ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 1, No.7 , January 2014
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Action recognition system based on human body tracking with depth images

M. Martínez-Zarzuela[1], F.J. Díaz-Pernas, A. Tejeros-de-Pablos[2], D. González-Ortega, M. Antón-Rodríguez

[1] University of Valladolid
Valladolid, Spain
marmar@tel.uva.es

[2] University of Valladolid
Valladolid, Spain
atejpab@ribera.tel.uva.es

Abstract

When tracking a human body, action recognition tasks can be performed to determine what kind of movement the person is performing. Although a lot of implementations have emerged, state-of-the-art technology such as depth cameras and intelligent systems can be used to build a robust system. This paper describes the process of building a system of this type, from the construction of the dataset to obtain the tracked motion information in the front-end, to the pattern classification back-end. The tracking process is performed using the Microsoft(R) Kinect hardware, which allows a reliable way to store the trajectories of subjects. Then, signal processing techniques are applied on these trajectories to build action patterns, which feed a Fuzzy-based Neural Network adapted to this purpose. Two different tests were conducted using the proposed system. Recognition among 5 whole body actions executed by 9 humans achieves 91.1% of success rate, while recognition among 10 actions is done with an accuracy of 81.1%.

Keywords: *Body-tracking, action recognition, Kinect depth sensor, 3D skeleton, joint trajectories.*

1. Introduction

Body tracking and action recognition are study fields that are nowadays being researched in depth, due to their high interest in many applications. Many methods have been proposed whose complexity can significantly depend on the way the scene is acquired. Apart from the techniques that use markers attached to the human body, tracking operations are carried out mainly in two ways, from 2D information or 3D information [1, 2].

On the one hand, 2D body tracking is presented as the classic solution; a region of interest is detected within a 2D image and processed. Because of the use of silhouettes, this method suffers from occlusions [1, 2]. On the other hand, advanced body tracking and pose estimation is currently being carried out by means of 3D cameras, such as binocular and Time-of-Flight (ToF) cameras. Within the ToF field, different techniques are utilized; the most commonly used are the extraction of features from depth images and the processing of a stick figure model (or skeleton) using depth information (Figure 1).

According to [3] the use of 3D information results advantageous over the pure image-based methods. The collected data is more robust to variability issues including vantage point, scale, and illumination changes. Additionally, extracting local spatial features from skeleton sequences provides remarkable advantages: the way an action is recognized by humans can be conveniently represented this way, since the effect of personal features is reduced, isolating the action from the user who performed it.

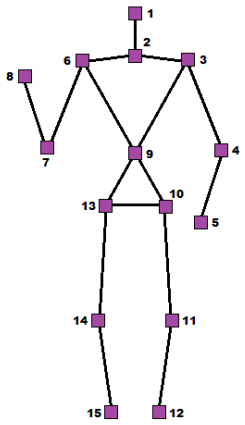

Fig. 1 Fifteen-joints skeleton model

As stated by [4], one of the main issues existing in action recognition is related to the extremely large dimensionality of the data involved in the process (e.g. noises in 2D videos, number of joints of 3D skeletons…).

This derives in problems such as an increase in computational complexity and makes more difficult the extraction of key features of the action. Current methodologies use four stages to solve this: feature extraction, feature refinement to improve their discriminative capability (such as dimension reduction or feature selection), pattern construction and classification. Then, by studying the temporal dynamics of the body performing the movement, we can decode significant information to discriminate actions. To that extent, a system able to track parts of the human body for the whole duration of the action is needed.

However, many action recognition approaches do not utilize 3D information obtained from consumer depth cameras to support their system. The existence of state-of-the-art technology offering positions of the human body in all three dimensions calls for the emergence of methodologies that use it to track this information for a variety of purposes such as action recognition. In addition to this, an algorithm that is accurate and computationally efficient is needed.

