Viewpoint Adaptation of 2D Human Poses using Autoencoders

A. De Toro-Castro[1[0009−0005−7681−0282]], M. Lupión[2[0000−0001−7697−8062]], F. Cruciani[2[0000−0002−1870−0203]], V. González-Ruiz[1[0000−0001−6495−4856]], J.F. Sanjuan[1[0000−0002−2874−0903]], and P.M. Ortigosa[1[0000−0001−6514−6543]]

[1] Dpt. of Informatics, ceiA3, Univ. of Almería, 04120, Almería, Spain
adc144@inlumine.ual.es {vruiz,jsanjuan,ortigosa}@ual.es
[2] School of Computing, Ulster University, BT15 1ED, Belfast, UK
{m.lupionlorente,f.cruciani}@ulster.ac.uk

Abstract. Robust, domain-invariant pose transformation remains a central challenge in building reliable gesture-based interfaces. One of the main limitations of current systems is their sensitivity to variations in camera viewpoint and signer orientation, which often require large-scale, multi-view datasets to ensure generalization. Collecting and annotating such datasets can be costly and impractical, especially for applications that aim to operate in diverse, unconstrained environments such as smart homes or ambient intelligence systems.
This paper explores a modular approach to address this issue by introducing an intermediate transformation stage that translates pose coordinates with respect to viewpoint. Specifically, we propose a synchronised multi-camera setup during training, in which a dedicated module based on a convolutional autoencoder with skip connections learns to adapt skeletal data captured from arbitrary camera angles into a consistent, front-facing representation. The network jointly processes normalized 2D keypoints and their absolute positions, merging these features in the bottleneck to produce heatmaps of the transformed pose. Once trained, this module enables inference from a single arbitrary viewpoint by projecting the observed pose into the canonical view space. By decoupling viewpoint adaptation from downstream tasks, new viewpoints can be incorporated without retraining the core models, simply by updating the transformation component.
We evaluate our approach using two newly recorded datasets captured simultaneously from frontal and top-down cameras. Preliminary results show that the proposed autoencoder effectively transforms poses from a top-view perspective into a canonical frontal representation, achieving low reprojection errors even on unseen pose variations.

Keywords: Deep Learning · Human Pose Estimation · Autoencoder

1 Introduction

In smart environments, deep learning systems for recognizing activities, performing gait analysis [2] or detecting falls based on pose [7] data have gained

attraction due to their robustness to variations in lighting, background, and user appearance. By leveraging skeletal keypoints extracted from RGB or depth data, such systems can focus on the structural and temporal dynamics of signing gestures without relying heavily on visual textures or colour information. This makes them especially suitable for deployment in these environments, where non-intrusive and efficient gesture-based interfaces are essential.

Despite their advantages, pose-based systems often struggle with a fundamental challenge: viewpoint sensitivity [3]. Since the same pose can appear drastically different when observed from varying camera angles, models trained on pose data from a single view frequently fail to generalise to unseen orientations. Addressing this limitation usually requires large-scale multi-view setups [5], which are impractical in real environments.

To solve this problem, the use of the homography is proposed as a solution in some works. The homography matrix aims to transform a perspective using more than four points, which are mapped into the points found in the other perspective [4]. However, the homography is able to transform the image if the user is in the same plane. Nevertheless, in naturalistic conditions, the user can be present in the scene, so the static homography matrix is not enough to transform the pose adequately.

To address this limitation, we propose a modular approach that explicitly decouples viewpoint adaptation from the specific recognition task. Rather than training an specific model that uses the human pose directly on pose data from multiple views, we introduce a transformation module trained to normalize pose observations into a consistent, canonical front-facing representation. The core of the transformation module is an autoencoder, used to transform source upper 2D keypoints into destination frontal 2D keypoints, generating the heatmaps of the transformed keypoints. Since every perspective has a different transformation, the autoencoder makes use of the normalized 2D pose and the original user location in the image to be able to transform the perspective. Training is conducted using a synchronised multi-camera setup, where pairs of poses from different viewpoints serve as inputs and targets. Once trained, the transformation module enables inference from a single arbitrary viewpoint by projecting the observed pose into the canonical view space.

