applied sciences

MDPI

Article

Discrete HMM for Visualizing Domiciliary Human Activity Perception and Comprehension

Ta-Wen Kuan [1,*], Shih-Pang Tseng [2], Che-Wen Chen [3], Jhing-Fa Wang [1,3] and Chieh-An Sun [3]

[1] School of Information Science and Technology, Sanda University, Shanghai 201209, China; jameswangjf@gmail.com
[2] Software and Big Data School, Changzhou College of Information Technology, Changzhou 213164, China; tsengshihpang@ccit.js.cn
[3] Department of Electrical Engineering, National Cheng-Kung University, Tainan 701401, Taiwan; kfcmax300@gmail.com (C.-W.C.); jayan0616@gmail.com (C.-A.S.)
* Correspondence: dwguan@sandau.edu.cn

Abstract: Advances in artificial intelligence-based autonomous applications have led to the advent of domestic robots for smart elderly care; the preliminary critical step for such robots involves increasing the comprehension of robotic visualizing of human activity recognition. In this paper, discrete hidden Markov models (D-HMMs) are used to investigate human activity recognition. Eleven daily home activities are recorded using a video camera with an RGB-D sensor to collect a dataset composed of 25 skeleton joints in a frame, wherein only 10 skeleton joints are utilized to efficiently perform human activity recognition. Features of the chosen ten skeleton joints are sequentially extracted in terms of pose sequences for a specific human activity, and then, processed through coordination transformation and vectorization into a codebook prior to the D-HMM for estimating the maximal posterior probability to predict the target. In the experiments, the confusion matrix is evaluated based on eleven human activities; furthermore, the extension criterion of the confusion matrix is also examined to verify the robustness of the proposed work. The novelty indicated D-HMM theory is not only promising in terms of speech signal processing but also is applicable to visual signal processing and applications.

Keywords: discrete HMM; human activity comprehension; pose recognition; confusion matrix; autonomous AI

Citation: Kuan, T.-W.; Tseng, S.-P.; Chen, C.-W.; Wang, J.-F.; Sun, C.-A. Discrete HMM for Visualizing Domiciliary Human Activity Perception and Comprehension. *Appl. Sci.* **2022**, *12*, 3070. https://doi.org/10.3390/app12063070

Academic Editor: Antonio Fernández-Caballero

Received: 12 February 2022
Accepted: 15 March 2022
Published: 17 March 2022

1. Introduction

The trend of aging societies has attracted much attention globally, as an increasing elderly population is accompanied with gradually decreasing long-term care and nursing manpower. Advances in autonomous AI have led to the feasibility of developing AI robots for elderly care. To investigate AI-based autonomous robots for elder care, visual human activity recognition by AI machines plays a preliminary role for perceiving a user's mental and physical states as well as their daily life activities, etc. For example, elderly people without mobility may use summoning gestures to summon a robot for a drink of water, follow-me gestures to monitor their activities, or stop gestures to control a robot's movements. To reach such a goal, this work proposed discrete hidden Markov models (D-HMMs) for visual human activity recognition of eleven daily home activities, using a binocular camera with a Kinect v2 RGB-D sensor to extract embedded skeleton information and to further recognize human sequential poses to understand the implications for specific human activities.

Human activity recognition [1] has generally been categorized into two approaches, i.e., vision based [2–4] and sensor based [5–7]. The vision-based approach to human activity recognition utilizes computer vision methodology to reach a compromise between privacy consideration and light dependency and to report initial approaches for human activity

recognition [5]. The sensor-based approach to human activity recognition has also been categorized into optical and non-optical methods [8,9]. The optical method applies a binocular camera with an RGB-D sensor with active or passive markers to sense a person, whereas the non-optical method uses inertial measurement units (IMUs) or a magnetic system, for more complex environments [10]. For feature extraction, frame-based vision systems process frames or sequences using visual sensors in two individual steps, where the outcome is highly dependent on the quality of the captured frame [8]. Conventional computer vision algorithms process each informational frame individually, regardless of the noisy frame chip for performance efficiency, merely keeping the change rate in the scene similar to the frame rate, wherein practical applications are sometimes without any movement or in moments of fast or noisy motions. In addition, [11] reported that the performance of human activity recognition was also affected by several sources of variation, including, perspective, anthropometry, execution rate, personal style, etc.

HMMs are numerical methods that focus on representing and learning the sequential and temporal characteristics in activity sequences for human activity recognition. Such dynamic models provide simple and efficient models for learning; however, performance decreases if activities become complex for the Markov assumption [12–15]. To overcome this shortcoming, explicit duration HMMs [16], segmental HMMs [17], and layered HMMs [18,19] have been investigated to extend HMMs and to enhance the capability of distinguishing among classes.

In this work, discrete HMMs [20–22] are used for human activity recognition of eleven proposed daily activities in a home scenario. An RGB-D camera is used to capture a total of 25 skeletal joints along the entire human body skeleton, in which a set of feature points can be extracted for classification of the different poses; however, in this work, only ten skeleton joints are required for pose comprehension. Secondly, the relative positions of the ten selected skeletal joints are transformed from a Cartesian coordinate system into a spherical coordinate system for robustness on position invariance. Thirdly, the features are quantized into vectors, such that the sequential observations can be treated as a type of pose in terms of an observation sequence. Eventually, human activity recognition is performed through a discrete hidden Markov model (D-HMM) to build the codebook of trained pose sequences and to score the maximal posterior probability for target prediction. The confusion matrix with extension criteria is also examined in detailed experiments, to understand the misclassified rates for eleven poses in terms of the true classes with the corresponding predicted classes. The D-HMM theory seems promising for speech signal processing, on which D-HMM herein is successfully applicable on visual comprehension.

