Unsupervised View-Invariant Human Posture Representation

Faegheh Sardari\textsuperscript{1}  
faegheh.sardari@bristol.ac.uk  
Björn Ommer\textsuperscript{2}  
ommer@uni-heidelberg.de  
Majid Mirmehdi\textsuperscript{1}  
m.mirmehdi@bristol.ac.uk

\textsuperscript{1} Department of Computer Science  
University of Bristol  
UK  
\textsuperscript{2} Ludwig Maximilian University of Munich  
and IWR & HCI, Heidelberg University  
Germany

Abstract

Most recent view-invariant action recognition and performance assessment approaches rely on a large amount of annotated 3D skeleton data to extract view-invariant features. However, acquiring 3D skeleton data can be cumbersome, if not impractical, in in-the-wild scenarios. To overcome this problem, we present a novel unsupervised approach that learns to extract view-invariant 3D human pose representation from a 2D image without using 3D joint data. Our model is trained by exploiting the intrinsic view-invariant properties of human pose between simultaneous frames from different viewpoints and their equivariant properties between augmented frames from the same viewpoint. We evaluate the learned view-invariant pose representations for two downstream tasks. We perform comparative experiments that show improvements on the state-of-the-art unsupervised cross-view action classification accuracy on NTU RGB+D by a significant margin, on both RGB and depth images. We also show the efficiency of transferring the learned representations from NTU RGB+D to obtain the first ever unsupervised cross-view and cross-subject rank correlation results on the multi-view human movement quality dataset, QMAR, and marginally improve on the state-of-the-art supervised results for this dataset. We also carry out ablation studies to examine the contributions of the different components of our proposed network. Our code is available at https://github.com/fsardari/U-VI.

1 Introduction

RGB based deep learning approaches such as [2, 10, 13, 18, 28, 30, 41, 46] have shown an impressive performance in human action recognition and performance assessment. However, as stated in [43], the performance of such approaches drops significantly when they are applied on data that come from unseen viewpoints. To tackle this problem, a simple solution would be to train a network on data from multiple views [43]. However, in practice, capturing a labelled dataset of different views is cumbersome and rare - but two example cases are the commonly used NTU [38] and the recent health-related QMAR [37] datasets, where the granularity of the labelling is still coarse at action class level and at overall performance score level respectively. Ideally, a wholly view-invariant approach would be trained on data

© 2021. The copyright of this document resides with its authors.  
It may be distributed unchanged freely in print or electronic forms.
from as few views as possible and be able to perform well on a single (unseen) view at inference time. We shall use the NTU and QMAR datasets in our downstream tasks for our proposed view-invariant pose representation method towards this ideal.

Most current view-invariant action classification approaches are based on supervised learning, with both training and testing commonly carried out using skeleton data, such as [14, 16, 27, 48, 49]. Others like [6, 7, 31, 36] train on both RGB and 3D joint annotations to facilitate testing using RGB images alone. However, all these approaches rely on a significant amount of annotations (and sometimes camera parameters) during training, the provision of which is expensive and difficult in in-the-wild scenarios. Only very few works, such as [37, 43, 45] deal with training from RGB images only. Moreover, the authors are aware of no other unsupervised view-invariant RGB-only action classification or assessment study, and only of one such work based on RGB and depth [21].

In this paper, we propose a representation learning approach to disentangle canonical (view-invariant) 3D pose representation and view-dependent features from either an RGB-based 2D Densepose human representation map [26] or a depth mask image without using 3D skeleton annotations or camera parameters. We design an auto-encoder comprising two encoders and a decoder. The first is a view-invariant 3D pose encoder that learns 3D canonical pose representations from an input image, and the second is a viewpoint encoder that extracts rotation and translation parameters, such that when they are applied on the canonical pose features, it would result in view-dependent 3D pose representation which are fed into the decoder to reconstruct the input image. To train the network, we impose geometrical and positional order consistency constraints on pose representation features through novel view-invariant and equivariance losses respectively. The view-invariant loss is computed based on the intrinsic view-invariant properties of pose features between simultaneous frames from different viewpoints, while the equivariance loss is computed using the equivariant properties between augmented frames from the same viewpoint. After training, the 3D canonical pose representations can be used for downstream tasks such as view-invariant action classification and human movement analysis. Fig. 1 shows the proposed view-invariant pose representation learning framework and its application on a view-invariant downstream task.

