Temporal 3D Human Pose Estimation for Action Recognition from Arbitrary Viewpoints

Mohamed Adel Musallam, Renato Baptista, Kassem Al Ismaeil, Djamila Aouada
Interdisciplinary Center for Security, Reliability and Trust
University of Luxembourg, 29, Avenue JF Kennedy, L-1855 Luxembourg
{mohamed.ali, renato.baptista, kassem.alismaeil, djamila.aouada}@uni.lu

***Abstract*—This work presents a new view-invariant action recognition system that is able to classify human actions by using a single RGB camera, including challenging camera viewpoints. Understanding actions from different viewpoints remains an extremely challenging problem, due to depth ambiguities, occlusion, and large variety of appearances and scenes. Moreover, using only the information from the 2D perspective gives different interpretations for the same action seen from different viewpoints. Our system operates in two subsequent stages. The first stage estimates the 2D human pose using a convolution neural network. In the next stage, the 2D human poses are lifted to 3D human poses, using temporal convolution neural network that enforces the temporal coherence over the estimated 3D poses. The estimated 3D poses from different viewpoints are then aligned to the same camera reference frame. Finally we propose to use a temporal convolution network based classifier for cross-view action recognition. Our results show that we can achieve state of art view-invariant action recognition accuracy even for the challenging viewpoints by only using RGB videos, without pre-training on synthetic or motion capture data.**

***Index Terms*—View-invariant, human action recognition, monocular camera, pose estimation.**

I. Introduction

In computer vision, action recognition is to decipher an action component from videos scenes. The problem of understanding human actions is very challenging due to the variation in human motion, appearance, environmental settings, and camera viewpoints. Moreover, most of the actions are dynamic and have a temporal aspect, and hence may not typically be recognized by simple attention to single moments in time. The variation of camera viewpoints has emerged as one of the main challenges of action recognition where the system receives video acquisitions from different camera locations. Considering this, the goal is to recognize actions independently of the camera location, which is commonly denoted as view-invariant action recognition.

Most of the current action recognition methods assume that the subject performing the action is facing the camera [1], [2], [3], [4]. Subsequently, the performance of these methods drops while facing real-world scenarios, where camera positioning as well as the human body orientation may vary. With the introduction of RGB-D sensors, some works tackled the subject of view-invariant action recognition by directly using the 3D information provided in real-time by the depth sensors [5]. However, the usage of such sensors is not recommended in real-world scenarios due to two main limitations. First, the estimation of 3D skeletons is limited to a specific range. Second, these sensors are highly affected by lighting conditions. To overcome these limitations, recent works proposed to use only RGB information to achieve view-invariant action recognition [6], [7]. However, in scenarios where the camera is placed in challenging locations, street surveillance where cameras are usually placed on top of buildings, understanding of actions is not always possible, even for humans.

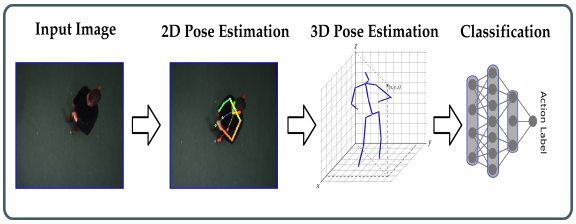

Fig. 1. High level overview of the proposed view-invariant action recognition system using only RGB images.

In this work, we focus on the case of view-invariant action recognition from challenging camera viewpoints by only using RGB information. Recently, few methods have been proposed to tackle the case of view-invariant action recognition directly from RGB images [8], [9]. These methods are based on the concept of knowledge transfer where 3D synthetic data are generated using Motion Capture system (MoCap), then the generated data are projected into various viewpoints. These methods further require 2D features as input, which are not view-invariant features by definition and do not further cover the radial motion information of the human body shape, Trajectory Shape Descriptor (TSD) [1], [2], [8], [9]. It has been shown that view-invariant action recognition can achieve better results by mapping the human subjects, from the image plane of the acquisition system to a 3D space [6], [7], [10]. This is possible due to the recent advances of deep neural networks in estimating the 3D human pose while using a single RGB camera [11], [12]. Nowadays, it is possible to estimate the human joint locations from a single RGB image, either in the 2D [13], [14] or in the 3D space [11], [12]. In [15], Martinez et al. showed that decoupling the 3D pose estimation from the 2D joint locations estimation followed by "lifting" to 3D space,