The remaining sections of this paper are organized as follows: in Section 2, we describe our proposed method for a human activity comprehension framework, including feature extraction, pose sequences for human activity presentation, and noise elimination; in Section 3, we explain the experiments and results in terms of a confusion matrix with corresponding derivation for detailed results; finally, in Section 4, we discuss conclusions and future work.

2. Human Activity Comprehension Framework

2.1. Overview

For human activity recognition, we use an RGB-D sensor of 3D skeleton joints [23–27] to construct a database in terms of skeletonized human activities that occur in daily life. The proposed framework, namely human activity comprehension (HAC), is elucidated as follows: First, the features are extracted from ten selected skeleton joints of entire user poses using the RGB-D sensor. After preprocessing of key element skeleton joints, the vector quantization encoding procedure extracts the features as the codebook for the HMMs, of which the hidden Markov model of unknown sensing human behavior is scored to target the specific human activity through the maximum posterior probability. The details are described in the following sections.

2.2. Human Activity Comprehension Module

For an autonomous AI home robot, visual comprehension of user activities is a preliminary step to prepare for interacting with the user. Therefore, in the proposed work, the robot vision automatically comprehends the user's activity through a prior training set of known poses. There are two advantages to this approach, which include letting the HAC module acquire a more compact representation of the sequences, and secondly, overcoming the challenges of speed variations associated with different user activities.

To reach the goal of the HAC module, four factors are carefully inspected in this work. First, the human profile images from the depth of the RGB-D sensor yield information from a total of 25 skeleton joints, wherein the joints represent different poses in the frames; however, our findings show that the information from only 10 skeleton joints is adequate for pose detection, since not all joints are equally informative due to the intrinsic noise of the sensor and the peculiarities of the human body. Secondly, the relative positions among skeletal points are transformed from a Cartesian coordinate system into a spherical coordinate system. The relative positions are extracted by the differences between positions of joints, wherein the differences should not be influenced by the position and orientation of the sensor or height/weight variations in user images in the spherical system. Thirdly, the features are quantized into vectors, since the feature set in each frame can be regarded as an observation symbol, such that the sequential observation from each frame in a video can be treated as a type of pose in terms of an observation sequence, and then, saved as the specific pose in the codebook. Finally, human activity recognition is performed through discrete hidden Markov models (D-HMMs) built from the sequences of training poses. A discrete HMM is trained for each activity on the poses for extracting as the observed symbols. Next, each activity sequence is encoded as a sequence of pose vocabularies for learning the HMMs, and then, the input testing sequence is estimated by scoring the models. Figure 1 shows the proposed D-HMM-based HAC framework.

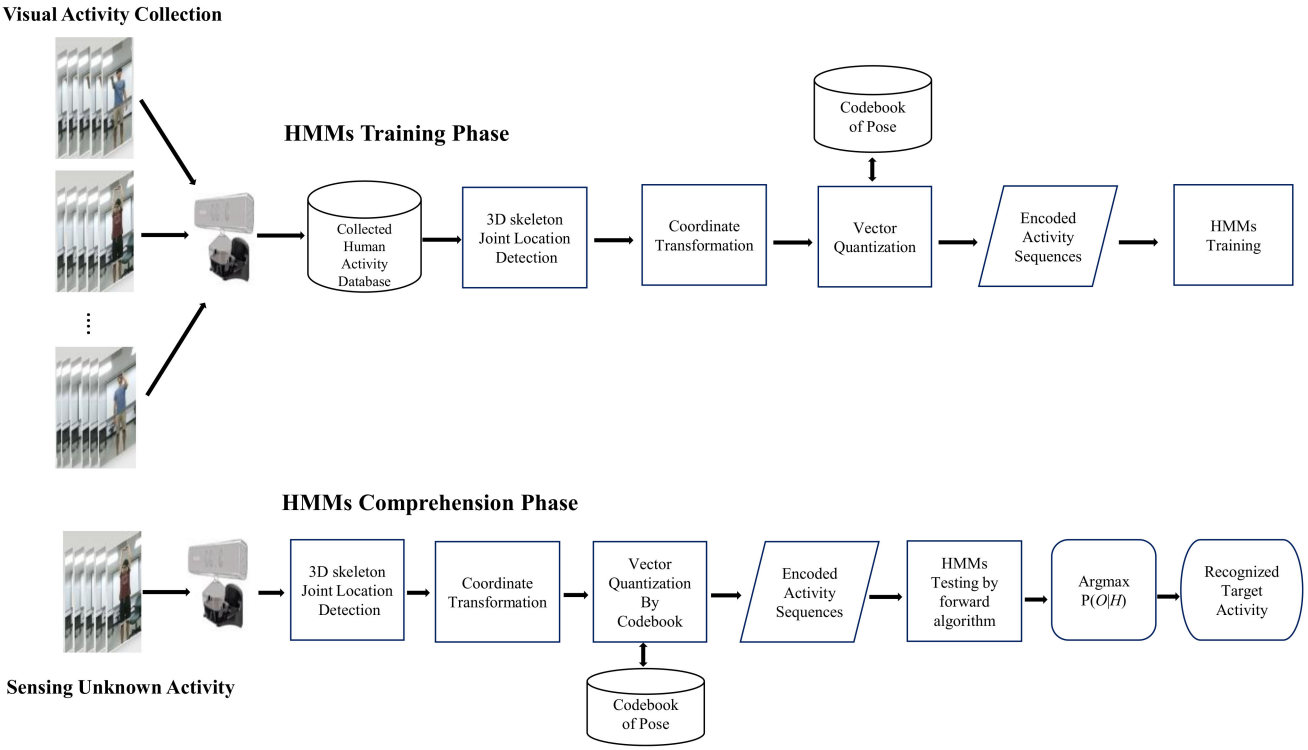

Figure 1. The proposed discrete HMM human activity comprehension framework.