Our key contributions can be summarized as follows: (i) we propose a novel unsupervised method that learns view-invariant 3D pose representation from a 2D image without using 3D skeleton data and camera parameters. Our view-invariant features can be applied

Figure 1: Left: the proposed network learns to disentangle canonical 3D human pose representations and view-dependent features through simultaneous frames from different views and augmented frames from the same view. Right: the unsupervised learned canonical pose representation can be used for downstream tasks.
directly by downstream tasks to be resilient to human pose variations in unseen viewpoints, unlike unsupervised 3D pose estimation methods such as [3, 4, 11, 15, 33, 42] which obtain view-specific 3D pose features, and require camera parameters and further steps to align their view-specific features in a canonical space, (ii) we introduce novel view-invariance and equivariance losses that impose on the network to preserve geometrical and positional order consistency of pose features - these losses can benefit the training process in other pretext tasks that exploit landmark representation, (iii) we evaluate the performance of learned pose features on two downstream tasks that demand view-invariancy and achieve state-of-the-art unsupervised cross-view action recognition accuracy on the NTU RGB+D standard benchmark dataset for RGB and depth images at 74.8% and 67.5% respectively, and for the first time we obtain unsupervised cross-view and cross-subject rank correlation results for human movement assessment scores on the QMAR dataset, while exceeding its supervised state-of-the-art results, (iv) we perform ablation studies to explore the impact of our loss functions on our proposed model.

2 Related Works

We now consider the more recent related works that deal with unsupervised pose representation and view-invariancy, in particular in relation to our chosen downstream tasks.

Unsupervised Pose Representation – There are several recent examples of RGB-based unsupervised learning approaches to 3D pose estimation, such as [3, 4, 11, 15, 33, 42]. Authors in [3, 4, 11] extract unsupervised pose features from 2D joints generated from RGB data. For example, Chen et al. [3] train a network through a 2D-3D consistency loss, computed after lifting 2D pose to 3D joints and reprojecting 3D onto 2D. Dundar et al. [11] disentangle pose and appearance features from an RGB image by designing a self-supervised auto-encoder that reconstructs an input image into foreground and background with the constraint that the appearance features remain consistent temporally while the pose features change. Honari et al. [15] also relies on temporal information and factorizes the pose and appearance features in a contrastive learning manner. In [33], the authors design a network to encode 3D pose features by predicting a different viewpoint of the input image, but there is no restriction to generate the same pose representation for the simultaneous frames. All these methods are view-specific and do not generate the same (i.e. canonical) 3D pose features for different viewpoints, so they cannot be applied to unseen-view downstream tasks, and camera parameters and extra steps are needed to map their view-specific output into a canonical view. Our proposed method learns view-invariant pose representation from the input image such that it can be applied directly to unseen-view tasks, such as action recognition. Rhodin et al. [34] use both labelled and unlabelled data to estimate canonical 3D pose. They train a network that maps multiple views into a canonical pose through mean square error, but as using only this constraint may generate random features without any positional order consistency, they also use a small subset of 3D pose annotations to enhance the output. However, our proposed approach achieves positional order consistency without utilizing any labels. Note, positional order consistency is important as it enables us to leverage temporal aspects of corresponding body joints to handle video-based downstream tasks.