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.

gives lower error rate compared to end-to-end methods. Thus, in this work we propose a 3D skeleton-based approach for view-invariant action recognition where we first estimate the 2D locations of the human joints and then we lift the human joint locations from the 2D space to the 3D space. This is done by only using the RGB images and without any MoCap or synthetic data. Figure. 1 shows a high level overview of the the proposed view-invariant action recognition system.

We summarize our contributions as follows. First, we propose a new 3D skeleton based action recognition system that is able to tackle the challenges caused by extreme camera viewpoints. We further use the TCN approach [12] in order to lift from the 2D space to the 3D space where the temporal information of the skeleton is considered. Hence, adding the temporal coherence to the estimated 3D skeletons which in turns reduces the noisy joints estimates as compared to the per-frame 3D pose estimation methods [11]. Finally, and in line with the temporal coherence of the 3D skeleton estimates, we also propose to use the TCN approach at the level of the action classification task. To evaluate the proposed system, experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [16] are realized.

This paper is organized as follows: Section II presents the problem definition. In Section III we describe the proposed approach, followed by the experimental results in Section IV. Finally, Section V concludes the paper.

II. PROBLEM DEFINITION

In this section, we formulate the problem of view-invariant action recognition.

Let $V_p = \{I_{p,1}, \cdots, I_{p,N}\}$ and $V_q = \{I_{q,1}, \cdots, I_{q,N}\}$ be two sequences of RGB images corresponding to the same action label but with different and arbitrary camera viewpoints p and q capturing the same scene, where N is the total number of frames. Subsequently, the goal of this work is to estimate a function $f(\cdot)$ such that we achieve view-invariant action recognition,

$$f(V_p) = f(V_q) = y, \tag{1}$$

where y corresponds to the action label. The objective of the function $f(\cdot)$ is to map a sequence of RGB images V_p to its corresponding label y,

$$\begin{aligned} f : \mathbb{R}^{M \times N} &\mapsto Y = \{1, \cdots, \ell\}, \\ V_p &\mapsto y, \end{aligned} \tag{2}$$

where M is the image dimension, and ℓ is the total number of action labels. Considering two arbitrary camera viewpoints V_p and V_q with $p \neq q$, $f(\cdot)$ is considered to be view-invariant if and only if (1) is verified.

In order to estimate the function $f(\cdot)$, we propose a two-step based approach. First, we estimate the human body joint locations in the image plane (2D skeleton), and secondly, we lift the human joint locations from the image plane to the corresponding 3D human joints position (3D skeleton). The estimated 3D skeleton is then used to model the human motion. In fact, by lifting the action space from 2D to 3D, we obtain an action representation in 3D space that enables us to have a better understanding of the human motion and behavior. Hence, it provides better features to design a view-invariant action recognition system as compared to directly working in the 2D space.

III. PROPOSED APPROACH

In this section, we describe the main components of the proposed two-step 3D skeleton based view-invariant action recognition system.

A. 2D Human Pose Estimation

Given an RGB image I, the goal is to map the information related to the human body present in the image I to the corresponding 2D locations of the human joints. In other words, we want to extract the 2D skeleton by applying the mapping function $g(\cdot)$ to the image I, such that

$$S_{2D} = g(I) \quad \text{with } S_{2D} \in \mathbb{R}^{2J}, \tag{3}$$

where S_{2D} represents the estimated 2D skeleton, in a vector representation, with J joints. In order to estimate the 2D skeleton from an RGB image, we use the state-of-art approach AlphaPose [13] approach as function $g(\cdot)$. Then, for a given video sequence V_p, the function $g(\cdot)$ is applied frame by frame resulting in a sequence of 2D skeletons $\Psi_{p,2D}$,