This paper is focused on the work field of action recognition using a 3D skeleton model. This skeleton is computed by tracking the joints of the human performing the movement. The emergence of novel consume-depth cameras, such as Kinect, enables convenient and reliable extraction of human skeletons from action videos. The motivation for this paper is to design an accurate and robust action recognition system using a scheme that takes into account the trajectory in 3D of the different parts of the body. As a result, our system is more reliable than those systems in which only thresholds are taken into account when determining if a body part has performed a movement. The movement is classified into a set of actions by means of a neural network adapted for this purpose. To that extent, we track the moving body using the Kinect hardware and OpenNI/NITE software to extract 3D information of the position of the different human joints along time.

2. 3D skeleton-based approaches

2.1 Action recognition system

The usage of depth cameras has allowed many new developments in the field of action recognition. In [5], depth images are processed with a GPU filtering algorithm, introducing a step further in Real Time motion capture. A pose estimation system based on the extraction of depth cues from images captured with a ToF camera is designed in [6]. They manage simultaneous full-body pose tracking and activity recognition. Another tracking method is used in [7], where local shape descriptors allow classifying body parts within a depth image with a human shape. The results of this point detector show that an articulated model can improve its efficiency. In [8], it is presented the construction of a kinematic model of a human for pose estimation. Key-points are detected from ToF depth images and mapped into body joints. The use of constraints and joint retargeting makes it robust to occlusions. In [9], human poses are tracked using depth image sequences with a body model as a reference. To achieve greater accuracy, this method hypothesizes each result using a local optimization method based on dense correspondences and detecting key anatomical landmarks, in parallel. Other approaches to skeleton tracking use multi-view images obtained from various cameras [10, 11].

In [12] challenge of recognizing actions is accounting for the variability that appears when arbitrary cameras capture humans performing actions, taking into account that the human body has 244 degrees of freedom. According to the authors, variability associated with the execution of an action can be closely approximated by a linear combination of action bases in joint spatio-temporal space. Then, a test employing principal angles between subspaces is presented to find the membership of an instance of an action.

Apart from ToF cameras, nowadays there exist several low-cost commodity devices that can capture the user's motion, such as the Nintendo Wiimote, Sony Move, or the Microsoft Kinect camera. This last device allows perceiving the full-body pose for multiple users without having any marker attached to the body. The mechanism used in this camera device is based on depth-sensing information obtained by analyzing infrared structured light, to process a 3D image of the environment. It is capable of capturing depth images with a reasonable precision and resolution (640x480 px 30fps) at a commercial cost.

Around this hardware, a community of developers has emerged and there is a wide variety of tools available to work with Kinect. A commonly used framework for creating Kinect applications is OpenNI [13]. OpenNI software has been developed to be compatible with commodity depth sensors and, in combination with PrimeSense's NITE middleware, is able to automate tasks for user identifying, feature detection, and basic gesture recognition. These applications vary from gaming and rehabilitation interfaces, such as FAAST [14] to

Augmented Reality software. However, the most up-to-date Kinect tools are provided in the Microsoft Kinect SDK [15], released more recently. A previous study to determine the advantages of Kinect SDK and OpenNI was carried out in [16]. It was finally concluded that, despite it lacked some minor joints, OpenNI offers an open multiplatform solution with almost the same functionality for Kinect-based applications.

In [3], a methodology for working with the Microsoft Kinect depth sensor is provided. They explain an efficient way to make use of the information about the human body (i. e. relationship between skeleton joints). To build their system, they assume that the distance function based on the kinetic energy of the motion signals shows enough discriminative power. Nevertheless, it is a nontrivial problem to define geometric (semantic) relationships that are discriminative and robust for human motions.

Taking Kinect skeleton data as inputs, [4] propose an approach to extract the discriminative patterns for efficient human action recognition. 3D Skeletons are converted into histograms, and those are used to summarize the 3D videos into a sequence of key-frames, which are labeled as different patterns. Then, each action sequence is approached as a document of action ”words”, and the text classification concept is used to rank the different heterogeneous human actions, e.g. "walking", "running", "boxing". However, this technique does not take into account most of the spatial structure of the data.

