Computer Vision and Image Understanding 148 (2016) 136–152

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A comparative study of pose representation and dynamics modelling for online motion quality assessment

Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi*, Sion Hannuna, Massimo Camplani, Tilo Burghardt, Ian Craddock

Faculty of Engineering, University of Bristol, Bristol BS8 1UB, United Kingdom

ARTICLE INFO

Article history:
Received 28 March 2015
Accepted 28 November 2015

Keywords:
Human motion quality
Human motion assessment
Continuous-state HMM motion analysis
Motion abnormality detection

ABSTRACT

Quantitative assessment of the quality of motion is increasingly in demand by clinicians in healthcare and rehabilitation monitoring of patients. We study and compare the performances of different pose representations and HMM models of dynamics of movement for *online* quality assessment of human motion. In a general sense, our assessment framework builds a model of normal human motion from skeleton-based samples of healthy individuals. It encapsulates the dynamics of human body pose using robust manifold representation and a first-order Markovian assumption. We then assess deviations from it via a continuous online measure. We compare different feature representations, reduced dimensionality spaces, and HMM models on motions typically tested in clinical settings, such as gait on stairs and flat surfaces, and transitions between sitting and standing. Our dataset is manually labelled by a qualified physiotherapist. The continuous-state HMM, combined with pose representation based on body-joints' location, outperforms standard discrete-state HMM approaches and other skeleton-based features in detecting gait abnormalities, as well as assessing deviations from the motion model on a frame-by-frame basis.

1. Introduction

Modelling and analysing human motion have been subject to extensive research in computer vision, in terms of feature extraction [1], action representation [2,3], action recognition [4,5], and abnormality detection [6]. While such works mostly apply to the challenging tasks of motion and action detection and recognition, only a few manage to provide a quantitative assessment of human *motion quality*. Such assessment aims at quantifying the motion quality from a functional point of view by assessing its deviation from an established model. This has potential use in many scenarios, for example, in sport applications [7], and for physiotherapists and medics [8], who may, for example, estimate the normality of human movement, possibly relative to a specific age group, or to quantify the evolution of their mobility during rehabilitation with respect to a personalized, preoperative model. Interestingly, physiotherapists assess human motion by *visually observing a person's ability to perform vital movements*, such as walking on a flat surface, sitting down, and gait on stairs, by rating the deviation from a normal movement using standard scores [9,10]. These well established scores are subjective and are insufficient to effectively monitor patients on a regular basis, as they can only be used by well-trained specialists and thus require the patients to be evaluated in clinical practices. Automated motion quality assessment can help in obtaining a more quantitatively accurate and temporal (inter-person and intra-person) comparative measure. It would also be essential for continuous assessment outside of a clinic, for example for use in the home for health and rehabilitation monitoring.

In addition to providing an overall score of 'normality', an *online* assessment measure can provide an immediate estimation of what parts of the motion deviate from normal, towards a more detailed understanding of the quality of the motion. The nature of online measures also enables assessing the motion before it has completed, thus allowing to trigger alerts, such as fall prevention in cases of unusually unstable gait.

This paper details and evaluates a method, first introduced in [11], for online estimation of the quality of movement from Kinect skeleton data, and presents its application to clinic-related movement types. To enable such an online assessment, a few challenges have been dealt with: (1) motion-related features are extracted from skeleton data and compacted into a lower-dimensional space to produce a simpler and more appropriate representation of pose, (2) a statistical model of human motion, that encapsulates both the appearance and the dynamics of the human motion, is learnt from training data of multiple individuals, suitable for periodic and nonperiodic motions,

* Corresponding author.Fax: +441179545209.
E-mail address: majid@cs.bris.ac.uk (M. Mirmehdi).

http://dx.doi.org/10.1016/j.cviu.2015.11.016

(3) an online quantitative assessment of motion is obtained by reference to the learnt model, which evaluates deviations in both appearance and dynamics on a frame-by-frame basis.

In [11], we proposed a framework in order to address these challenges, where we extracted 3D joint positions as a low-level feature, reduced their dimensionality while capturing their non-redundant information using a modified diffusion maps manifold method (challenge 1 above), modelled human movement with respect to a custom-designed statistical model (challenge 2), and evaluated the movement from an online measure based on the likelihood of the new observation to be described by such a model (challenge 3).

This paper updates and expands the work in [11], providing more thorough comparative evaluations of its framework and a comprehensive assessment of its individual modules, with the following additions: (a) in order to both demonstrate the versatility of our framework and further evaluate it, we apply our method to a variety of movement types, both periodic and non-periodic. (b) We show that the statistical model we introduced in [11] is in fact a continuous-state HMM, and we put it in perspective with more conventional variations of general HMM-based models. In particular, we compare their respective suitability to the task of capturing the dynamics of movements. (c) We assess what is the optimal pose representation for our HMM-based model of dynamics. First, as well as the joint position feature extracted from the skeleton data, we propose and compare against additional possible low-level skeletal features as some are more suitable for certain HMM models and for describing certain motions. Second, we investigate the optimum number of dimensions required in the manifold representations for describing the various low-level features. We also evaluate how the optimal pose representation varies with motion type. (d) We investigate whether the use of full-body information is beneficial for building pose representations, in particular for movements that are traditionally studied using partial-body information such as joints in the analysis of gait. (e) We propose a new online measure for quality assessment, and we compare it with the measure presented in [11].

Evaluation is performed on clinic-related motions of gait on stairs, walking on a flat surface, and transitions between sitting and standing – actions that are particularly relevant to the assessment of lower-extremity injuries. On the basis of testing on the dataset released in [11], a variety of common lower-extremity injuries are included in the test sequences. The groundtruth is labelled by a qualified physiotherapist.

Next, a review of the existing literature is provided in Section 2. Section 3 describes the framework for assessing the quality of a movement from skeleton data, introducing four variations of HMM techniques that are tested on our dataset. The experimental results are presented in Section 6, followed by a discussion and conclusion.

2. Related work

To consider the state-of-the-art, we now review related works on robust feature extraction from skeleton data, building a model of human motion from training data, and motion abnormality detection and quality of motion assessment, from both computer vision and clinical points of view.

2.1. Skeleton data from the depth sensors

A large number of studies have attempted to efficiently extract features from RGB images for analysing human actions, e.g. see [3], but RGB data is highly sensitive to view-point variations, human appearance, and lighting conditions. Recently, depth sensors have helped to overcome some of these limitations. Two commercially available devices are the Microsoft Kinect and Asus Xmotion, for which the depth is computed from structured light[1]. These sensors have become popular for modelling and analysing human motion, for example in [12], Uddin et al. extracted features from depth silhouettes using Local Directional Patterns and applied Principal Component Analysis (PCA) to reduce the dimensionality of their data. More commonly, motion analysis works exploit skeleton information derived from depth. Using random forests, 3D human skeletons are estimated at each frame from depth data by the Microsoft Kinect SDK [13] (for 20 joints) or by the OpenNI SDK [14] (for 15 joints). A human body pose can be well-represented as a stick figure made up of rigid segments connecting body joints. In this work, we use the OpenNI SDK to estimate skeleton data, as illustrated in Fig. 1. We focus next on methods that are based on skeleton features and refer the reader to a recent survey on non-skeleton features in [2].

Existing skeleton-based approaches have either used the full set of joints for general action recognition [15–18] or a subset chosen depending on the specific action/application [8,19]. In [19], only hips, knees, ankles and feet joints were used for detecting abnormal events during stair descent. The method in [8] used feet joints along with the projection of hand and torso joints for evaluating musculoskeletal disorders on patients who suffer from Parkinson's Disease (PD). To avoid action-specific approaches, we use the full set of joints along with dimensionality reduction techniques, explained next.

Robust feature extraction from skeleton data

A variety of low-level features have been used to represent the skeleton data: body joint locations [20], body joint velocities [17], body joint orientations [16], relative body joint positions [18], rigid segment angles [21] and transformations (rotations and translations) between various body segments [15]. Some of these proposed features may be more suitable for describing certain motions than others, e.g. the relative position and orientation between head and foot may provide sufficient description for the 'sitting' motion for some applications.

The high dimensionality of full-body skeleton data contains redundant information when modelling human motion, as will be demonstrated in Section 3.2. It is thus possible to employ dimensionality reduction methods to capture the intrinsic body configuration of the input data. It is common to apply linear PCA for dimensionality reduction in appearance modelling, however, human motion represented by skeleton data is highly non-linear and the mapping between the original data space and the reduced space is better described by non-linear mapping. Non-linear manifold learning methods have therefore been exploited for human motion recognition [22], such as locality preserving projections (LPP) [23] and isometric feature mapping (ISOMAP) [24].

While these approaches achieve dimensionality reduction for non-linear data, they are not necessarily unerring in handling outliers and/or very noisy data. The estimated skeleton will often be noisy. In fact the Kinect's skeleton pose estimation has mostly been trained for poses required for a gaming platform [25]. In case of occlusion or self-occlusion, the positions of joints are only roughly estimated (Fig. 1d, e). Furthermore, we are using the Kinect on a non-planar surface which does lead to less efficient skeleton proposals from the device. Some motion analysis approaches, such as [8], convolved the feature subspace with a Gaussian filter to achieve temporal smoothness. Others re-trained the pose estimator, e.g. for sign-interpreted gesture recognition [26].

Reducing the dimensionality of noisy data is still a challenging problem. Gerber et al. [27] introduced an extension of Laplacian Eigenmaps to cope with noisy input data, but such representation depends on the density of the points on the manifold, which may not be suitable for non-uniformly sampled data, such as skeleton data.

[1] Microsoft Kinect 2, released in 2014, uses time-of-flight technology.

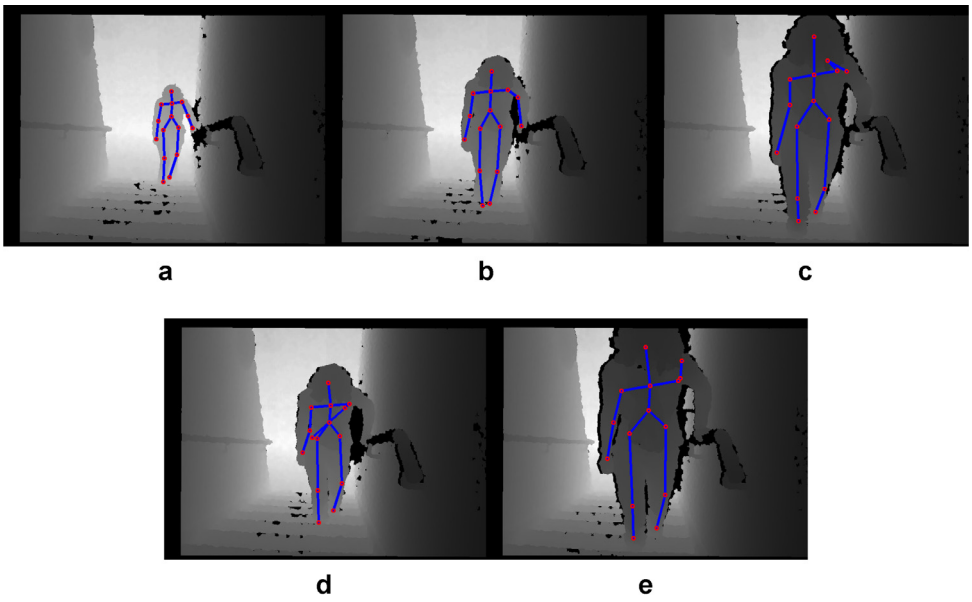

Fig. 1. RGB-D data and skeletons at bottom, middle, and top of the stairs ((a) to (c)), and examples of noisy skeletons ((d) and (e)).

2.2. Human motion modelling

Human motion (e.g. walking, jumping, sitting, kicking) typically consists of one or more body-part configurations that occur in a predefined order and could be periodic (e.g. walking, waving). A model of human motion thus often incorporates the related body-part configurations as well as temporal modelling of transitions and durations of these configurations. In the literature, there have been various approaches for the modelling of human motion. Only a few works model motions in order to assess their quality, while the majority build motion models for supervised recognition of actions (i.e. classifying the motion into a set of predefined labels). The modelling requirements may differ between these two tasks, for example in the sensitivity to modelling motion and sub-motion durations. Nevertheless, we review here the main works on modelling human motion regardless of their application.

Motion can be analysed by providing spatio-temporal features to a classifier. In [19], such features were extracted from lower body joints to train a binary classifier in order to distinguish abnormal motions from normal. These features can be made up of 3D XY-Time volumes computed from RGB [28] and depth images [29]. However, spatio-temporal volume representations are not suitable for online analysis as the motion analysis can only take place once the full motion sequence is observed.