In the remainder of this paper, Section 2 reviews previous approaches for viewpoint-invariant pose representation and transformation. Section 3 presents the proposed modular architecture, detailing the design of the perspective transformation autoencoder and the multi-camera training setup. Section 4 describes the recording procedure of the custom datasets and reports the quantitative and qualitative results of our method, including a comparison with a classical homography-based transformation. Finally, Section 5 summarises the main contributions of this work and outlines potential directions for future research.

2 Related works

A persistent challenge in pose-based systems is their sensitivity to camera viewpoint. Actions or signs observed from different angles can project to significantly different 2D keypoint configurations, reducing model generalisation and limiting real-world applicability. Early approaches to this issue relied on planar geometric transformations, such as homographies or affine transforms [4]. While computationally inexpensive, these methods assume coplanarity and are insufficient for the depth variation typical of natural human motion.

To overcome these limitations, more recent research has explored both geometric and learning-based strategies for achieving viewpoint invariance. From a geometric perspective, normalization techniques have been applied to align poses into canonical orientations. In the context of 2D data, authors in [**?**] proposed learning a shared latent space where sequences of human motion from different viewpoints are mapped to a common representation. Similarly, authors in [8] addressed cross-view gait recognition by learning feature transformations that align pose descriptors across perspectives. These techniques show that latent representations can mitigate view-related variance, although they often require substantial multi-view training data. These methods propose the training of a model that takes into account the different perspectives, being constrained by the number of perspectives included during training. With respect to the 3D space, authors in [1] reviewed a range of 3D pose normalization methods, including those based on aligning body axes, re-centering joints, and projecting to consistent planes, although these approaches often rely on depth information or full 3D reconstructions, which are not straightforward using single-camera scenarios.

More recently, view-invariant representation learning has emerged as a prominent solution. Authors in [9] proposed a view-adaptive recurrent neural network that learns to internally rotate skeletons to a more informative view before classification. This idea of integrating viewpoint normalization into the learning pipeline was extended in [6], who used a view-normalization GAN to generate front-facing pose sequences conditioned on arbitrary viewpoints. Nevertheless, these methods rely on reconstructing the entire pose but using 3D keypoints, which cannot be accurately extracted in single-camera scenarios.

3 Methodology

This section describes the proposed modular viewpoint adaptation approach, detailing the overall system architecture, data acquisition setup, and the design of the perspective transformation module.

3.1 General Architecture

The proposed system is based on a gesture recognition model trained using frontal-view data. This model is designed to classify predefined gestures in order

to trigger corresponding alert messages within a smart home environment. From each frontal image, the MediaPipe framework is employed to extract 2D pose keypoints of the person. These keypoints are subsequently normalized by centering the subject within a fixed reference frame, ensuring consistency in scale and position.

In our approach, we extend this setup by incorporating a complementary module that captures the scene from a top-down perspective. In this configuration, the normalized keypoints provided by MediaPipe, together with their absolute positions in the original top-view image, are processed by a transformation module responsible for mapping them to the frontal-view reference frame. The resulting transformed 2D keypoints are used to reconstruct an equivalent frontal pose representation, which is then fed into the gesture recognition model.

This architecture enables viewpoint-invariant gesture recognition and is illustrated in Figure 2. At inference time, the system conditionally applies the transformation module depending on the active camera viewpoint.

