mathbb{R}^{3J}$. Hence, a noisy joint estimate is directly reflected on some of the dimensions of the observed vector. One typical solution for removing noise and redundancy is to contract the data to a lower dimensional space [17]. On the contrary, in this paper, we expand the data to a higher dimensional space. The main motivation for expanding the data is to disentangle explanatory factors that are obscured by noisy joint estimates. Consequently, the parameters of the expansion function are learned directly from the training dataset. Expansion of an observed skeleton is defined as follows

$$\tilde{\mathbf{x}} = \tanh(\mathrm{W}\mathbf{x} + \mathrm{b}), \tag{3}$$

where W is a $k \times 3J$ matrix with $k \gg 3J$, b is a bias vector in a k-dimensional space, and the $\tilde{\mathbf{x}}$ denotes the expanded pose estimate.

Temporal model and action labeling: The temporal dependency between the sequential data points is modelled using layers of LSTM units [18]. An LSTM is a gated recurrent neural network that models temporal dependency as a stationary process. Although it has several components, we herein will refer to the integrated computational unit as LSTM. Subsequently, given an expanded input data $\tilde{\mathbf{x}}$, we estimate hierarchical latent variables by layering LSTM units one on top of another, see Fig. 2. Consequently, the inferred latent space from the i^{th} pose estimate is given as

$$h_i^L = \mathrm{LSTM}(\tilde{\mathbf{x}}_i), \tag{4}$$

where L denotes the index of the last LSTM layer. Finally, an action label from a set Ψ, is assigned to a sequence as

$$\tilde{\psi} = \arg\max_{\psi \in \Psi}(\tanh(\mathrm{W}h_n^L + \mathrm{b})), \tag{5}$$

where n is the index of the last pose estimate. The connection weights and biases of the overall network (temporal model and data expansion) are trained together by minimizing the cross-entropy between the predicted and the given probability of an action label via back-propagation and back-propagation through time [16].

3. EXPERIMENTS

In this section, the experimental setup is presented along with the obtained results. For the evaluation of the proposed approach, our experiments are conducted on the Northwestern-UCLA Multiview Action3D dataset [19] denoted as *NW-UCLA*.

3.1. NW-UCLA dataset

NW-UCLA dataset is one of the most challenging RGB-D based datasets in multi-view action recognition. It consists of 1494 videos of 10 action classes (*pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry*) performed by 10 subjects. Each action can be repeated from 1 to 6 times per subject. These actions are captured simultaneously from 3 different viewpoints and both RGB and depth modalities are provided along with the corresponding estimated 3D skeleton sequences. For our experiments, we follow the splitting protocol suggested in [19], where two viewpoints are used for training and the third one for testing.

3.2. Experimental setup and implementation details

For the estimation of 3D skeleton sequences from RGB videos, we used the pre-trained VNect model[1], which provides a 3D skeleton estimate with 20 joints per frame. In addition, the skeleton sequences were temporally aligned using zero padding in an automated way.

In our experiments, we set the batch size to 2 considering the small size of NW-UCLA dataset. Moreover, using cross-validation, the optimal learning rate is set to 0.0002 and the number of epochs is chosen to be 300. The implementation of our architecture is based on PyTorch[2] using 128 hidden units per layer.

[1] http://gvv.mpi-inf.mpg.de/projects/VNect/

[2] https://pytorch.org/

3.3. Experimental Results

To validate the effectiveness of the proposed RGB-based view-invariant model, we consider three scenarios: (1) Evaluation with and without expansion unit; (2) Comparison of VNect poses against RGB-D provided skeletons; and (3) Comparison against state-of-the-art.

3.3.1. Evaluation with and without expansion unit

The expansion unit is the first layer of our proposed network, as shown in Fig. 1 and Fig. 2. This unit is mostly responsible for removing noise and redundancy from the input skeleton sequences. We run the experiments using both cases and the results are presented in Table 1. In this case, we used the provided RGB-D skeleton data and the viewpoints 1 and 2 for training and the viewpoint 3 for testing. Our proposed approach with the incorporation of the expansion module achieves 83.4% accuracy. The latter is 3.5% higher than the one reported without the utilization of this specific module mostly because of the dimensionality expansion. In this case, the abstraction of the data description increases. Thus, noisy joint estimates have lower contribution to the representation.

Method	Accuracy
No expansion + LSTM	79.9
Expansion + LSTM	83.4

Table 1. Accuracy of recognition (%) on the NW-UCLA dataset considering the cases where the expansion module is present and not present. The results are obtained using viewpoints 1 and 2 for training and viewpoint 3 for testing.

3.3.2. Comparison of VNect poses against RGB-D estimated skeletons

To evaluate the reliability of VNECT poses, we conduct experiments using the provided RGB-D skeleton sequences as input to the proposed network. In this scenario, we also use the viewpoints 1 and 2 for training and viewpoint 3 for testing. The reported accuracy in Table 2 using the provided RGB-D skeletons is 83.4% which is 3.8% lower than the accuracy achieved with the use VNect-provided poses. Although VNect generates 3D skeleton data from RGB data, it shows robustness to partial self-occlusions compared to RGB-D sensors.

3.3.3. Comparison against state-of-the-art approaches

In Table 3, the obtained results of our approach are presented and compared against some state-of-the-art approaches. Our network performance outperforms RGB-based approaches by more than 10%. Indeed, our approach reaches 79.9% of accuracy on NW-UCLA dataset against 69.4% using NKTM [10].

