To appear in the 14$^{th}$ International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, Czech Republic (25-27 February 2019).

# A View-invariant Framework for Fast Skeleton-based Action Recognition Using a Single RGB Camera

Enjie Ghorbel, Konstantinos Papadopoulos, Renato Baptista, Himadri Pathak, Girum Demisse, Djamila Aouada, Björn Ottersten
*Interdisciplinary Centre for Security, Reliability and Trust, Luxembourg*
{*f_author, s_author*}*@uni.lu*

Keywords: View-invariant, human action recognition, monocular camera, pose estimation.

Abstract: View-invariant action recognition using a single RGB camera represents a very challenging topic due to the lack of 3D information in RGB images. Lately, the recent advances in deep learning made it possible to extract a 3D skeleton from a single RGB image. Taking advantage of this impressive progress, we propose a simple framework for fast and view-invariant action recognition using a single RGB camera. The proposed pipeline can be seen as the association of two key steps. The first step is the estimation of a 3D skeleton from a single RGB image using a CNN-based pose estimator such as VNect. The second one aims at computing view-invariant skeleton-based features based on the estimated 3D skeletons. Experiments are conducted on two well-known benchmarks, namely, IXMAS and Northwestern-UCLA datasets. The obtained results prove the validity of our concept, which suggests a new way to address the challenge of RGB-based view-invariant action recognition.

## 1 INTRODUCTION

Understanding human motion from a video represents a fundamental research topic in computer vision due to the diversity of possible applications such as video surveillance (Baptista et al., 2018), Human-Computer Interaction (Song et al., 2012), coaching (Lea et al., 2015), etc. A huge number of methods have been proposed and have proven their ability to efficiently recognize human actions as reflected in these two surveys (Aggarwal and Xia, 2014; Poppe, 2010). Usually, it is important to note that classical approaches assume ideal conditions. For example, in (Wang et al., 2011; Wang and Schmid, 2013; Fernando et al., 2015), the subject performing the action is considered to be facing the camera. However, in a real-world scenario, camera positioning as well as human body orientation can vary, and consequently affect the recognition task if the used method does not take into account the viewpoint variability. In fact, viewpoint invariance represents one of the most important challenges in human action recognition. Solving view-invariance requires relating a given acquisition of the subject to its 3D representation. While it is a simple task with RGB-D cameras, it is less obvious using RGB cameras, which only provide 2D information, and no explicit 3D. The development of low-cost RGB-D cameras has made possible the real-time extraction of 3D information via depth maps and skeletons. This has significantly boosted the research on viewpoint invariant action recognition (Haque et al., 2016; Hsu et al., 2016; Xia et al., 2012). However, the disadvantages of RGB-D based approaches are tied to RGB-D sensors. First, the estimation of an acceptable depth map and skeleton is limited within a specific range. Second, RGB-D cameras show a high sensitivity to external lighting conditions, making outdoor applications potentially challenging. Both of these reasons restrict their applicability in real-world scenarios such as in video surveillance.
There is therefore a need to solve the view-invariance problem using RGB cameras. Among the most successful state-of-the-art approaches are methods based on knowledge transfer (Gupta et al., 2014; Rahmani and Mian, 2015). To ensure view-invariance, these methods find a view-independent latent space where the features are mapped and then compared. To achieve that, they use 3D synthetic data computed by fitting cylinders to real data captured with a Motion Capture (MoCap) system, and by projecting them to various viewpoints.

The aforementioned approaches make use of trajectory shape descriptors (Wang et al., 2011). These descriptors are, by definition, not view-invariant. In-

deed, motion shape in 2D can only be described as points on the image grid; therefore, any radial motion information is mostly lost. In addition, some actions include similar motion patterns from different body parts, which can negatively impact the classification (Papadopoulos et al., 2017).

In this paper, instead of relying on a set of 2D projections of synthetic data, we propose to augment 2D data by a third component. Motivated by the very recent encouraging progress on pose estimation from a single RGB image (Pavlakos et al., 2017; Mehta et al., 2017; Yang et al., 2018), we introduce a novel way of approaching the viewpoint invariant action recognition problem using a single 2D or RGB camera. Our approach consists in estimating human 3D poses from 2D sequences, then directly using this 3D information with a robust 3D skeleton descriptor. Using 3D skeleton-based descriptors makes the approach fully view-invariant, since they involve 3D points for describing the body structure. Such descriptors have been proven robust in multiple scenarios (Xia et al., 2012; Yang and Tian, 2012). The main advantages of this framework are its simplicity and its low computation time thanks to the use of a high-level representation. In order to validate it, we propose to use the Convolutional Neural Network (CNN)-based approach referred to *VNect* for the estimation of 3D skeletons from 2D videos (Mehta et al., 2017). The VNect system was selected over related ones (Pavlakos et al., 2017; Mehta et al., 2017; Yang et al., 2018), because of its real-time performance and its ability to ensure temporal coherence. Two different view-invariant skeleton-based descriptors are used to test this framework, namely, *Kinematic Spline Curves* (KSC) (Ghorbel et al., 2018; Ghorbel et al., 2016) and *Lie Algebra Representation of body-Parts* (LARP) (Vemulapalli et al., 2014). Finally, the experiments are conducted on two different cross-view action recognition benchmarks: the Northwestern-UCLA (Wang et al., 2014) and the IXMAS (Weinland et al., 2006) datasets.