2.3. Feature Extraction

The RGB-D sensor has built-in functions for 3D positions of 25 skeleton joints, including head, neck, spine-shoulder, shoulder, spine-mid, elbow, spine-base, wrist, hand, hand tip, thumb, hip, knee, ankle, and foot for the human profile shown in Figure 2. Assume the real-world 3D positions of each skeleton joint, i, can be represented by the camera reference system for each frame, t, of a sequence shown in Equation (1) as:

$$j_i(t) = (x_i(t), y_i(t), z_i(t)) \quad (1)$$

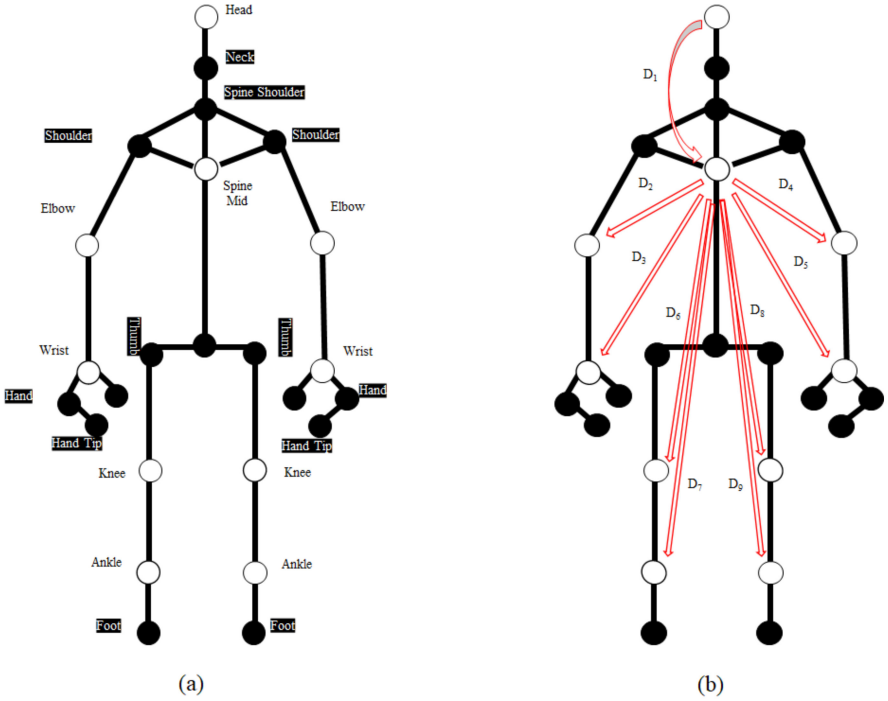

Figure 2. Twenty-five defaulted joints of black-or-white circle points are provided by a third-party sensor in (**a**), Only ten white circle points were selected for the relative position connection and distance calculation by nine red lines shown as D1, D2 … … D9, from centroid (Spine Mid) to the other 9 joints in (**b**), The remaining 15 unselected points labeled on the black circles points show no affection during activity detection in (**b**).

The original point ($x_i = 0$, $y_i = 0$, and $z_i = 0$), in three dimensions, is located at the center of the binocular camera, wherein x_i is pointed toward the sensor's left, y_i is pointed up, and z_i is pointed in front of the sensor. Let N_i be the number of skeleton joints, which consists of the posture of the skeleton at frame t in $3N_i$ dimension, as represented in Equation (2):

$$u(t) = [j_1(t)j_2(t)j_3(t) \ldots\ldots j_{Nj}(t)] \quad (2)$$

For an activity sequence composed of N_f frames, N_f feature vectors are extracted that can be built as a feature matrix for the whole sequences as Equation (3):

$$A^{mat} = \left[u(1), u(2), \ldots\ldots, u\left(N_f\right)\right] \quad (3)$$

Then, this matrix is represented as the variations of the joint positions over time. Each size of the feature matrix is defined as $3A_i^{mat} \times A_f^{mat}$.

To effectively apply the joint information for pose detection, our findings show that only 10 of the 25 joints are sufficiently beneficial for the proposed work, owing to the intrinsic noise of the sensor and the peculiarities of the human body, particularly, the

remaining unselected joints which are trivial, reflecting the dynamic variation of the activity. For example, the poses of some trivial joints are redundant (e.g., wrists and ankles) due to adjacent joints (e.g., hands and feet), in addition, some joints are completely irrelevant for human activity recognition (i.e., neck and hip). Therefore, the remaining set of joints are chosen for human activity recognition and categorized as: head, spine-mid, elbows (left and right), hands (left and right), knees (left and right), and ankles (left and right). The set J is defined as Equation (4):

$$J = \{j_i | i = 1, \ldots\ldots\ldots, 10\} \tag{4}$$

Relative joint position is used because it is an intuitive and efficient way to represent human motions. Consider, for example, the action "victory gesture", which can be interpreted as "arms raised up over the shoulder" and is significantly useful to be characterized by the relative positions. However, several essential factors are carefully considered when representing the local position for each joint, owing to the fact that the feature descriptor should be invariant in terms of the position, the orientation of the sensor, and the variance in height and weight among different people. Raptis et al. [28] showed that using relative positions between joints rather than using absolute positions originated at the sensor were less dependent on perspectives.

For each joint i in the set, the relative positions are extracted by the difference between the positions of joint i and spine-mid (herein treated as the center point of ten selected skeleton joints) as Equations (5) and (6):

$$D_k = j_1 - j_i \tag{5}$$

$$D_k(t) = (x'_k(t), y'_k(t), z'_k(t),) \tag{6}$$

where $i \in J$: $i \neq 1$, and each relative position D_k implied the 3D Cartesian coordinate system at t-th as Equation (7):

$$D = \{D_k | k = 1, \ldots\ldots\ldots 9\} \tag{7}$$

where k is the index of the relative positions.