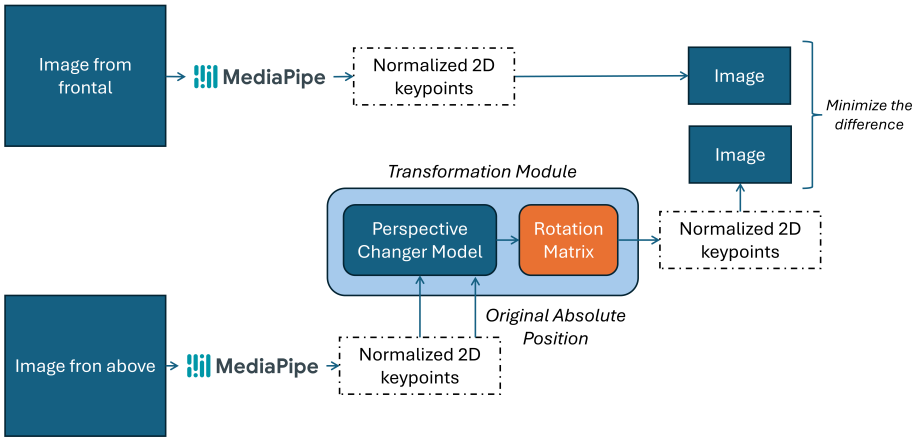


Fig. 1: General architecture of the system.

3.2 Recording Setup and Keypoints Extraction

To enable the training of the top-view transformation module, the recording setup was designed to capture synchronized images from two viewpoints. The original gesture recognition system relied on a frontal camera, with users performing gestures directly facing the device. For this work, an additional camera was mounted on the ceiling, positioned vertically above the scene to simulate a top-down perspective. Both cameras were connected to the same server and time-synchronized to capture paired frames at approximately 30 FPS, ensuring temporal alignment between the two viewpoints.

For each frame pair, the MediaPipe Pose model was applied to extract human pose information in the form of 17 3D keypoints. These landmarks include the shoulders, elbows, wrists, hips, pinky, thumb, index finger, nose, and eyes. This set of keypoints provides a compact yet informative representation of the upper body and hands, which are the most relevant regions for recognizing gestures.

To ensure geometric consistency across users and viewpoints, the extracted keypoints were normalized within a fixed 100×100 reference frame. This process involved cropping the region around each detected person and rescaling the keypoints so that the body remained centered and size-invariant. Such normalization mitigates inter-person variability in body proportions and position within the image, resulting in a uniform representation well-suited for the subsequent reorientation and gesture classification stages.

The inclusion of head-related landmarks (nose and eyes) is particularly important, as they provide cues about head orientation and gaze direction. These additional points enhance the ability of the autoencoder to accurately estimate the spatial configuration of the subject and improve the generation of a consistent front-facing pose, even under partial occlusions or variations in body posture.

3.3 Perspective Transformation Module

The perspective transformation module is implemented as an autoencoder that jointly processes the normalized pose from the upper perspective, and the 3D location of these keypoints. By trying to reproduce the desired output (all the keypoints except those from the face), it learns how to predict a canonical front-facing representation of the observed pose. The encoder architecture can be seen in 2.

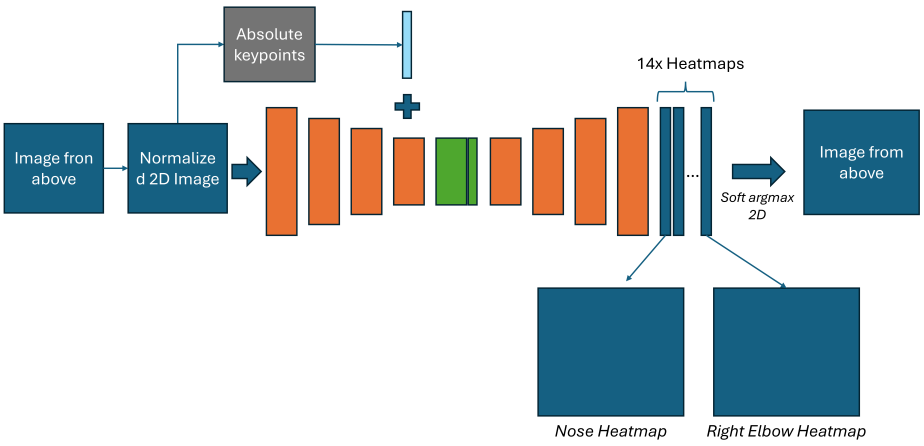


Fig. 2: General architecture of the autoencoder neural network.