The main contributions of this paper may be summarized as follows:

- A novel framework for fast view-invariant human action recognition using a single RGB camera.
- Comparison of two different view-invariant skeleton-based descriptors integrated into the proposed framework.
- Extensive experimental evaluation on two well-known datasets and a deep analysis of the obtained results.

The remainder of the paper is organized as follows: In Section 2, relevant state-of-the-art approaches are summarized. Section 3 presents the proposed framework and details the used skeleton-based descriptors. The experimental evaluation is described in Section 4, along with a thorough discussion of the results. Finally, Section 5 concludes this work and presents directions for future works.

# 2 RELATED WORK

As mentioned in Section 1, invariance to viewpoint represents a major challenge in action recognition. Viewpoint invariant human action recognition can be categorized into two main classes: RGB-D and RGB based approaches as overviewed below.An extensive review may be found in the recent survey by Trong et al. (Trong et al., 2017).

## 2.1 RGB-D-based methods

The emergence of RGB-D cameras has importantly facilitated the task of viewpoint invariant action recognition thanks to the availability of 3D information (Hsu et al., 2016; Rahmani et al., 2014). Indeed, RGB-D cameras provide depth images that may be directly used for defining view-invariant descriptors.

Depth images only provide partial 3D information. In the context of action recognition, human 3D skeletons estimated from depth images are considered to be a more complete high level 3D representation, which is view-invariant by nature. In addition, with the rapid development of dedicated algorithms to estimate skeletons from depth maps such as (Shotton et al., 2013), numerous view-invariant skeleton-based approaches have been proposed. One of the pioneering works has been introduced in (Xia et al., 2012), where a descriptor encoding a histogram of 3D joints was proposed. Nevertheless, since the absolute position of joints is used, these features are sensitive to anthropometric variability. To resolve this issue and preserve view-invariance, some approaches proposed to describe actions using the distance between joints. For instance, in (Yang and Tian, 2012), actions are depicted using a novel descriptor called *eigenjoints*. The latter is computed by applying Principal Component Analysis (PCA) on the spatial and temporal Euclidean distances between joints.
To cope with viewpoint variability and increase accuracy, other approaches have modeled human actions using more sophisticated geometric tools. In (Evangelidis et al., 2014), authors proposed a novel view-invariant representation by introducing a descriptor based on the relative position of joint quadruples. Also, Vemulapalli et al. suggested a new representa-

tion called Lie Algebra Representation of body-Parts (LARP) by computing the geometric transformation between each pair of skeleton body-parts (Vemulapalli et al., 2014).

The presented descriptors are implicitly unaffected by the viewpoint variability as they are defined using invariant features such as the distance between joint, angles, transformation matrices, etc. Nevertheless, since the 3D skeleton contains the full 3D information, an alignment pre-processing can be simply applied before undertaking the descriptor computation. For example, we cite the work of (Ghorbel et al., 2018), where the motion has been modeled by computing and interpolating kinematic features of joints. In this case, the *Kinematic Spline Curves* (KSC) descriptor is not view-invariant by nature; thus, the skeletons are initially transferred to a canonical pose.

Although these representations have shown their effectiveness in terms of computation time and accuracy, they are hardly applicable in various scenarios, since the skeletons are estimated using RGB-D cameras. Indeed, the skeleton estimation accuracy decays in the presence of a non-frontal view (Rahmani et al., 2016) due to self-occlusions. Furthermore, as mentioned in Section 1, RGB-D cameras require specific conditions to optimally work such as outdoor environment, closeness to the camera, moderate illumination, etc. As a result, RGB-D based human action recognition has limited applications.

## 2.2 RGB-based methods

Very recent efforts have been made to propose view-invariant human action recognition methods using a monocular RGB camera. However, the challenge is that RGB images do not explicitly contain 3D information and consequently traditional descriptors, such as the Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Motion Boundary Histograms (MBH) (Dalal et al., 2006), are highly affected by the introduction of additional views (Presti and Cascia, 2016). Thus, some RGB-based methods have been specifically designed to overcome viewpoint variation (Gupta et al., 2014; Li et al., 2012; Zhang et al., 2013; Wang et al., 2014; Li and Zickler, 2012; Rahmani and Mian, 2015; Weinland et al., 2006; Lv and Nevatia, 2007).

One way of approaching the problem is to match one viewpoint to another using geometric transformation as in (Weinland et al., 2007; Lv and Nevatia, 2007). However, this category of methods which are usually based on 3D examplars require the use of labeled multi-view data. Another way consists in designing spatio-temporal features which are insensitive to viewpoint variation (Li et al., 2012; Parameswaran and Chellappa, 2006; Rao et al., 2002). However, their discriminative power has been shown to be limited (Rahmani and Mian, 2015).