2.2 Datasets

In order to develop a recognition system, it is necessary to have a set of movements to learn and classify. In this section, some of the existing datasets that use depth information are briefly described and a self-made movement database constructed using a consumer depth camera is explained. As described in [17], datasets can be classified according to different criteria such as the complexity of the actions, type of problem or source. Heterogeneous gestures are those who represent natural realistic actions (e.g. jumping, walking, waving…).

According to [2], one of the datasets used to evaluate both 2D and 3D pose estimation and tracking algorithms is the HumanEva database [18]. Their intention is to become a standard dataset for human pose evaluation. Videos are captured using a single setup for synchronized video and ground-truth 3D motion, using high-speed cameras and a marker-based motion acquisition system. HumanEva-I dataset consists of four subjects performing six predefined actions with a certain number of repetitions each. These actions are: walking, jogging (slow running), gesture (“hello” and “goodbye”), throw/catch (different styles), boxing, combo (sequence of walking, jogging and balancing).

Some datasets, recorded using ToF cameras include depth information. The ARTTS 3D-TOF database [19] offers datasets of faces for detection, heads for orientation, and gestures for recognition. The gestures dataset is composed by nine different simple actions, such as push and resize, performed by different users. Other example is the SZU Depth Pedestrian Dataset [20], which contains depth videos of people standing and walking for pedestrian detection. In [6], a database to evaluate a pose estimation system was proposed, since a synchronized dataset of ToF captured images was not available online. This dataset contains ten activities conducted by ten subjects; besides, each of the movements was recorded 6 times: clapping, golfing, hurrah (arms up), jumping jack, knee, bends, picking something up, punching, scratching head, playing the violin and waving.

Ganapathi et al. built a database composed of 28 real-world sequences with annotated ground truth [5]. These streams of depth images vary from short sequences with single-limb motions to long sequences such as fast kicks, swings, self-occlusions and full-body rotations. Actions can have different length, occlusions, speed, active body parts and rotation about the vertical axis. Depth information is captured with a ToF camera, and ground-truth is collected using 3D markers. The scope of this extensive dataset is to evaluate performance on tracking systems.

Apart from using ToF cameras, another way to obtain 3D information is by capturing the video via a commercial Kinect sensor, which offers a better resolution. An RGBD dataset captured with Kinect is provided by [21], containing specific daily life actions such as answering the phone or drinking water. However, to the best of our knowledge a Kinect dataset which contains heterogeneous motions such as running, jumping, etc., cannot be found. Moreover, several other approaches using 3D skeletons, such as some of the aforementioned, are also evaluated with this kind of movements.

With this in mind and taking as a reference a set of simple actions, we built our own corpus to test the proposed action recognition system and compare it with other related approaches. In order to perform complete action recognition, the desired dataset should contain full-body natural motions, not based on interfaces or for specific recognition tasks (dance moves…), performed by

different people, as in [18]. It consists of ten independent actions represented by nine different users. As it is shown in Figure 2, the actions are: bending down, jumping jack, jumping, jumping in place, running, galloping sideways, hop-skip, walking, waving one hand and waving both hands. This database will be further extended to include more users and action types. The scene is composed of a cluttered background inside an office, although this scenario does not need to be maintained.

Fig. 2 Example of different movements and users

3. System description

After presenting state-of-the-art 3D-based recognition systems and datasets, this section describes our approach derived from this study. Figure 3 depicts the organization of the proposed system.

In our approach for action recognition, an OpenNI/NITE wrapper is utilized to access the kinematic skeleton. OpenNI is an API that can be used to interact with the information provided by the Kinect technology. One of its main features includes body tracking of one or more users in front of the camera, which involves detecting a body model and offering the coordinates of some of its joints. OpenNI, in conjunction with NITE [22], supports the detection and tracking of a kinematic skeleton. As shown in Figure 1, in the case of Kinect this skeleton consists of fifteen joints: head, neck, left shoulder, left elbow, left hand, right shoulder, right elbow, right hand, torso, left hip, left knee, left foot, right hip, right knee and right foot.