First, the decoder part is composed of a series of convolutional blocks that progressively extract high-level spatial features from the input image, while preserving relevant local information through skip connections. Each block consists of two consecutive convolutional layers with LeakyReLU activations, batch normalization, and dropout for regularisation. Downsampling is achieved via max-pooling operations, resulting in progressively reduced spatial dimensions and an increased number of feature maps.

In parallel to the image stream, two additional branches encode pose-related information: (i) the *normalized 2D keypoints*, which correspond to the skeletal keypoints extracted from the top-view image and scaled to a 100×100 reference frame, and (ii) the *absolute 3D keypoints*, which include the depth estimation of each joint provided by MediaPipe. These two branches are concatenated and projected through a fully connected layer, reshaped into a 6×6 feature map, and merged with the image feature stream at the bottleneck of the encoder.

The decoder part mirrors the encoder structure using a series of upsampling layers combined with convolutional operations. Skip connections between encoder and decoder blocks allow the recovery of spatial details lost during downsampling. At each stage, the decoder progressively increases the spatial resolution while refining the feature maps. The final output is a set of 14 heatmaps of size 100×100, each representing the 2D predicted location of a keypoint in the canonical view. It is important to note that the output is composed with heatmaps because a direct regression of the 2D keypoints from the last convolutional layers did not yield good results in preliminary experiments, due to the spatial relations between these.

This design enables the module to learn a mapping between poses observed from arbitrary viewpoints and a standardised front-facing representation. By leveraging both raw image features and skeletal information, the network can robustly infer the geometric transformation even under variations in user position, scale, and camera perspective.

4 Experimental Results

This section presents the evaluation of the proposed perspective transformation module. We report both quantitative and qualitative results to analyse the model's performance on different datasets.

4.1 Dataset

For this work, two entirely new datasets were recorded specifically to train and evaluate the proposed perspective transformation module. Both datasets were captured using a dual-camera setup composed of a frontal camera and a top-down camera mounted on the ceiling, time-synchronized to record paired frames at approximately 30 FPS.

The first dataset, used for training and validation, consists of a single participant moving sequentially to each of the four corners of the room while keeping

both arms raised. At the end of the sequence, the participant moves to the center of the room and performs similar arm movements. The total duration of this recording is approximately three minutes. After acquisition, the dataset was shuffled and split to create the training and validation sets. The training datasets contains a total of XX images, and the validation dataset XX images.

The second dataset was reserved exclusively for testing. In this recording, the same participant remains in the center of the scene and performs abrupt pose changes in several directions, varying the depth and position of the arms and hands. The aim of this dataset is to evaluate the model's ability to generalise to unseen pose variations and depth configurations that were not present during training. This test recording also has a total duration of approximately three minutes, containing a total of XX images.

Preliminary experiments confirmed that when a dataset contains very similar or repetitive poses, the model can easily learn a direct transformation with minimal error. However, in this work, we deliberately designed a testing dataset that is significantly different from the training and validation data. This setup enables a more challenging and realistic evaluation of the generalisation capability of the proposed transformation module.

4.2 Perspective Changer Model

The proposed perspective transformation model achieved its best performance at epoch XX, after approximately XX minutes of training on an NVIDIA A30 GPU. The training process was conducted using mean squared error (MSE) as the objective function. Quantitative results in terms of MSE and average pixel error are summarised in Table 1.