Supervised View-Invariant Action Recognition and Performance Assessment – To deal with view-invariancy, most action recognition methods are based on 3D skeleton joints, such as [14, 16, 17, 22, 31, 36, 48, 49]. For example, Zhang et al. [48] present a dual-stream network, one LSTM and one CNN, and fuse the results to predict the action label.
Both streams include a view adaptation network estimating the transformation parameters of skeleton data to a canonical view, followed by a classifier. In general, methods that rely on skeleton annotations must rely on fulsome 3D joint representations which are difficult to come by in in-the-wild scenarios. Recently a few works have developed view-invariant action recognition or analysis approaches from RGB-D images, such as [8, 37, 43, 45]. Varol et al. [43] deploy multi-view synthetic videos for training their network to perform action recognition given novel viewpoints, but still use 3D pose annotations to produce the synthetic data, while the newly generated videos would have to be also labelled by experts if they were to be used for specialist applications such as healthcare.

In view-invariant action performance assessment, we are aware of only one study where Sardari et al. [37] investigated a supervised model to assess and score the quality of movement in subjects simulating Parkinson and Stroke symptoms by evaluating canonical spatio-temporal trajectories derived from body joint heatmaps.

**Unsupervised View-Invariant Action Recognition** – There are also only a relatively few unsupervised deep learning approaches that challenge view-invariant action recognition, e.g. [5, 21]. For instance, Li et al. [21] introduce an RGB-D based auto-encoder network that extracts unsupervised view-invariant spatio-temporal features from a video sequence. The proposed network is trained to reconstruct two simultaneous source and target view sequences from a given source view video. This method requires both RGB and depth data for training. Cheng et al. [5] propose a 3D skeleton-based unsupervised approach by using motion prediction as a pretraining task to learn temporal dependencies for long video representation with a transformer-based auto-encoder.

### 3 Proposed Method

Our aim is to learn view-invariant 3D pose representation from 2D RGB or depth images without relying on 3D skeleton annotations and camera parameters. Our method leverages on geometric transformation amongst different viewpoints and the equivariant property of human pose. The proposed auto-encoder includes a view-invariant pose encoder $E_\circ$, a viewpoint encoder $E_\odot$, and a decoder $D$ arranged as shown in Fig. 2. $E_\circ$ learns 3D canonical pose features from a given image which can be either an RGB-based 2D Densepose human representation map [26] or a depth mask image. As the extracted pose features are canonical, they are mapped into a specific viewpoint using the parameters obtained through encoder $E_\odot$ before being passed to $D$ to allow the decoder to reconstruct the input image. The network optimises through four losses to generate its view-invariant representation.

**Model Architecture and Formulation** – The view-invariant pose encoder $E_\circ$ learns 3D canonical pose features $P_{\circ}^I = E_\circ(I)$ given image $I \in \mathbb{R}^{3 \times W \times H}$ where $P_{\circ}^I \in \mathbb{R}^{3 \times N}$ and $N$ refers to the number of 3D pose features. $E_\odot$ estimates the viewpoint parameters $(R^I, T^I) = E_\odot(I)$, i.e. rotation $R^I = (\theta_x, \theta_y, \theta_z)$ and translation $T^I = (t_x, t_y, t_z)$. These viewpoint parameters are applied on the canonical pose features $P_{\circ}^I$ to transfer them into a specific viewpoint $P_{\odot}^I$, such that $P_{\odot}^I = R^I \cdot P_{\circ}^I + T^I$ where $P_{\odot}^I \in \mathbb{R}^{3 \times N}$. Then, the decoder reconstructs the input, $\tilde{I} = D(P_{\odot}^I)$. The network’s purpose is therefore that it learns to extract the same canonical 3D pose features for simultaneous frames from different viewpoints while maintaining equivariance for the pose features from their augmented frames (shifted in position - see details under Equivariance Loss below) from the same viewpoint. We train the proposed network by combining four losses, view-invariant $L_{\text{invar}}$, equivariance $L_{\text{equiv}}$, and two reconstruction losses $L_{\text{rec1}}$ and $L_{\text{rec2}}$. 
View-Invariant Loss – We start with two simultaneous frames \((I^v_k, I^w_k)\) from different views \(v\) and \(w\) of the same scene from their corresponding video sequences at current frame \(k\). These are passed to encoders \(E\circ\) and \(E\triangleleft\) to extract the canonical 3D pose features \(P^I_{\phi k} = E\circ(I^\phi_k)\) and viewpoint parameters \((R^\phi_k, T^\phi_k) = E\triangleleft(I^\phi_k)\), for \(\phi \in \{v, w\}\).

