3D Visual-based Human Motion Descriptors: A Review

Margarita Khokhlova
Le2i UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté, France
margarita.khohklova@u-bourgogne.fr

***Abstract*—This paper aims to provide a comprehensive reference source on depth-based human motion descriptors. Motion description is a challenging problem which became popular with recent advances in 3D computer vision. Our goal is to introduce the main trends in human 3D motion descriptor design and evaluation next to presenting a review of recent methods belonging to three application categories: action recognition, gesture recognition and gait assessment.**

***Keywords*-human motion description; motion 3D; action recognition; gesture recognition; gait assessment.**

I. Introduction

Visual or image descriptors are short feature vectors of the salient regions of images and video sequences. Descriptors aim to describe the visual characteristics of present objects such as shape, color, texture or motion and may replace an image as the input to a classifier.

Specially designed and tailored descriptors can also represent motion, which is an essential part of algorithms in rather diverse applications, such as the human activity recognition, gait recognition and analysis, motion tracking, 3D scene reconstruction to name a few examples.

With the advent of low-cost 3D sensing cameras and continued efforts in advanced point cloud processing, 3D perception has gained more importance in the vision domain. 3D sensing devices not only provide the user with a general projection of the 3D world to a 2D image plane as regular cameras, but also acquire the 3D geometry, or depth. The depth images deliver natural surfaces which can be exploited to capture geometrical features of the observed scene with a rich descriptor. Compared with conventional color videos, the additional depth information in RGBD data helps to adjust for different lighting conditions, remove background noise and simplify intra-class motion variations. Therefore, in general, RGBD-based descriptors outperform the RGB-based ones [1].

The standard output of a 3D sensor is a point cloud, which contains points on the scene next to their XYZ coordinates and an optional color tuple. A descriptor specifically designed for a point cloud video sequence could serve as an important basis for motion characterization.

The information extracted from 3D point clouds is predominantly comprised of the shape, color (or intensity) and the spatial relation between cloud points. Shape descriptors are the most popular 3D descriptors for point clouds, and amongst them normal-based descriptors are the most widely used.

Widely applied three dimensional shape descriptors include the Normal Aligned Radial Features (NARF) [2] and the Spin image descriptor [3] whereas for two dimensional planes HOG3D [4], SURF and SIFT [5] are common. The latter may be used to represent a motion trajectory when applied to individual consecutive frames [6]. Many have been implemented by the Point Cloud Library [7].

Finally, 3D motion, sometimes referred to as 4D, combines 3D spatial scenes though time. This data may be described via the calculation of the so-called Motion or Scene Flow, which are build of from 3D velocity vectors. Dense motion flow is information rich, but is computationally intensive to determine and carries a significant memory and storage burden. Thus in recent years many alternative algorithms have been proposed that aim at reducing this burden. This paper aims to review and categorize this existing family of methods, specifically for applications in human motion analysis.

Motion analysis using depth data was extensively reviewed in [8]. However, in this work researchers were focused primarily on activity recognition methods applied to RGBD video sequences omitting (3D) motion descriptors. Therefore, we assumed a need for a specialized review dedicated to 3D descriptors applied to human motion, divided into the following subcategories: Motion Recognition, Gesture Recognition, and Movement Analysis and specifically, Gait recognition.

The remainder of this article is organized as follows. Section 2 presents the requirements and evaluation criteria for motion descriptors. In the second part of Section 2 we introduce the most popular 3D datasets for motion descriptor evaluation. Further in Section 3, 4 and 5 we talk about the common trends and representative methods to describe the motion for 3 applications: action recognition, gesture recognition and gait analysis. We summarize the main approaches and point out their interesting and novel parts. Finally, Section 6 concludes the paper, where we highlight the main existing trends in motion description for point clouds sequences and propose our outlook for the future.

II. Motion descriptor evaluation

A. Motion descriptor characteristics

Common requirements for 3D descriptors are invariance to transformations of the target object, user-independence, robustness to noise and clutter and storage efficiency (i.e., compactness). We propose to compare existing descriptors by a combination of the following characteristics, which highlight descriptor specifics and give an idea in which scenarios their performance is optimal:

Application: Corresponds to the particular purpose for which the descriptor is designed.

