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

Renato Baptista, Enjie Ghorbel, Konstantinos Papadopoulos, Girum Demisse, Djamila Aouada, Björn Ottersten
{firstname.lastname}@uni.lu

SIGCOM Research Group / Computer Vision Team

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.

Problem Definition

Given two sets of RGB images of different viewpoints from the same action $V_p = \{I_{p,1}, \cdots, I_{p,N}\}$ and $V_q = \{I_{q,1}, \cdots, I_{q,N}\}$, the goal is to estimate $f(\cdot)$ such that we achieve view-invariance action recognition,

$$f(V_p) = f(V_q) = \psi. \quad (1)$$

$f(\cdot)$ maps a set of RGB images V_p to its label ψ (human action),

$$f: \mathbb{R}^{\mathrm{M}\times\mathrm{T}} \rightarrow \Psi = \{1, \cdots, l\}, \quad (2)$$

$(V_p) \rightarrow \psi$, where M is the image dimention and T temporal information.

Considering two arbitrary viewpoints V_p and $V_q, p \neq q$, $f(\cdot)$ is said to be view-invariant if and only if

$$f(V_p) = f(V_q) = \psi. \quad (3)$$

Proposed Approach

1. 3D Human Pose Estimation From RGB Images

We use VNect [1] approach as $g(\cdot)$ to estimate the 3D skeleton with J joints, such that

$$X_p = g(V_p), \text{where } X_p = \{\mathbf{x}_{p,1}, \cdots, \mathbf{x}_{p,N}\}, \text{ with } \mathbf{x}_{p,i} \in \mathbb{R}^{3J}\ \forall\ i = 1, \cdots, N. \quad (4)$$

2. Data Alignment

Given two 3D skeletons from different viewpoints $\mathbf{x}_{p,i}$ and $\mathbf{x}_{q,i}$, the goal is to estimate $\mathbf{R}$ by

$$\arg\min_{\mathbf{R}} \left\|\mathbf{x}_{p,i} - \mathbf{R}\mathbf{x}_{q,i}\right\|_2^2, \text{ having a closed-form solution:}$$
$$\widetilde{\mathbf{R}} = \mathrm{V}\mathrm{U}^T, \text{ where } \mathrm{U}\Sigma\mathrm{V}^T = \mathbf{x}_{p,i}\mathbf{x}_{q,i}^T. \quad (5)$$

3. Pose Sequence Modelling

Network weights $\mathrm{W} = \begin{bmatrix}\mathrm{W_e}\\ \mathrm{W_d}\end{bmatrix}$ and bias vector $\mathbf{b} = \begin{bmatrix}\mathbf{b}_e\\ \mathbf{b}_d\end{bmatrix}$ are learned together, where W_e and $\mathbf{b}_e$ are the weights and bias learned in the expansion unit, while W_d and $\mathbf{b}_d$ are learned in the LSTM block. (6)

➢ Data Expansion

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

W_e: $k \times 3J$ matrix with $k \gg 3J$; $\mathbf{b}_e$: bias vector $\in \mathbb{R}^k$.

➢ Temporal Modeling and Action Labeling (7)

$$h_i^L = \mathrm{LSTM}(\tilde{\mathbf{x}}_i), \text{ then}$$
$$\tilde{\psi} = \arg\max_{\psi \in \Psi} \left(\tanh(\mathrm{W}_d h_n^L + \mathbf{b}_d)\right),$$

n: index of the last pose estimate; L: index of the last LSTM layer;

$\tilde{\psi}$: predicted action label.