Normalization on the detected features is essential for scale-invariant consideration, for example, the distance and the target rotation between the user and the sensor would be variant for feature extraction. To overcome such a phenomenon, the transformation from the Cartesian coordination system into the Spherical system is explored. As shown in Figure 3, we convert the relative position D_k from Cartesian (x, y, z) into Spherical (r, θ, φ) 3D coordination. The polar angle θ is between the zenith direction and D_k (the differences between Spine-Mid point and the other points). The azimuthal angle φ is the signed angle measured from the azimuth reference direction to the orthogonal projection of D_k on the reference plane. Note that γ is the length of D_k, which was not considered in the proposed work; for the difference between user and sensor, γ has been normalized. Therefore, we use the spherical representation $f_k(t)$ treated as the relative position, where the detailed function is shown in Equations (8)–(10):

$$f_k(t) = (\theta_k(t), \varphi(t)) \tag{8}$$

$$\theta_k(t) = arccos\left(\frac{z'_k(t)}{\sqrt{(x'_k(t))^2 + (y'_k(t))^2 + (z'_k(t))^2}}\right) \tag{9}$$

$$\varphi(t) = arctan\left(\frac{y'_k(t)}{x'_k(t)}\right) \tag{10}$$

where k indicates the kth relative position and t is the tth frame. The feature set of a skeleton on k relative positions in a frame is defined as Equation (11):

$$F = \{f_k | k = 1, \ldots\ldots\ldots, k\} \tag{11}$$

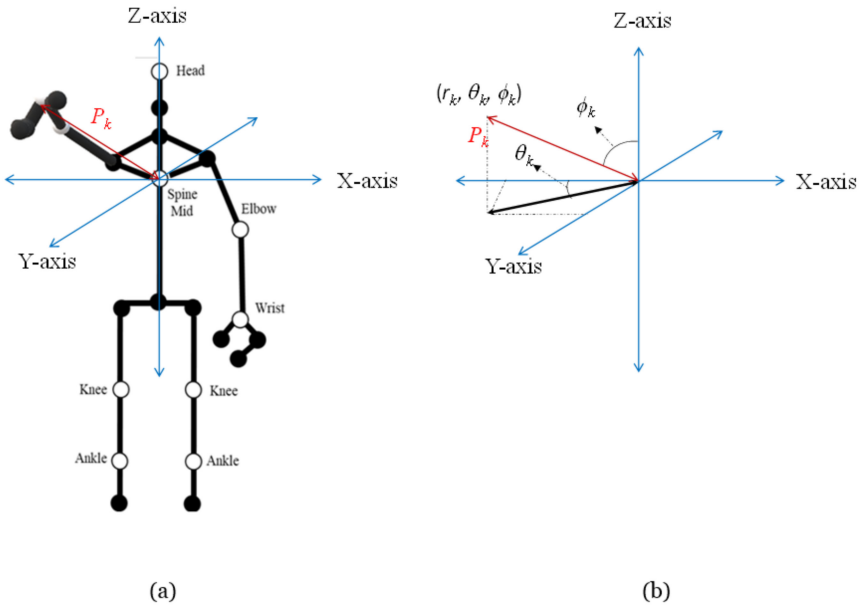

Figure 3. (**a**) Example of the relative position of the skeletonized summon pose in the Cartesian coordinate system; (**b**) transformation from the Cartesian into the Spherical coordinate system with corresponding azimuthal angle θ_k and polar angle φ_k.

2.4. Pose Sequences for Human Activity

To estimate a specific human activity from the sequential frames in a video clip, the feature set in each frame is treated as an observation symbol, such that an identical activity can be regarded as the composition of the sequential symbols, herein named pose sequences; for example, the summoning pose activity is interpreted by five sequential poses, as shown in Figure 4. However, the raw video contains numerous frames in which some noisy clips might influence the corrective pose predictions, and thus, lead to a worse performance. To overcome such a case, the features are clustered into sets in terms of pose classification through vector quantization in order to downsize the number of poses, and accordingly, given the feature set of each frame (F_1, F_2, F_3, ... F_N), vector quantization is applied to separate N observations into k sets.

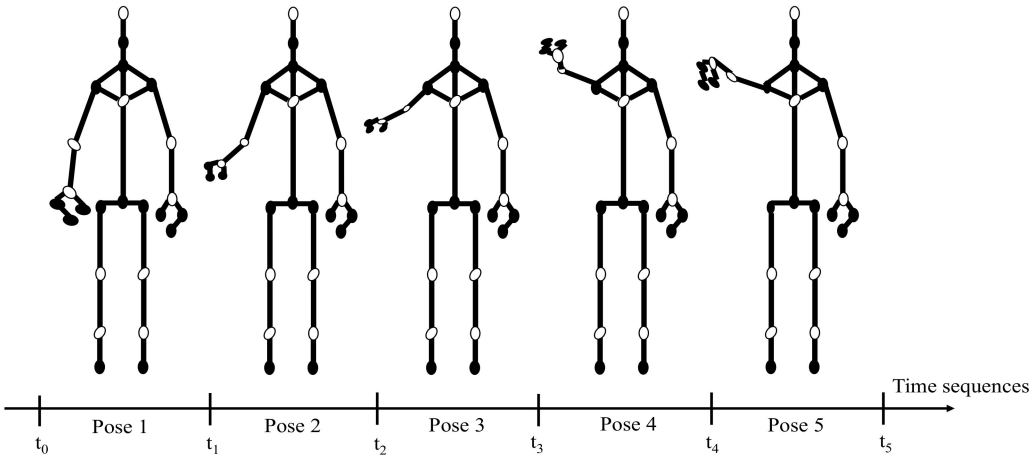

Figure 4. Example of the skeletonized "summon-pose" activity sequences divided into five sequential poses by time index accordingly.