Locality: Descriptor can be local (regional) or global based on the features it captures. If the descriptor is local, it is applied to selected points of interest, or in a local support region surrounding a basis point of an object. In this case, before applying the descriptor, potentially interesting points should be identified, preferably, automatically. Rarely researchers try to select the most representative points by applying a descriptor to all the data points of the model and comparing the values to find 'rare' ones [9]. Commonly algorithms may be split in a detector and a descriptor part, where the detector locates regions that are perceived to be interesting or stable and the descriptor encodes a region of interest around the location of the feature. Using motion descriptors for local features generally provides one with invariance to geometric transformations and gives a reduction of the perturbations caused by variations in scale, rotation, and viewpoint. Using local descriptors also allows to reduce the computational complexity of the algorithms. However, local descriptors ignore the global spatial structure information of the scene.

Global descriptor is applied to all points or to points selected by using a linear sampling scheme and is capable of capturing global spatial structure and affiliation.

Analogous to the descriptors for 2D images, descriptors applied locally are more common due to the fact that the cost of globally computing descriptors is usually high.

Dimensionality: Usually directly connected to the size and representativeness of a descriptor. Dimensionality shows how many values are stored in each descriptor array. Dimensionality of a motion descriptor is rarely a significant issue unless the descriptors for very long video sequences should be stored or the rich motion flow based descriptor is used. In this case the general storage burden is predominantly determined by the number of descriptors and not their individual size.

View-invariance and scale-invariance: View-invariance of a descriptor entails its robustness against changes in view-point. Many approaches for 3D shape retrieval and identification transform an object of interest into a canonical pose. Translating the center of gravity of the object into the origin and normalizing the area/volume or radius of the bounding circle/sphere/parallelepiped etc. This operation guarantees view-invariance to a reasonable extend.

Scale-invariance is the invariance of the descriptor for the objects size or the distance of the camera to the object. In Local Spatio-Temporal (LST) descriptors it can be, for example, the fixed support region of the feature point which is not adapted to the linear perspective view variations.

Accuracy: Characteristic which shows how well the proposed descriptor performed for a given task and if it corresponds to the task at hand. In this work, we state the accuracy as reported by the researchers and also the theoretical evaluation of the method proposed. When available, we report the results obtained on a benchmark dataset. However, we should mention that even when an identical dataset is used, the validation method used by each work might differ from the others [10], so a direct comparison is not always possible.

Computational complexity : The calculation resources required by an algorithm. Sometimes computational complexity is reported by the authors or could be approximated by the specifics of the algorithm.

Classifier: The algorithm that implements the recognition task in the application. The use of a different classifier can actually change the final recognition score significantly, however, with a set of discriminative features even a very simple classification method like linear regression can show good results. It is very common to use the Bag-of-Words model for classification tasks ([11], [12], [13], [14], [6]) and in this case the model defines the used classifier.

Dataset: Motion descriptors reviewed in this work were proposed for a particular application and evaluated using task corresponding test data. The choice of a dataset obviously affects the reported accuracy of an algorithm. Datasets will be evaluated on their difficulty and their data scope.

B. Datasets for the motion descriptors

When proposing a new algorithm, its performance usually needs to be evaluated in comparison to existing ones for a given application. Construction and annotation of a new database is often a long and arduous process, therefore it is preferable to do only do this when an existing benchmark can not be used.

We propose a list of popular Benchmark datasets, most widely used for 3D human motion descriptor evaluation next to their specifics and designation for a particular task.

MSR Action3D: This is the most used RGBD human action-detection and recognition dataset [11], [15], [16], [17], [18]. Mainly because this is one of the first RGBD datasets capturing motions (dated 2010) and it contains the biggest amount of different actions. The MSR Action3D Dataset [12] consist of 20 action types performed by 10 subjects 2 or 3 times. The actions are: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw an x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. The resolution of the video is not very high, namely 320x240. The data was recorded with a depth sensor similar to the Kinect device. It is a challenging dataset to test an algorithm on and to compare the results with earlier proposed methods.

MSRGesture3D: This dataset was captured by a Kinect device. There are 12 dynamic American Sign Language gestures performed 2-3 times by 10 people. The hand sign language represents a concentrated entity in motion, combining both the overall movement of hands and the inner variability of fingers. The hand portion (above the wrist) is segmented. An alternative to this dataset with a less number of gestures is the Sheffield Kinect Gesture (SKIG) dataset [19], which captures forearm gestures under different hand poses, background and illumination variations from 6 subjects. This makes SKIG more applicable to a real-life scenario than the MSR Gesture 3D dataset. However, the latter remains the most widely used for gesture recognition research [20], [21], [22].