$$\Psi_{p,2D} = \{g(I_{p,1}), \cdots, g(I_{p,N})\}. \tag{4}$$

To infer the human action in a video, we first build a 2D model for the human body in every image of the video. In order to improve the results obtained with AlphaPose for the estimation of the 2D skeleton, the human body has to be detected accurately in the image plane. For this task, the pre-trained "You Only Look Once" (YOLO) object detection network [17] is used. This is done in order to obtain the bounding box containing the human subject. Consequently, this information is provided along with the image to the AlphaPose [13] resulting in the estimation of the 2D human joints in the respective image.

B. 3D Human Pose Estimation and Data Alignment

For 3D pose estimation we build on the state-of-art approaches that formulate the problem as a 2D pose estimation followed by lifting to 3D space [15], [12]. In such approaches, the low dimensional representation like 2D pose, represented by a set of joints, can be discriminative enough to estimate the 3D pose with high accuracy [18]. Such decoupling of the problem reduces the difficulty of the task at hand, and gives the possibility of human supervision on the estimated 2D poses, prior to 3D pose estimation. These two stages-approaches have been proven to be more accurate than end-to-end approaches [11], [19], [20].

However, estimating 3D pose from individual frames leads to temporally incoherent estimation, where independent error from each frame leads to unstable 3D pose estimation over the video sequence. Thus, in our work we follow the same approach proposed by Pavllo et al. [12] where they use a fully

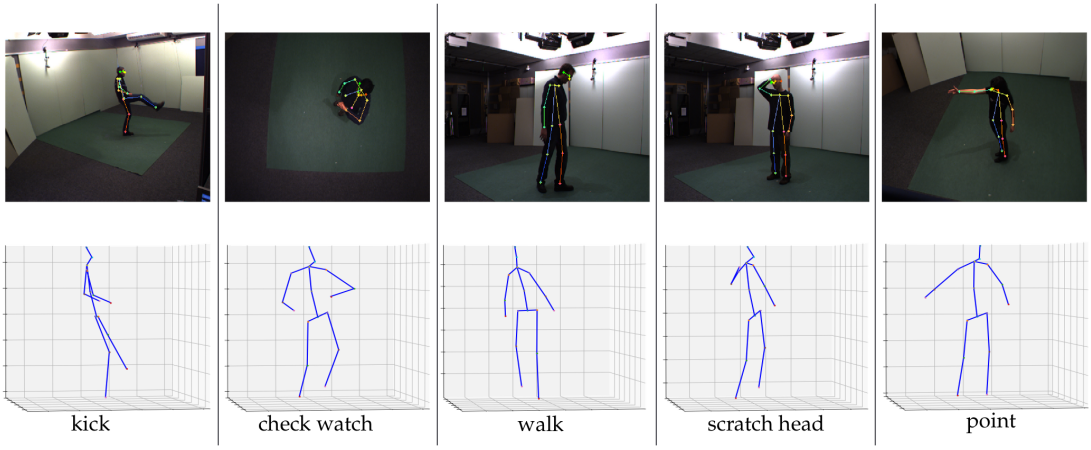

Fig. 2. Human pose estimation of different actions acquired from different camera viewpoints. First row illustrates the 2D skeleton estimates along with the RGB images, while in the bottom row, the corresponding estimated 3D skeletons are shown.

convolutional architecture that performs temporal convolutions over 2D joints in order to estimate the 3D skeleton in a video.