Each frame \(k\) has a distinct translation parameter, while the rotation is the same for all the frames of a sequence captured from the same viewpoint. Thus, if we estimate the rotation parameters from two random frames \(I^v_m\) and \(I^w_n\) from corresponding sequences and views instead, the network should still retrieve the view-specific pose features. We use this constraint to prevent the model leaking any pose information through \(E\triangleleft\) and force it to concentrate on only the viewpoint parameters. Hence, with a probability of 0.5, we randomly select the frame to predict the rotation parameters for the two views,

\[
R^v = \begin{cases} \begin{align*} R^w_k & \text{if } r < 0.5 \\ R^w_m & \text{else} \end{align*} \end{cases} \quad \text{and} \quad R^w = \begin{cases} \begin{align*} R^w_k & \text{if } r < 0.5 \\ R^w_n & \text{else} \end{align*} \end{cases}, \quad r \in U(0, 1),
\]

(1)

where \(U(0, 1)\) denotes a uniform distribution returning a number between 0 and 1. As we assume \(E\circ\) encodes the same canonical 3D pose features for \(I^v_k\) and \(I^w_k\), then swapping their pose features while their viewpoint features are retained, the network has to still be able to reconstruct them. Thus, the view-invariant loss is obtained by

\[
\tilde{I}^v_k = D(P^I_{\circ k}^{\phi}) \quad \text{where} \quad P^I_{\circ k}^{\phi} = R^\phi_k \cdot P^I_{\circ} + T^\phi_k,
\]

(2)

\[
\tilde{I}^w_k = D(P^I_{\circ k}^{\phi}) \quad \text{where} \quad P^I_{\circ k}^{\phi} = R^\phi_k \cdot P^I_{\circ} + T^\phi_k,
\]

(3)

\[
\mathcal{L}_{\text{invar}} = \sum_{\phi \in \{v, w\}} \text{MSE}(I^\phi_k, \tilde{I}^\phi_k),
\]

(4)

where \(\text{MSE}\) indicates the mean square error.

However, computing only \(\mathcal{L}_{\text{invar}}\) is not enough to learn the view-invariant pose features, and \(E\circ\) still has to reconstruct the simultaneous frames even without swapping their canonical pose features, otherwise the network learns to only assign random latent codes for canon-
ical pose features, so we introduce $L_{\text{rec}1}$ as

$$L_{\text{rec}1} = \sum_{\phi \in \{v, w\}} \text{MSE}(\hat{I}_{k}^{\phi}, \tilde{I}_{k}^{\phi}),$$

where $\tilde{I}_{k}^{\phi} = D(P_{\odot}^{I_{k}^{\phi}})$ with $P_{\odot}^{I_{k}^{\phi}} = R^{\phi} \cdot P_{\odot}^{I_{k}} + T_{k}^{\phi}$ for $\phi \in \{v, w\}$.

**Equivariance Loss** – The effect of this loss is to help teach the network to preserve the positional order of the pose components. For example, if the $i^{th}$ dimension of the latent variable indicates the right shoulder of a subject, it should be consistent for all the images. We assume that the proposed network generates consistent order of pose features, and $x$ and $y$ axes of view-specific 3D pose space are the same as the $x$ and $y$ directions of the 2D images, so when $I_{k}^{v}$ and $I_{k}^{w}$ shift by some pixels in the $x$ and $y$ directions, then all components of the view-specific pose $P_{\odot}^{I_{k}^{\phi}}$ would shift similarly. Hence, we propose an equivariance loss computed from augmentations of $I_{k}^{v}$ and $I_{k}^{w}$, where the augmented images, $\hat{I}_{k}^{v}$ and $\hat{I}_{k}^{w}$, represent positional changes of the human subject in the scene, for example by $c_{1}$ and $c_{2}$ pixels respectively, i.e.

$$L_{\text{equiv}} = \sum_{\phi \in \{v, w\}, j \in \{1, 2\}} \text{MSE}(P_{\odot}^{\hat{I}_{k}^{\phi}} + c_{j}, P_{\odot}^{\hat{I}_{k}^{\phi}}),$$

where $P_{\odot}^{\hat{I}_{k}^{\phi}} = E_{\odot}(\hat{I}_{k}^{\phi})$ and $P_{\odot}^{\hat{I}_{k}^{\phi}} = R^{\phi} \cdot P_{\odot}^{I_{k}} + T_{k}^{\phi}$ for $\phi \in \{v, w\}$.