The most popular RGB-based approaches are *knowledge transfer*-based methods. The idea of knowledge transfer for view-invariant action recognition is to map features from any view to a canonical one by modeling the statistical properties between them. For instance, Gupta et al.introduced a novel knowledge transfer approach using a collection of data containing unlabeled MoCap sequences (Gupta et al., 2014). Dense motion trajectories from RGB sequences are matched to projections of 3D trajectories generated from synthetic data (cylinders fitted to MoCap data). However, the number of these projections is finite, which means that not every viewing angle is represented. In addition, it is highly possible that different but similar-looking (from a specific angle) 2D motion patterns are incorrectly matched, since the 2D descriptors used in this context are view-dependent.

In (Rahmani and Mian, 2015), dense motion trajectories (Gupta et al., 2014) are computed using synthetic data similar to (Gupta et al., 2014), and represented using a codebook. A histogram is then built in order to be used as a final descriptor. This particular method is robust even when the testing view is completely different from the training views. This is due to the fact that the introduced Non-Linear Transfer Model (NKTM) allows the approximation of non-linear transformations. Despite their efficiency, the two methods proposed in (Rahmani and Mian, 2015) and in (Gupta et al., 2014) rely on 2D-based descriptors that are not invariant to viewpoint changes.

# 3 PROPOSED FRAMEWORK FOR RGB-BASED VIEW-INVARIANT ACTION RECOGNITION

In this section, we present the proposed framework to perform a fast view-invariant human recognition from a single RGB camera. Inspired by the advances in human pose estimation and the performance of skeleton-based approaches, we propose to first generate 3D human skeletons from a monocular RGB camera based on the recently introduced CNN-based approaches. Then, the extracted skeletons are used to compute skeleton-based features. Figure 1 illustrates the proposed pipeline. In what follows, we detail the different steps of this pipeline.

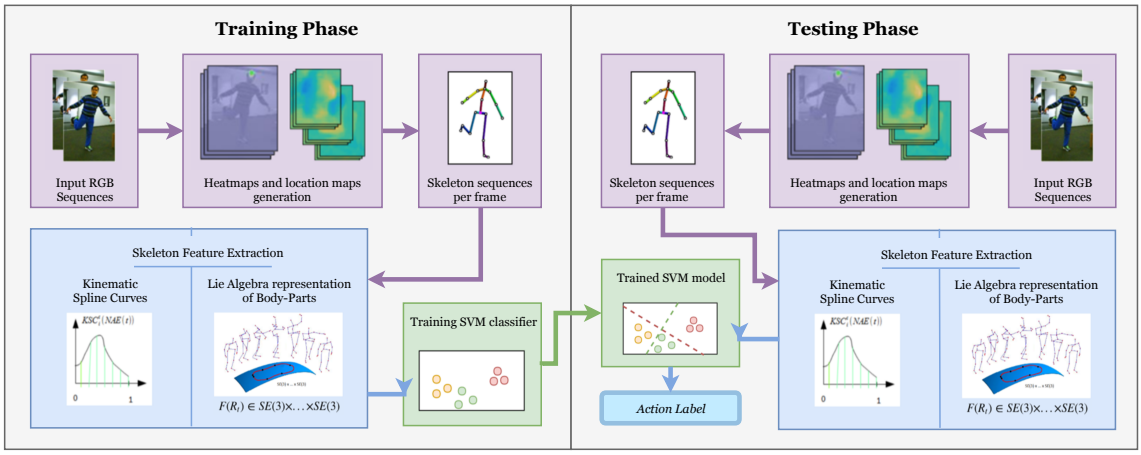


Figure 1: Overview of the proposed pipeline for fast and view-invariant human action recognition from a monocular RGB image: in both the training phase and the testing phase, skeletons are extracted from RGB images using the heatmaps and locations maps generated by the VNect algorithm (Mehta et al., 2017). Then, based on the estimated skeleton, skeleton features are computed e.g., LARP and KSC. Finally, in order to train a model of classification and use it to recognize actions, linear SVM is used.

## 3.1 Pose estimation from a monocular RGB image

Given a sequence of RGB images $\mathbf{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_t, \cdots, \mathbf{R}_N\}$, where $N$ is the total number of frames, the goal of a pose estimation algorithm $f(\cdot)$ is to extract a 3D skeleton composed of $n$ joints. We denote the sequence of extracted skeletons by $\mathbf{P}(t) = [\mathbf{P}_1(t), \mathbf{P}_2(t), ..., \mathbf{P}_n(t)]$ such that

$$\mathbf{P}(t) = f(\mathbf{R}_t), \tag{1}$$

where $f(\cdot)$ is a function that maps a single RGB image to an estimated representation of the human pose in three dimensions and where $t \in [1, \cdots, N]$ denotes the frame index. Lately, with the recent success of Deep Neural Networks, numerous methods have been proposed to estimate 3D skeletons from a single RGB image (Bogo et al., 2016; Tekin et al., 2016). However, the resulting 3D estimated skeletons are neither temporally stable nor computed online (Mehta et al., 2017). Meanwhile, VNect proposed in (Mehta et al., 2017) addresses both of these issues effectively.