Given a sequence of 2D skeletons from an arbitrary camera viewpoint p, $\Psi_{p,2D} = \{S_{2D_1}, \cdots, S_{2D_N}\}$, the goal is to lift the 2D skeleton sequence into the 3-D space. To that end, we need to estimate a function $h(\cdot)$, which maps a 2D skeleton sequence to its corresponding 3D skeleton sequence, such that

$$\Psi_{p,3D} = h(\Psi_{p,2D}). \tag{5}$$

The function $h(\cdot)$ may be estimated using the work in [12]. Consequently, the 3D skeleton sequence can be obtained from an arbitrary viewpoint video sequence V_p by using the combination of the functions $g(\cdot)$ and $h(\cdot)$, such that

$$\Psi_{p,3D} = h(\,\{g(I_{p,1}), \cdots, g(I_{p,N})\}\,). \tag{6}$$

Examples of the 3D skeletons estimates from different camera viewpoints for different actions are shown in Figure 2.

Due to the variation of how every subject in the dataset preforms the action and where he/she localizes himself/herself in the room, the estimated 3D skeletons of the same action can be oriented differently from one subject to another. As our focus is to infer the action, we hence normalize the estimated 3D skeletons of the same action in order to relate them to the same reference frame. To that end, we follow the same normalization process proposed in [21] to eliminate the anthropometric variability.

At each instant t there is a 3D skeleton pose S_{3D_t} corresponds to the 3D position of a set of J joints. The position of each joint j at time t_k is then denoted by $S_{3D_{t,j}} = [j_{x(t)}, j_{y(t)}, j_{z(t)}]$. The initial position of the human hip joint j_{hip} is assumed to be the reference joint. Thus, the hip joint coordinates for each 3D skeleton are subtracted from each joint coordinates of the corresponding 3D skeleton resulting in a normalized 3D skeleton, $\widetilde{S}_{3D_t} = [j_1 - j_{hip}, j_2 - j_{hip}, \cdots, j_{17} - j_{hip}]$.. Then, the normalized 3D skeletons of the same action are aligned with respect to the same reference coordinate system by estimating a transformation matrix. This is done by considering the first 3D skeleton in the sequence as the rest state where the subject does not preform any action. This state is common in every sequence and for every subject. Then, we consider the first pose of one of the sequences as a reference pose. We further optimize a set of transformation matrices between the reference pose and the first poses of the rest of the sequences from different viewpoints of the same action. Finally, each 3D skeleton sequence is aligned to the reference 3D skeleton using the corresponding estimated transformation matrix.

C. Temporal Modeling for 3D poses

In the context of action recognition, the temporal information of the action plays a major role in producing robust predictions with less sensitivity to noise. Different methods have been introduced to model the temporal evolution using deep neural networks [22], [18]. More recently Pavllo et al. [12] presented a fully convolutional architecture that performs temporal convolutions over the 2D human joints in order to estimate 3D poses from videos. The main idea behind is the usage of TCN, where dilated convolutions are applied along the temporal axis of the action represented by the 2D skeletons, and hence producing the desired 3D skeletons.

Considering the same concept adopted for the 3D skeleton estimation, we propose to use the TCN based model presented in [23] to learn the temporal features directly from the 3D skeletons and predicting the intended action by the human subject. The 3D skeletons provide information about the behavior of the human subject and the different body movements over time. This is further represented by 17 joints locations for each 3D skeleton over time as shown on the left side of the Figure 3.

TCN is a variation of convolutional neural network for sequence modelling tasks. Compared to traditional Recurrent Neural Networks (RNNs), TCN offers more direct high-bandwidth access to past and future information. This allows TCN to be more efficient to model the temporal information of the input data with fixed size [24]. TCN can be causal; meaning that there is no information "leakage" from future to past, or non-causal where past and future information is considered. The main critical component of the TCN is the