Motion can also be seen as a sequence of body-part configurations. Dynamic Bayesian networks, such as Hidden Markov models (HMMs) and their variations, are the most popular generative models for sequential data and have been successfully used as probabilistic models of human motion, e.g. human gait [16,30,31]. In HMMs, each hidden state is associated with a collection of *similar* body poses and a transition model encapsulates sequences of body-part configurations. The most common HMM model is one that uses a fixed number of discrete states, known as the *classical* HMM, along with a discrete observation model. This has been used to recognise 10 basic actions in [16], and to classify motions between normal and abnormal in [12]. Continuous HMMs, which also use discrete states but continuous observation models such as a mixture of Gaussians, were used to recognise 22 actions in [31] and to distinguish normal from abnormal motions in [32]. Particularly in [32], optical flow features, together with feet position and velocity, were used to detect abnormalities during stairs descent from RGB data. The model uses 10 hidden states with full-covariance Gaussian mixture emissions and random initialisation of the EM algorithm. A single extra state with high covariance, low mixture proportion, and low transition probabilities were added for regularisation.

Apart from classical HMMs, extensions of HMMs introducing more flexible models have been widely applied. A hierarchical HMM (HHMM) was used in [33] along with a time-varying transition probability. Three-level hierarchies were implemented representing composite actions, primitive actions and poses respectively. In [34], a factored-state HHMM was used to define each state as a hierarchy of two-levels for each action and tested on a dataset of 4 basic actions.

For periodic motions, a cyclic HMM was tried on 4 basic actions in [35]. HMM variations that model state durations are frequently applied in activity recognition where temporal dependencies can be found. For example, Duong et al. [36] modelled the duration of each atomic action within an activity using a Coxian distribution, and thus modelled the activity by an HMM with explicit state durations. To the best of our knowledge, *HMM modelling of state duration* has never been applied to the modelling of human motion.

Online motion models

Depending on the application, the analysis can be either run offline incorporating data across the motion sequence, or processed online analysing an incoming frame before the entire motion is complete. Online motion models are important for scenarios such as surveillance, healthcare, and gaming. Most HMM-based motion modelling approaches mentioned above require temporal segmentation, and therefore are restricted to offline processing. The work in [33] dealt with online gesture recognition using a hierarchical HMM. To achieve online recognition, the method extended the standard decoding algorithm to an online version using a variable window [37], since the Viterbi algorithm cannot be directly applied to online scenarios.

Nowozin and Shotton [38] developed an online human action recognition system by introducing action points for precise temporal anchoring of human actions. Recently, works based on incremental learning have been applied to human motion analysis. In [39], an incremental covariance descriptor and on demand nearest neighbour classification were used for online gesture recognition. Instead of using incremental features, the work in [40] proposes a general framework via nonparametric incremental learning for online action

recognition which can be applied to any set of frame-by-frame feature descriptors.

2.3. Quality of motion assessment

We define the *quality of motion* as a continuous measure of the ability of the person to perform the motion when compared to a reference motion model. Such a model represents the normal range of motions (we simply refer to normal motion in the rest of the article.) for the relevant population group, or it could be a personalized model, and can be used to assess rehabilitation or pathological deterioration in mobility of humans for healthcare purposes. For example, quantitatively assessing the ability to balance on one leg following a knee replacement surgery could be used to track a person's rehabilitation. Similarly, Parkinson's patients' ability to stand up from a sitting position deteriorates with time, and continuous assessment of this functionality is needed to evaluate the progress of the disease [8]. The number of works targeting quality of motion are rare, with most attempting to perform abnormality detection as binary classification. Thus, we first briefly review some abnormality detection methods, and then focus on the small number of works on quantifying the degree of abnormality in human motion. We finish by presenting the current clinical approach for analysing motion quality.

2.3.1. Abnormality detection

Abnormality detection methods build a binary classifier to discriminate between normal and abnormal instances. Two main approaches exist, those that assume prior knowledge of expected abnormalities, and those that do not. In the first approach, the work of [19] used two support vector machine (SVM) binary classifiers that recognised normal and abnormal motions respectively, based on space-time features. The approach was tested on stairs descent and ascent motions, and it labelled normal and abnormal motions (e.g. fall or slip) from the classifier with the strongest response. Similarly, the work of [12] trained two HMMs on normal and abnormal gaits. Classification was also based on comparing the likelihood of the test sequence using both of these HMM models. No clear definition of 'abnormal' was provided in [12], and abnormalities encompassed a wide range of anomalies.

Abnormal motions may be highly variant and difficult to define a priori. Most abnormalities are rare and difficult to capture during training. The second approach, where there is no prior knowledge of abnormalities, predicts them as variations from the model of normal motion, built solely from regular/normal examples. This approach thus aims to quantitatively estimate the dissimilarity from the normal model—a kind of novelty detection. While this is a sensible compromise, the motion model needs to capture as much variation of normal motion examples as possible to avoid high false negative rates.

In [41], hierarchical appearance and action models were built for normal movements to detect abnormalities from RGB silhouettes in a home environment. For both hierarchies, appearance and action, the intra-cluster distance within a node was used to set a threshold for abnormalities.

The work that is most closely related to ours is [32] which used a single HMM for detecting abnormalities during stairs descent from RGB (only) data. The HMM was trained on sequences of normal 'descending stairs' motion, and a threshold on the likelihood was selected to detect abnormal sequences. Their results showed their system can successfully detect nearly all anomalous events for data captured in a controlled laboratory environment, but is highly reliant on accurate feet tracking.

2.3.2. Quality assessment of motion

Quality assessment focuses on calculating a discrete or continuous score that measures the match between a motion and a pre-trained model. Wang et al. [8] presented a method for quantitatively evaluating musculoskeletal disorders of patients who suffer from PD. One motion cycle from the training data was selected as a reference, and all other cycles were aligned to the reference for encoding the most consistent motion pattern. The method was tested for walking, as well as standing up, motion on PD and non-PD subjects. Results demonstrated that the method is able to quantify a clinical measurement which reflects a subject's mobility level. However, the specific features used (step size, arms and postural swing levels, and stepping time) make it difficult to generalise to other motions.

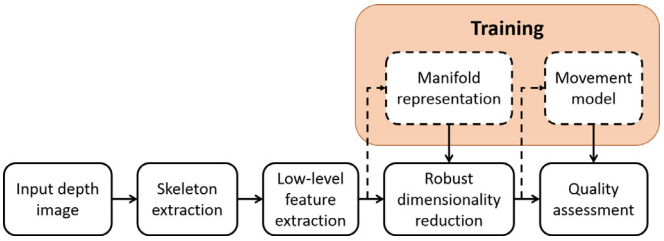

Fig. 2. Proposed pipeline for movement quality assessment: the dashed lines denote a learning phase that is performed off-line to create the two models represented by the dashed rectangles.

In a recent work on action assessment from RGB data, presented in [7], the quality assessment was posed as a supervised non-linear regression problem. The method provided a feedback score on how one performs in sports actions, particularly diving and figure skating, by comparing a test sequence with the labelled scores provided by coaches. Training a regression model required a relatively large number of labelled data points covering the spectrum of possible feedback scores.

In [11], we proposed a continuous measure of motion quality, computed online, as the log-likelihood of a continuous-state HMM model. To the best of our knowledge, [11] is the first and only work to address the problem of online quality assessment.

3. Proposed methodology

In this section, we describe our pipeline for assessing the quality of motion from skeleton data, as illustrated in Fig. 2. Skeleton data are first obtained from the OpenNI SDK [14]. Then, a low-level feature extraction stage (Section 3.1) determines a descriptor from the skeleton data. This is followed by a dimensionality reduction step that is made less sensitive to noise and non-linear manifold learning (Section 3.2). In the reduced space, the significant and non-redundant aspects of the pose and the dynamics of the motion are expected to be preserved. A model of the motion is then learnt off-line from instances of 'normal motion' (Section 3.3). The quality of movement is assessed by measuring the deviation of a new observation from the learned model (Section 5).

This pipeline was first presented in our previous work [11] where only one possible low-level skeleton feature and one possible motion model were discussed. Here, we introduce and compare different low-level features, and we assess our motion modelling method with respect to more traditional discrete HMM-based models.

3.1. Skeleton data representation

Skeleton data are view-invariant[2] and depth information alleviates the effect of human appearance differences and lighting variations. As a first step, we apply an average filter over a temporal window for each joint position independently in order to compensate for the high amount of noise typically found in OpenNI skeletons.

[2] Although the performance of the OpenNI SDK skeleton tracker suffers severely when the subject is not facing the camera.

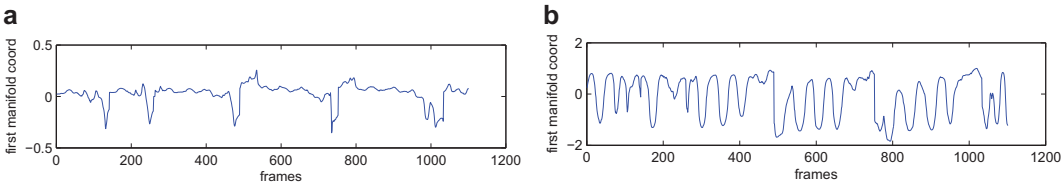

Fig. 3. First dimension of gait data in reduced space, using JP low-level feature. (a) original diffusion maps according to [27], (b) robust diffusion maps according to [11].

Given J joints, where $J = 25$ or $J = 15$ for skeletons from the Microsoft Kinect2 SDK or OpenNI SDK respectively, and a pose $\hat{C} = [\hat{c}_1, \ldots, \hat{c}_J]^T \in \mathbb{R}^{3J\times 1}$ comprising smoothed 3D positions $\hat{c}_i$ in J, a normalised pose $C = g(\hat{C})$ is computed to compensate for global translation and rotation of the view point, and for scaling due to varying heights of the subjects. The normalising function $g(\cdot)$ could be Procrustes alignment or other alignment approaches depending on which feature is in use. Let F^t be the low-level skeleton feature at time t. Using features that previously appeared in works such as [16–18,20], we scrutinize four possible alternative feature descriptors for normalised pose:

1. Joint Positions (JP): concatenate and vectorise 3D coordinates $\hat{c}_i$ of all the joints at time t, to give features $F^t = C^t$.
2. Joint Velocities (JV): concatenate and vectorise the 3D velocities of all the joints, to give features $F^t = C^t - C^{t-1}$.
3. Pairwise Joint Distances (PJD): Given 3D positions of a normalised pose, we calculate a $J \times J$ Euclidean distance matrix between all pairs of joints where $d_{ij} = \|c_i - c_j\|$. Since this is a symmetric matrix with zero entries along the diagonal, we obtain a $J(J-1)/2$ feature vector $F^t = [d_{12}, \ldots, d_{1J}, d_{23}, \ldots, d_{(J-1)J}]^T$. The pairwise joint distances give unique coordinate-free representation of the pose kinematics.
4. Pairwise Joint Angles (PJA): The Kinect skeleton of the human body consists of $J-1$ line segments connecting pairs of neighbouring joints. Assuming the segment e_i connects two joints J_i and J_{i+1}, the Euler angle between two segments is computed as $\rho_{ij} = \arccos(\frac{e_i^T \cdot e_j}{\|e_i\|\|e_j\|})$. Our feature vector F^t is a $(J-1)(J-2)/2$ vector that consists of all the Euler angles for all segments, such that $F^t = [\rho_{12}, \ldots, \rho_{1(J-1)}, \rho_{23}, \ldots, \rho_{(J-2)(J-1)}]^T$. Concatenating all the Euler angles between any two body segments captures the full 3D angles between body parts.

In the rest of this paper, unless specified otherwise, these four feature descriptors are computed using all 25 or 20 body-joints from Kinect2 SDK or OpenNI SDK, respectively, and so represent the whole skeleton.

3.2. Robust manifold learning

As previously noted, skeleton data is highly redundant for modelling motion and does not represent its true complexity. To reduce the dimensionality of the low-level feature F_i, we select a non-linear manifold learning method - diffusion maps - which is a graph-based technique with quasi-isometric mapping Φ, from original higher space $\mathbb{R}^N$ to a reduced low-dimensional diffusion space $\mathbb{R}^n$, where $n \ll N$. Given a training set $\mathbf{F}$, where $F_i \in \mathbf{F}$, the method is capable of recovering the underlying structure of a complex manifold, has robustness to noise, and is efficient to implement when compared to conventional non-linear dimensionality reduction methods [42].

Building diffusion maps requires computing a weighted adjacency matrix W with the distances between neighbouring points weighted by a Gaussian kernel G:

$$w_{i,j} = G\big(F_i, F_j\big) \tag{1}$$

The optimal mapping Φ is obtained from the eigenvalues δ and the corresponding eigenvectors φ of the Laplace-Beltrami operator L [42],

$$\Phi(F_i) \mapsto [\delta_1\varphi_1(F_i), \ldots, \delta_n\varphi_n(F_i)]^T, \tag{2}$$

retaining the first n eigenvectors (corresponding to the first n eigenvalues). An approximation of the operator L is computed, following [43], from the matrix W. However, skeleton data can suffer from a relatively large amount of noise, and outliers, especially when parts of the body are occluded. In[11], we proposed a modification of the original diffusion maps by adding an extension similar to that proposed in [27] for Laplacian eigenmaps. We modified the entries of the adjacency matrix as

$$w_{ij} = (1-\beta)G(F_i, F_j) + \beta\mathcal{I}(F_i, F_j)$$

$$\text{with } \mathcal{I}(F_i, F_j) = \begin{cases} 1, & F_i \in \mathcal{K}_i \text{ or } F_j \in \mathcal{K}_j \\ 0, & otherwise \end{cases}, \tag{3}$$

where $\mathcal{K}_i$ is a set of neighbours of F_i, and $\mathcal{I}(\cdot)$ is an indicator function with the weighting factor β that was introduced in [27]. The indicator function avoids disconnected components in Laplacian eigenmaps, thus reducing the influence of outliers.