As in (Mehta et al., 2016; Pavlakos et al., 2017), VNect makes use of CNN models. However, authors select a smaller architecture, Residual Netoworks (ResNet) to achieve real-time performance. More importantly, it is based on a network architecture with fewer parameters, hence inference can be done in a computationally efficient manner. This CNN pose regression allows the estimation of 2D and 3D skeletons using a monocular RGB camera. To that aim, for each joint $j$, the network is trained to estimate a 2D heatmap $\mathbf{H}_j$ of body parts along with joint location maps in each of the three dimensions, which we denote as $\mathbf{X}_j, \mathbf{Y}_j, \mathbf{Z}_j$. The position of each joint $j$ is therefore estimated by extracting the maximum values from the location maps of the associated heatmap $\mathbf{H}_j$.

The network is trained by considering the weighted $L_2$ norm difference between estimated joint location and the ground truth– the cost is summed over each dimension. For instance, the loss of predicting location $x_j$, is given as

$$\text{Loss} = \|\mathbf{H}_j^{GT} \odot (\mathbf{X}_j - \mathbf{X}_j^{GT})\|_2, \tag{2}$$

where $GT$ refers to the Ground Truth and $\odot$ indicates the Hadmord product.

The network is pre-trained using the annotated 3D and 2D human datasets (Ionescu et al., 2014; Mehta et al., 2016; Andriluka et al., 2014).

In order to ensure temporal coherence, the estimated joint positions are later smoothed. This is of great importance in our case since our goal is to recognize actions.

## 3.2 Feature extraction

Using the estimated skeletons, we propose to independently integrate two different view-invariant skeleton-based methods: LARP (Vemulapalli et al., 2014) and KSC (Ghorbel et al., 2018). In (Vemulapalli et al., 2014), the used features are view-invariant by nature, while in (Ghorbel et al., 2018), a skeleton alignment pre-processing is realized. In the following two subsections, we describe both LARP and KSC.

#### 3.2.1 Lie algebra representation of body-Parts (LARP)

In (Vemulapalli et al., 2014), an efficient skeleton-based action recognition approach is introduced. The approach is based on describing the geometric relationship between different coupled body segments. Let $\mathbf{S}(t) = (\mathbf{P}(t), \mathbf{E}(t))$ be a set of skeleton sequences $\mathbf{P}(t)$ with $n$ joints, and $m$ rigid-oriented body parts $\mathbf{E}(t)$. The skeleton sequence are described in (1), while the rigid-body parts are defined as $\mathbf{E}(t) = \{\mathbf{e}_1(t), \mathbf{e}_2(t), \cdots, \mathbf{e}_m(t)\}$. Each body part $\mathbf{e}_1(t)$ is assigned a 3D local coordinate system. Then, between each couple of local coordinate systems attached to the body-parts $\mathbf{e}_i(t)$ and $\mathbf{e}_j(t)$, a 3D rigid transformation matrix $\mathbf{T}_{i,j}(t)$ is defined as:

$$\mathbf{T}_{i,j}(t) = \begin{bmatrix} \mathbf{Q}_{i,j}(t) & \mathbf{t}_{i,j}(t) \\ 0 & 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{Q}_{m,n}$ is a $3 \times 3$ rotation matrix and $\mathbf{t}_{i,j}(t)$ a three-dimensional translation vector.

To completely encode the geometric relation between $\mathbf{e}_m$ and $\mathbf{e}_n$, both $\mathbf{T}_{m,n}$ and $\mathbf{T}_{n,m}$ are estimated. Subsequently, a sequence of skeletons varying over time is represented as $\mathbf{C}(t) = [\mathbf{T}_{1,2}(t), \mathbf{T}_{2,1}(t), ..., \mathbf{T}_{n,m}(t), \mathbf{T}_{m,n}(t)]$.The set of rigid transformation matrices define a direct product of non-Euclidean observation space called the Special Euclidean group SE(3). As a result, each representation of a skeleton is a point and skeleton sequence is a curve in $\mathrm{SE}(3)^{2C_m^2}$, with $C_m^2$ denoting the combination operation. Classification of the observed curves is done on the tangent space of the identity matrix, using Support Vector Machine (SVM) algorithm. Note that, a preliminary point matching is necessary to achieve temporal alignment which, in (Vemulapalli et al., 2014), is achieved via dynamic time warping and Fourier temporal pyramid representation. The use of 3D rigid transformation matrices between body-parts as features ensures the view-invariance, since they are independent from the view of acquisition.

#### 3.2.2 Kinematic spline curves (KSC)

This second skeleton-based representation has been introduced in (Ghorbel et al., 2018) and is mainly characterized by its compromise between computational latency and accuracy. To do that, the chosen components are carefully selected to ensure accuracy and computational efficiency. The descriptor is based on the computation of kinematic values, more specifically joint position $\mathbf{P}(t)$, joint velocity $\mathbf{V}(t)$ and joint acceleration $\mathbf{A}(t)$.