MSR Daily Activity 3D: Also captured using a Kinect device, it includes 16 activities performed by 10 subjects 2 times, in a standing and a sitting position. The actions are: drink, eat, read a book, call a cellphone, write on a paper, use a laptop, use a vacuum cleaner, cheer up, sit still, toss a paper, play a game, lie down on a sofa, walk, play a guitar, stand up, sit down. There is a sofa in the scene. RGB and depth channels are recorded, and also the skeleton joint positions are extracted. However, the RGB channel and depth channel are recorded independently, so they are not strictly synchronized. The dataset is more challenging than MSR Action3D, because it represents natural everyday activities, which are harder to distinguish. MSR Daily Activity 3D is a good choice to evaluate a real-life scenario dedicated application and compare the results with other algorithms [14], [16], [15].

Berkeley MHAD: The Berkeley Multimodal Human Action Database (MHAD) [23] is a complete and general purpose dataset which consist of temporally synchronized and geometrically calibrated data from an optical motion capture system, multi-baseline stereo cameras from multiple views, depth sensors, accelerometers and microphones. The dataset contains 11 actions performed by 7 male and 5 female subjects (in the range 23-30 years of age except for one elderly subject) 5 times. The total recording time is 82 minutes, which makes this dataset one of the biggest by the amount of video sequences it contains. The specified set of actions comprises of the following: (1) actions with movement in both upper and lower extremities, e.g., jumping in place, jumping jacks, throwing, etc., (2) actions with high dynamics in upper extremities, e.g., waving hands, clapping hands, etc. and (3) actions with high dynamics in lower extremities, e.g., sit down, stand up. Berkeley MHAD is popular [14], [6], probably due to the fact it allows to perform a multi-modal analysis of human motion and is easy to use.

TUM GAID: For depth based gait recognition and assessment, this challenging multimodal recognition database [24] was proposed in 2014. This database simultaneously contains RGB video, depth and audio. The database contain 305 individual gait captures, acquired in different weather conditions and in a different context, i.e: the person walks normally or is wearing a backpack or coating shoes (some persons performed all the actions and some only a subset). Other databases for depth gait analysis are available, however, TUM GAID is the most cited and currently used. It is the only database which allows for multi-modal gait recognition using video, depth and audio features along with different acquisition conditions.

Dataset	Action number	Person number	Calibration	Annotation	Total seq
MSR Action 3D	20	10	no	skeleton joints	567
MSR Gesture 3D	12	10	no	segmentation	336
MSR Daily Activity 3D	16	10	no	skeleton joints	320
Berkeley MHAD	11	12	yes	temporal synchronization	660
TUM GAID	3	305	no	metadata	3370

Table I: Popular 3D Video datasets and their characteristics

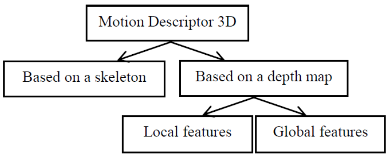

Figure 1: Motion descriptors classification

III. Motion descriptors for action recognition

Motion description is an essential part of human activity recognition. From a computational perspective, actions are best defined as four-dimensional patterns in space and in time. Methods which are able to discriminate the class of action being performed based on analysis of the video sequence combine a motion descriptor with a classifier. Most of the work on human action recognition published up to today relies on information extracted from 2D images and videos. However, with the availability of affordable depth sensors, this research area enlarged considerably with new studies dedicated to 3D. For a detailed review on action recognition, the works [25], [26] could be referred. These reviews, however, are not particularly oriented on RGBD-based methods as ours.

Motion descriptors can be categorized with respect to various criteria. A general classification feature-based scheme is shown in Figure 1. Further in this work we introduce 9 different motion descriptors for action recognition representing popular strategies and choices made by researchers in this field.

A very common approach for human action recognition is to track the human joints from the depth maps [27], [16]. A simple yet effective example of this approach is the depth-based algorithm by Xia et al. [11]. A view-invariant posture representation was devised using histograms of 3D joint locations (HOJ3D) within a modified spherical coordinate system. The positions of the joints in time form the 3D spatial histogram. HOJ3D were re-projected using LDA and clustered into k posture visual words. The temporal evolutions of these visual words were modeled by a discrete hidden Markov model (HMM).