Fig. 3 illustrates the first dimension of the dimensionality reduced JP data for gait, clearly indicating that the original diffusion maps could not capture the intrinsic cyclic nature of the gait, while the robust diffusion maps method better captures the periodicity of the walking cycles.

Mapping testing data - The Nyström extension [44] extends the low dimensional representation computed from a training set to new samples, by evaluating the mapping of a new data F^t as

$$\Phi'_k(F^t) = \sum\nolimits_{F_i\in\mathbf{F}} L(F^t, F_i)\varphi_k(F_i) \tag{4}$$

with $\Phi'_k(F^t)$ the k^{th} component of $\Phi'(F^t)$, $k = 1 \ldots n$. The operator $L(F^t, F_i)$ is obtained in the same fashion as in [43], but based on our new definition of w_{ij} with the added indicator function $\mathcal{I}(\cdot)$. We use this mapping $O = \Phi'(F^t)$ as our high-level feature for building a motion model.

3.3. Human motion modelling

HMM-based methods can efficiently represent temporal dynamics of motion, and later in Section 5, we show how they naturally can be applied to motion quality assessment. The term 'continuous HMM' is often used to refer to models where the observation vector is continuous in $\mathbb{R}^n$ [45,46]. As the observation space is continuous in our case, all the models presented next are in fact 'continuous HMMs', but we use only 'HMM' for brevity.

Four variations of an HMM-based motion model are explained next in order of complexity and novelty of usage for human motion modelling. Their main characteristics are summarised in Table 1.

Notation. We use the following notation throughout the section. Suppose M is the number of possible states denoted $\mathcal{S} = \{S_1 \ldots S_M\}$, where the state at time t is $q_t \in \mathcal{S}$. The $M \times M$ transition matrix is $A = \{a_{ij}\}$, where $a_{ij} = P\big(q_t = S_j | q_{t-1} = S_i\big)$, and let $\pi = \{\pi_i\}$ be an initial state distribution, where $\pi_i = P(q_1 = S_i)$. The observation

Table 1
Characteristics of the four HMM models.

Model	State type	Modelling of time information	(Continuous) Observation model	Transition model
λ_a	Discrete	None	GMM	Transition matrix learnt using the Baum–Welch method
λ_b	Discrete	Explicit state duration		
λ_c	Discrete, manually defined	Explicit through the manual definition of the states	SVM classifier	
λ_d	Continuous	Implicit within the internal state	Parzen estimates of PDFs	Analytical

probability distribution is denoted by $B = \{b_j(O_t)\}$, where $b_j(O_t) = P(O_t|q_t = S_j)$, $j = 1 \ldots N$ is the probability of observing O_t when in state S_j. For continuous observations, the observation probability $b_j(O_t)$ is defined as a probability density function (PDF). Here, different continuous observation models are used for the four HMMs that we now introduce.

3.3.1. Classical HMM

We refer to an HMM with continuous observation densities and finite number of discrete hidden states as a *'classical'* HMM, in line with [45]. A classical HMM has three basic elements which can be written in a compact form as $\lambda_a = \{A, B, \pi\}$. In our implementation, Gaussian mixture models are used as the observation model:

$$b_j(O_t) = \sum_{i=1}^{I} c_{ji}\mathcal{N}(O_t; \mu_{ji}, \sigma_{ji}) \tag{5}$$

with I the number of components in the mixture, $\sum_{i=1}^{I} c_{ji} = 1$, and $c_{ji} \geq 0$. Such HMMs are trained by maximising the probability of the observation sequences given by the model, $\lambda_a{}^* = \arg\max_{\lambda_a} P(O|\lambda_a)$, and is solved by the *Baum-Welch* method. In testing, the likelihood of a new sequence, given the trained model, is calculated as,

$$P(O|\lambda_a) = \sum_{q_1,\ldots,q_T} \pi_{q_1} P(O_1|q_1) \prod_{t=2}^{T} P(O_t|q_t) P(q_t|q_{t-1}) \tag{6}$$

using the forward algorithm. The '*classical*' HMM is a parametric model, as the number of states M needs to be decided a priori, or optimised based on an evaluation set.

3.3.2. HMM with explicit state duration density

When modelling human motion, we note that the time elapsed at each body-pose configuration can be indicative of the quality of motion. For example, freezing during the walking cycle is highly indicative of deteriorating functional mobility, e.g. in Parkinson's and stroke patients. In classical HMMs, the state duration, i.e. the time elapsed between transiting to a state and transiting out of it, is not modelled and they would have difficulty discriminating the evolution of the body motion through time.

To overcome the problem, and keep the semantic meaning in the latent states while dealing with the lack of transition between them, explicitly modelling the state duration can help to address the problem [45]. A state duration model can be built as $D = \{P(d|S_1) \ldots P(d|S_M)\}$, where the state duration for each state S_j is modelled by the probability density $P(d|S_j.)$. We implement this probability with a Poisson distribution $P(d|S_j) = P(d; \theta_j) = \frac{e^{-\theta_j}\theta_j^d}{d!}$, where θ_j is the mean duration of state S_j. By this definition, the likelihood of a state duration observation d_{q_r} at time t depends only on the current state q_r and is independent of the duration of the previous state.

The probabilities in the trained HMM model are thus expanded to $\lambda_b = \{A, B, \pi, D\}$, with B implemented as in (5). Again λ_b is a parametric model with a discrete number of states M as its parameter. The likelihood of the observed sequence $O = \{O_1 \ldots O_T\}$ given the trained model is calculated as,

$$\begin{aligned} P(O|\lambda_b) = \sum_{r=1}^{R} \sum_{q_1,\ldots,q_r} \sum_{d_1,\ldots,d_r} \pi_{q_1} P(d_1|q_1) P(O_1, \ldots, O_{d_1}|q_1) \\ \prod_{i=2}^{r} P(q_i|q_{i-1}) P(d_i|q_i) P(O_{\sum_{k=1}^{i-1} d_k+1}, \ldots, O_{d_i}|q_i) \end{aligned} \tag{7}$$

where $P(O_i, \ldots, O_j|q) = \prod_{k=i}^{j} P(O_k|q)$, and r is the number of different states reached during the sequence, restricted to a minimum $R = \lceil \frac{T}{D} \rceil$ in case of a maximum state duration D. As with classical HMMs, the likelihood of a sequence can be obtained using the forward algorithm.

3.3.3. HMM with a discriminative classifier

Classical HMM has been employed efficiently when the motion can be broken into distinct sub-motions [46,47]. However, some motions can not be automatically divided into such sub-motions by the training of the conventional Gaussian mixture-based observation model, and require uniformly splitting the motion cycles in training sequences into M manually defined states. For a smooth motion (e.g. walking), such splitting of the motion cycle may lead to poor discrimination between the states when training the observation model. To avoid this, the traditional Gaussian mixture-based observation model could be replaced by a discriminative classifier which is trained to discriminate the poses of one state from another.

Given a set of extracted features from the training data, the objective is to build a suitable classifier which better discriminates the data. In this work, SVMs as large margin classifiers are used, although other classifiers could also be employed. Combining SVMs with HMMs has been previously applied, e.g. in speech recognition [48] and facial action modelling [49], where the posterior class probability is approximated by a sigmoid function [50]. We employ this hybrid classification method for our observation model, following [49] where the multi-class SVM is implemented using one-versus-one approach. In total, $M(M-1)/2$ SVMs are trained for the pairwise classification representing all possible pairs out of M classes. For each SVM, pairwise class probability $\alpha_{ij} = P(S_i|S_i \text{ or } S_j, O_t)$ is calculated using Platt's method [51]. Such pairwise probabilities are transformed into posterior probabilities as,

$$P(q_t = S_j|O_t) = 1/\left[\sum_{j=1, j\neq i}^{M} \frac{1}{\alpha_{ij}} - (M-2)\right]. \tag{8}$$

The continuous observation probabilities $b_j(O_t)$ are formed by the posterior probabilities using Bayes' rule,

$$b_j(O_t) \propto P(q_t = S_j|O_t)/P(q_t = S_j). \tag{9}$$

Similar to classical HMMs, the discriminative approach is parametric and relies on the number of states M. The model $\lambda_c = \{A, B, \pi\}$ does not differ from λ_a in training or testing, but the observation model is based on the discriminative classifier.

3.3.4. Continuous-state HMM

In the previous model λ_c, the hidden state represents the proportion of motion completion at the current frame, which is by nature continuous. Thus in [11], we proposed a statistical model that described continuous motion completion, as an approach that is highly suited to motion quality assessment. This model is in effect a continuous-state HMM, and we represent it here from that perspective. Continuous-state HMMs have been widely used in signal processing in general, for example in [52] where a continuous-state HMM model of deforming shapes was implemented for monitoring crowd movements.

We introduced in [11] the continuous variable X with value $x_t \in [0, 1]$ to describe the progression of motion, i.e. the proportion of motion completed at frame t which linearly increases from 0 at the start of the motion to 1 at its end. For periodic motions, x_t is analogous to the motion's phase, and increases within one cycle of the motion, and then resets to 0 for the next cycle. The hidden state of our continuous-state HMM is then $q_t = x_t$.

The crucial advantage of using this continuous state variable is that the motion does not have to be discretized into a number of segments, the model is non-parametric, and the problem of choosing an optimal M becomes irrelevant. However, the infinite number of possible states makes the commonly used approaches for training an HMM and evaluating an observation sequence impractical since these algorithms are based on integrating over a finite number of possible states. Thus, novel algorithms were introduced in [52,53], e.g. based on particle filtering. Our model differs from these HMMs, both in the definition of the observation model and state transition probabilities, and in the algorithms used to perform the training and evaluation.

In our continuous-state HMM, the observation model is the PDF $b_{x_t}(O_t) = f_{O_t}(O_t|q_t = x_t)$. We learn this probability from training data as

$$f_{O_t}(O_t|q_t = x_t) = \frac{f_{O_t,x_t}(O_t, x_t)}{f_{x_t}(x_t)}\,, \tag{10}$$

using a Parzen window estimator. The kernel bandwidth of the estimator is a parameter of this method that we set empirically so as to avoid over-smoothing of the PDFs. Learning the observation model requires knowing or estimating x_t for the training data. For simplicity, we assume that our training data represents motions with uniform dynamics (i.e. uniform speed within motion or motion cycle), and we compute x_t proportional to time. An example observation model PDF is shown in Fig. 4 for the motion of ascending stairs.

We define the transition model A analytically as the PDF

$$f_{x_t}(x_t|x_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\Delta x_t - v\Delta\tau_t}{\sigma}\right)^2}\,, \tag{11}$$

where $\Delta x_t = x_t - x_{t-1}$, τ_t is the time at frame t, and $\Delta\tau_t = \tau_t - \tau_{t-1}$. This transition model thus assumes proportionality between the proportion of motion completion x and time τ. v is the speed of the motion and is estimated as

$$v = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta x_i}{\Delta\tau_i}\,, \tag{12}$$

so that the model adapts to different motion speeds. During training, v is computed for the complete motion or motion cycle. When evaluating a test sequence, v is computed within a sliding window in order to handle sequences with non-constant speeds, although its values are kept within empirically determined limits for a normal movement. The size of the window will be discussed later in this section. The standard deviation σ in (11) modulates the constraint that Δx is proportional to $\Delta\tau$. Its choice has been determined empirically so as to enforce a strong constraint when evaluating the probability of a sequence ($\sigma_{eval} = 10^{-3}$), and a weaker constraint ($\sigma_{est} = 7e^{-3}$) when estimating x_t. This relaxation of the proportionality constraint when estimating x_t aims at increasing flexibility of the model to describe motion dynamics that deviate from normal due to significant speed variations. Note that such abnormal motions would still be penalised by significantly lower probabilities $P(O|\lambda_d)$ due to the lower σ_{eval}.

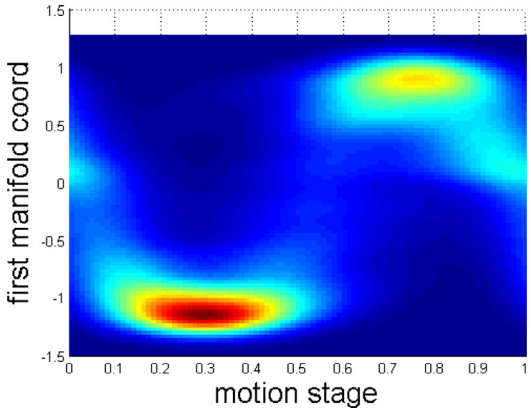

Fig. 4. Example of PDF that defines the observation model in model λ_d. The plot shows the marginal of the PDF for the first manifold dimension.