Figure 2: Frame samples from the Northwestern-UCLA dataset: an example is given for each viewpoint $V_1$, $V_2$and $V_3$

The key idea of this approach is to define a kinematic curve of a skeleton sequence as

$$\mathbf{KF}(t) = [\mathbf{P}(t), \mathbf{V}(t), \mathbf{A}(t)]. \tag{4}$$

Subsequently, a kinematic curve can be reparameterized such that it is invariant to execution rate using a novel method called Time Variable Replacement (TVR) (Ghorbel et al., 2018). As its name indicates, this method consists in changing the variable time by another variable that is less influenced by the variability in execution rate. It be written as

$$\mathbf{KF}(\phi(t)) = [\mathbf{P}(\Phi(t)), \mathbf{V}(\Phi(t)), \mathbf{A}(\Phi(t))]. \tag{5}$$

The new parameter $\phi$ is constrained to be bijective, increasing with respect to $t$, and have a physical rate-invariant meaning. In our case, we use the Pose Motion Signal Energy function proposed in (Ghorbel et al., 2018) to define $\phi$. Subsequently, in order to obtain a meaningful descriptor, the discrete data point samples $\mathbf{KF}(\phi(t))$ are interpolated using a cubic spline interpolation, then, uniformly sampled. Finally, the classification is carried out using a linear SVM. It is important to note that the computation of this descriptor includes also skeleton normalization and skeleton alignment steps making it respectively invariant to anthropometric and viewpoint changes. The alignment is carried out by estimating a transformation matrix between each skeleton and a canonical pose.

## 4 EXPERIMENTS

The proposed pipeline is tested on two different cross-view human action recognition benchmarks: the Northwestern-UCLA Multiview Action3D (Wang et al., 2014) denoted by N-UCLA and the INRIA Xmas Motion Acquisition Sequences dataset (Rahmani and Mian, 2015) denoted by IXMAS.

## 4.1 Datasets

### 4.1.1 Northwestern-UCLA dataset

The Northwestern-UCLA dataset consists of videos captured by using 3 different Kinect sensors from different viewpoints. Thus, this dataset contains in total 3 modalities: RGB images, depth maps and skeleton sequences and includes 10 action classes: *pick with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw* and *carry*. Each action class is repeated by 10 subjects from 1 to 6 times. The main challenge of this dataset is that it contains very similar actions such as *pick with one hand* and *pick up with two hands*. Figure 2 illustrates examples from this benchmark.

### 4.1.2 IXMAS dataset

This dataset is captured using 5 synchronized RGB-cameras placed in 5 different viewpoints: four from the side and one from the top of the subject. IXMAS dataset is constituted from 11 different action categories: *check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick* and *pick up*. This dataset is challenging since it contains complex viewpoints leading to self-occlusions. Such viewpoints are illustrated in Figure 4 (top row).

## 4.2 Experimental settings and implementation details

All the experiments were run on an i7 Dell Latitude laptop with 16GB RAM and implemented in *Matlab*. For both datasets, we follow the same experimental protocol used in (Rahmani and Mian, 2015). For the case of the Northwestern dataset, two viewpoints are used for the training and the third for the testing. In total, 3 experiments are performed. Moreover, each test on IXMAS dataset involves every combination of viewpoint pairs for training and testing, resulting in 20 experiments in total.

In this work, we consider two types of experiments: *VNect+KSC* and *VNect+LARP*. VNect+KSC refers to our framework combined with the KSC descriptor, while VNect+LARP denotes our framework merged with the LARP descriptor. We compare our framework with the recent RGB-based methods denoted in the rest of the paper by Hanklets (Li et al., 2012), Dis- criminative Virtual Views (DVV) (Li and Zickler, 2012), AND-OR Graph (AOG) (Wang et al., 2014), Continuous Virtual Pat (CVP) (Zhang et al., 2013), Non-linear Circulant Temporal Encoding (nCTE) (Gupta et al., 2014) and Non-linear Knowledge Transfer Model (NKTM) (Rahmani and Mian, 2015).

| {Source} \| {Target} | {1,2} \| 3 | {1,3} \| 2 | {2,3} \| 1 | Mean |
|---|---|---|---|---|
| Hankelets (Li et al., 2012) | 45.2 | - | - | - |
| dvv1 (Li and Zickler, 2012) | 58.5 | 55.2 | 39.3 | 51.0 |
| CVP (Zhang et al., 2013) | 60.6 | 55.8 | 39.5 | 52.0 |
| AOG (Wang et al., 2014) | 73.3 | - | - | - |
| nCTE (Gupta et al., 2014) | 68.8 | 68.3 | 52.1 | 63.0 |
| NKTM (Rahmani and Mian, 2015) | 75.8 | 73.3 | 59.1 | 69.4 |
| VNect+LARP (ours) | 70.0 | 70.5 | 52.9 | 64.47 |
| VNect+KSC (ours) | **86.29** | **79.72** | **66.53** | **77.51** |

Table 1: Accuracy of recognition (%) on the Northwestern-UCLA dataset: We report the accuracy obtained for each test (when two viewpoints are used for training (Source) and one viewpoint for testing (Target)) and the average accuracy for the three tests (Mean).