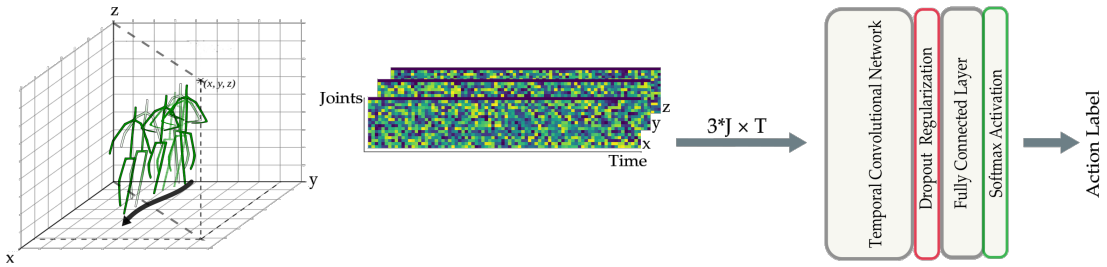

Fig. 3. Proposed temporal classification model for view-invariant action recognition. After estimating the 3D skeletons, we use the temporal 3D joints information as input to the deep neural network. Such network consists of a TCN model as temporal feature extractor, followed by a fully connected layer with softmax activation for the multi-class classification.

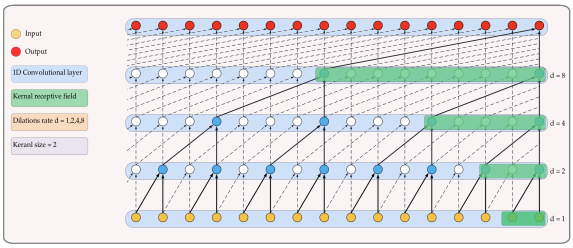

Fig. 4. Example of the TCN model with kernel size $k = 2$ and dilation rate $d = [1, 2, 4, 8]$.

dilated convolution [25] layer, which allows to properly treat temporal order and handle long-term dependencies without an explosion in model complexity. Figure 4 illustrates TCN with different dilation factors. For simple convolution, the size of the receptive field of each unit - block of input which can influence its activation - can only grow linearly with the number of layers. In the dilated convolution, the dilation factor d increases exponentially at each layer. Therefore, even though the number of parameters grows only linearly with the number of layers, the effective receptive field of units grows exponentially with the layer depth. The dilated convolution of a 1D signal s with a kernel of size k and dilation factor d is defined as:

$$(k *_d s)_t = \sum_{\tau=-\infty}^{\infty} k_\tau \cdot s_{t-d\tau}.$$

Convolutional models enable parallelization over both the batch and the time dimension while RNNs cannot be parallelized over time [23]. Also the path of the gradient between output and input has a fixed length regardless of the sequence length, which mitigates the vanishing and exploding gradients which has a direct impact on the performance of the RNNs [23]. Architectures with dilated convolutions have been successfully used for audio generation in Wavnet network [26], semantic segmentation [27] , machine translation [28], and 3D pose estimation [12]. As stated in [23], TCNs generally outperform most of the commonly used networks such as Long Short-Term Memory (LSTM) [29] or Gated Recurrent Unit (GRU) [30] for different tasks.

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental setup along with the obtained results. In order to evaluate the proposed approach, we conducted experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS) multi-view dataset [16].

A. IXMAS Dataset

IXMAS dataset is dedicated for the task of multi-view action recognition. This dataset is captured using 5 synchronized RGB cameras that are placed in 5 different locations. Such locations include four cameras placed on the side and one camera placed on top of the subject. IXMAS dataset consists of 11 different actions performed 3 times by 11 actors. The action categories are: *check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick and pick up*. This is a challenging dataset due to the fact that it contains an extreme camera location, where it is placed on top of the subject, which leads to self-occlusions. Examples of the camera viewpoints presented in this dataset are shown in Figure 2 top row.