$L_{\text{equiv}}$ is computed based on the view-specific pose features while we can also benefit from the reconstruction of the augmented frames to improve on the pose representation, so we introduce $L_{\text{rec}2}$ as.

$$L_{\text{rec}2} = \sum_{\phi \in \{v, w\}} \text{MSE}(\tilde{I}_{k}^{\phi}, \hat{I}_{k}^{\phi}),$$

where $\tilde{I}_{k}^{\phi} = D(P_{\odot}^{I_{k}^{\phi}})$. The total loss is computed as

$$L_{\text{total}} = \alpha \cdot L_{\text{invar}} + \beta \cdot L_{\text{equiv}} + \gamma \cdot (L_{\text{rec}1} + L_{\text{rec}2})$$

We determine the weights empirically to be $\alpha = 1.0, \beta = 0.001$, and $\gamma = 1.0$. After training the proposed network to learn the 3D canonical pose features, $E_{\odot}$ is used for our example view-invariant downstream tasks. Next, in Section 4, we outline our proposed method to model temporal aspects of the canonical pose features for each downstream task.

### 4 Experiments

**Datasets** – Capturing multi-view datasets requires elaborate set-ups and is inevitably time-consuming and potentially quite expensive. Hence, very few multi-view action classification or movement assessment datasets exist. We present results on two existing datasets, NTU RGB+D and QMAR.

NTU RGB+D [28] is the main benchmark dataset for view-invariant action recognition, including 60 action classes performed by 40 subjects. NTU (for short) contains 17 different environmental settings captured by three cameras from three viewpoints. Two standard
protocols are used to evaluate the performance of view-invariant action recognition methods on NTU, cross-view (CV) and cross-subject (CS). In CV, different views are adopted for training and testing, while in CS different subjects are engaged for training and testing. We followed both protocols by using the same training and testing sets as in [38] for both pretext and downstream tasks.

QMAR [29] is the only RGB+D multi-view dataset known to the authors for quality of movement assessment. It comprises 38 subjects captured from 6 different views, three frontal and three sides. The subjects were trained by an expert to simulate four Parkinson and Stroke movement tests: walking with Parkinson (W-P), walking with Stroke (W-S), sit-to-stand with Parkinson (SS-P), and sit-to-stand with Stroke (SS-S), and the movements were annotated to determine severity of the abnormality scores. For evaluation, we followed [29, 37] and obtained Spearman’s rank correlation (SRC) results under CV and CS protocols.

Our proposed auto-encoder can learn unsupervised 3D pose representations without using any action labels. To encapsulate the temporal element of the action recognition downstream task, we added a two-layer bidirectional gated recurrent unit (GRU) followed by one FC layer after our view-invariant pose encoder $E_\odot$, and trained it on

<table>
<thead>
<tr>
<th>Module</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_\odot$</td>
<td>{C2(3×3,64), BN, ReLU} × 2, MP(2×2), {C2(3×3,128), BN, ReLU} × 2, MP(2×2), {C2(3×3,256), BN, ReLU} × 2, {C2(3×3,512), BN, ReLU} × 1, {C2(3×3,512), ReLU} × 1, {FC(1024), ReLU}, {FC(512), ReLU}, {FC(3×70)}</td>
</tr>
<tr>
<td>$E_d$</td>
<td>{C2(5×5,128), BN, ReLU} × 2, MP(7×7), {C2(5×5,256), BN, ReLU} × 2, {FC(512), ReLU, Drp}, {FC(6)}</td>
</tr>
<tr>
<td>$D$</td>
<td>{FC(16×16×512), ReLU, Drp}, {C2(3×3,256), BN, ReLU} × 2, {CT2(3×3,128), BN, ReLU} × 2, {CT2(3×3,64), BN, ReLU} × 2, {CT2(3×3,3), BN, ReLU} × 2, tanh</td>
</tr>
</tbody>
</table>