As suggested in [3, 23], instead of using edges or regions within the images for recognition, a coordinate system of some parts of the human body is utilized. The coordinates represent the position of the points that match the main joints of the human model. With these points, a stick figure is drawn on the original image, overlaid in the silhouette of the person.

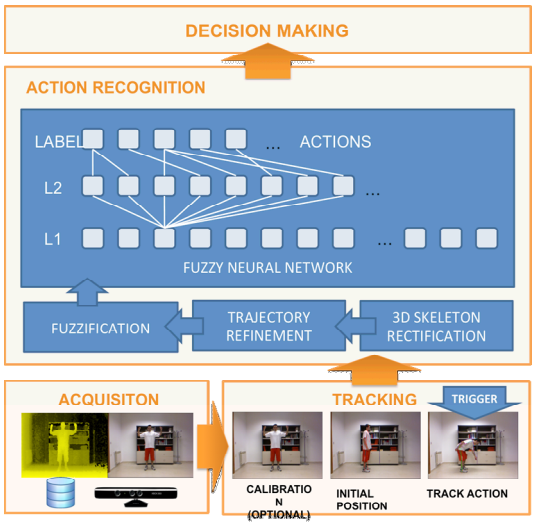

Fig. 3 An action video is acquired live using a Kinect camera or via the action dataset, and the human body is tracked from the beginning of the movement to the end. The obtained 3D skeleton is processed to build motion patterns which are introduced to the action classifier in order to determine which action is being performed.

3.1 Skeleton pre-processing

Once body information is obtained, it is necessary to break it down into motion data in order to represent an action. Aggarwal et al. [24] describe a series of steps to model an action recognition system. To represent the trajectory of a joint, the coordinate system takes the center of the image as the origin and then positive and negative values are assigned to x and y coordinates. Coordinate z is represented with a positive value, taking the position of the camera as the origin. The value of these measures is offered in millimeters.

Once the movement starts, the system stores the body coordinates. In each frame, the system takes from the coordinate matrix extracted from OpenNI the x-y-z coordinates of each joint. This coordinate information is centered with respect to the position of the torso in the first frame of the movement. This step allows reducing variability with respect to the initial position and redundancy of the coordinate values.

3.2 Pattern Generation

Once an action is tracked, the segmented sequence is represented into a motion matrix containing the x-y-z normalized coordinates of the whole movement along

time, $\vec{p}_j(t) = (x_j, y_j, z_j), j \in [1,15]$. These motions signals are rectified with respect to the position of the user's torso in the first frame, since we consider that the action can be executed anywhere in the scene. Then, this tracking block is processed into an action pattern with the purpose to feed an intelligent system.

After studying several motion representation schemes [25], it was inferred that using the coordinate points directly in a time sequence results in enormous vectors when modeling a motion pattern. In order to transform the raw trajectory information into a manageable representation, different compression techniques were analyzed [26]. The way we reduce dimensionality of joint trajectories is by applying the Fourier transform. As stated in [27], this technique is used in many fields, and allows representing time-varying signals of different duration. Also, by keeping only the lowest components, the data can be represented without losing highly relevant information, smoothing the original signal. This also eliminates possible existing noise. It is also indicated in [27] that selecting the first five components of the Fourier transform, the main information is always preserved.

Unlike in the aforementioned paper, our system not only follows the trajectory of a single point, but of a whole skeleton composed by fifteen joints. When building the motion pattern, the FFT trajectories of each point, $\widetilde{P}_j = FFT(\vec{p}_j(t)) = (\widetilde{X}_j, \widetilde{Y}_j, \widetilde{Z}_j)$, are assembled together to compose a single vector. In order to reduce even more the dimension of the final motion pattern, only the main joints are introduced into the final system. It was determined that the joints to select were head and limbs, discarding neck, shoulders, hips and torso. This is because the range of movements of these joints is quite limited, and also we are subtracting the initial torso position to represent a relative motion.