Joint-based methods are popular, but will fail if the initial joints were estimated wrongly, which is still an issue. Moreover, if a very fine action is to be recognized (for example, a gesture), the joint-based methods lack precise information on shape and movement. For this reason, low-level attributes in depth images often outperform more high-level representations [15].

In 2010 Li et al.[12] proposed a new depth-map based method based on local features. They use an expandable graphical model to explicitly model the temporal dynamics of the actions and propose to use a bag of 3D points extracted from the depth map to model the postures. To select the points, they project the depth map onto the orthogonal Cartesian planes and further sample a specified number of points at equal distance along the contours of the projections. Selected points are then clustered in order to obtain salient postures. A Gaussian Mixture Model is used to globally model the postures by the distribution of points, and an action graph is constructed from the training samples to encode all actions that need to be recognized. This method gives better results than the 2D silhouette based action recognition, and the final descriptor is very compact. However, the cross-subject activity recognition results reported are low due to the fact that the proposed sampling scheme is view dependent. Secondly, the descriptors looses spatial context information between interest points, which could be a problem when using the method in a real life scenario.

Simplified motion-flow based descriptors could also be used [28], [13]. Munaro at et al. [29] proposed a global descriptor which takes the direction and magnitude of motion of every body part into account. For that, they first identify the human on the point cloud and then center a 3D grid around it. This grid divides the space around the person into a number of cubes. The flow information (direction and force) as a mean and a summary of motion vectors extracted from each cube is used as a motion description. Motion flow is calculated by using the KD-search algorithm and the color distance in HSV space. A single descriptor is concatenated for every video sequence. The published results are good, however, the descriptor is not view-independent and the task of the aligning the video frames in time is not addressed. Moreover, we can imagine the color information to be not of a great use when person tracked has solid color clothes making it hard to establish point correspondences based on color similarity.

Hadfield et al. [13] propose a novel local motion descriptor for RGBD video sequences. The descriptor encodes the 3D orientation of flow vectors around established interesting points extracted from evenly spaced regions. The nature of the local motion field is described using a spherical histogram in the velocity domain: the contribution of each flow vector to the histogram is weighted based on the magnitude of the flow vector. To remove a 3D rotation ambiguity and to make the descriptors completely consistent, the invariance to camera roll is encoded and the direction of the motions of flow vectors within the sub-region histograms is made rotationally-invariant. An interesting approach is also to perform PCA on the local region of the motion field, which finally leads to a descriptor which is invariant to all 3 types of camera viewpoint change, next to being robust to outlier motions. To obtain a global descriptor, the video sequence is divided into space-time blocks, each of which is encoded independently to provide the final description of the sequence. In this article it was proved that normalization and adaptation of the features so that they are scale and view point invariant improves the overall recognition of the system. However, the method is used for local interest points, discarding many relational information about the movement. Sequence time block-division schemes can also lead to wrong results in the recognition.

A more advanced HON4D (histogram of oriented 4D surface normals) descriptor [15] is analogous to the histogram of gradients in color sequences and extends the histogram of normals in static point clouds. In order to construct HON4D, the 4D space (XYZ, t) is initially quantized (in order to get the grid representation) using a regular 4D extension of a 2D polygon, namely, a 600-cell Polychoron. Then the normal to the surface in the 4D space represented as the set of points is computed and normalized by the length. To form a 120-dimensional motion descriptor, researchers compute the corresponding distribution of 4D surface normal orientation for each bin. Videos are divided in parts and final descriptor is a concatenation of individual parts descriptors. The results show that the motion descriptor outperforms several earlier descriptors [18], [16], however, they do not take into account the movement in the y and x direction, building their descriptor based on the change of depth.