Several concerns regarding vector quantization (VQ) [29–32] in this work are addressed below, given a vector source with its known statistical properties, such as F in Equation (11), given a distortion measure, and given the number of codevectors to find a codebook with minimized average distortion. In VQ, the set of quantized values for the vectors is a codebook. By quantizing each component of the source vectors, i.e., the feature set F of the frame, they can be subsequently substituted from a carefully chosen set and saved as an index set to yield a much more compact representation of the frame. The detailed steps of VQ are elucidated as follows:

Step 1: Generate the training vectors from the video dataset. The training vectors are defined as in Equation (12):

$$T = \{F_n | n = 1, 2, \ldots\ldots, N\} \tag{12}$$

where N is the number of training vectors, i.e., the number of frames. Let M be the number of codevectors and let C represent the codebook in Equation (13) as follows:

$$C = \{C_m | m = 0, 1, \ldots\ldots, M-1\} \tag{13}$$

Each codevector is k-dimensional, similar to the training vectors. Assume all training vectors to be a cluster C_0, i.e., codebook size M = 1 and codeword C_0 = 1. Then, find the k-dimensional cluster centroid as codevectors in Equation (14) as:

$$c_0 = \frac{1}{N} \sum_{n=1}^{N} F_n \tag{14}$$

Step 2: Double the current codebook size M to 2 M by splitting each cluster into two using Equation (15):

$$\begin{cases} c_i^{(l)} = c_i(1+\varepsilon) \\ c_{M+i}^{(l)} = c_i(1-\varepsilon) \end{cases} \tag{15}$$

where M is the size of the current codebook, c_i is the centroid of the ith cluster C_i, and ε is a splitting parameter vector in k-dimension. In this work, we set ε = 0.001 for each dimension; l is the iteration index.

Step 3: Classify each k-dimensional sample F of the training feature vectors into one of the clusters at each iteration by the k-nearest neighbor (KNN) approach. For n = 1, 2, 3, . . . , N, find the minimal value using Equation (16):

$$\|F_n - c_m^{(l)}\|^2 \tag{16}$$

Step 4: Update the codeword, i.e., symbol O_i, of each cluster C_m by computing new cluster centers in Equation (17):

$$c_m^{(l+1)} = \frac{\sum_{Q(F_n)=m} F_n}{\sum_{Q(F_n)=m} 1} \tag{17}$$

where m = 0, 1, . . . , $M-1$ at the (l + 1)th iteration.

Step 5: Set $l = l + 1$. The average distortion of the lth iteration is given by Equation (18):

$$D(l) = \frac{1}{N_k} \sum_{n=1}^{N} \|F_n - c_{Q(F_n)}^{(l)}\|^2 \tag{18}$$

If | D (l) – D (l c 1) | is lower than the chosen threshold, go to Step 6, otherwise, go back to Step 3.

Step 6: If the codebook size M is equal to the codebook size as expected, terminate the iteration, otherwise, go back to Step 2.

2.5. Noise Elimination

To increase the HAC accuracy recognition in terms of frame data, for pose sequence extraction, data smoothing for noise elimination is applied to remove the trivial poses or behaviors in the frame data, for noise data significantly influencing the entire activity recognition performance. Our findings show that some poses give the lower confidence for activity recognition. For example, as shown in Figure 5, the paradigms of unwanted pose frames are remarkably lower than the confidence in training the sequential-pose model. To eliminate such noise data, the threshold δ is proposed and set to examine the adjacent frames by using Equation (19):

$$\|F_i - F_{i-1}\|^2 \geq \delta \tag{19}$$

where F_i represents the feature vector of ith frame and δ is a threshold.

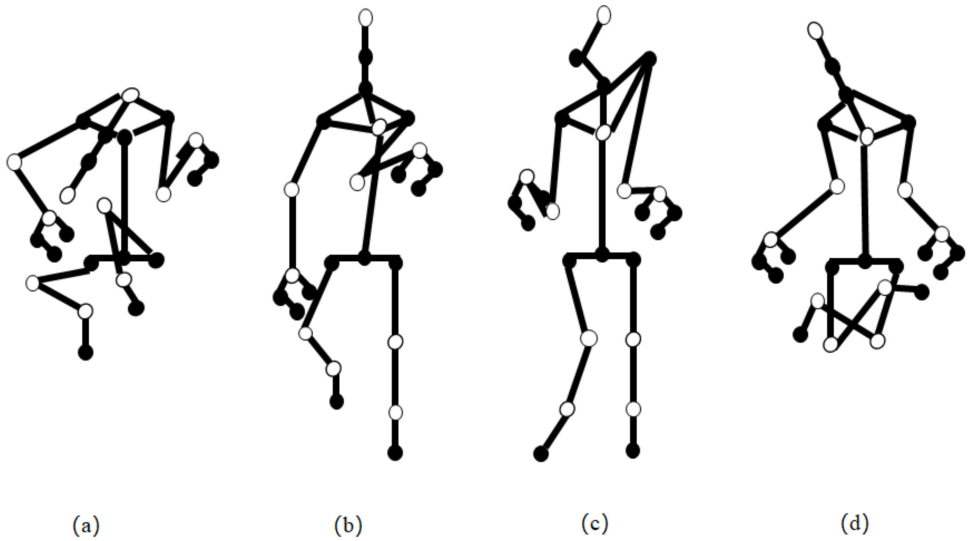

Figure 5. Four examples (**a**–**d**) of noisy skeletonized poses in collected activity sequences, which would be removed prior to encode the poses for discrete HMM on human activity estimation.