Small values for the FFT may add non-relevant information whereas greater values may eliminate important information within those first components. A sub-vector of nine elements is constructed using the real and imaginary parts of the first five components of the FFT. The motion pattern is then represented by assembling the Fourier sub-vectors of each coordinate of each point, that is 9 FFT elements (0…5) x 3 coordinates (x,y,z) x 9 joints (k) = 243, as depicted in Eq. (1).

$$P = \bigcup_{k=1}^{9} \begin{pmatrix} \widetilde{X}_{k_0}, \mathrm{Re}\{\widetilde{X}_{k_1}\}, \mathrm{Im}\{\widetilde{X}_{k_1}\}, \ldots, \mathrm{Im}\{\widetilde{X}_{k_4}\}, \\ \widetilde{Y}_{k_0}, \mathrm{Re}\{\widetilde{Y}_{k_1}\}, \mathrm{Im}\{\widetilde{Y}_{k_1}\}, \ldots, \mathrm{Im}\{\widetilde{Y}_{k_4}\}, \\ \widetilde{Z}_{k_0}, \mathrm{Re}\{\widetilde{Z}_{k_1}\}, \mathrm{Im}\{\widetilde{Z}_{k_1}\}, \ldots, \mathrm{Im}\{\widetilde{Z}_{k_4}\} \end{pmatrix} \quad (1)$$

3.3 Action recognition

The core of the system is in charge of carrying out the action recognition tasks. Henceforth, we will use the term class to represent each kind of movement. Therefore, input patterns of the same class represent the same actions performed by various users or various repetitions of the same user.

The proposed system uses a neural network capable of learning and recognizing action patterns *P*. This neural network is based on the ART (Adaptive Resonance Theory) proposed by Grossberg et al. [28], thus includes interesting characteristics like fuzzy computation and incremental learning. Due to the nature of Fuzzy logic, action patterns *P* need to be normalized altogether, adjusting their values to the range of [0, 1], before they can be introduced to the neural network for learning or testing. Moreover, patterns are complement-coded to enter the neural network. As a consequence, the action recognition network works with very large information patterns of 486 elements.

The neural network is comprised of an input layer L1, an output layer L2 and a labeling stage. L1 and L2 are connected through a filter of adaptive weights w_j. Similarly, L2 is linked to the labelling stage through L_{jl} weights. Before neural network training takes place, weights w_j are initialized to 1 and L_{jl} to zero.

The action patterns *P* activate L1 using complement coding $I = (P, 1\text{-}P)$. Each input *I* activates every node *j* in L2 with an intensity T_j, through adaptive weights w_j (j=0,…, N). Activation is computed using Eq. (2), where |.| is the L1 norm of the vector, ^ is the fuzzy AND operator ($(p \wedge q)_i = \min(p_i, q_i)$).

$$T_j = \frac{|I \wedge w_j|}{(0.05 + |w_j|)} \quad (2)$$

In contrast to the ARTMAP neural network [28], a fast winner tracking node mechanism is included, so that only those committed nodes in L2 that are linked to supervision label l would participate in the competition process. This way, a mismatch between the supervision label and the label of the label of winning node is avoided.

Also, this mechanism favors a parallel implementation of the system and speeds-up the training process.

In layer L2 takes place a winner-take-all competition, and a winner node J so that T_J=max(T_j) is selected. Then, the resemblance between the action pattern, I_I, and the adaptive winner node, w_J is measured using a match function in Eq. (3)

$$\frac{|I_I \wedge w_J|}{|I_I|} \geq threshold \qquad (3)$$

In case this criterion is not satisfied, a new node is promoted in L2 and linked through the adaptive weights L_{jl}, to the supervision label l. In case the criterion is satisfied, the network enters in resonance and the input vector, I_I, is learnt according to Eq. (4).

$$w_J = \frac{1}{2}[w_J + (I_I \wedge w_J)] \qquad (4)$$

During classification, weights do not change and the winning node in layer L2 will determine the predicted label, corresponding to weight L_{jl} =1.