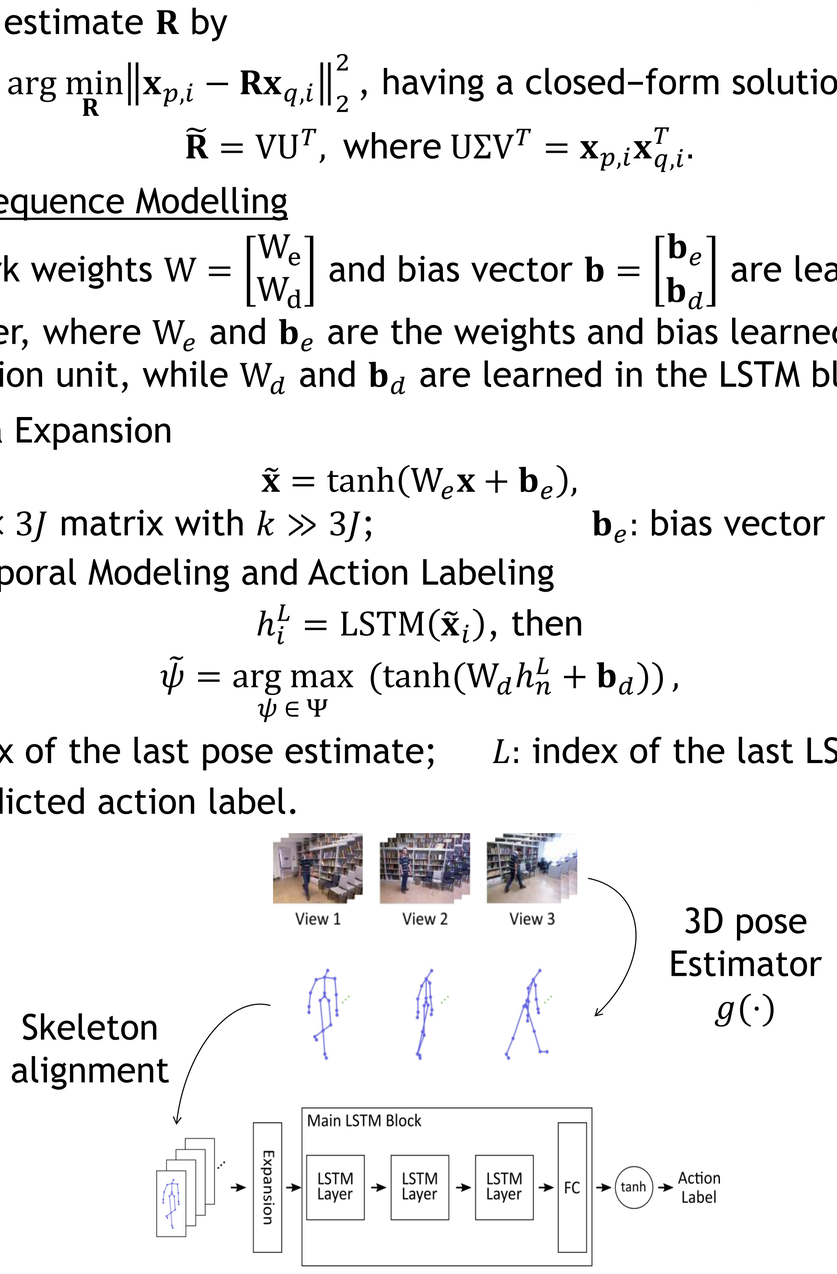

Figure 1 - Overview of the proposed approach

Results

Northwestern-UCLA Multiview Action 3D Dataset [2]:

- 3 modalities: RGB, depth and 3D skeleton
- 10 actions
- 10 subjects with 1 to 6 repetitions
- 3 different camera viewpoints

Table 1 - Comparison with state-of-the-art (%) – RGB-based approaches

{Source} | {Target}	{1,2} | 3	{1,3} | 2	{2,3} | 1	Mean
nCTE [3]	68.8	68.3	52.1	63.0
NKTM [4]	75.8	73.3	59.1	69.4
R-NKTM [5]	78.1	-	-	-
VE-LSTM (proposed)	**87.2**	**82.1**	**70.4**	**79.9**

Ablation Study

Table 2 - No expansion vs. Expansion unit

Method	Accuracy (%)
No expansion + LSTM	79.9
Expansion + LSTM	83.4

Table 3 - RGB-D-based 3D skeletons vs. CNN-based 3D skeletons

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

Conclusion

We proposed a novel view-invariant action recognition approach using a single RGB camera via 3D human pose estimation. 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;
2) a multi-layer LSTM for modelling the temporal dependency.

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.

References

[1] - D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.P. Seidel, W. Xu, D. Casas and C. Theobalt. "*VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera*", 2017, vol. 36.

[2] - J. Wang, X. Nie, Y. Xia, Y. Wu and S. Zhu, "*Cross-view action modeling, learning and recognition*" in CVPR, 2014, pp. 2649–2656.

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

[4] - H. Rahmani and A. Mian, "*Learning a nonlinear knowledge transfer model for cross-view action recognition*" in CVPR 2015, IEEE.

[5] - H. Rahmani, A. Mian and M. Shah, "*Learning a deep model for human action recognition from novel viewpoints*", PAMI, vol. 40, no. 3, pp. 667- 681, 2018.

Acknowledgements

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 Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947. We thank Oyebade Oyedotun for the fruitful discussions.

STARR

Luxembourg National Research Fund