Method	Accuracy
Expansion + LSTM	83.4
VNect + Expansion + LSTM (VE-LSTM)	87.2

Table 2. Accuracy of recognition (%) on the NW-UCLA dataset using the provided RGB-D skeletons and the estimated skeletons from VNect. The results are obtained using viewpoints 1 and 2 for training and viewpoint 3 for testing.

{Source} | {Target}	{1,2}| 3	{1,3}| 2	{2,3}| 1	Mean
Hankelets [20]	45.2	-	-	-
DVV [21]	58.5	55.2	39.3	51.0
CVP [22]	60.6	55.8	39.5	52.0
AOG [23]	73.3	-	-	-
nCTE [9]	68.8	68.3	52.1	63.0
NKTM [10]	75.8	73.3	59.1	69.4
R-NKTM [24]	78.1	-	-	-
VE-LSTM (ours)	**87.2**	**82.1**	**70.4**	**79.9**

Table 3. Accuracy of recognition (%) on the NW-UCLA dataset. The reported results are obtained using two viewpoints for training and the remaining one for testing. *Source* indicates the viewpoints used for the training step, while *Target* specifies the testing viewpoint.

4. CONCLUSION

In this paper, we proposed a novel view-invariant action recognition approach using a single RGB camera. This is achieved by using a 3D human pose estimator from RGB images. The estimated 3D poses are used for computing a view-alignment rotation between observations. Subsequently, an LSTM based network is proposed in order to estimate the temporal dependency between noisy skeleton pose estimates. To that end, we proposed two main components: (1) a feed-forward network for expanding the data to a high-dimensional space; and (2) a multi-layer LSTM for modelling the temporal dependency. Experimental results show the superiority of our approach when compared to existing methods. Also, the 3D skeleton estimates using VNect show higher accuracy compared to the ones provided by Kinect, showing robustness to possible occlusions that may appear on the RGB images. As future work, we intend to investigate in more detail the noise introduced by the estimated skeletons over time, as well as the impact of adding challenging viewpoints.

5. REFERENCES

[1] David Mumford, "Pattern theory: a unifying perspective," in *Fields Medallists' Lectures*, pp. 226–261. World Scientific, 1997.

[2] Xiaodong Yang and Ying Li Tian, "Eigenjoints-based

action recognition using naive-bayes-nearest-neighbor," in *CVPRW*, 2012, pp. 14–19.

[3] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *CVPRW*, 2012, pp. 20–27.

[4] Girum G Demisse, Konstantinos Papadopoulos, Djamila Aouada, and Björn Ottersten, "Pose Encoding for Robust Skeleton-Based Action Recognition," in *CVPRW*, 2018, pp. 188–194.

[5] Enjie Ghorbel, Rémi Boutteau, Jacques Boonaert, Xavier Savatier, and Stéphane Lecoeuche, "Kinematic Spline Curves: A temporal invariant descriptor for fast action recognition," *IVC*, vol. 77, pp. 60–71, 2018.

[6] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *CVPR*, 2014, pp. 588–595.

[7] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.

[8] Enjie Ghorbel, Konstantinos Papadopoulos, Renato Baptista, Himadri Pathak, Girum Demisse, Djamila Aouada, and Björn Ottersten, "A view-invariant framework for fast skeleton-based action recognition using a single rgb camera," in *VISAPP*, 2019.

[9] Ankur Gupta, Julieta Martinez, James J. Little, and Robert J. Woodham, "3D Pose from Motion for Cross-View Action Recognition via Non-linear Circulant Temporal Encoding," in *CVPR*. jun 2014, IEEE.

[10] Hossein Rahmani and Ajmal Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *CVPR*. jun 2015, IEEE.

[11] Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, and Björn Ottersten, "Enhanced trajectory-based action recognition using human pose," in *ICIP*, 2017, pp. 1807–1811.

[12] Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, and Björn Ottersten, "A Revisit of Action Detection using Improved Trajectories," in *ICASSP*, 2018.

[13] Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3D Human Pose Estimation Using Transfer Learning and Improved CNN Supervision," *CoRR*, vol. abs/1611.09813, 2016.

[14] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *CVPR*, 2017, pp. 1263–1272.

[15] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt, "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera," 2017, vol. 36.

[16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[17] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik, "Dimensionality reduction: a comparative," *JMLR*, vol. 10, pp. 66–71, 2009.

[18] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, 2014, pp. 2649–2656.

[20] Binlong Li, Octavia I Camps, and Mario Sznaier, "Cross-view activity recognition using hankelets," in *CVPR*. IEEE, 2012, pp. 1362–1369.

[21] Ruonan Li and Todd Zickler, "Discriminative virtual views for cross-view action recognition," in *CVPR*, 2012, pp. 2855–2862.

[22] Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, Shuang Liu, and Cunzhao Shi, "Cross-view action recognition via a continuous virtual path," in *CVPR*, 2013, pp. 2690–2697.

[23] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, 2014, pp. 2649–2656.

[24] Hossein Rahmani, Ajmal Mian, and Mubarak Shah, "Learning a deep model for human action recognition from novel viewpoints," *PAMI*, vol. 40, no. 3, pp. 667–681, 2018.