Table 1: The proposed auto-encoder’s modules – All modules are 2D. $C2(d \times d, ch)$: $d \times d$ convolution filters with $ch$ channels, $CT2$: transposed convolution filters, $MP$: max pooling, $BN$: batch normalization, $FC(O)$: FC layer with $O$ outputs.

To select the 3D canonical pose feature size $P_l^d \in \mathbb{R}^{3\times N}$, we used cross-validation and evaluated the total loss $\mathcal{L}_{total}$ in Eq. 8 for $N$ in the range between 40 and 190 with a step-size of 30. The lower bound was inspired by motion capture systems that use 39 markers, and the upper bound was selected based on Rhodin et al. [33] who set their latent code size at $3 \times 200$. As shown in Table 2, the average $\mathcal{L}_{total}$ cross-validation results on the NTU dataset for both CV and CS protocols is best when $N = 70$, hence our 3D canonical pose feature size is set at $3 \times 70$.

Action Classification – Our proposed auto-encoder can learn unsupervised 3D pose representations without using any action labels. To encapsulate the temporal element of the action recognition downstream task, we added a two-layer bidirectional gated recurrent unit (GRU) followed by one FC layer after our view-invariant pose encoder $E_\odot$, and trained it on
Table 2: Optimising $P_I^E$ - Average $\mathcal{L}_{total}$ cross-validation results on NTU for different canonical pose size ($3 \times N$).

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Input</th>
<th>Supervised (%)</th>
<th>Unsupervised (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>scratch</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CV CS</td>
<td></td>
</tr>
<tr>
<td>Shuffle &amp; Learn [25]</td>
<td>ECCV 2016</td>
<td>Depth</td>
<td>- -</td>
<td>40.9 46.2</td>
</tr>
<tr>
<td>Luo et al. [24] CVPR 2017</td>
<td>VGG + ConvLSTM</td>
<td>Depth</td>
<td>- -</td>
<td>53.2 61.4</td>
</tr>
<tr>
<td>Vyas et al. [44] ✓ ECCV 2020</td>
<td>3D CNN + LSTM</td>
<td>Depth</td>
<td>- -</td>
<td>37.7 42.3</td>
</tr>
<tr>
<td>Li et al. [21] ✓ NeurIPS 2017</td>
<td>2D ResNet + ConvLSTM</td>
<td>Depth</td>
<td>- -</td>
<td>60.4 72.7</td>
</tr>
<tr>
<td>Ours ✓</td>
<td>2D ResNet + LSTM</td>
<td>Depth</td>
<td>76.7 75.9</td>
<td>82.5 78.8</td>
</tr>
<tr>
<td>Ours ✓</td>
<td>2D CNN + GRU</td>
<td>Depth</td>
<td>77.0 70.3</td>
<td>78.2 73.8</td>
</tr>
</tbody>
</table>

|                              |          |       | fine-tune       |                  |
|                               |          |       | CV CS           |                  |
| Scatter & Learn [25]          | ECCV 2016| Depth | - -             | 86.3 82.3        |
| Luo et al. [24] CVPR 2017     | VGG + ConvLSTM | RGB | - - | 66.5 67.6 |
| Vyas et al. [44] ✓ ECCV 2020  | 3D CNN + LSTM | RGB | 29.2 36.6 | 49.3 55.5 |
| Li et al. [21] ✓ NeurIPS 2017| 2D ResNet + ConvLSTM | RGB | 66.5 67.6 | 78.2 73.8 |
| Ours ✓                       | 2D ResNet + LSTM | RGB | 77.0 70.3 | 83.6 78.1 |
| Ours ✓                       | 2D CNN + GRU | RGB | 74.8 68.3 | 74.8 68.3 |

Table 3: Action classification accuracy on NTU for RGB and depth based representation learning approaches. The ✓ symbol highlights view-invariant methods. Our unsupervised results were obtained after freezing $E_\circ$’s parameters during the downstream task while the supervised results were obtained by both fine-tuning $E_\circ$ and training it from scratch. The best and the second-best results are in Bold and underline respectively.

fixed-size 16-frame input sequences with the cross-entropy loss function. Similar to [25], we subsampled the sequences such that every sequence was divided into 16 segments and one random frame was selected amongst all frames of each segment.