A notion of hierarchy can be successfully employed in 3D motion descriptors. Kong et al. [30] improve the algorithm [15] using kernel descriptors alongside with the surface normals. They present a 3D gradient kernel descriptor which is a low-level depth sequence descriptor with the ability to capture detailed information by computing a pixel-level 3D gradient. The descriptor captures the change in shape of the 3D surface in time in the following way: Firstly, the 3D normals are computed and they are projected to a learned set of compact KPCA basis vectors [31]. The kernel is measures the similarity between orientations of the gradient for corresponding pixels from different patches of a video. A Bag-of-words method is used to build a hierarchical structure upon the low-level patch features to produce mid-level feature vectors. EMK [32] is employed over the output of the 3D kernel descriptor. The method achieves state-of-the-art performance for action recognition using depth data, but it is computationally expensive. It is also not explained how exactly the correspondences between pixels in different frames of the video are estimated.

Spatio-temporal features based descriptors gained lots of attention recently [14], [6]. Yang Xiao et al. [33] proposed a 3D trajectory shape descriptor for unconstrained RGBD video. To extract the 3D dense trajectory feature, the candidate feature points are densely sampled in each RGB frame and tracked using optical flow first. Then, by mapping the 2D positions of the RGB trajectory points to the depth map, motion information along the depth direction can be intuitively captured to form 3D trajectories. This method is a good illustration of how to enrich the 2D optical flow. To obtain the global representation of RGBD videos, researchers combine several earlier approaches: Motion Boundary Histogram along the depth direction and trajectory shape descriptors [34] encoded using Fisher Vector [35]. It is however, not justified if the estimations of the 3D flow using the 2D and depth information gives good results in general.

Zhang et al. [14] propose discriminative and robust LST features named 4-D color-depth (CoDe4D) that incorporate both intensity and depth information acquired from RGBD cameras. The feature detector constructs a saliency map through applying independent filters in the XYZt dimension to represent texture, shape and pose variations, and selects its local maxima as interest points. A multichannel orientation histogram adaptive MCOH descriptor applies a 4-D support region, which is adaptive to linear perspective view changes, on each interest point. Then, image gradients of color-depth patches within the support region are computed and quantized using a spherical coordinate-based method to form a final feature vector. The method is interesting, however, the use of the color and texture information in the descriptor can be a disadvantage when a general real-case intra-person scenario application is aimed for. In this case it is proposed to tune the weighting parameters of the descriptor in order to weigh the depth information more and provide different weighting schemes for the datasets tested.

With the recent advances in human activity recognition, researchers are also addressed a challenging task of a group action recognition. Znang et al. [6] propose to use LST features which they call Adaptive Human-Centered (AdHuC) features. As in the previous method, their features are adapted to depth. To incorporate spatio-temporal and color-depth information in XYZt space researchers use a cascade of three filters: a pass-through filter to encode cues along the depth dimension, a Gaussian filter to encode cues in XY space, and a Gabor filter to encode time information. Then the color and depth cues are fused to form a saliency map. Local maximums on this map are then the LST features. The spatial-color-time based descriptor is then calculated for each point. The HOG3D [4] descriptor is modified in order to incorporate multichannel information. The final descriptor has a histogram form and concatenated of the per-channel descriptors.

IV. Motion descriptors for gesture recognition

Nowadays computer applications require new ways of interaction, especially within the growing Virtual Reality domain. For that reason, human-computer interaction and particularly, gesture recognition, became a very popular field of research in the last few years.

A gesture can be defined as a physical movement of the hands, arms, face and body with the intent to convey information or meaning [36]. Research in hand gesture recognition aims to design algorithms that can identify explicit human gestures. Gesture recognition dedicated motion descriptors are similar to the activity recognition ones with the difference that the descriptors should be capable of capturing very fine movements. Hand gestures are more difficult to recognize than body gestures due to the fact that the motions are more subtle, there are more degrees of freedom and serious occlusions occurs between the fingers. A detailed recent review on the advances in this area can be found in [37]. In this review, all the specifics of gesture recognition task are discussed along with the proposed algorithms. However, this paper tends to capture all the steps of gesture recognition process from detection up to classification and is not particularly dedicated to dynamic gesture recognition using 3D motion descriptors. The major part of the methods reviewed do not exploit the 3D point cloud based gesture recognition, which became particularly popular after 2012. For this reason, we also include in an overview of several motion descriptors applied to the gesture recognition on 3D data.

Similarly to full body motion description, many different types of visual features have been proposed for hand gestures. Early works uses 2D information and builds descriptors for the 2D silhouette of a hand. Due to the ambiguity of 2D data, the accuracy of such methods was not high. The latest dynamic gesture recognition methods use depth information and their 2D counterparts.