2.6. Discrete HMMs for Activity Recognition

The discrete HMM approach herein is used to estimate the human activity given by the observation sequence $X = (o_1, o_2, \ldots, o_T)$ to score the parameters $\lambda = (A, B, \pi)$ by maximizing the probability $P(X \mid \lambda)$ for target estimation, where A is the initial state probability, B is the state transition probability, and π is the state observation probability. Accordingly, each activity with the extracted poses sequences trained by DHMM treated as the observed symbols, of which the discrete time sequences implied the output of a Markov process whose states cannot be observed directly. Consequently, each activity sequence is encoded as a sequence of pose vocabulary to learn the HMMs. Once the HMMs are trained, they can then predict the input testing sequences through the maximal posterior probability.

In most cases, a very short time condition in micro perspective in a system can be observed as the one of the system's N states, denoted as $S = \{S_1, S_2, \ldots, S_N\}$, and the state at time t is denoted as q_t. Given the set of prior probabilities $\pi = \{\pi_i\}$ in Equation (20):

$$\pi_i = P[q_1 = S_i],\ 1 \leq i \leq N \tag{20}$$

where π_i is the probability of S_i of the initial state in a state sequence; the probability of transition from the state S_i to the state S_i, i.e., the corresponding probability, $A = \{a_{ij}\}$ in Equation (21):

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i],\ 1 \leq i, j \leq N \tag{21}$$

Assume M is the number of different observation symbols in a state, the individual symbol set is $V = \{v_1, v_2, \ldots, v_M\}$, and the observation symbol probability distribution in state j and $B = \{b_j(k)\}$ in Equation (22):

$$b_j(k) = P\left[v_k att | q_t = S_j\right], 1 \le j \le N, 1 \le k \le M \quad (22)$$

The overall parameter set of a D-HMM model is the triplet in Equation (23):

$$\lambda = (A, B, \pi) \quad (23)$$

Once the HMMs trained, they estimated the input testing sequence via the maximal posterior probability by Equations (24) and (25):

$$decision = argmax\{L_i\}, 1 \le i \le M \quad (24)$$

$$L_i = P(O|HMM_i) \quad (25)$$

where L_i is the likelihood of ith HMM and HMM_i is the input testing sequences.

3. Experimental Results

3.1. Dataset and Activity Implications

In this work, to validate the performance of the proposed D-HMM-based HAC framework prior to autonomous AI vision interaction with user domiciliary, we assess its performance on eleven daily life activities, including sitting, standing, walking, drinking water, calling up, reading, stretch, akimbo, follow-me, summoning, and stopping. For example, in elderly care, "slow-go" or "no-go" elderly people with aging mobility might be assisted by homecare robots, such that users can use a summoning gesture to call for a robot's help and a follow-me or stopping gesture to control the actions of the robot mobilities. Akimbo might be perceived by a robot as a user probably implying a spirit of competition, a strong sense of attack, or an enterprising spirit. Sitting, standing, or reading activities may be recognized by a robot as implying that the user is in a state of relaxation or concentration, whereas the stretch pose, generally performed as an unconscious behavior, implies the user is in a sleepy state. Images of the paradigms of the eleven activity poses with corresponding pose skeletons are shown in Figure 6.

Prior to validating the proposed work, eleven different human activities are filmed and collected into a database through a single, fixed binocular camera with an RGB-D sensor. The distance between the front of the camera and the participants ranged from 4 to 11 feet to yield RGB images with 1920 × 1080 resolution and depth information with 512 × 424 resolution, with skeleton joints at 30 fps for each pose sequence. In order to further verify the robustness of the proposed D-HMM-based HAC, each activity is filmed for five participants, five times. Particularly, to inspect the robustness of the drinking water activity, left and right hands, respectively, each person was filmed two times. In total, 275 samples of sequences with 16,172 frames were collected into a database. The number of frames for each sample video spanned between 33 to 156 frames. The sampling RGB image with corresponding skeleton clips in the dataset are shown in Figure 6. Note that the RGB images are just for illustration, and only the skeletonized joint information is used for activity recognition. The evaluation platform featured a 3.6 GHz Intel Core i7 CPU laptop and implemented by C#.