To summarise, the continuous-state HMM, first proposed in a different formulation as a statistical model in [11], is defined by $\lambda_d = \{A, B, \pi\}$ where A is defined analytically and B is estimated from training data. The initial state distribution π is uniform to enable evaluation from any point in the motion.

Similarly to finite state HMMs, the likelihood of a sequence of observations $O = \{O_1 \dots O_T\}$ under model λ_d is an integration over all possible values for the hidden states

$$\begin{aligned} P(O|\lambda_d) &= \int_{\{x_1,\dots,x_T\}} f_{O,x_1,\dots,x_T}(O, x_1, \dots, x_T) \\ &= \int_{\{x_1,\dots,x_T\}} f_{x_1}(x_1) f_{O_1}(O_1|x_1) \prod_{i=2}^{T} f_{O_i}(O_i|x_i) f_{x_i}(x_i|x_{i-1}). \end{aligned} \tag{13}$$

The derivation of (13), that exploits Markovian properties, can be found in [11].

Such an integral over an infinite number of possibilities is impractical to compute. The approximation we present next allows reducing (13) to a more easily solvable form. From our definition of the transition model in (11), given a value x_{t-1} of variable X at frame $t-1$, its value x_t at frame t follows a normal distribution around $x_{t-1} + v\Delta_{\tau_t}$ with standard deviation σ. In the ideal case of a perfectly normal motion, σ should tend to 0 and the normal distribution would tend to a Dirac distribution. For σ small enough, that is to say for a strong enough constraint on the evolution of X during the motion, we can use the approximation $\sigma \approx 0$, which leads to

$$P(O|\lambda_d) \approx f_{x_1}(\hat{x}_1) f_{O_1}(O_1|\hat{x}_1) \prod_{i=2}^{T} f_{O_i}(O_i|\hat{x}_i) f_{x_i}(\hat{x}_i|\hat{x}_{i-1})\,. \tag{14}$$

The notation $\hat{x}_i$ highlights that this value is the most likely for X at frame i given x_{i-1} and Δ_{τ_i}, i.e. $\hat{x}_i = x_{i-1} + v\Delta_{\tau_i}$.

When computing $P(O|\lambda_d)$ using this approximation, the values $\hat{x}_i$ need to be estimated. This can be done by maximising their likelihood

conditional on the sequence of observations:

$$\begin{aligned}\{\hat{x}_1,\ldots,\hat{x}_T\} &= \arg\max_{x_1,\ldots,x_T} f_{x_1,\ldots,x_T}(x_1,\ldots,x_T|O)\\ &= \arg\max_{x_1,\ldots,x_T} \frac{f_{O,x_1,\ldots,x_T}(O,x_1,\ldots,x_T)}{f_O(O_1,\ldots,O_t)}\\ &= \arg\max_{x_1,\ldots,x_T} f_{x_1}(x_1) f_{O_1}(O_1|x_1)\prod_{i=2}^{T} f_{O_i}\\ &\times (O_i|x_i) f_{x_i}(x_i|x_{i-1})\ . \end{aligned} \tag{15}$$

In our implementation, this estimation is performed using unconstrained nonlinear optimisation. Similar to the estimation of v, and for the sake of efficiency, we estimate $\{\hat{x}_1,\ldots,\hat{x}_T\}$ within a window of dynamic width ω_t, to encompass the frames for which $\hat{x}_t$ has not yet converged. This strategy is based on the empirical observation that the estimated value $\hat{x}_i$ at a previous frame i does not change significantly after a few iterations. In practice, we consider $\hat{x}_i$ to have converged when its change is less than 10^{-3} for 2 consecutive iterations.

4. Comparison of HMM models

The four HMMs introduced above attempt to describe motion by capturing the dynamics of body poses. A key aspect of the models is the relation of their hidden state q_t with these body poses and with time. In models λ_a and λ_b, a direct association between q_t and body pose ensues from the training of the Gaussian mixture-based observation model that groups similar body poses into distinct states. For models λ_c and λ_d, the internal state is associated with sub-motions, i.e. distinct phases of the motion, and these sub-motions tend to have characteristic body poses. Note that in this last case, the states might not have distinctive body poses. For example, in walking, the body goes through similar poses at various points in time within one cycle. An examination of the relation between the hidden states, and both body poses and motion phases or time, provides an insight into the respective effectiveness of the models at describing motions and their dynamics. We now perform this analysis for the case of gait motion on stairs.

Fig. 5 plots the various states corresponding to the training data in different colours, in a graph that represents both time/motion phase (horizontal axis) and the first dimension of the high-level feature O, i.e. body pose (vertical axis). In model λ_a, the states are predominantly separated in the domain of body poses, and many of them span the same temporal regions. This lack of separation of the states in the temporal domain limits their ability to discriminate the stages of the motion. As another consequence, transition between different states is not necessary for motion evolution. This may lead to poor modelling of the dynamics of the motion, as will be shown in Section 6 where freezes of gait often cannot be detected by model λ_a. Note in Fig. 5 that increasing the number of states M does not significantly improve the description of dynamics as the additional separation is predominantly in the domain of body pose O than in the motion phase/time domain.

The explicit modelling of state duration in model λ_b addresses the problem of state stagnation in model λ_a. Although the possible states are still badly separated in the temporal domain, as seen in Fig. 5, the explicit modelling of state duration enables model λ_b to better describe the dynamics of motions, and in particular to detect freezes of gaits.

Another way of addressing the issues of model λ_a is to define the hidden states as corresponding to distinct temporal regions, by manually dividing a motion uniformly into equal-length segments. This is the strategy used in model λ_c. Note that, depending on the type of motion, several of the resulting states may correspond to similar body poses. This is for example the case of gait, as discussed earlier and illustrated in Fig. 5 where several distinct states are located in the same region of the embedded space. Consequently, as mentioned in Section 3.3c, the observation model produced by the classical HMM training algorithm may be poorly discriminative, and requires to be replaced by a more robust classifier. It should be stressed that the number of possible states significantly impacts the ability of such a model to represent the temporal dynamics of the motion. Indeed, in a model with too few possible states, the probability of staying in a well populated state may be higher than transiting to the next one, resulting in the same state stagnation problem than in model λ_a. On the other hand, when the number of possible states is too high, the body poses of distinct states may become too similar and overcome the discrimination power of the classifier, leading to a reduction in performance. This is illustrated in Fig. 6(c), where the best ROC curves are obtained for 15–30 states, while deteriorating quickly when the state is less than 10 or higher than 40. Further, discriminative classifiers, such as SVMs, cannot naturally handle unknown observations, and would therefore not clearly attribute a state to an unusual observed body pose.

We note in model λ_c that an increase in the number of states (while remaining within the "discriminative zone" of the classifier) leads to a better representation of the dynamics of the motion. Model λ_d extends this idea by having a continuous state, thus imposing an infinite number of possible states. Its observation model does not rely on a discriminative classifier, but instead it exploits non-parametric estimations of conditional PDFs, as explained in Section 3.3d. When two or more significantly different states are equally probable given an observation, as for example in the gait model of Fig. 4, model λ_d relies on the relative rigidity of its transition model to handle these ambiguities.

5. Quality assessment measures

Using any one of our four models trained on normal motion sequences, one can detect anomalies in new observations and assess the quality of the motion based on the likelihood of the new observation to be described by the model. An online assessment of the motion, computed on a frame-by-frame basis, would be desirable for triggering timely alerts when the observed motion drops below a threshold in its level of normality. A straightforward way of obtaining an online measure would be to compute the likelihood $P(O|\lambda_i)$ within a sliding window. However, this strategy may prove to be difficult to apply, as the choice of window size requires a delicate compromise between a sufficient number of frames, in order to capture and analyse the dynamics of the movement, and a small enough window so as to preserve the instantaneous properties of an online measure. Moreover, this window size would have to be adjusted for each type of motion, and also for instances of a motion performed at significantly different speeds.

To overcome these problems, we propose a dynamic measure

$$\mathcal{M}_t = \log P(O_t|O_1,\ldots,O_{t-1},\lambda_i)\ , \tag{16}$$

that is the log-likelihood of the current frame given the previous frames and the model. For models λ_a, λ_b, and λ_c, $P(O_t|O_1,\ldots,O_{t-1},\lambda_i)$ may be simply computed as $\frac{P(O|\lambda_i)}{P(O_1,\ldots,O_{t-1}|\lambda_i)}$ using two calls to the forward algorithm. In the case of model λ_d, this measure can only be obtained after the convergence of x_t, and $P(O_t|O_1,\ldots,O_{t-1},\lambda_i)$ may be calculated using the approximation of (14) as $f_{O_t}(O_t|\hat{x}_t) f_{x_t}(\hat{x}_t|\hat{x}_{t-1})$.

In [11], we proposed a similar online measure, that instead of waiting for the convergence of x_t, integrated $P(O_t|O_1,\ldots,O_{t-1},\lambda_i)$ over the dynamic sliding window of size ω_t which was defined for model λ_d in Section 3.3d., in order to account for the updated values

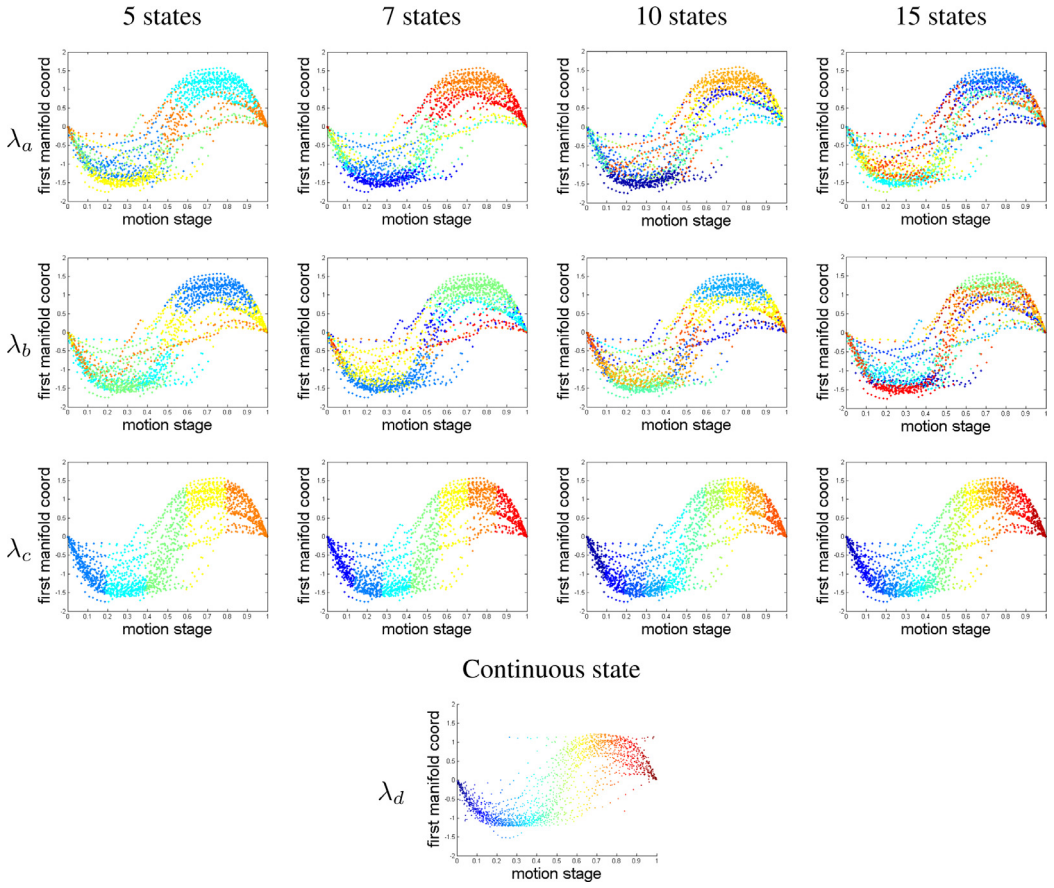

Fig. 5. States defined in models λ_a (top row), λ_b (2nd row), λ_c (3rd row), and λ_d (bottom row). For the discrete models (λ_a-λ_c), colours denote different states, while for the continuous model (λ_d) continuous colour gradient is used based on the value of the internal state.