Model-based methods are very popular. Researchers often use the positions or the rotation angles of the joints from the skeleton structure of the fingers as the visual features [38]. An alternative approach is to use some form of geometric features extracted from a depth sequence [21]. For example, a shape silhouette can be used as a descriptor [39] or the cell occupancy information [40] similar to [17] can be used as feature. This results in approaches which are less dependent on separate segmentation and tracking algorithms.

Recently, 3D dynamic gesture recognition methods similar to action recognition ones are starting to avoid human body models and focus more on depth-based ones.

Cirujeda and Binefa [21] propose to use a Covariance matrix composed of the selected features from a 3D depth video sequence frame. Their descriptor doesn't use the absolute features themselves, but exploits representations of complex interactions between variations of 3D features in the spatial and temporal domain. It helps to make the descriptor robust to inter-subject and intra-class variations. The feature vector is the result of experimenting with several low-level cues and includes information about depth itself combined with other coarse observations such as first and second image derivatives, gradient magnitude, curvature and temporal information. The idea of their descriptor is to measure how several variables change together, capturing the intrinsic correlation between distributions of the involved cues. The final sequence descriptor is a concatenation of the three scene-wise covariances in its vectorized form. The descriptor captures the global motion patterns. The method is easily generalizable for action recognition and outperforms [20], [15]. It is also independent of the sequences length and from the cluttered background, however, it doesn't explicitly use the information about the movement (i.e. force and direction), which still can be useful in action recognition.

Descriptor	Year	Locality	Dimensionality	View-invariance	Accuracy, %	Complexity	Classifier	Dataset
HOJ3D [11]	2012	local	mid($\approx$ 1008)	yes	mid (78.97*)	low	HMM	MSR Action3D, custom (10 actions).
Bag of 3D points [12]	2010	local	low	no	low (74.7**)	low	NN	Custom, 20 actions, complex combinations.
3D grid-based descriptor[29]	2013	global	mid ($\approx$ 5760)	no	mid (87.4**)	low	NN	IAS-Lab Action Dataset (15 actions, 12 person).
HON4D [15]	2013	global	low ($\approx$ 120)	yes	high (88.89*)	low	SVM	MSR Action 3D, MSR Gesture 3D, MSR Actionv 3D Pairs, MSR Daily Activity 3D.
3D Flow descriptor[13]	2014	local	low($\approx$ 144)	yes	high(36.9**)	high	SVM	Hollywood 3D (14 actions, multi-cam setup).
3D kernel descriptor [30]	2015	hierachical	high ($\approx$ 13824)	yes	high (92.73*)	high	SVM	MSR Action 3D, MSR Action 3D Pairs, and MSR Gesture 3D.
3D Trajectories [33]	2014	global	low ($\approx$ 96)	no	mid (29.76**)	low	SVM	Hollywood 3D.
CoDe4D [14]	2016	local	high ($\approx$ 21600)	yes	high (86**)	high	SVM	Berkeley MHAD, $ACT4^2$, MSR Daily Action 3D, UTK Action3-D.
AdHuC [6]	2015	local	low	yes	high (85.7*)	high	SVM	Berkeley MHAD and $ACT4^2$.

Table II: Motion descriptors for action recognition. **Reported accuracy is for MSR Action 3D dataset when available as reported by authors. **Accuracy reported for other dataset.*

Recently Ohn-Bar and Trivedi [1] proposed a combination of global low-level spatio-temporal features for gesture recognition in naturalistic driving settings. Their feature set is combining other earlier features: motion history image (MHI) [41] extension and HOG features. Several descriptors are tested along with different fusion schemes to establish the better-performing ones. Moreover, only compact descriptors are used, so the resulting one is small in dimensionality and fast to compute. For this work, depth and color data descriptors were extracted separately and their performance was compared. The best combination is Extended HOG2 + Extended MHI paired with DTM descriptor. This work shows that the approach to use a descriptor on color and depth data separately and fuse them in the final step of the algorithm works very well.

V. Motion descriptors for gait analysis

Lately, the subject of gait recognition and analysis from 3D data became popular. Gait is a manner of walking on a solid substrate. Observation of gait can provide early diagnostic clues for a number of movement disorders such as Parkinson's disease, cerebral palsy, stroke, arthritis, chronic obstructive pulmonary disease and many others. A general review on the subject of gait analysis is [42] and [43] overview all the recent advances of skeleton-based gait recognition. In our review, we provide the information on the use of motion descriptors in 3D gait assessment and recognition tasks.