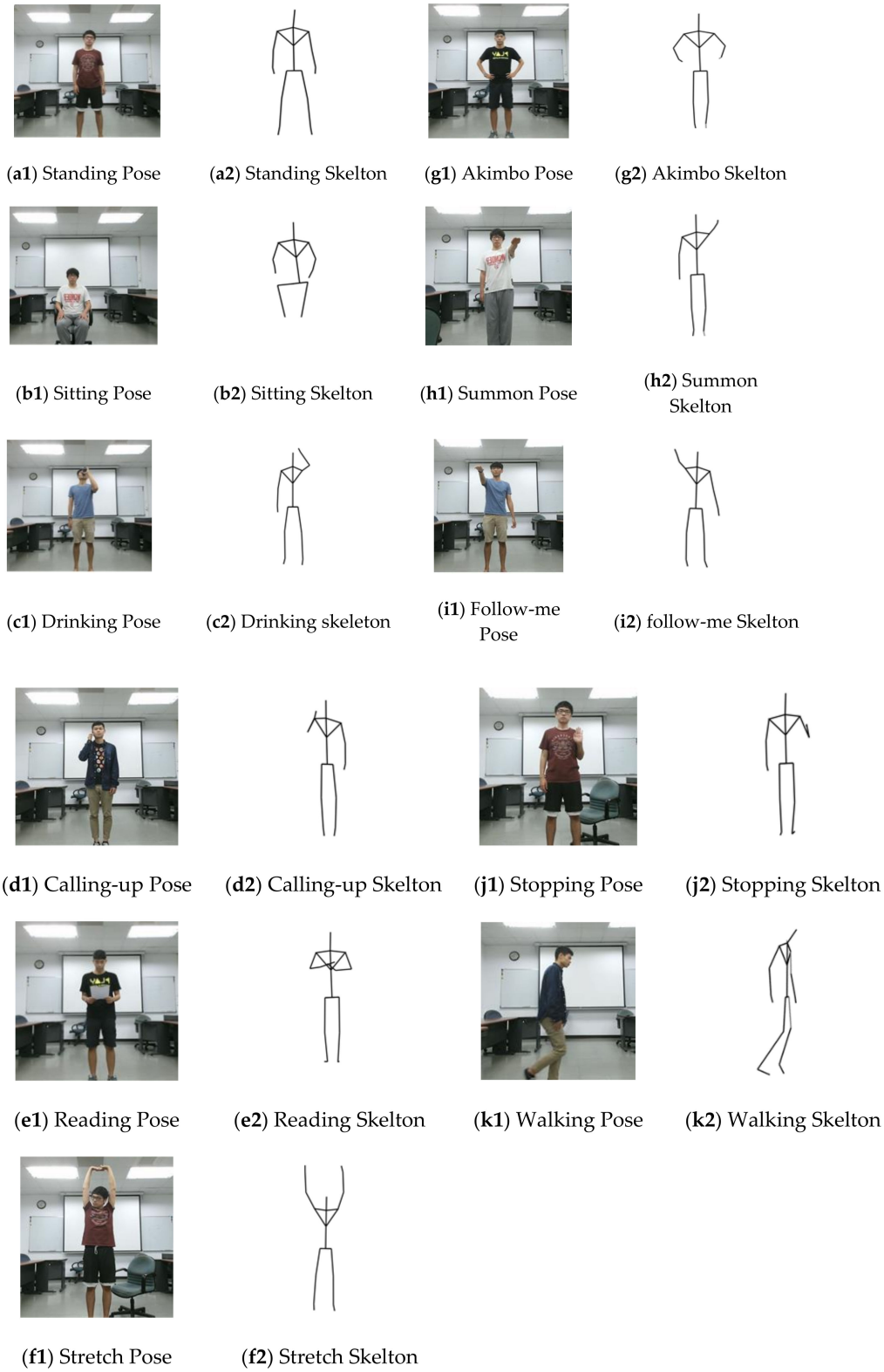

Figure 6. Sampling images of the assessing eleven poses in activities snaps with corresponding skeletons from videos of the collected database. Activities labeled from top to bottom, left to right, including (**a1**,**a2**) standing, (**b1**,**b2**) sitting, (**c1**,**c2**) drinking water, (**d1**,**d2**) calling up, (**e1**,**e2**) reading, (**f1**,**f2**) stretch, (**g1**,**g2**) akimbo, (**h1**,**h2**) summon, (**i1**,**i2**) follow-me, (**j1**,**j2**) stop gesture, and (**k1**,**k2**) walking poses.

3.2. Experimental Evaluation

Confusion Matrix Assessment

The proposed work is validated by using the collected 275 samples of sequences with 16,172 frames for the five-fold cross-validation experiments, wherein cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations, treated as phases in terms of testing and training datasets, whereas the number of clusters in vector quantization is set as $K = 128$ and the number of states in HMMs is set as $N = 5$. The average accuracy reached 95.64%, wherein some discriminative poses, including standing, sitting, stop gesture, and follow-me, achieved 100% recognition accuracy. However, the walking, drinking, and follow-me poses could be easily misclassified, and therefore, they led to a lower recognition accuracy of 88%. To detail the misclassified rates for 11 poses in terms of the true classes with the corresponding predicted classes, a confusion matrix was derived, as shown in Table 1.

In Table 1, by observing the drinking pose, in which 8% is misclassified as the calling-up pose. The aforementioned drinking pose is trained by left and right hands, respectively, such that the right-hand calling-up skeleton joints might be easily classified as the right-hand drinking skeleton joints, as shown in Figure 6(c2) of the drinking skeleton joints and Figure 6(d2) of the calling-up skeleton joints. In particular, 4% in the reading pose, 8% in the stretch pose, 4% in the summoning pose, and 8% in the walking pose, with the corresponding skeleton joints, as shown in Figure 6, are also partially misclassified as the drinking pose. The raised-hand pose has similar skeleton joint characteristics as the calling-up, the stretch, and the summoning poses, which leads to the misclassified rates, whereas the reading and walking poses give the perspective paradox on the misclassified rate, which is needed for further analysis. In addition, the akimbo pose is also partially misclassified in 4% as the follow-me pose.

As an extension of the confusion matrix in Table 1, the derivative criteria, including the conditional positive, conditional negative, true positive, true negative, false positive, and false negative with F_1 scores for the confusion matrix on the proposed 11 poses, are also examined and are shown in Table 2, where the F_1 score measures the test's accuracy, by considering both the precision p and the recall r of the test to compute the score. The proposed D-HMMs with ground true data were validated in the previous section; then, a real-world validation was conducted by using ten participants to test the proposed work. Each participant performed 11 activities ten times in terms of differentiated acting speed and personal action style. Once the individual-person case was evaluated, the multiple-person case was further verified to entirely validate the robustness of the proposed work in the cases of single user and multiple users. The results are shown in Figure 7. In most cases, the classified rate of a single person is generally superior to the multiple person case; however, the drinking pose and the follow-me gesture in the multi-person case performed better classified rates than the single-person case, probably due to insufficient data.

Table 1. The normalized confusion matrix for the proposed 11 poses activities, wherein each row represents an instance of the true class, whereas each column represents an instance of the predicted class, and the values of the diagonal elements represent the degree of correctly predicted classes.