## 4.3 Results and discussion

The results on the Northwestern-UCLA dataset are reported in Table 1 and prove that our method (VNect+KSC) outperforms state-of-the-art methods. Indeed, an increase of around 8% compared to the most competitive approach can be noted (NKTM (Rahmani and Mian, 2015)). Moreover, Figure 3 shows that for almost all action classes, VNect+KSC outperforms nCTE(Gupta et al., 2014) and NKTM(Rahmani and Mian, 2015). On the other hand, despite the fact that VNect+LARP shows a lower accuracy by 5% compared to NKTM, this approach stands among the best performing ones, showing promising results.

The results for the IXMAS dataset are presented in Tables 2 and 3. Our proposed approach (VNect+KSC) achieves the third best mean recognition accuracy, achieving 58.12% (against 72.5% for NKTM (Rahmani and Mian, 2015) and 67.4% for NCTE(Gupta et al., 2014)). However, as depicted in Table 2, for every viewpoint pair, our approach shows a competitive performance, except for the ones which include viewpoint $V_4$. For example, tests $2 \mid 0$ and $2 \mid 3$ outperform earlier works and respectively reach an accuracy of 85.2% and 88.5%, while tests $0 \mid 4$ and $2 \mid 4$ present very low results (respectively 15.5% and 16.4%). This poor performance is the result of erroneous and noisy skeleton estimation coming from the pose estimator. Figure 4 illustrates an example of the extraction of skeletons from different viewpoints using VNect. This figure highlights the fact that all skeletons are visually coherent except for the one extracted from $V_4$ which represents the top viewpoint. The presence of self-occlusions in $V_4$ is crucial for the performance of VNect, since it makes the skeleton estimation by nature more challenging. Nevertheless, this constraint can be generalized to other approaches, affecting their performance, as well. By investigating

| {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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hankelets (Li et al., 2012) | 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 |
| DVV (Li and Zickler, 2012) | 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 |
| CVP (Zhang et al., 2013) | 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 |
| nCTE (Gupta et al., 2014) | **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 |
| NKTM (Rahmani and Mian, 2015) | 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** |
| VNect+LARP (ours) | 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 |
| VNect+KSC (ours) | 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 |

Table 2: Accuracy of recognition (%) on the IXMAS dataset: the different tests are detailed. Each time, one viewpoint is used for training (Source) and another one for testing (Target).

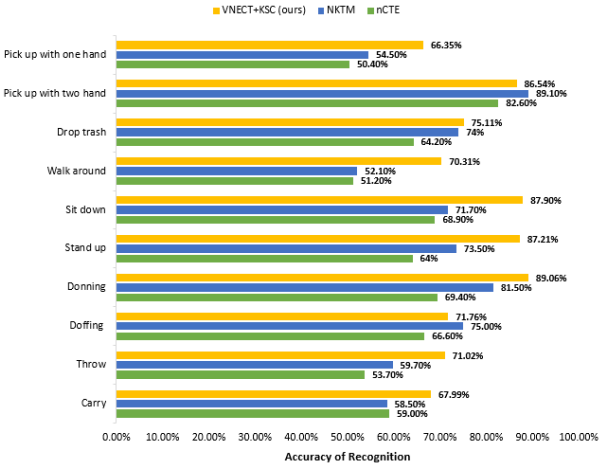


Figure 3: Action recognition accuracy for each action on the Northwestern-UCLA dataset: comparison of our method with NKTM(Rahmani and Mian, 2015) and nCTE(Gupta et al., 2014)

more on this question, we discovered that VNect is not trained on extreme viewpoints such as $V_4$. Thus, we underline a very interesting research issue to study in the future.

For this reason, we propose to evaluate the proposed concept by keeping in mind that the current version of VNect is not adapted yet to the estimation of skeletons from top views. Thus, we compute the average accuracy by ignoring the tests where $V_4$ has been considered. The results reported in Table 3 show that our approach competes with state-of-the-art by achieving 83.03% of recognition. It shows the second highest accuracy after NKTM (Rahmani and Mian, 2015) approach (reaching 84.46%) with only 1% of difference.

#### 4.3.1 RGB-based skeletons vs. RGB-D-based skeletons

In order to compare the quality of skeletons extracted from VNect compared to the ones provided by RGB-D cameras for the task of action recognition, we propose to compute the KSC descriptor using both the VNect-generated skeletons and the RGB-D skeletons.