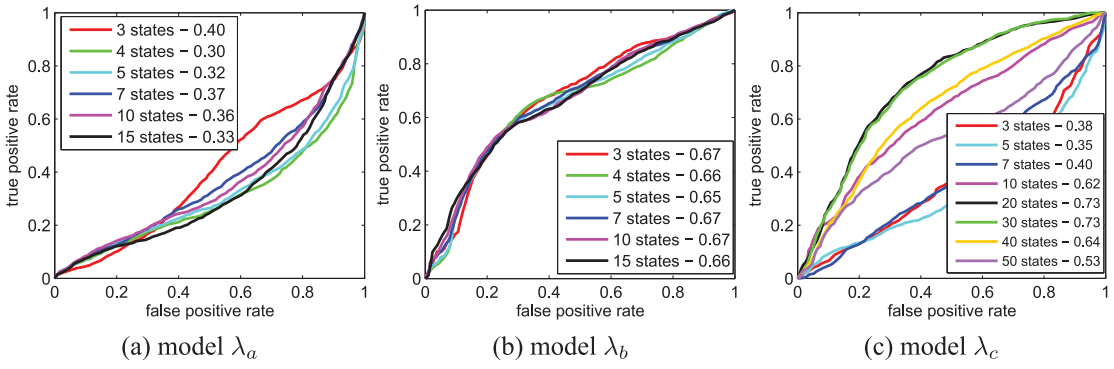

Fig. 6. Frame classification accuracy for gait on stairs: ROC curves using our online measure $\mathcal{M}_{\omega_t}$ for different number of states for feature type JP.

of x_t that are re-estimated within the window:

$$\begin{aligned}\mathcal{M}_{\omega_t} &= \sum_{j=t_{min}}^{t} \log P\big(O_j|O_1,\ldots,O_{j-1},\lambda_i\big)\\ &= \log \prod_{j=t_{min}}^{t} P\big(O_j|O_1,\ldots,O_{j-1},\lambda_i\big)\\ &= \log \prod_{j=t_{min}}^{t} \frac{P(O_1,\ldots,O_j,\lambda_d)}{P(O_1,\ldots,O_{j-1},\lambda_i)}\\ &= \log \frac{P(O_1,\ldots,O_t,\lambda_d)}{P(O_1,\ldots,O_{t_{min}-1},\lambda_i)}\\ &= \log P\big(O_{t_{min}},\ldots,O_t|O_1,\ldots,O_{t_{min}-1},\lambda_i\big)\,,\end{aligned} \tag{17}$$

with t_{min} the first frame of the sliding window. Thus $\mathcal{M}_{\omega_t}$ can be seen as the log-likelihood of the sliding window given the previous observations. This conditionality in the probability alleviates the effect of the window size that we discussed earlier. In our experiments, for convenience and efficiency, we limit ω_t to a maximum of 15 frames, although it rarely goes above 10 frames. For models λ_a, λ_b, and λ_c, the forward algorithm does not require the estimation of x_t as it sums probabilities over all possible states, so the value of ω_t cannot be determined automatically. Instead, we set it to a constant value ω, and we explore the influence of its choice on the results in Section 6, where we shall also compare our two online measures $\mathcal{M}_t$ and $\mathcal{M}_{\omega_t}$.

In addition to these two measures of dynamics quality, we also proposed in [11] a measure of pose quality, computed independently for each frame as:

$$\mathcal{M}_{\text{pose}} = \log f_{O_i}(O_i)\,. \tag{18}$$

6. Experimental evaluation

To demonstrate the performance of the motion quality analysis framework, we analysed the motions of walking on a flat surface, gait on stairs, and transitions between sitting and standing, which are particularly critical for rehabilitation monitoring in patients with musculoskeletal disorders, disease progression in PD patients, and for many others. For the analysis of such motions, we compared different low-level features, dimensions of the manifold embedding, and motion models, as proposed in Sections 3.1–3.3 respectively. We also investigated whether full-body information is consistently needed for all tested movement types. We tested gait on stairs on the dataset SPHERE-staircase2014 (first introduced in [11]) as well as two new datasets SPHERE-Walking2015 and SPHERE-SitStand2015 for the assessment of gait on a flat surface and of sitting and standing movements respectively[3] The datasets were used to perform abnormality detection by applying the online measures $\mathcal{M}_t$ (Eq. 16) and $\mathcal{M}_{w_t}$ (Eq. 17), both on a frame-by-frame basis and for the whole sequence.

6.1. Datasets

SPHERE-Staircase2014 dataset [11] – This dataset includes 48 sequences of 12 individuals walking up stairs, captured by an Asus Xmotion RGB-D camera placed at the top of the stairs in a frontal and downward-looking position. It contains three types of abnormal gaits with lower-extremity musculoskeletal conditions, including freezing of gait (FOG) and using a leading leg, left or right, in going up the stairs (i.e. LL or RL respectively). All frames have been manually labelled as normal or abnormal by a qualified physiotherapist. We used 17 sequences of normal walking from 6 individuals for building the model and 31 sequences from the remaining 6 subjects with both normal and abnormal walking for testing.

SPHERE-Walking2015 dataset – This dataset includes 40 sequences of 10 individuals walking on a flat surface. This dataset was captured by an Asus Xmotsion RGB-D camera placed in front of the subject. It contains normal gaits and two types of abnormal gait, simulating , under the guidance of a physiotherapist, stroke and Parkinson disease patients' walking. We used 18 sequences of normal walking from 6 individuals for building the model, and 22 sequences from 4 other subjects with both normal and abnormal gaits for testing. The testing set includes 5 normal, 8 Parkinson, and 9 Stroke sequences.

SPHERE-SitStand2015 dataset – This dataset includes 109 sequences of 10 individuals sitting down and standing up in a home environment. Since the Asus Xmotion RGB-D camera is unable to track the skeleton for movements that cause self-occlusions, the data was captured using a Kinect 2 camera instead. It contains normal and two types of abnormal motions, including (a) restricted knee and restricted hip flections and (b) freezing. We used 9 sequences of normal movement from 8 individuals for building each sitting and standing model, and 91 sequences from two other subjects with normal and abnormal movements for testing, including 31 normal and 12 abnormal sitting, and 36 normal and 12 abnormal standing. The abnormal sequences comprise 4 samples of each abnormality type.

In the following experiments, we first compare the methods on the SPHERE-Staircase2014 dataset. Then, we show that the methods can be extended to other types of human motion, both periodic and non-periodic, using the SPHERE-Walking2015 and SPHERE-SitStand2015 datasets.

6.2. Parameter setting

Number of states – Three of the motion models (λ_a, λ_b and λ_c) are parametric, expecting the number of states M to be identified in advance. It is commonly known that classical HMM models are sensitive to the number of states. To select the appropriate number, we plotted our results as ROC curves of frame classification accuracy using our online measure on all test sequences for different numbers of states. Fig. 6 shows the ROC curves together with their area under the curve (AUC) values when using feature JP. Both λ_a and λ_b models seem insensitive to the number of states, especially when $M \geq 5$. The performance of motion model λ_c is highly sensitive to the number of states with significantly improved performance for $10 < M < 40$. As discussed in Section 4, this is as expected, since walking cycles are uniformly divided into several states, and fewer states may lead to high probabilities of self-transitions which would then fail to explain the temporal evolution of the motion. On the other hand, having a relatively larger value of M may cause difficulty in discriminating data, thus leading to poor recognition results.

Table 2
Optimal number of states for each low-level feature for each discrete-state HMM (models λ_a, λ_b, λ_c) and motion type. For the continuous-state HMM (model λ_d), the number of states is undefined and hence the parameter is not applicable (N/A). For gait on flat surface, sitting, and standing motions, only models λ_c and λ_d were evaluated.

Motion	Motion model	JP	JV	PJD	PJA
Gait on stairs	λ_a	3	4	3	3
	λ_b	3	3	4	4
	λ_c	20	20	20	15
Walking on a flat surface	λ_c	15	5	5	7
Sitting	λ_c	10	7	7	5
Standing	λ_c	15	15	5	7
All	λ_d	N/A	N/A	N/A	N/A

To choose the optimal number of states, the model with the maximal value of AUC was selected. We followed the same process to obtain the optimal number of states for low-level features JV, PJD, and PJA for each of the discrete-state HMMs, as summarised in Table 2 Model λ_d does not require optimizing the number of states, since its hidden variable is continuous.

Temporal window size – ω_t is also a parameter for models λ_a, λ_b, and λ_c (see Section 5). We investigated the effects of different temporal window sizes on the detection accuracy when computing $\mathcal{M}_{\omega_t}$. We chose the optimal settings (feature type and the number of states) that provided the best results for each of the models and tested with different temporal window sizes set to 1, 5, 10, 15, 20 and 25 frames. This test was not performed for model λ_d as ω_t is set dynamically for that model (see Section 3.3d. for details).

As shown in Table 3, the best results for model λ_a, λ_b and λ_c for gait on stairs were obtained with a temporal window size of 15 frames, although smaller number of frames, such as 5 or 10, are not far in performance. Selecting too small a size of window may allow the noise to prevent capturing the abnormality of a frame, while too large a window may include both abnormal and normal frames within the window and would thus fail to detect the abnormality.

Choice of online measure – As discussed earlier, models λ_a, λ_b, and λ_c obtained the best results when computing measure $\mathcal{M}_{\omega_t}$ with a temporal window size of 15 frames. The measure $\mathcal{M}_t$ is equivalent to $\mathcal{M}_{\omega_t}$ at a window size of 1 frame (as in the 1st column of Table 3). The often worse results achieved with $\mathcal{M}_t$ for models λ_a, λ_b, and λ_c were caused by errors obtained from unsmoothed likelihoods between frames, while with $\mathcal{M}_{\omega_t}$, the likelihoods were smoothed by a temporal window.

Table 4 reports AUC values in the case of gait on stairs, and shows that model λ_d did not suffer from unsmoothed likelihoods and obtained its best results with $\mathcal{M}_t$, due to the time averaging delaying the detections of $\mathcal{M}_{\omega_t}$. For other motion types such as sitting and standing, where the scores are averaged over the full sequences (see Section 6.5), this timely detection of abnormal events is less important and both measures perform comparatively.

[3] To be released to the public domain soon.

Table 3
AUC values at different temporal window sizes for different models of the gait on stairs motion, in each case using the optimal feature and the optimal number of states.

Motion model	Temporal window size					
	1 frame	5 frames	10 frames	15 frames	20 frames	25 frames
λ_a	0.60	0.62	0.64	**0.66**	0.65	0.64
λ_b	0.70	0.71	0.72	**0.73**	0.68	0.66
λ_c	0.68	0.72	0.74	**0.75**	0.73	0.71

Table 4
AUC results for gait on stairs movement for different skeleton representations (low-level features and manifold dimensions) for each of the four models, using $\mathcal{M}_{\omega_t}$ with optimal ω_t for the discrete models and both online measures ($\mathcal{M}_t$ / $\mathcal{M}_{\omega_t}$) for the continuous model.

Motion model	Feature	Manifold dimension n				
		1	2	3	4	5
	JP	0.37	0.35	0.39	0.35	0.40
$\boldsymbol{\lambda_a}$	JV	**0.64**	**0.60**	**0.66**	**0.66**	**0.63**
	PJD	0.38	0.36	0.43	0.46	0.46
	PJA	0.43	0.52	0.53	0.58	0.56
	JP	0.64	0.62	0.65	0.67	0.68
$\boldsymbol{\lambda_b}$	JV	0.60	0.58	0.62	0.65	0.64
	PJD	**0.66**	**0.70**	**0.73**	**0.70**	**0.70**
	PJA	0.56	0.61	0.61	0.61	0.64
	JP	0.71	0.71	0.73	0.73	0.73
$\boldsymbol{\lambda_c}$	JV	0.62	0.64	0.64	0.62	0.62
	PJD	**0.72**	**0.75**	**0.75**	**0.75**	**0.75**
	PJA	0.57	0.58	0.64	0.63	0.61
	JP	**0.81 / 0.75**	0.81 / 0.74	**0.82 / 0.75**	**0.81 / 0.74**	**0.81 / 0.74**
$\boldsymbol{\lambda_d}$	JV	0.65 / 0.65	0.60 / 0.60	0.60 / 0.61	0.59 / 0.59	0.59 / 0.58
	PJD	0.74 / 0.70	**0.83** / **0.80**	0.74 / 0.65	0.62 / 0.54	0.61 / 0.50
	PJA	0.57 / 0.57	0.63 / 0.61	0.67 / 0.63	0.69 / 0.64	0.66 / 0.65

6.3. Gait on stairs

6.3.1. Comparison of different skeleton representations

To select the most effective representation for the skeleton data, we applied the four low-level features introduced in Section 3.1 while we varied the dimensionality n of the manifold between 1 and 5. The frame classification ROC curves and their AUC in Fig. 7 show the performance accuracy for each different skeleton representation and model. All the curves are plotted using their optimal number of states, as stated in Table 2, and their optimal window size ω_t. Table 4 reports the AUC values obtained by each composition of low-level features, dimensionality values for n, and motion models. Values obtained using both online measures $\mathcal{M}_t$ and $\mathcal{M}_{\omega_t}$ are provided for model λ_d. Only measure $\mathcal{M}_{\omega_t}$ with the optimal window size ω_t was used for models λ_a, λ_b, and λ_c, since for these three models, $\mathcal{M}_t$ is equivalent to $\mathcal{M}_{\omega_t}$ with a non-optimal window size of 1 frame.

As observed in Fig. 7 and Table 4, for model λ_a, the JV feature provided significantly better results than the JP, PJA and PJD features. This can be explained by the fact that the joint velocities are calculated based on two consecutive frames, and hence the feature can capture a significant extent of the dynamics of the motion, counterbalancing the difficulty of model λ_a's ability in describing the motion's dynamics; the other types of features only consider the current frame. For model λ_b, there was no remarkably significant variation in the results for the different features, however, the PJD feature performed best across all dimensions. For model λ_c, again PJD provided the best outcome in all dimensions. Further, Table 4 shows that for all the best results of the three discrete models λ_a, λ_b, and λ_c, the accuracy does not depend strongly on the dimensionality of data. In summary, we chose the JV feature for model λ_a and PJD feature for models λ_b and λ_c with the first 3 manifold dimensions as the optimum skeleton representation for these three models, as highlighted in Table 4.