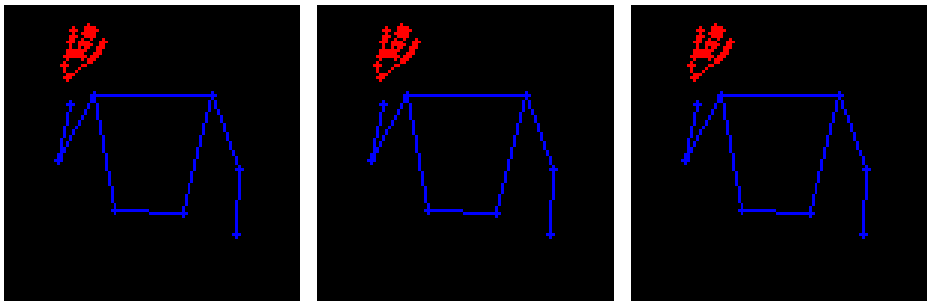
The results indicate that the autoencoder successfully learns to map the upper-body keypoints from the top-view input into a canonical frontal perspective. The low error values obtained on both the training and validation sets confirm the model's ability to reconstruct the frontal view with high spatial accuracy. Notably, when evaluated on the unseen test dataset, the model achieves an average error of less than XX pixels, despite the gestures in this dataset being substantially different from those observed during training.

These findings demonstrate that the model is capable of generalising to unseen viewpoints and pose variations to a certain extent. Nevertheless, the inclusion of a more diverse set of poses and movements in the training data would likely further enhance its generalisation ability, particularly for complex or abrupt changes in depth and limb positioning.

Set	**MSE Loss**	**Avg. Error (px)**
Train	0.00132	1.14
Validation	0.00145	1.21
Test	0.00148	1.22

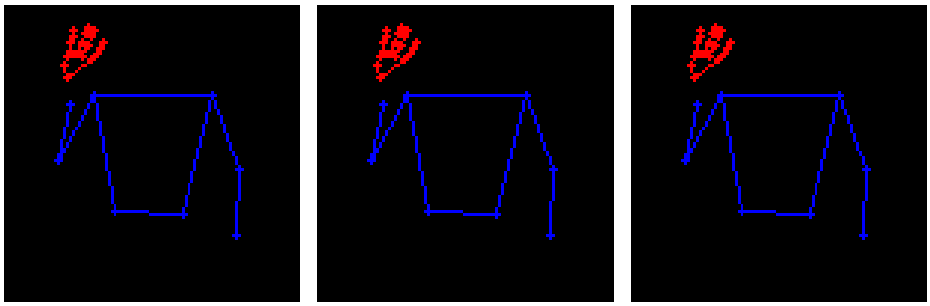
Table 1: Final evaluation metrics (image size: 100×100 px).

Figure 3 illustrates an example of the perspective transformation on the validation dataset, while Figure 4 presents the corresponding transformation on the test dataset. For clarity, in the validation example, we selected the perspective most similar to the one used in the test set to emphasise the differences between both datasets. As shown, the autoencoder successfully learns to transform the validation samples with minimal error, confirming the effectiveness of the proposed architecture. Although the transformation on the test dataset is not flawless, it still produces a representation that is considerably closer to the frontal view compared to the original top-down perspective, highlighting the model's ability to generalise to unseen poses.



Frontal Top-view Transformed Frontal

Fig. 3: Sample of transformation using the autoencoder in the validation dataset.



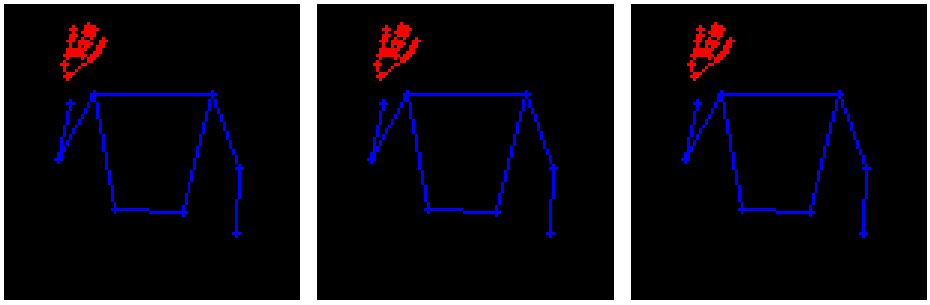
Frontal Top-view Transformed Frontal

Fig. 4: Sample of transformation using the autoencoder in the test dataset.