4. Evaluation

Two tests were carried out against the prototype of the action recognition system. For each, different movements from the dataset were used; the first one, called Basic, involved the simplest movements (bending down, jumping, jumping in place, walking and waving one hand). The evaluation was carried out with the cross-correlation leave-one-out technique. This involves nine experiments, each one consisting in learning the actions of eight users and performing the test with the remaining subject.

Figure 4 shows the accuracy rates of the set of tests with the Basic dataset, resulting quite satisfactory. Movements corresponding to jumping and walking are slightly confused. The recognition system achieves a classification success of 91.1%. Table 1 shows the corresponding confusion matrix, which shows the specific mistakes made by the system.

Once the results of this test were obtained, the inclusion of new movements in the dataset was determined in order to carry out further experiments. The second set of tests was carried out including videos for jumping jack, running, galloping sideways, hop-skip and waving both hands.

One of the advantages of our system is that it can be trained to learn new actions and the neural network does not forget previous trained examples. For this test for example, it was only necessary to teach the system to include the new movements that appear in the larger set. Figure 5 depicts the accuracy rates for this case.

Fig. 4 Accuracy rates of the Basics test

Table 1: Confusion matrix of the Basic test

Basic	*Bend*	*Pjump*	*Jump*	*Walk*	*Wave1*
Bend	9				
Pjump		7		2	
Jump			9		
Walk		2		7	
Wave1					9

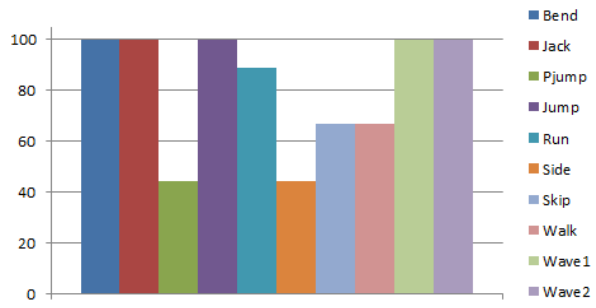

Fig. 5 Accuracy rates of the Complete test

Table 2 presents the confusion matrix for the more complex test. Although movements such as jumping jack and waving both hands are successfully introduced into the system, jumping forward and galloping sideways are misclassified into the running category. Skipping and walking actions show also low accuracy in their recognition results. The average success rate in this case is 81.1%. These results allow drawing some important conclusions about the proposed system.

Table 2: Confusion matrix of the Complete test

Complete	*Bend*	*Jack*	*Pjump*	*Jump*	*Run*	*Side*	*Skip*	*Walk*	*Wave1*	*Wave2*
Bend	9									
Jack		9								
Pjump			4		3			2		
Jump				9						
Run					8	1				
Side			1		4	4				
Skip			2		1		6			
Walk			1		1		1	6		
Wave1									9	
Wave2										9

4.1 Discussion

The system is robust recognizing actions that do not involve very fast transitions and those that involve different body parts. However, the recognition rate is decreased when including more actions that do not follow these premises.

By understanding how the whole system works, it can be determined the reason of these results. When two joints follow similar trajectories in the 3D coordinate system, the components in the action pattern are closer and the action can be misclassified. This is the reason of the errors that occur among movements such as walk, run, skip and jump, when the tracked points share similar paths. It is also necessary to take into account that truncating the trajectory signal into the Fourier dimension eliminates part of the information, smoothing the differences between similar signals.

When compared to other approaches that use joint trajectories, the results are analogous. In [4] the accuracy ratio is similar, and their confusion matrices also show misclassification in jumping-waling actions. Authors in [12] also state the difficulty of distinguishing walking and running tracked actions using their approach, lowering the hit rate.

An advantage of our system against other tracking and recognition approaches is that the original motion signals are recoverable, since the preprocessing of the motion patterns can always be inverted. This system does not need the actions to follow any specific rhythm either.