For model λ_d, although the best result in Table 4 was for the PJD feature in 2D, the JP feature performed best in the majority of the cases and still obtained results very close to feature PJD's best outcome, even when based on only the 1st manifold dimension. The superiority of the JP feature over PJA and PJD in 1D may be understood by considering the PDFs of their observation models (depicting the path of normality of motion), plotted in the first column of Fig. 8. The normality path for JP is more constrained, i.e. narrower, than for PJA and PJD, thus the accepted variance around the normality path is smaller, making the model more discriminative. In the case of JV, although the normality path is as narrow as for JP (in the 1st dimension), the results were the least performing of the four features when considered across all the dimensions. We attribute this to the incompatibility of using absolute speeds in the low-level feature at the same time as relative speeds in the HMM modelling where variable v attempts to normalise the motions speeds.

When using more than one dimension, the accuracy remained high for the JP feature, due to the PDF in these dimensions also having small variances around the normality path. For PJA, the use of more dimensions (up to 4) improved the results by combining their respective discriminative powers, but adding the 5th dimension failed to contribute further gains. However, when a dimension had particularly low discriminative power, its impact on the results of the model was negative. For example, the third dimension of the PJD feature (bottom right plot of Fig. 8), and the second dimension of the JV feature (middle plot of the second row), did not exhibit a clear preferred normality path.

Although the best AUC was obtained by the PJD feature in a 2D manifold, it was only marginally higher than for JP in a 3D manifold, and the ROC curve for JP indicates consistently better performance than that of PJD's (see Fig. 7(d)). Hence, to conclude, we chose the JP feature for model λ_d with 3 manifold dimensions as the optimum skeleton representation (keeping consistency on all four models).

The average processing time (in milliseconds per frame) for building high-level features are 1.18, 1.14, 10.06 and 29.32 for JP, JV, PJD and PJA features, respectively. The experiments were performed using Matlab on a workstation with an Intel I7-3770S CPU 3.1GHz processor and 8GB RAM. The number of dimensions of the manifold does

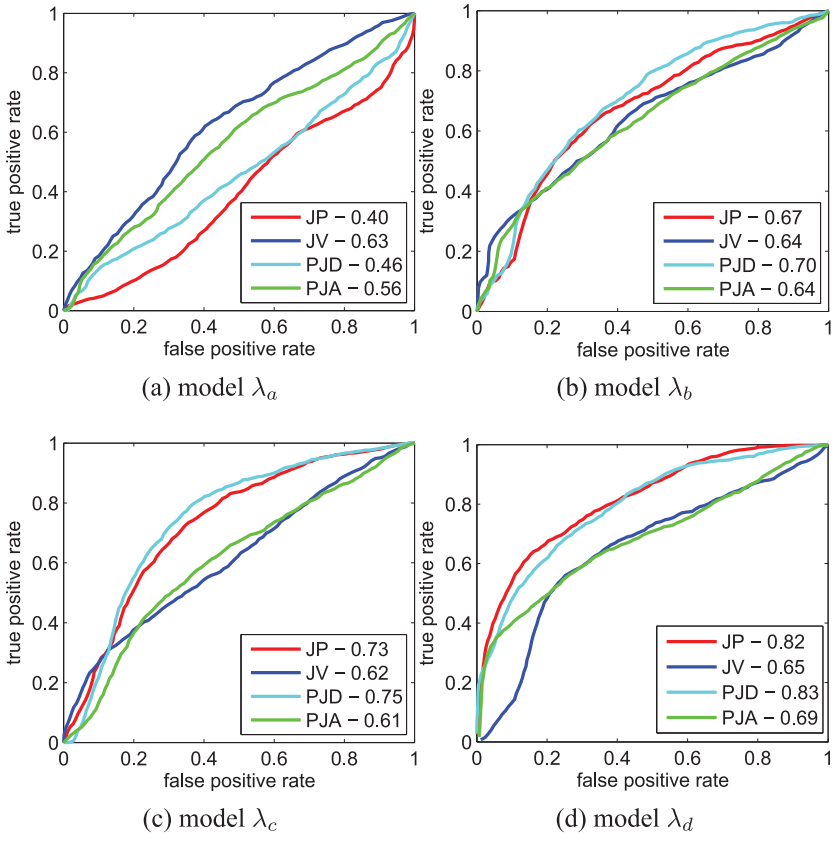

Fig. 7. Comparison of different skeleton representations (low-level features with their respective optimal manifold dimensionality) for models (a) λ_a, (b) λ_b, (c) λ_c, and (d) λ_d, at abnormal frame detection for the gait on stairs movement. The plots are for the optimal state numbers (see Table 2) and online measure for each model: $\mathcal{M}_{\omega_t}$ with $\omega_t = 15$ for models λ_a, λ_b, and λ_c, and $\mathcal{M}_t$ for model λ_d.

not affect the processing time, since its selection is performed after generating the manifold space.

6.3.2. Comparison of the motion models

We evaluated and compared the ability of each model to detect various abnormalities in the sequences under optimal parameter settings. Abnormal frames were detected when the measure of normality, $\mathcal{M}_{\omega_t}$ or $\mathcal{M}_t$, dropped below a threshold. Returning to Fig. 7, it shows the true positive rate against false positive rate at different threshold values. It is clear that model λ_d performed better than the other models at detecting abnormal frames.

Significantly, when an expert, e.g. a physiotherapist, observes a patient, he/she anticipates a disruption in the normal cycle of gait. This would be before it could reasonably be identified by an automated system. This is an artefact of using frame by frame labelling, especially for RL and LL events. When the expert notes a minimal reduction in the speed of the swinging leg, he/she anticipates that the heel strike will not take a place at 'normal' position. Hence, the expert classifies all of the frames leading up to that point as abnormal. However, in terms of the pose trajectory along the manifold, the motion is normal, other than a very subtle reduction in speed. Our approach is robust to subtle changes in gait velocity as this is present in normal gait as well. Thus, we provide an alternative measure by detecting the abnormality based on the whole event. This motion analysis is still online, since abnormal events are detected as new frames are being acquired, without having to wait for the full sequence to be available. We first eliminated noise in the frame classification by removing isolated clusters of less than 3 normal or abnormal frames. Then, we defined an abnormal event as succession of (at least) 3 consecutive abnormal frames.

We counted as true positive (TP) detections any event that had at least three frames detected as abnormal, while false negatives (FN) were events with less than three detected frames. False positive (FP) detections were either detected events that did not intersect by at least three frames with a true abnormal event, or normal periods between abnormal events that had all their frames classified as abnormal. The abnormatity event classification results are illustrated in Fig. 9 and Table 5. Fig. 9 presents precision and recall values when varying the threshold on the frame classification measure $\mathcal{M}_{\omega_t}$ or $\mathcal{M}_t$, all other parameters being set optimally for each motion model. Note that this is not the usual Precision against Recall (PR) plot for event detection, since the threshold we are varying here is not on the measure of likelihood of abnormal event, but on a measure of likelihood of abnormal frame, hence, the unusual aspect of the plot. Defining a measure of the likelihood of an abnormal event is not in the scope of this study, but will be the focus of our future work.

For each model, the point closest to the top-right corner of the plot (indicated with a square) was chosen as the best precision-recall compromise, and its corresponding measure threshold was used to obtain the results reported in Table 5. As observed in the table, although models λ_a and λ_b are able to detect all the abnormal events, the very high number of wrongly detected events (FPs) makes the models impractical. Model λ_c shows it is able to detect most abnormal events with only two missed detections, while model λ_d gives the fewest errors (FP+FN).

The average processing time (in milliseconds per frame) of each motion model are 15.99, 16.27, 30.16 and 153 for λ_a, λ_b, λ_c and λ_d, respectively. These numbers are computed when using the optimal manifold dimensions, optimal low-level feature and the corresponding optimised number of states for each model. Note that models λ_a, λ_b, and λ_c have been implemented using an optimized toolbox, while

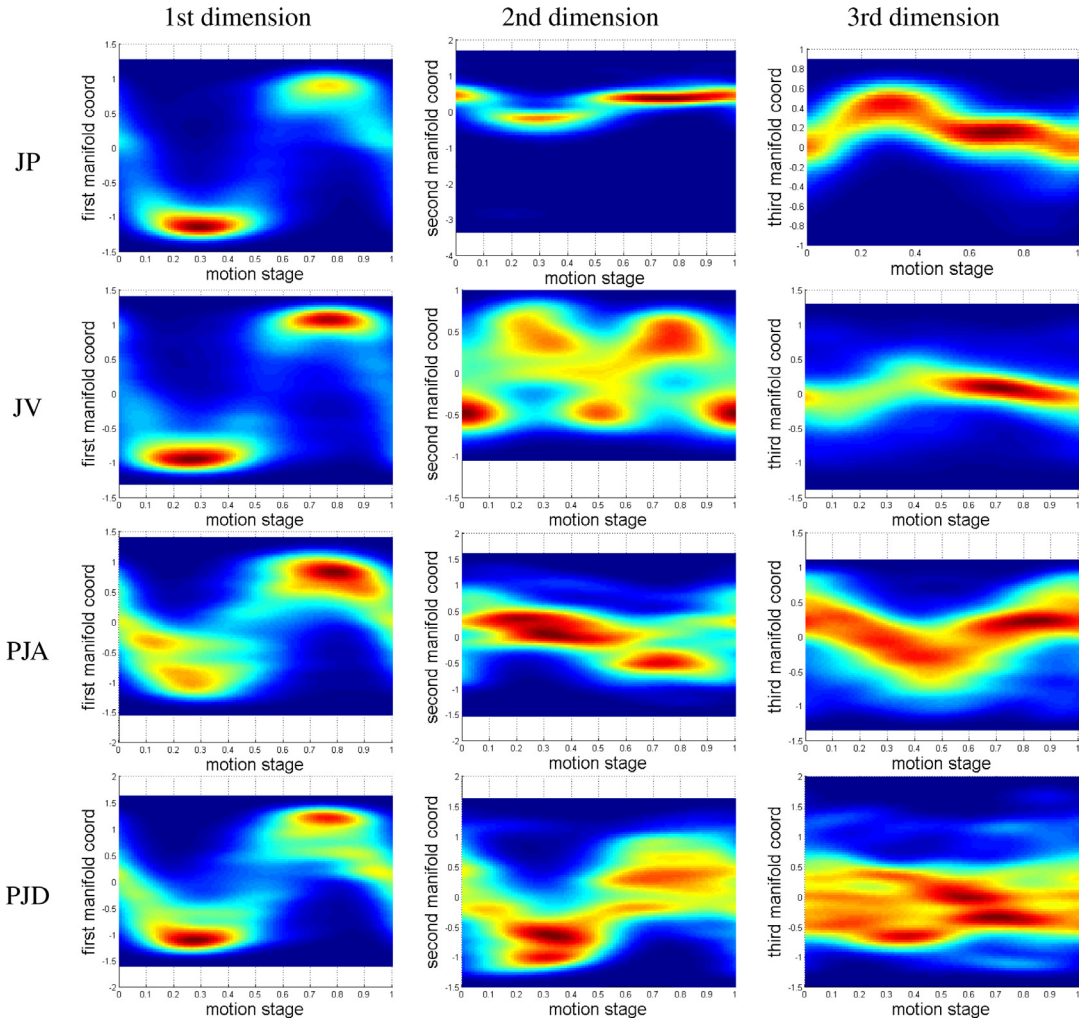

Fig. 8. Marginals of the observation model PDFs of model λ_d over the first three manifold dimensions, for the gait on stairs movement and each low-level feature. These PDFs depict the path of normality of motion, with warmer colours indicating more likely states.

Table 5
Detection rate of abnormal events in the gait on stairs scenario for best Precision-Recall results in each model.

Type of sequence	No. of abnormal events	λ_a			λ_b			λ_c			λ_d		
		Precision = 0.63 Recall = 1			Precision = 0.67 Recall = 1			Precision = 0.75 Recall = 0.95			Precision = 0.84 Recall = 0.87		
		TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
Normal	0	0	2	0	0	2	0	0	4	0	0	4	0
RL	25	25	12	0	25	8	0	25	7	0	23	1	2
LL	22	22	16	0	22	17	0	20	1	2	16	1	6
FOG	13	13	5	0	13	7	0	13	7	0	13	4	0
Total	60	60	35	0	60	30	0	58	19	2	52	10	8

λ_d has a non-optimized Matlab implementation. For all models, there is no significant additional cost by using extra dimensions.

λ_a and λ_b were found to be significantly worse in distinguishing normal and abnormal movements, thus in the rest of the article, the results from these models are not presented.