Descriptors for gait recognition commonly include the biometrics parameters, because intra-person variability is no longer an issue as in the case of action recognition. Motion information is the part of information used to describe a gait pattern. For this reason usually the motion descriptors used for 3D gait recognition are more simple and compact. Despite the fact that RGBD cameras are popular for gait assessment tasks, it is quite common to use 2D projections in order to obtain a gait descriptor. Moreover, a vast majority of modern gait recognition and analysis methods perform a 3D-2D transformation of the depth sequence to form a final gait descriptor [44], [45], [46] or use 2D sensors directly [47]. The most well-known 2D gait descriptor is a Gait Energy Image [48], which is basically the average silhouette over one gait cycle. It was lately upgraded to a Gait Energy Volume (GEV)[49] by using information obtained from 3 Kinect sensors. GEV is derived by averaging all the voxel volumes over a gait cycle.

Similar to action and gesture recognition, 3D gait recognition methods can be categorized as methods based on skeleton joints [50], [51] (model-based) and methods based on depth images [52] (model-free). Skeleton-based methods are similar to analogue methods for action recognition: descriptors are based on the spatial and temporal position of the human skeleton joints or a human body model is used. Depth based gait features work on detailed information about shape and depth variation of a walking individual and do not require a model fitting.

Kwolek [44] propose a view independent motion-based algorithm for gait recognition using a multi-camera setup. They use particle swarm optimization for full-body motion tracking. A 3D human-body model is also proposed in order to improve the results. The final descriptor is a gait signature composed of the dynamic distances between joints projected to a 2D plan and evaluated through time of a single gait cycle. This approach is interesting because it uses the multi-camera setup in order to fit a 3D human model, however, the accuracy of the 3D model is limited due to the use of 2D images without depth information.

Tang et al. [45] introduce a 2.5D voxel gait model that includes only a one-side surface portion of the human body. A 2.5 gait model corresponding to a gait cycle is obtained from several Kinect depth frames. View-invariance is obtained by simply rotating the 2.5 gait model and synthesizing obtained views. The final descriptor is a color 2D image based on a combination of Gaussian and mean curvature [53] of the point cloud data. The method shows good results and avoids the high computational cost of 3D gait modeling, however, 2.5D gait model cannot address the problems of the lack of robustness to covariates such as different appearance due to various clothes etc.

Lim et al. [52] propose real-time model-based gait tracking and analysis method using a depth image sensor installed on a robotic walker. The particle filter is adapted to the depth camera video sequences to obtain the spatio-temporal gait parameters. Segmented leg regions of the point cloud are also tracked using particle filtering improved by implementing a simple harmonic motion model. To simplify the problem, each particle represents the predicted leg model part. Spatio-temporal parameter data can be deduced from the tracked leg pose parameters. This methods proposes a computationally effective gait analysis method suited for clinical gait assessment, however, a specific setup is considered and the segmentation scheme proposed might not work in the case of a person with a movement disorder.

VI. Conclusion

This survey reveals the progress made in the last 6 years in the field of 3D motion descriptors for point cloud data. It is clear that 3D motion descriptors are developing towards more general and efficient descriptors, however, automatic motion analysis and classification remains an open problem. General applicable motion description algorithms are in trend. The latest approaches are compact, transformation invariant to the target object and robust to noise.

According to published results, approaches that model spatial and temporal statistics holistically for point cloud data show less promising results than LST feature points and projection based methods.

The main issue for many methods remains the time to extract the features from point cloud sequences. The features with the best recognition performance are often costly to compute. A detailed comparison of the time complexity of several popular descriptors can be found in [1].

Methods based on joint estimation usually provides compact and meaningful descriptors, and with the advances in skeleton joints recognition have great potential as well. Model-fitting methods remain popular for gait recognition and gesture recognition but for action recognition the main focus has shifted towards model-free methods.

Issues that must be addressed in future work in our opinion are: integration of all the cues for the better performance; computationally less expensive solutions; temporal alignment for the classification stage.

References

[1] E. Ohn-Bar and M. M. Trivedi, "A comparative study of color and depth features for hand gesture recognition in naturalistic driving settings," in *IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 845–850.