Results obtained on the Northwestern-UCLA dataset are reported in Table 4. Skeleton-RGB-D and skeleton-VNect refer to the results obtained by applying respectively the KSC descriptor to the skeletons provided by the Kinect and the skeletons provided by the VNect. The reported results show that action recognition can be more robust using VNect-generated skeleton sequences. In fact, using VNect skeletons, the mean accuracy increased by 7.4% compared to the utilization of the provided RGB-D skeleton sequences. The reason for that is the fact that the extraction of skeletons from RGB-D cameras is less accurate when the human body is not totally visible. With the variation of human body orientation with respect to the camera, self-occlusions occur, impacting negatively the skeleton estimation.

| {Source} \| {Target} | Mean with $V_4$ | Mean without $V_4$ |
|---|---|---|
| Hankelets (Li et al., 2012) | 56.4 | 61.41 |
| DVV (Li and Zickler, 2012) | 38.2 | 36.2 |
| CVP (Zhang et al., 2013) | 42.2 | 49.60 |
| NCTE (Gupta et al., 2014) | 67.4 | 80.45 |
| NKTM (Rahmani and Mian, 2015) | **72.5** | **84.46** |
| LARP-VNect (ours) | 31.50 | 45.91 |
| KSC-VNect (ours) | 58.12 | 83.03 |

Table 3: Average accuracy of recognition (%) on the IXMAS dataset: the first value (Mean with $V_4$) reports the average of all the tests done, while the second value (Mean without $V_4$) computes the average of all texts excepting the ones involving $V_4$.

#### 4.3.2 LARP vs. KSC

The results performed on the Northwestren dataset as well as on the IXMAS dataset show the superiority of KSC descriptor for viewpoint action recognition when combined with VNect skeletons. Indeed, KSC+VNect presents an average accuracy of 77.51% against 64.47% for VNect+LARP on the Northwestern UCLA dataset. On IXMAS dataset, KSC outperforms LARP, as well, by achieving an average accuracy of 83.03% against 58.12% when ignoring $V_4$ and of 45.91% against 31.5% when considering it. The interpretation of this result lies on the fact that KSC+VNect is less sensitive to noise than LARP.

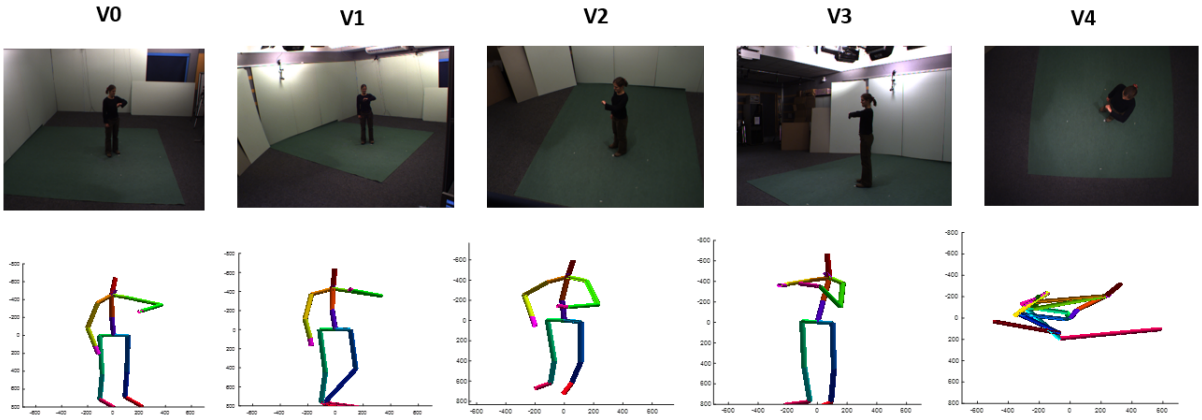


Figure 4: Illustration of skeleton extraction from the IXMAS dataset using VNect system: it can be noted that for the four first views ($V_0, V_1, V_2, V_3$), the quality of the estimated is visually acceptable. However, the quality of the last view $V_4$ is completely biased. This fact is confirmed by our experiments.

| {Source} \| {Target} | {1,2}\| 3 | {1,3}\| 2 | {2,3}\| 1 | Mean |
|---|---|---|---|---|
| skeleton-RGB-D | 80.5 | 72.6 | 61.0 | 71.1 |
| skeleton-VNect | **86.3** | **79.7** | **66.5** | **77.5** |

Table 4: Accuracy of recognition (%) on the Northwestern dataset using the KSC descriptor: the performances obtained when using the skeletons provided by RGB-cameras and the ones extracted using VNect algorithm are compared. We report the accuracy obtained for each test (when two viewpoints are used for training and one viewpoint for testing) and the average accuracy (Mean).

| Method | Training + Testing |
|---|---|
| AOG* (Wang et al., 2014) | 1020 |
| NCTE* (Gupta et al., 2014) | 612 |
| NKTM* (Rahmani and Mian, 2015) | 38 |
| VNect+KSC | 6 |

Table 5: Computation time in minutes on the Northwestren dataset by using $V_1$ and $V_2$ for training and $V_3$ for testing. All the reported computation time includes descriptor calculation. *The reported values for AOG (Wang et al., 2014), NCTE (Gupta et al., 2014), NKTM (Rahmani and Mian, 2015) have been reported from the paper (Rahmani and Mian, 2015) and therefore the computation time has not been computed on the same computer.