6.3.3. Selection of body joints

The results we present are produced using all body joints. This strategy allows our method to be applied to any motion type, as will be shown next. We also believe that, even though some motions may intuitively seem sufficiently represented using selected body joints – such as lower body joints in the case of walking – the exploitation of full body information may add beneficial information on the overall balance of the person. We demonstrate this by performing the analysis using different subsets of body joints in the case of gait on stairs using the JP low-level feature, a 3D manifold and model λ_d. We first use lower body joints only, then in a second test use Orthogonal Marching Pursuit (OMP) to select the low-level features that are most relevant for deriving the high level features. Fig. 10 shows that the high level features reconstruction error is dramatically reduced using the 9 most significant low-level features, and does not improve significantly using more of them. Therefore, in our second test we use the 9 most significant low-level features selected by OMP and summarized in the first row of Table 6. Note that these features correspond to both legs and arms data.

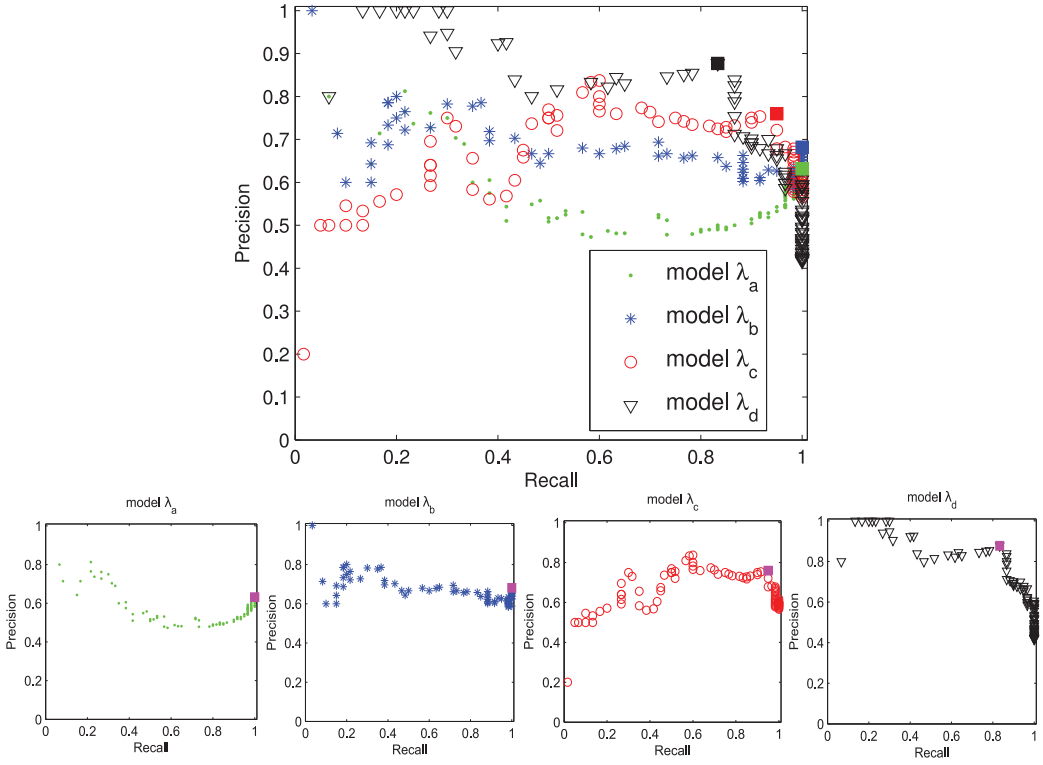

Fig. 9. Upper: Precision and recall values for event detection in the gait on stairs scenario when varying the threshold on frame classification, plotted for the best parameter setting for each motion model. Bottom: Split of the scatter plot into four, for better visualisation.

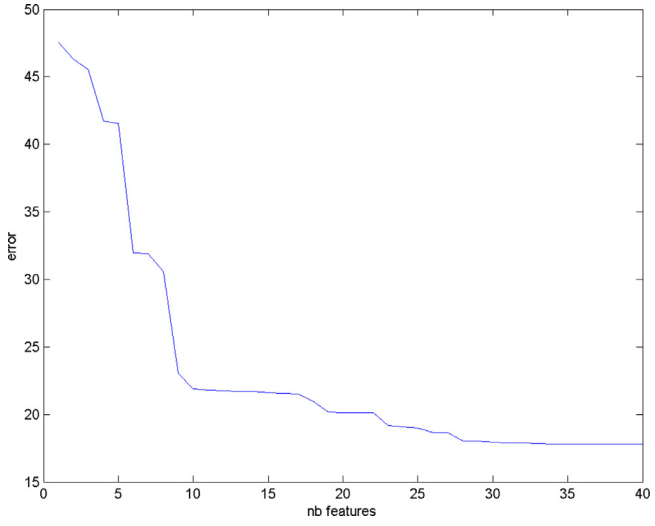

Fig. 10. Selection of low-level features using the Orthogonal Marching Pursuit: high level feature reconstruction error as a function of the number of low-level features.

The ROC curves obtained for frame classification, and the precision and recall values for abnormal event detection, are shown for both tests in Fig. 11. The AUC values are reported in the second row of Table 6. Our first observation is that, although the best results are obtained using the 21 lower body features, with AUC of 0.74 and 0.77 using $\mathcal{M}_t$ and $\mathcal{M}_{\omega_t}$ respectively, the only 9 features selected by OMP, and that mix lower and upper body information, are very close with AUC of 0.71 for both measures. Secondly, the lower joints results are significantly worse than the best result of using all body joints that had a AUC of 0.82 using $\mathcal{M}_t$. We conclude from these two observations that upper body joints contain information that may contribute significantly to the analysis of gait and that should not be discarded.

Table 6
Low-level features used in the feature selection tests, and AUC results using both online measures ($\mathcal{M}_t$ / $\mathcal{M}_{\omega_t}$).

	Lower body joints	OMP selection	Full body
Low-level features	xyz torso xyz left hip xyz right hip xyz left knee xyz right knee xyz left foot xyz right foot (21 features)	z left hand y left elbow y left foot y right hand y right foot z left foot x left hand x right elbow z right foot (9 features)	All joint coordinates (45 features)
AUC	0.74 / 0.77	0.71 / 0.71	0.82 / 0.75

6.4. Walking on a flat surface

The abnormal sequences in the SPHERE-Walking2015 dataset differ from the previous gait on stairs ones in that all frames are abnormal. The continuous scoring of our method is a particularly useful feature in this case, while its frame-by-frame analysis ability is less relevant. Therefore, to test the performance of different models on this dataset, one overall continuous score is provided for each sequence. In order to assess the ability of this score to discriminate abnormal from normal movements for each model, we compute the AUC of the ROC curves of sequence classification accuracy. Note that these AUCs are different to the ones used in Section 6.3 for per-frame classification accuracy.

We show the results of models λ_c and λ_d in Table 7 using different low-level features and manifold dimension n. The table shows that for both models, feature JP provides a good representation of the data that can discriminate the normal and abnormal walking movements. Features PJA and PJD for model λ_c, and JV and PJD for model λ_d, also yield very good results.

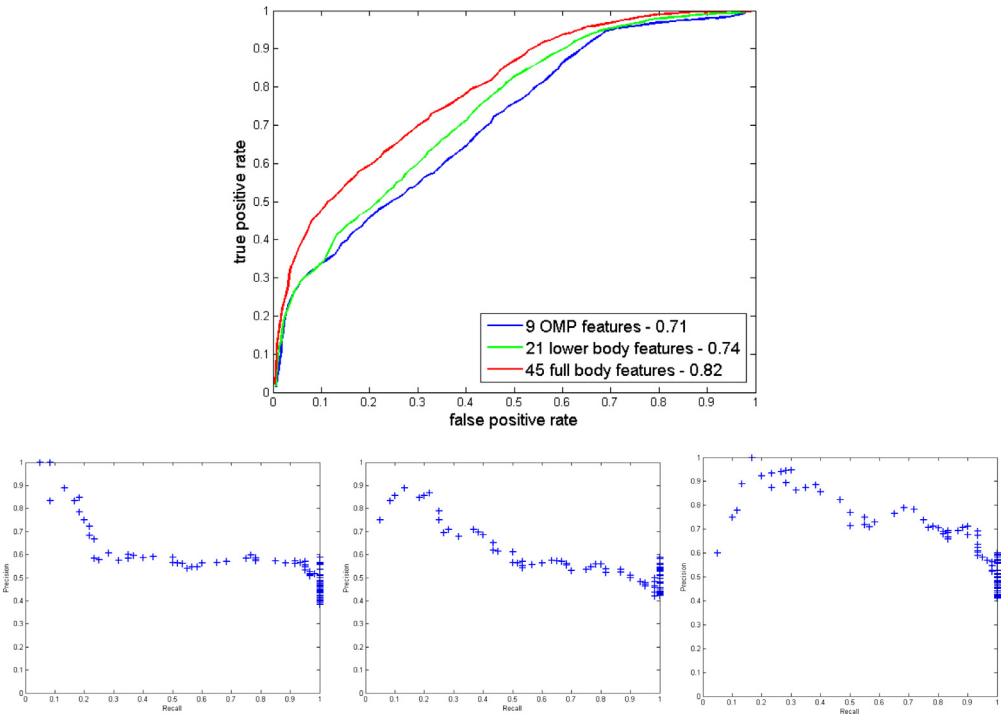

Fig. 11. ROC curves of frame classification (top) and precision and recall values for abnormal event detection (bottom) using the lower body joints (left), the 9 low-level features selected by OMP (middle), and all body joints (right).

Table 7
AUC results in the case of the walking on a flat surface motion for different skeleton representations and measures for models λ_c ($\mathcal{M}_t$) and λ_d ($\mathcal{M}_t/\mathcal{M}_{\omega_t}$).

Motion model	Feature	Manifold dimension n 1	2	3	4	5
λ_c	JP	0.96	1.00	0.99	1.00	1.00
	JV	0.93	0.79	0.86	0.95	0.85
	PJA	0.91	0.98	1.00	0.96	0.98
	PJD	1.00	1.00	1.00	1.00	1.00
λ_d	JP	0.96 / 1.00	0.99 / 1.00	0.95 / 1.00	0.99 / 1.00	0.93 / 1.00
	JV	0.95 / 0.98	0.88 / 0.86	0.87 / 0.95	0.91 / 0.95	1.00 / 1.00
	PJA	0.89 / 0.91	0.82 / 0.88	0.91 / 0.96	0.94 / 0.96	0.94 / 0.96
	PJD	0.96 / 1.00	0.91 / 0.96	0.92 / 0.95	0.89 / 0.93	0.93 / 0.98

Table 8
AUC results in the case of the sitting movement for different skeleton representations and measures using models λ_c ($\mathcal{M}_t$) and λ_d ($\mathcal{M}_t/\mathcal{M}_{\omega_t}$).

Motion model	Feature	Manifold dimension n 1	2	3	4	5
λ_c	JP	0.82	0.96	0.99	0.97	0.97
	JV	0.88	0.87	0.85	0.83	0.77
	PJA	0.59	0.67	0.68	0.63	0.73
	PJD	0.80	0.92	0.93	0.92	0.86
λ_d	JP	0.99 / 1.00	0.99 / 0.99	0.98 / 0.99	0.97 / 1.00	0.95 / 1.00
	JV	0.69 / 0.73	0.70 / 0.79	0.72 / 0.70	0.71 / 0.66	0.61 / 0.59
	PJA	0.67 / 0.67	0.61 / 0.56	0.68 / 0.68	0.65 / 0.70	0.62 / 0.66
	PJD	0.42 /0.47	0.77 / 0.79	0.76 /0.81	0.86 / 0.92	0.81 / 0.85

Table 9
AUC results in the case of the standing movement for different skeleton representations and measures using models λ_c ($\mathcal{M}_t$) and λ_d ($\mathcal{M}_t/\mathcal{M}_{\omega_t}$).

Motion model	Feature	Manifold dimension n 1	2	3	4	5
λ_c	JP	0.88	0.98	1.00	0.98	0.99
	JV	0.88	0.98	0.95	0.85	0.71
	PJA	0.83	0.74	0.84	0.86	0.83
	PJD	0.90	0.84	0.97	0.95	0.97
λ_d	JP	0.85 / 0.86	0.44 / 0.50	0.56 / 0.62	0.85 / 0.76	0.88 / 0.75
	JV	0.92 / 0.95	0.86 / 0.84	0.92 / 0.97	0.91 / 0.97	0.88 / 0.91
	PJA	0.34 / 0.39	0.38 / 0.51	0.59 / 0.56	0.66 / 0.64	0.59 / 0.60
	PJD	0.82 / 0.87	0.95 / 0.95	0.84 / 0.84	0.85 / 0.86	0.78 / 0.77

For model λ_d, we note that the advantage of $\mathcal{M}_t$ over $\mathcal{M}_{\omega_t}$ is not as obvious as in Section 6.3. This may be due to the averaging of the scores over the full sequence, which makes a timely detection of abnormal events less relevant. The results obtained for this movement are overall more satisfactory than in Section 6.3 with gait on stairs. We explain this by the easier challenge of this test (whole sequence vs per-frame analysis), linked to the abnormality type.