B. Implementation Details

In order to detect the human subject present in the image, we follow the same steps introduced in the AlphaPose approach [13]. We use the pretrained YOLOv3 SPP [17] object detection network, with spatial pyramid pooling to pool and concatenate the multi-scale local region features. Thus, the network can learn the object features more comprehensively. The output of the object detection network is a set of bounding boxes around the human subjects (output from the YOLO network). In order to guarantee that the entire person region will be extracted, the detected bounding boxes containing the human proposals are extended by 30% along both the height and width directions. Every detected human proposal is then cropped and passed to the human pose estimator. We use the Cascaded Pyramid Network (CPN) [31] 2D pose estimator to estimate all the 2D skeleton sequences. The estimated 2D skeleton sequences are further passed through a TCN network in order to obtain the corresponding 3D skeleton sequences. By using a TCN network we preserve the temporal coherence present in the 2D sequences which leads, in turn, to improving the quality of the estimated 3D skeletons. Lastly, the 3D skeleton information is then used for the action classification part, where again a TCN network is used as shown in Figure 3. The following parameters are considered: kernel size = 6; dilation's rate = [1, 2, 4, 8, 16, 32]; number of stacks of residual blocks

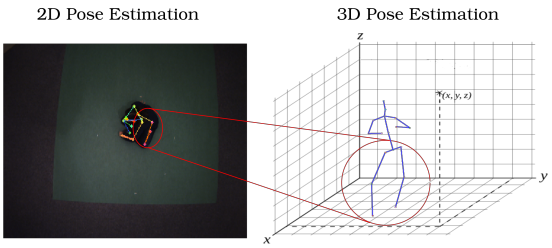

Fig. 5. Example of an erroneous 3D skeleton estimate.

= 2; Adaptive Moment Estimation (ADAM) optimizer with learning rate = 0.001; and epochs = 50.

C. Results and Discussion

In this section we evaluate the quality of the obtained results qualitatively and quantitatively. Overall and in most of the cases, the obtained 3D skeleton estimates from an RGB video sequence resemble the human motion and its structure. Furthermore, the 3D skeleton estimates from different camera viewpoints are very close to each other, which is a desired behavior for the action recognition task. Figure 2 illustrates the 3D skeleton estimates from the IXMAS dataset (bottom row). However, in some cases, the lower body parts (legs, knees and ankles) of the 3D skeletons were not correctly estimated. Note that, most of the cases happened when considering the extreme camera viewpoint where the camera is located on top of the subject. This is quite challenging due to self-occlusions that may happen. Even for the human perception it is hard to understand how the lower body part of the subject present in the image behaves. Figure 5 illustrates some cases where the 3D skeleton estimate is not correct. It is clear that the lower body part of the 3D skeleton does not correspond to the subject's pose seen in the RGB image. Even though wrong 3D skeleton estimates may happen, yet and in general, the 3D skeleton estimates are good enough for conducting the view-invariant action recognition task using RGB images.

Table I presents the action recognition results conducted on the IXMAS dataset. We follow the same cross-view protocol as in [7], [8], [9], where all sequences of the same viewpoint are used for training; then, during testing, a different viewpoint is used. Thus, for each training viewpoint, the other 4 different viewpoints are tested sequentially. The obtained results show that our method outperforms the other methods when comparing the average action classification on the IXMAS dataset. Moreover, we note that for the cases where the challenging viewpoint is considered, our method has the highest classification accuracy. The reason for this major improvement, is that the 3D skeleton estimates from the challenging viewpoint are actually better estimated using our proposed strategy when compared to the other methods. Looking at the average column in Table I, we see that we achieve the highest accuracy for the action classification task. This is achieved by only using the RGB information as input to the proposed approach, and without any additional information provided by a MoCap systems or knowledge transfer from synthetic data. In addition, the reported results are coherent regarding all testing scenarios, obtaining the highest accuracy for the majority of the testing scenarios. In order to better visualize this consistency.

Our results imply that building the 3D skeleton as a human motion model by decoupling the pose estimation into two steps (2D skeleton estimation followed by the lifting to the 3D space preserving the temporal coherence) provides a better understanding of the human action acquired from arbitrary viewpoints. This can be noted specially for the case where the camera is placed on top of the subject ("cam4"), where our method is achieving the highest scores.