5. Conclusions

This study presented a body tracking and action recognition system for heterogeneous actions using consumer depth cameras, i.e. Kinect. To test the system, also an action dataset was created from scratch. Since our system is not only designed to recognize gestures but to track general actions, the actions that are contained in the dataset represent whole body movements. The results obtained show that the proposed system is capable of performing action recognition for heterogeneous movements with a high success rate. When the base set of five actions is used, a mean accuracy of 91.1% is achieved. By adding another group of similar but more complex actions, the mean hit rate is 81.1%.

Our system is able to track a human body by means of a commodity depth camera, and software that provides a set of trajectories. Taking these trajectories as features to represent an action has been proved to be robust and reduce the variability that different actors and repetitions may introduce into the motion signals. Then, the intelligent network used to classify the actions is capable of incremental learning, so new actions can be included into the system once deployed. It is also remarkable that the processing applied to the motion signals is reversible, so the original movement can be recovered.

A possible future work would include the extension of the dataset, with more varied actions and poses with trajectories in the third dimension. We will consider the possibility of opening the dataset for collaboration, so that new videos could be submitted using an online webpage. Also, since the system has been designed as a highly parallelizable algorithm, GPU parallel coding techniques

may be speed-up computational expensive operations, such as pattern processing and neural network recognition.

References

[1] Poppe R (2007) Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108: 4–18.
[2] Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding 115: 224–241.
[3] Raptis H.H.M, Kirovski D (2011) Real-Time Classification of Dance Gestures from Skeleton Animation. Proceedings of the 2011 ACM SIGGRAPH, ACM Press pp. 147–156.
[4] Chen H.L.F, Kotani K (2012) Extraction of Discriminative Patterns from Skeleton Sequences for Human Action Recognition. RIVF International Conference, IEEE 2012 pp. 1–6.
[5] Ganapathi V, Plagemann C, Koller D, Thrun S (2010) Real Time Motion Capture Using a Single Time-Of-Flight Camera. Computer Vision and Pattern Recognition pp. 755-762.
[6] Schwarz L.A, Mateus D, Castañeda V, Navab N (2010) Manifold Learning for ToF-based Human Body Tracking and Activity Recognition. Proceedings of the British Machine Vision Conference 80: 1-11.
[7] Plagemann C, Ganapathi V, Koller D, Thrun S (2010) Real-time identification and Localization of Body Parts from Depth Images. IEEE International Conference on Robotics and Automation pp. 3108-3113.
[8] Zhu Y, Dariush B, Fujimura K. Kinematic self retargeting: A framework for human pose estimation. Computer Vision and Image Understanding 114: 1362-1375.
[9] Zhu Y, Fujimura K (2010) A Bayesian Framework for Human Body Pose Tracking from Depth Image Sequences. Sensors 10: 5280-5293.
[10] Yániz C, Rocha J, Perales F (1998) 3D Part Recognition Method for Human Motion Analysis. Proceedings of the International Workshop on Modelling and Motion Capture Techniques for Virtual Environments.
[11] Chen D, Chou P, Fookes C, Sridharan S (2008) Multi-View Human Pose Estimation using Modified Five-point Skeleton Model.
[12] Sheikh M.S.Y, Sheikh M (2005) Exploring the Space of a Human Action. International Conference on Computer Vision 2005, IEEE 1: 144–149.
[13] OpenNI, Plataform to promote interoperability among devices, applications and Natural Interaction (NI) middleware. Available: http://www.openni.org. Accessed 2013 Apr 21.
[14] Suma E.A, Lange B, Rizzo A, Krum D.M, Bolas M (2011) FAAST: The Flexible Action and Articulated Skeleton Toolkit. Virtual Reality Conference pp. 247-248.
[15] Microsoft Kinect SDK, Kinect for Windows SDK. Available: http://www.microsoft.com/en-us/kinectforwindows/. Accessed 2013 Apr 21.
[16] Martínez-Zarzuela M, Díaz-Pernas F.J, Tejero de Pablos A, Perozo-Rondón F, Antón-Rodríguez M, González-Ortega D (2011) Monitorización del cuerpo humano en 3D mediante tecnología Kinect. SAAEI XVIII Edition pp. 747-752.
[17] Chaquet J.M, Carmona E.J, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding 117(6): 633-659.
[18] Sigal L, Balan A.O, Black M.J. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision 87(1-2): 4-27.
[19] ARTTS 3D-TOF Database. Available: http://www.artts.eu/3d_tof_db.html. Accessed 2013 Apr 21.
[20] Shenzhen University (SZU) Depth Pedestrian Dataset. Available: http://yushiqi.cn/research/depthdataset. Accessed 2013 Apr 21.
[21] Ni B, Wang G, Moulin P (2011) RGBD-HuDaAct: A color-depth video database for human daily activity recognition. International Conference on Computer Vision Workshops (ICCV Workshops), 2011 IEEE pp. 1147-1153.
[22] NITE, Algorithmic infrastructure for user identification, feature detection and gesture recognition. Available: http://www.primesense.com/solutions/nite-middleware/. Accessed 2013 Apr 21.
[23] Guo Y, Xu G, Tsuji S (1994) Understanding Human Motion Patterns. Proceedings of International Conference on Pattern Recognition Track B: 325-329.
[24] Aggarwal J.K, Park S. Human Motion: Modeling and Recognition of Actions and Interactions. 3D Data Processing, Visualization and Transmission pp. 640-647.
[25] Kulic D, Takano W, Nakamura Y (2007) Towards Lifelong Learning and Organization of Whole Body Motion Patterns. International Symposium of Robotics Research pp. 113-124.
[26] Gu Q, Peng J, Deng Z (2009) Compression of Human Motion Capture Data Using Motion Pattern Indexing. Computer Graphics Forum 28(1): 1-12.
[27] Naftel A, Khalid S (2006) Motion Trajectory Learning in the DFT-Coefficient Feature Space. IEEE International Conference on Computer Vision Systems pp. 47-54.
[28] Carpenter G, Grossberg S, Markuzon N, Reynolds J, Rosen D (1992) Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks 3(5): 698-713.