Fig. 12 highlights the potential of our continuous scores to help differentiate between the two types of abnormality (Parkinson and Stroke) in our SPHERE-Walking2015 dataset. Fig. 12a shows that the dynamics score $\mathcal{M}_t$ can successfully differentiate normal gaits from both types of abnormalities, while Fig. 12b shows that the pose score may also help in distinguishing Parkinson's from stroke gaits. Indeed, Parkinson sequences tend to have lower pose scores than stroke sequences, due to their pose being consistently abnormal throughout the sequence (blue curve in Fig. 13), while the pose in stroke sequences vary periodically between strongly abnormal and nearly normal within each gait cycle (red curve in Fig. 13). This result denotes a clear potential of our method for clinical applications, which will be further assessed in future works.

6.5. Sitting and standing

As in Section 6.4, in the SPHERE-SitStand 2015 dataset the sequences are either fully normal or fully abnormal, thus an overall score is provided for each sequence to assess its overall abnormality level. Tables 8 and 9 show the sequence-wise AUC values of the sitting and standing movements, respectively, obtained by the different pose representations and motion models. For the sitting motion, both

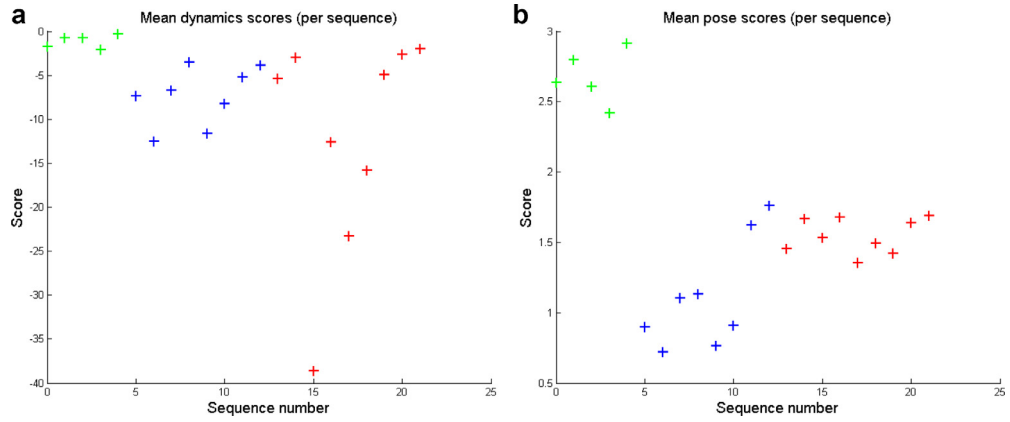

Fig. 12. Quality measures for each of the walking sequences: (a) dynamics measure $\mathcal{M}_t$, and (b) pose measure $\mathcal{M}_{\text{pose}}$ for normal sequences (green), Parkinson sequences (blue), and stroke sequences (red).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

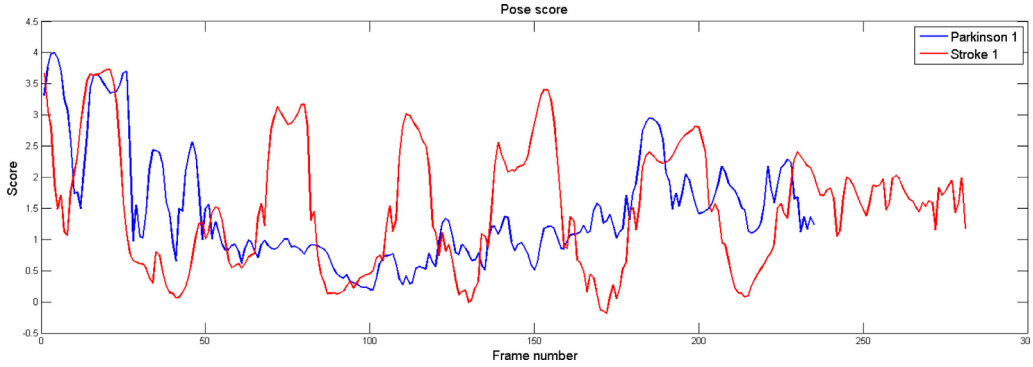

Fig. 13. Comparison of the pose measure $\mathcal{M}_{\text{pose}}$ in two examples of Parkinson (blue) and stroke (red) sequences. $\mathcal{M}_{\text{pose}}$ is consistently low in the Parkinson sequences, while it varies periodically in the stroke one.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

models λ_c and λ_d perform better with the JP feature. For the standing motion, model λ_c also performs better with JP, while model λ_d should use either JV or PJD. Both models perform similarly well at detecting abnormal sequences, with best AUCs of model λ_c at 0.99 and 1.00 for the sitting and standing motions respectively, and 1.00 and 0.97 for model λ_d.

7. Conclusion

In this work, we have studied the efficiency of different pose representations and HMM-based dynamics models for describing and assessing the quality of four motions used by clinicians to assess functional mobility. The results show that the continuous-state HMM is better suited for describing motion dynamics than classical, discrete-state HMMs when a frame-by-frame analysis is required. For globally analyzing whole sequences, both the continuous-state HMM and the classical (discrete-state) HMM with discriminative classifier performed well. Furthermore we have found that the adequacy of the pose representation to modelling pose variations plays a key role in the ability of the dynamics model to represent and discriminate the motion.

The proposed method provides a continuous score for assessing the level of abnormality of movements. We showed in this work that this score can generalise to various movement and abnormality types. Future work will include further assessing the clinical relevance of this continuous score by comparing it against manual scoring schemes that are routinely used in clinical practice.

Moreover, although the robust manifold helps to reduce the effects of noise, abnormal poses may be seen as noisy normal data instead of being properly represented and picked up as abnormal. The ability of our pose representation at discriminatingly representing abnormal poses should therefore be evaluated as part of future work. Training on a large variety of poses (both normal and abnormal) for building the pose manifold may address this possible limitation of our current pose representation.

Acknowledgments

This work was performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

References

[1] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comput. Graph. Vis. 3 (3) (2008) 177–280.
[2] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer, 2013, pp. 149–187.
[3] J. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput Surv. 43 (3) (2011).
[4] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[5] S.-R. Ke, H.L.U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, K.-H. Choi, A review on video-based human activity recognition, Computers 2 (2) (2013) 88–131.
[6] O. Popoola, K. Wang, Video-based abnormal human behavior recognition a review, IEEE Trans. Syst., Man, Cybern, Part C: Appl. Rev. 42 (6) (2012) 865–878.
[7] H. Pirsiavash, C. Vondrick, A. Torralba, Assessing the quality of actions, in: Proceedings of the Computer Vision–ECCV 2014, Springer, 2014, pp. 556–571.
[8] R. Wang, G. Medioni, C. Winstein, C. Blanco, Home monitoring musculo-skeletal disorders with a single 3d sensor, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2013, pp. 521–528.
[9] J.M. VanSwearingen, K.A. Paschal, P. Bonino, J.-F. Yang, The modified gait abnormality rating scale for recognizing the risk of recurrent falls in community-dwelling elderly adults, Phys. Therapy 76 (9) (1996) 994–1002.
[10] L. Wolfson, R. Whipple, P. Amerman, J.N. Tobin, Gait assessment in the elderly: a gait abnormality rating scale and its relation to falls, J. Gerontol. 45 (1) (1990).

[11] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, M. Mirmehdi, Online quality assessment of human movement from skeleton data, in: British Machine Vision Conference, BMVA press, 2014, pp. 153–166.
[12] M.Z. Uddin, J.T. Kim, T. Kim, Depth video-based gait recognition for smart home using local directional pattern features and hidden markov model, Indoor Built Environ. 23 (1) (2014) 133–140.
[13] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.
[14] OpenNI organization, OpenNI User Guide (November 2010) URL http://www.openni.org/documentation.
[15] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 588–595.
[16] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 20–27.
[17] A. Yao, J. Gall, L. Van Gool, Coupled action recognition and pose estimation from multiple views, Int. J. Comput. Vis. 100 (1) (2012) 16–37.
[18] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.
[19] G.S. Parra-Dominguez, B. Taati, A. Mihailidis, 3d human motion analysis to detect abnormal events on stairs, in: Proceedings of the International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, IEEE, 2012, pp. 97–103.
[20] M.E. Hussein, M. Torki, M.A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2466–2472.
[21] E. Ohn-Bar, M.M. Trivedi, Joint angles similarities and HOG2 for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2013, pp. 465–470.
[22] A. Elgammal, C.-S. Lee, The role of manifold learning in human motion analysis, in: Human Motion, Springer, 2008, pp. 25–56.
[23] L. Wang, D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Trans. Image Process. 16 (6) (2007) 1646–1661.
[24] J. Blackburn, E. Ribeiro, Human motion recognition using isomap and dynamic time warping, in: Human Motion–Understanding, Modeling, Capture and Animation, Springer, 2007, pp. 285–298.
[25] Microsoft corp. redmond wa. kinect for xbox 360.
[26] J. Charles, T. Pfister, D. Magee, D. Hogg, A. Zisserman, Upper body pose estimation with temporal sequential forests, in: Proceedings of the British Machine Vision Conference 2014, BMVA Press, 2014, pp. 1–12.
[27] S. Gerber, T. Tasdizen, R. Whitaker, Robust non-linear dimensionality reduction using successive 1-dimensional laplacian eigenmaps, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 281–288.
[28] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008., IEEE, 2008, pp. 1–8.
[29] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 2834–2841.
[30] Z. Liu, S. Sarkar, Improved gait recognition by gait dynamics normalization, IEEE Conf. Pattern Anal. Mach. Intell. 28 (6) (2006) 863–876.
[31] F. Lv, R. Nevatia, Recognition and segmentation of 3D human action using hmm and multi-class adaboost, in: Proceedings of the Computer Vision–ECCV 2006, Springer, 2006, pp. 359–372.
[32] J. Snoek, J. Hoey, L. Stewart, R. Zemel, A. Mihailidis, Automated detection of unusual events on stairs, Image Vis. Comput. 27 (1) (2009) 153–166.
[33] P. Natarajan, R. Nevatia, Online, real-time tracking and recognition of human actions, in: Proceedings of the IEEE Workshop on Motion and video Computing, IEEE, 2008, pp. 1–8.
[34] P. Peursum, S. Venkatesh, G. West, Tracking-as-recognition for articulated full-body human motion analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[35] H.L.U. Thuc, S.-R. Ke, J.-N. Hwang, P. Van Tuan, T.N. Chau, Quasi-periodic action recognition from monocular videos via 3D human models and cyclic hmms, in: Proceedings of the International Conference on Advanced Technologies for Communications, IEEE, 2012, pp. 110–113.
[36] T. Duong, D. Phung, H. Bui, S. Venkatesh, Efficient duration and hierarchical modeling for human activity recognition, Artif. Intell. 173 (7) (2009) 830–856.
[37] M. Narasimhan, P. Viola, M. Shilman, Online decoding of markov models under latency constraints, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 657–664.
[38] S. Nowozin, J. Shotton, Action points: a representation for low-latency online human action recognition, Techical Report MSR-TR-2012-68, Microsoft Research Cambridge, 2012.
[39] I. Kviatkovsky, E. Rivlin, I. Shimshoni, Online action recognition using covariance of shape and motion, Comput. Vis. Image Underst. 129 (2014) 15–26.
[40] R. De Rosa, N. Cesa-Bianchi, I. Gori, F. Cuzzolin, Online action recognition via nonparametric incremental learning, in: Proceedings of the British Machine Vision Conference, 2014, pp. 1–15.
[41] F. Nater, H. Grabner, L. Van Gool, Exploiting simple hierarchies for unsupervised human behavior analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, IEEE, 2010, pp. 2014–2021.
[42] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[43] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proc. Natl. Acad. Sci. USA 102 (21) (2005) 7426–7431.
[44] P. Arias, G. Randall, G. Sapiro, Connecting the out-of sample and pre-image problems in kernel methods, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[45] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[46] Z. Uddin, T.-S. Kim, Continuous hidden markov models for depth map-based human activity recognition, INTECH Open Access Publisher, 2011.
[47] J. Kwon, F.C. Park, Natural movement generation using hidden markov models and principal components, IEEE Trans. Syst., Man, Cybern., Part B: Cybern., 38 (5) (2008) 1184–1194.
[48] S. Krüger, M. Schafföner, M. Katz, E. Andelic, A. Wendemuth, Speech recognition with support vector machines in a hybrid system., in: Interspeech, 2005, pp. 993–996.
[49] M. Valstar, M. Pantic, Combined support vector machines and hidden markov models for modeling facial action temporal dynamics, in: Human–Computer Interaction, Springer, 2007, pp. 118–127.
[50] H.-T. Lin, C.-J. Lin, R. Weng, A note on platts probabilistic outputs for support vector machines, Mach. Learn. 68 (3) (2007) 267–276.
[51] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classif. 10 (3) (1999) 61–74.
[52] N. Vaswani, A.K. Roy-Chowdhury, R. Chellappa, " shape activity": a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection, IEEE Trans. Image Process. 14 (10) (2005) 1603–1616.
[53] M.J. Beal, Z. Ghahramani, C.E. Rasmussen, The infinite hidden markov model, in: Machine Learning, MIT Press, 2002, pp. 29–245.