1) Model selection: During the experiments, different network architectures were tested. The results lead to the conclusion that TCN based model outperforms the LSTM based models. This difference in performance can be directly attributed to the number of trainable parameters in each tested model, where for the LSTM model, the number of parameters is considerably higher. While LSTM based models perform well in a variety of tasks related to sequence modeling and temporal feature extraction, they are more complex and require more data to train. For this case, only 330 sequences are considered for training and testing.

Contrary to LSTM models, TCN based models have much less trainable parameters and can process longer sequences without an increase of the model complexity. Furthermore, using residual connection in TCN based models, does not cause the problem of vanishing gradients like LSTM when processing longer sequences. Also, with much lower difference between the training accuracy on one viewpoint and the testing accuracy on the other viewpoints. This implies that the TCN based model can be effectively used for the task of view-invariant action recognition.

2) Padding effect: During our experiments, we tested the padding methods used for sequence classification, such as pre and post padding, zero padding, and padding with duplicates of the last frame of the sequence. Inconsistent with previous deep learning based sequence classification methods, we observed that using the sequences without any padding or truncating, it improves the results by about $10\% - 20\%$.

V. CONCLUSION

In this paper, we proposed a new view-invariant action recognition system using only RGB information. This is achieved by estimating the 3D skeleton information of the subject present in the image. Furthermore, we decoupled the 3D human pose estimation problem into two steps: 1) per-image 2D human pose estimation; and 2) per-sequence 3D human pose estimation. In addition, we proposed to use the TCN for the action classification task, where the main advantage is preserving the temporal coherency present in the 3D skeleton sequence. Experimental results show that our approach achieves the highest view-invariant action recognition accuracy when considering the average off all testing scenarios. Moreover, for the majority of the testing cases, our approach is performing better than the existent works. Specially for the cases where the challenging viewpoint is

{Source} | {Target}	0| 1	0| 2	0| 3	0| 4	1| 0	1| 2	1| 3	1| 4	2| 0	2| 1	2| 3	2| 4	3| 0	3| 1	3| 2	3| 4	4| 0	4| 1	4| 2	4| 3	Avg.
Hankelets [32]	83.7	59.2	57.4	33.6	84.3	61.6	62.8	26.9	62.5	65.2	72.0	60.1	57.1	61.5	71.0	31.2	39.6	32.8	68.1	37.4	56.4
DVV [33]	72.4	13.3	53.0	28.8	64.9	27.9	53.6	21.8	36.4	40.6	41.8	37.3	58.2	58.5	24.2	22.4	30.6	24.9	27.9	24.6	38.15
CVP [34]	78.5	19.5	60.4	33.4	67.9	29.8	55.5	27.0	41.0	44.9	47.0	41.0	64.3	62.2	24.3	26.1	34.9	28.2	29.8	27.6	42.16
nCTE [8]	**94.8**	69.1	**83.9**	39.1	90.6	79.7	79.1	30.6	72.1	**86.1**	77.3	62.7	82.4	79.7	70.9	37.9	48.8	40.9	70.3	49.4	67.2
NKTM [9]	92.7	**84.2**	**83.9**	44.2	**95.5**	77.6	**86.1**	40.9	82.4	79.4	85.8	**71.5**	82.4	**80.9**	82.7	44.2	57.1	48.5	**78.8**	51.2	72.5
VNect+LARP [7]	46.6	42.1	53.9	9.7	50.6	37.5	47.3	10.0	43.4	33.0	53.6	11.8	51.2	37.8	53.6	9.1	10.9	8.7	10.9	7.9	31.48
VNect+KSC [7]	86.7	80.6	82.4	15.5	91.5	79.4	81.8	15.8	**85.2**	77.0	88.5	16.4	**83.0**	77.9	82.4	12.1	28.1	24.8	29.1	24.2	58.1
Ours	92.1	77.5	**83.9**	**58.1**	90.0	**80.6**	83.6	**56.9**	80.2	79.3	**90.6**	70.8	80.9	79.0	**89.0**	**55.1**	**68.4**	**55.4**	72.4	**58.8**	**75.13**