Mario Martínez Zarzuela was born in Valladolid, Spain; in1979. He received the M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2004 and 2009, respectively. Since 2005 he has been an assistant professor in the School of Telecommunication Engineering and a researcher in the Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. His research interests include parallel processing on GPUs, computer vision, artificial intelligence, augmented and virtual reality and natural human-computer interfaces.

Francisco Javier Díaz Pernas was born in Burgos, Spain, in1962. He received the Ph.D. degree in industrial engineering from Valladolid

University, Valladolid, Spain, in 1993. From 1988 to 1995, he joined the Department of System Engineering and Automatics, Valladolid University, Spain, where he has worked in artificial vision systems for industry applications as quality control for manufacturing. Since1996, he has been a professor in the School of Telecommunication Engineering and a Senior Researcher in Imaging & Telematics Group of the Department of Signal Theory, Communications, and Telematics Engineering. His main research interests are applications on the Web, intelligent transportation system, and neural networks for artificial vision.

Antonio Tejero de Pablos was born in Valladolid, Spain, in 1987. He received his M.S. in Telecommunication Engineering and M.S. in ICT Research from the University of Valladolid in 2012 and 2013 respectively. During his academic career he has collaborated on various projects with the Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. Now he works as a trainee researcher at NTT Communications Science Lab, in Japan. His research fields of interest are Action Tracking & Recognition, Augmented Reality based interfaces and General Purpose GPU Programming.

David González Ortega was born in Burgos, Spain, in 1972. He received his M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2002 and 2009, respectively. Since 2003 he has been a researcher in the Imaging and Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. Since 2005, he has been an assistant professor in the School of Telecommunication Engineering, University of Valladolid. His research interests include computer vision, image analysis, pattern recognition, neural networks and real-time applications.

Míriam Antón Rodríguez was born in Zamora, Spain, in 1976. She received her M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2003 and 2008, respectively. Since 2004, she is an assistant professor in the School of Telecommunication Engineering and a researcher in the Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. Her teaching and research interests include applications on the Web and mobile apps, bio-inspired algorithms for data mining, and neural networks for artificial vision.