[2] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Narf: 3d range image features for object recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, vol. 44, 2010.

[3] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[4] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference-BMVC*, 2008, pp. 275–1.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Proceedings of the 15th international conference on Multimedia*, 2007, pp. 357–360.

[6] H. Zhang, C. Reardon, C. Zhang, and L. E. Parker, "Adaptive human-centered representation for activity recognition of multiple individuals from 3d point cloud sequences," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1991–1998.

[7] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1–4.

[8] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and 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.

[9] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann, "Robust global registration," in *Symposium on geometry processing*, vol. 2, no. 3, 2005, p. 5.

[10] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset," *arXiv preprint arXiv:1407.7390*, 2014.

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

[12] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 9–14.

[13] S. Hadfield, K. Lebeda, and R. Bowden, "Natural action recognition using invariant 3d motion encoding," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 758–771.

[14] H. Zhang and E. Parker Lynne, "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *IEEE Transaction on Circuits Syst Video Technol*, vol. 26, no. 3, pp. 541–555, 2016.

[15] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[17] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2012, pp. 252–259.

[18] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1057–1060.

[19] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data." in *IJCAI*, vol. 1, 2013, p. 3.

[20] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in

IEEE Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 1975–1979.

[21] P. Cirujeda and X. Binefa, "4dcov: a nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences," in *2nd IEEE International Conference on 3D Vision*, vol. 1, 2014, pp. 657–664.

[22] C. Chen, B. Zhang, Z. Hou, J. Jiang, M. Liu, and Y. Yang, "Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features," *Multimedia Tools and Applications*, pp. 1–19, 2016.

[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 53–60.

[24] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 195–206, 2014.

[25] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[26] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[27] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3176–3183.

[28] S. Hadfield and R. Bowden, "Kinecting the dots: Particle based scene flow from depth sensors," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2290–2295.

[29] M. Munaro, S. Michieletto, and E. Menegatti, "An evaluation of 3d motion flow and 3d pose estimation for human action recognition," in *RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras*, 2013.

[30] Y. Kong, B. Satarboroujeni, and Y. Fu, "Hierarchical 3d kernel descriptors for action recognition using depth sequences," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, 2015, pp. 1–6.

[31] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[32] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *Advances in neural information processing systems*, 2009, pp. 135–143.

[33] Y. Xiao, G. Zhao, J. Yuan, and D. Thalmann, "Activity recognition in unconstrained rgb-d video using 3d trajectories," in *ACM SIGGRAPH Asia Autonomous Virtual Humans and Social Robot for Telepresence*, 2014, p. 4.

[34] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[35] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV*. Springer, 2010, pp. 143–156.

[36] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[37] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[38] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping." in *VISAPP (1)*, 2013, pp. 620–625.

[39] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1093–1096.

[40] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in *20th IEEE International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3105–3108.

[41] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[42] M. J. Nordin and A. Saadoon, "A survey of gait recognition based on skeleton mode l for human identification," *Research Journal of Applied Sciences, Engineering and Technology*, 2016.

[43] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[44] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3d gait recognition using spatio-temporal motion descriptors," in *Intelligent Information and Database Systems*. Springer, 2014, pp. 595–604.

[45] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5 d multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, 2014.

[46] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5 d gait biometrics using the depth gradient histogram energy image," in *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 399–403.

[47] M. Alotaibi and A. Mahmood, "Automatic real time gait recognition based on spatiotemporal templates," in *IEEE Systems, Applications and Technology Conference (LISAT)*, 2015, pp. 1–5.

[48] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[49] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.

[50] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using rgb-d camera," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1843–1856, 2014.

[51] M. Milovanovic, M. Minovic, and D. Starcevic, "Walking in colors: human gait recognition using kinect and cbir," *IEEE MultiMedia*, vol. 20, no. 4, pp. 28–36, 2013.

[52] C. D. Lim, C.-Y. Cheng, C.-M. Wang, Y. Chao, and L.-C. Fu, "Depth image based gait tracking and analysis via robotic walker," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 5916–5921.

[53] P. Tosranon, A. Sanpanich, C. Bunluechokchai, and C. Pintavirooj, "Gaussian curvature-based geometric invariance," in *IEEE. 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, vol. 2, 2009, pp. 1124–1127.