#### 4.3.3 Computation time and memory

The main advantage of our framework is its low computation time. The training plus testing process takes only 6 minutes, as presented in Table 5. This shows that our framework can be considered as a real-time system.
On the other hand, the proposed framework, when using VNect for the skeleton estimation step, requires to consume only 58.5MB of further memory which is comparable to the memory needed to store the learned R-NKTM and the general codebook (57MB) in (Rahmani and Mian, 2015) and which is significantly lower than the memory needed to store the samples (30 GB) in (Gupta et al., 2014).

## 5 CONCLUSION AND FUTURE WORK

In this work, a simple but original framework has been proposed to resolve the issue of cross-view action recognition based on a single monocular RGB camera. For this purpose, a novel concept aiming at augmenting 2D images by a third dimension is proposed taking advantage of the recent advances in 3D pose estimation from a monocular RGB camera and the effectiveness of skeleton-based descriptors. A 3D skeleton is first estimated from a single 2D image using a CNN-based approach. Then, a view-invariant skeleton-based method is applied to the estimated skeletons. To prove the validity of our framework, the recently introduced VNect system has been chosen to extract 3D skeletons from RGB images. After that, two different view-invariant skeleton-based approaches have been tested: KSC (Ghorbel et al., 2018) and LARP (Vemulapalli et al., 2014). The experiments on two datasets have shown the superiority of KSC when integrated into that framework. The obtained results are competitive with respect to recent state-of-the-art approaches on both datasets, except for the cases where an extreme viewpoint (the top viewpoint) is considered. This suggests that it would be important to extend 3D pose estimator to extreme viewpoints.

## ACKNOWLEDGEMENTS

This work was funded by the European Unions Horizon 2020 research and innovation project STARR under grant agreement No.689947, and by the National Research Fund (FNR), Luxembourg, under the project C15/IS/10415355/3D-ACT/Björn Ottersten.

## REFERENCES

Aggarwal, J. K. and Xia, L. (2014). Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Baptista, R., Antunes, M., Aouada, D., and Ottersten, B. (2018). Anticipating suspicious actions using a small dataset of action templates. In *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., and Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*. IEEE.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part II*, ECCV'06, pages 428–441, Berlin, Heidelberg. Springer-Verlag.

Evangelidis, G., Singh, G., and Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4513–4518. IEEE.

Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., and Tuytelaars, T. (2015). Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387.

Ghorbel, E., Boutteau, R., Bonnaert, J., Savatier, X., and Lecoeuche, S. (2016). A fast and accurate motion descriptor for human action recognition applications. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 919–924. IEEE.

Ghorbel, E., Boutteau, R., Boonaert, J., Savatier, X., and Lecoeuche, S. (2018). Kinematic spline curves: A temporal invariant descriptor for fast action recognition. *Image and Vision Computing*, 77:60–71.

Gupta, A., Martinez, J., Little, J. J., and Woodham, R. J. (2014). 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., and Fei-Fei, L. (2016). Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*, pages 160–177. Springer.

Hsu, Y.-P., Liu, C., Chen, T.-Y., and Fu, L.-C. (2016). Online view-invariant human action recognition using rgb-d spatio-temporal matrix. *Pattern recognition*, 60:215–226.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Lea, C., Hager, G. D., and Vidal, R. (2015). An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In *Applications of computer vision (WACV), 2015 IEEE winter conference on*, pages 1123–1129. IEEE.

Li, B., Camps, O. I., and Sznaier, M. (2012). Cross-view activity recognition using hankelets. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1362–1369. IEEE.

Li, R. and Zickler, T. (2012). Discriminative virtual views for cross-view action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2855–2862. IEEE.

Lv, F. and Nevatia, R. (2007). Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Mehta, D., Rhodin, H., Casas, D., Sotnychenko, O., Xu, W., and Theobalt, C. (2016). Monocular 3d human pose estimation using transfer learning and improved CNN supervision. *CoRR*, abs/1611.09813.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36.

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

Parameswaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *International Journal of Computer Vision*, 66(1):83–101.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1263–1272. IEEE.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990.

Presti, L. L. and Cascia, M. L. (2016). 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147.

Rahmani, H., Mahmood, A., Huynh, D., and Mian, A. (2016). Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443.

Rahmani, H., Mahmood, A., Huynh, D. Q., and Mian, A. (2014). Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer.

Rahmani, H. and Mian, A. (2015). Learning a non-linear knowledge transfer model for cross-view action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Rao, C., Yilmaz, A., and Shah, M. (2002). View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116.

Song, Y., Demirdjian, D., and Davis, R. (2012). Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):5.

Tekin, B., Rozantsev, A., Lepetit, V., and Fua, P. (2016). Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000.

Trong, N. P., Minh, A. T., Nguyen, H., Kazunori, K., and Hoai, B. L. (2017). A survey about view-invariant human action recognition. In *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. IEEE.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558.

Wang, J., Nie, X., Xia, Y., Wu, Y., and Zhu, S.-C. (2014). Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656.

Weinland, D., Boyer, E., and Ronfard, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7. IEEE.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 20–27. IEEE.

Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., and Wang, X. (2018). 3d human pose estimation in the wild by

adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1.

Yang, X. and Tian, Y. L. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 14–19. IEEE.

Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S., and Shi, C. (2013). Cross-view action recognition via a continuous virtual path. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2690–2697.