TABLE I

CROSS-VIEW ACTION RECOGNITION ACCURACY (%) ON THE IXMAS DATASET. EACH TIME, ONE VIEWPOINT IS USED FOR TRAINING (SOURCE) AND ANOTHER ONE FOR TESTING (TARGET). VIEWPOINT NUMBER 4 IS CONSIDERED AS A CHALLENGING VIEWPOINT, WHERE THE CAMERA IS PLACED ON TOP OF THE SUBJECT. VALUES IN BOLD REPRESENT THE BEST SCORE FOR THE CORRESPONDING EXPERIMENT.

considered. We also noticed that decoupling the problem of the 3D human poses estimation into two stages, it considerably improves the estimation of the 3D skeleton, even in cases where it is difficult for humans to understand the subject's behavior in the scene.

REFERENCES

[1] K. Papadopoulos, M. Antunes, D. Aouada, and B. Ottersten, "Enhanced trajectory-based action recognition using human pose," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1807–1811, IEEE, 2017.

[2] K. Papadopoulos, M. Antunes, D. Aouada, and B. Ottersten, "A revisit of action detection using improved trajectories," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2067–2071, IEEE, 2018.

[3] A. E. R. Shabayek, R. Baptista, K. Papadopoulos, G. Demisse, O. Oyedotun, M. Antunes, D. Aouada, B. Ottersten, M. Anastassova, M. Boukallel, S. Panëels, G. Randall, M. Andre, A. Douchet, S. Bouilland, and L. O. Fernandez, "Starr - decision support and self-management system for stroke survivors vision based rehabilitation system," in *European Project Space on Networks, Systems and Technologies - Volume 1: EPS Porto 2017,*, pp. 69–80, INSTICC, SciTePress, 2017.

[4] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action recognition by dense trajectories," in *CVPR 2011-IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, IEEE, 2011.

[5] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, IEEE, 2012.

[6] R. Baptista, E. Ghorbel, K. Papadopoulos, G. Demisse, D. Aouada, and B. Ottersten, "View-invariant action recognition from rgb data via 3d pose estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019*, 2019.

[7] E. Ghorbel, K. Papadopoulos, R. Baptista, H. Pathak, G. Demisse, D. Aouada, and B. Ottersten, "A view-invariant framework for fast skeleton-based action recognition using a single rgb camera," in *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, 25-27 February 2018*, 2019.

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2601–2608, 2014.

[9] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2458–2466, 2015.

[10] K. Papadopoulos, E. Ghorbel, R. Baptista, D. Aouada, and B. Ottersten, "Two-stage rgb-based action detection using augmented 3d poses," in *Computer Analysis of Images and Patterns* (M. Vento and G. Percannella, eds.), (Cham), pp. 26–35, Springer International Publishing, 2019.

[11] 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," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[12] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," *arXiv preprint arXiv:1811.11742*, 2018.

[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV*, 2017.

[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science*, p. 483–499, 2016.

[15] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017.

[16] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *CVIU*, vol. 104, no. 2-3, pp. 249–257, 2006.

[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[18] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "Meta-learning with temporal convolutions," *CoRR*, vol. abs/1707.03141, 2017.

[19] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[20] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: A weakly-supervised approach," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[21] E. Ghorbel, J. Boonaert, R. Boutteau, S. Lecoeuche, and X. Savatier, "An extension of kernel learning methods using a modified log-euclidean distance for fast and accurate skeleton-based human action recognition," *Computer Vision and Image Understanding*, 09 2018.

[22] K. Lee, I. Lee, and S. Lee, "Propagating lstm: 3d pose estimation based on joint interdependency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–135, 2018.

[23] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.

[24] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," 2017.

[25] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, pp. 286–297, Springer, 1990.

[26] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[28] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *arXiv preprint arXiv:1610.10099*, 2016.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[30] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014.

[31] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[32] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1362–1369, IEEE, 2012.

[33] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2855–2862, IEEE, 2012.

[34] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, "Cross-view action recognition via a continuous virtual path," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2690–2697, 2013.