All the representation learning methods on NTU that are compared to ours here can operate on either RGB or depth data for training and inference, except Li et al. [21] which requires both RGB and depth for its training stage. Providing like-to-like evaluations against these relevant methods is difficult since for all such techniques their method defines the nature of their backbone architecture, for example they extract spatio-temporal features while we learn pose representation, e.g. [44] uses 3D CNNs whereas ours is integrally a 2D design. In the case of [21] which applies a 2D ResNet with added ConvLSTM [39], we provide results with the closest possible backbone, comprising a 2D ResNet and an LSTM.

Table 3 shows that for the unsupervised scenario, our 2D CNN and GRU backbone significantly improves the state-of-the-art across CV and CS tests, at 74.8%, 68.3% for RGB, and 67.5%, 64.7% for depth data, respectively. The 2D ResNet + LSTM incarnation of our method also exceeds across the board on the state-of-the-art in unsupervised results on NTU, e.g. achieving 62.1% in almost direct comparison to [21]’s 40.7% for cross-view RGB inference. For the supervised learning case, we improve on all other works with depth data whether training from scratch or fine-tuning our network with best results at 82.5% and 78.8% on CV and CS protocols respectively, and attain very competitive results using RGB in comparison to the 3D CNN-based [44].

In Table 4, we report the results of recent state-of-the-art unsupervised pose representation methods that operate on 3D skeleton data. Yao et al. [47] perform better than our
method in CV mode and Cheng et al. [5]’s result is marginally better than ours in CS mode. These result vindicate our approach as a viable alternative to skeleton-based methods which are altogether more cantankerous to deal with in real-world applications than RGB or depth derived data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Unsupervised (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su et al. [40] CVPR 2020</td>
<td>GRU</td>
<td>76.1 (CV) 50.7 (CS)</td>
</tr>
<tr>
<td>Lin et al. [23] ACM Multimedia 2020</td>
<td>GRU</td>
<td>- (CV) 52.5 (CS)</td>
</tr>
<tr>
<td>Yao et al. [47] ICME 2021</td>
<td>GRU + GCN</td>
<td>79.2 (CV) 54.4 (CS)</td>
</tr>
<tr>
<td>Cheng et al. [5] ✓ ICME 2021</td>
<td>Transformer</td>
<td>72.8 (CV) 69.3 (CS)</td>
</tr>
<tr>
<td>Rao et al. [32] ✓ Information Sciences 2021</td>
<td>LSTM</td>
<td>64.8 (CV) 58.5 (CS)</td>
</tr>
</tbody>
</table>

Table 4: State-of-the-art action recognition accuracy results on NTU for skeleton-based representation learning approaches. The ✓ symbol highlights view-invariant methods. The best and the second-best results are in Bold and underline respectively.

**Ablation Study** – We ablate our losses to examine their impact on the learning of our pose features. Table 5 shows the unsupervised action classification accuracy on NTU as we drop each or both of $L_{\text{invar}}$ and $L_{\text{equiv}}$.

Table 5 shows that removing $L_{\text{equiv}}$ from the training process, our results for both CV and CS in both RGB and depth deteriorates. This verifies that positional order consistency is essential in both cases. We also observe that eliminating $L_{\text{invar}}$ causes our method’s performance to drop in all cases, except for the cross-subject case with depth as the input modality. The increase in performance in this scenario may be attributed to the removal of the extra geometrical constraints that are imposed on the features by the extra simultaneous frames through the presence of the $L_{\text{invar}}$ computation.