4.3 Comparison with the homography matrix

To quantitatively assess the effectiveness of the proposed autoencoder, we compared its keypoint reprojection performance against a classical homography-based transformation. For the homography approach, a homography matrix was computed using randomly selected frames from the test dataset, after which the remaining frames were transformed according to this matrix. This procedure was repeated five times to ensure a fairer evaluation.

Figure 5 depicts an example of the transformation obtained using the homography matrix on the test dataset. As can be observed, relying solely on a 2D planar transformation results in significantly poorer pose reconstruction compared to the autoencoder. Moreover, the homography matrix is only valid for a fixed plane corresponding to the user's position. When the user moves along the depth axis relative to the camera, recalibration of the homography is required—an impractical solution in realistic scenarios.



Frontal Top-view Transformed Frontal

Fig. 5: Sample of transformation using the homography in the test dataset.

5 Conclusions and future works

In this work, we proposed a viewpoint normalization module designed to transform 3D human poses captured from a top-down camera into a canonical front-facing view. By leveraging 3D keypoints extracted with the MediaPipe Pose model and estimating a rigid transformation via a neural network, our system achieves consistent pose alignment regardless of the camera perspective.

The proposed transformation module is implemented as a lightweight multilayer perceptron that predicts rotation and translation parameters on a per-sample basis. To ensure geometric consistency, the predicted rotation matrix is projected onto the special orthogonal group $SO(3)$ using singular value decomposition. This guarantees that the transformation preserves the relative structure of the body pose.

Our approach is modular and generalizable, making it compatible with any downstream task that relies on viewpoint-invariant pose representations, such as gesture recognition or action classification. Experimental results confirm that decoupling viewpoint normalization from gesture inference significantly improves recognition accuracy in multi-camera settings.

Future work will explore extending the transformation module to incorporate temporal context, enabling smooth pose transitions across video frames and improved robustness to occlusions and partial keypoint detections.

Acknowledgements This work has been funded by the projects R+D+i PID2021-123278OB-I00 and PDC2022-133370-I00 from MCI-N/AEI/10.13039/501100011033/ and ERDF funds; and the Department of Informatics of the University of Almería.

References

1. Aristidou, A., Lasenby, J.: Inverse kinematics techniques in computer graphics: A survey. Computer Graphics Forum **37**(6), 35–63 (2018)
2. Carneros-Prado, D., Dobrescu, C.C., Cabañero, L., Villa, L., Altamirano-Flores, Y.V., Lopez-Nava, I.H., González, I., Fontecha, J., Hervás, R.: Synthetic 3d full-body skeletal motion from 2d paths using rnn with lstm cells and linear networks. Computers in Biology and Medicine **180**, 108943 (2024)
3. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision. pp. 262–275. Springer (2008)
4. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2nd edn. (2003)
5. Lupión, M., Polo-Rodríguez, A., Medina-Quero, J., Sanjuan, J.F., Ortigosa, P.M.: 3d human pose estimation from multi-view thermal vision sensors. Information Fusion **104**, 102154 (2024)
6. Pan, Z., Li, Y., Shao, L.: View-normalization gan for skeleton-based action recognition. In: ACM Multimedia. pp. 1571–1579 (2021)
7. Song, L., Yu, G., Yuan, J., Liu, Z.: Human pose estimation and its application to action recognition: A survey. Journal of Visual Communication and Image Representation **76**, 103055 (2021). https://doi.org/10.1016/j.jvcir.2021.103055
8. Wang, C., Hu, X., Tan, T.: Cross-view gait recognition using deep feature transformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 316–323 (2012)
9. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View-adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2117–2126 (2017)