<table>
<thead>
<tr>
<th>Ours with</th>
<th>Depth CV(%)</th>
<th>CS(%)</th>
<th>RGB CV(%)</th>
<th>CS(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{\text{rec}1} + L_{\text{rec}2}$</td>
<td>35.4</td>
<td>32.1</td>
<td>35.6</td>
<td>34.1</td>
</tr>
<tr>
<td>$L_{\text{equiv}} + L_{\text{rec}1} + L_{\text{rec}2}$</td>
<td>64.1</td>
<td>65.5</td>
<td>69.1</td>
<td>64.9</td>
</tr>
<tr>
<td>$L_{\text{invar}} + L_{\text{rec}1} + L_{\text{rec}2}$</td>
<td>59.6</td>
<td>52.5</td>
<td>70.3</td>
<td>63.3</td>
</tr>
<tr>
<td>$L_{\text{invar}} + L_{\text{equiv}} + L_{\text{rec}1} + L_{\text{rec}2}$</td>
<td>67.5</td>
<td>64.7</td>
<td>74.8</td>
<td>68.3</td>
</tr>
</tbody>
</table>

Table 5: Ablation studies on different combinations of losses used in the unsupervised learning process. The best and the second-best results are in Bold and underline respectively.

**Human Movement Analysis** – Here, we aim to study the efficiency of the learned representation on NTU for quality of movement scoring on QMAR. As in the action recognition task, we added a two-layer bidirectional GRU followed by one FC layer on top of $E_{\text{C}}$ to deal with temporal analysis. The size of the FC layer is equal to the maximum score for a movement type. However, for movement quality assessment, we require to analyse every single frame of a sequence, so we cannot apply any subsampling strategies for this task. We followed [37] to divide each video sequence into non-overlapping 16-frame video clips. Our network was trained on a random 16-frame clip through the cross entropy loss function, and for inference, all 16-frame clips of a video sequence were processed. The score for a sequence was estimated by averaging the outputs of the last FC layer, as in [37]. Our work offers the first ever unsupervised results on QMAR, and for further direct comparison, we also show the results of our implementation of [37]. Sardari et al. [37] who introduced QMAR present the only other supervised view-invariant results on this dataset. We also provide the performance of two other architectures, taken from [37].

As shown in Table 6, our unsupervised human movement analysis results on QMAR
Table 6: Spearman’s rank correlation (SRC) between predicted scores and ground truth labels for cross-view and cross-subject analysis on different actions of QMAR dataset. I3D was pretrained on Kinetic-400 [19], and the ✓ symbol highlights view-invariant methods. The best supervised and unsupervised results are in **Bold** and the second-best supervised results are underlined.

outperforms Li et al. [21], reaching an average SRC of 0.54 and 0.58 for CV and CS respectively. These are broadly already competitive to the supervised results on QMAR, particularly when compared against the Kinetic-400 [19] pretrained, deep I3D network. Finally, the supervised version of our method, where we fine-tune our network weights after transferring the weights learnt through NTU training, exceeds Sardari et al. [37]’s performance on average and achieves 0.72 and 0.70 for CV and CS respectively.

5 Conclusion

Most current view-invariant action recognition and performance assessment approaches are based on supervised learning and rely on a large number of 3D skeleton annotations. In this paper, we dealt with these through an unsupervised method to learn view-invariant 3D pose representation from a 2D image. Our experiments show that not only can our learned pose representations be applied on unseen view videos from the same training data, but it can also be used in different domains. Our unsupervised approach is particularly helpful in applications where the use of multi-view data is essential and capturing 3D skeletons is challenging, e.g. in healthcare rehabilitation monitoring at home or in the clinic.

In the pretext stage of our model, we require synchronised multi-view frames to learn view-invariant 3D pose representations. For future work, we will investigate extracting view-invariant pose features from a single view or non-synchronized frames to allow learning to become a simpler process for application to any suitable dataset.

Acknowledgements

The 1st author is funded by a University of Bristol Overseas Postgraduate Student Scholarship.
References