		Predicted Classes										
		Standing Pose	**Sitting Pose**	**Drinking Pose**	**Talking Pose**	**Reading Pose**	**Stretch Pose**	**Akimbo Pose**	**Summon Pose**	**Follow-me Pose**	**Stopping Pose**	**Walking Pose**
True Classes	Standing Pose	1	0	0	0	0	0	0	0	0	0	0
	Sitting Pose	0	1	0	0	0	0	0	0	0	0	0
	Drinking Pose	0	0	0.92	0.08	0	0	0	0	0	0	0
	Calling up Pose	0	0	0.08	0.92	0	0	0	0	0	0	0
	Reading Pose	0	0	0.04	0	0.96	0	0	0	0	0	0
	Stretch Pose	0	0	0.08	0	0	0.92	0	0	0	0	0
	Akimbo Pose	0	0	0	0	0	0	0.96	0	0.04	0	0
	Summon Pose	0	0	0.04	0	0	0	0	0.96	0	0	0
	Follow-me Pose	0	0	0	0	0	0	0	0	1	0	0
	Stopping Pose	0	0	0	0	0	0	0	0	0	1	0
	Walking Pose	0	0	0.08	0	0	0	0	0	0.04	0	0.88

Table 2. The derivate criteria from the confusion matrix in Table 1 for the proposed 11 poses.

	P (Condition Positive)	**N (Condition Negative)**	**TP (True Positive)**	**TN (True Negative)**	**FP (False Positive)**	**FN (False Negative)**	**TPR (True Positive Rate)**	**TNR (True Negative Rate)**	**FNR (False Negative Rate)**	**FPR (False Positive Rate)**	**ACC (Accu_Racy)**	**F1 Score**
Standing Pose	1	0	1	0	0	0	1	0	0	1	1	1
Sitting Pose	1	0	1	0	0	0	1	0	0	1	1	1
Drinking Pose	0.92	0.08	0.92	0.08	0.08	0.92	0.5	0.5	0.5	0.5	0.5	0.65
Talking Pose	0.92	0.08	0.92	0.08	0.08	0.92	0.5	0.5	0.5	0.5	0.5	0.65
Reading Pose	0.96	0.04	0.96	0.04	0.04	0.96	0.5	0.5	0.5	0.5	0.5	0.66
Stretch Pose	0.92	0.08	0.92	0.08	0.08	0.92	0.5	0.5	0.5	0.5	0.5	0.65
Akimbo Pose	0.96	0.04	0.96	0.04	0.04	0.96	0.5	0.5	0.5	0.5	0.5	0.66
Summon Pose	0.96	0.04	0.96	0.04	0.04	0.96	0.5	0.5	0.5	0.5	0.5	0.66
Follow-me Pose	1	0	1	0	0	0	1	0	0	1	1	1
Stopping Pose	1	0	1	0	0	0	1	0	0	1	1	1
Walking Pose	0.92	0.08	0.92	0.08	0.08	0.92	0.5	0.5	0.5	0.5	0.5	0.65

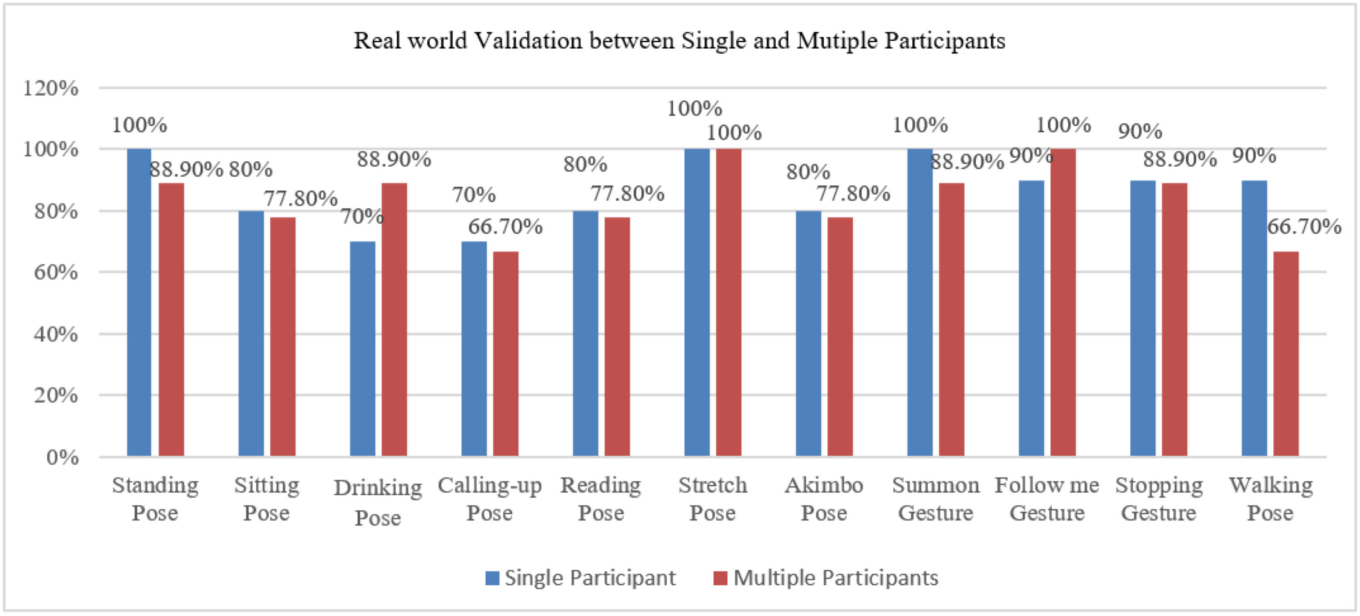

Figure 7. Real-world validation in terms of single and multiple participants performing 11 poses.

4. Conclusions

In this paper, we propose a human activity comprehension framework based on a discrete HMM approach. The spatio-temporal sequences of eleven collected visualized human activities with corresponding skeletal information are extracted to yield pose features, and efficient performance of human activity recognition is achieved through coordination transformation, vector quantization, and HMMs. The evidential framework, in the experiments, indicates that the proposed framework has a better capability to manage the data in terms of uncertainty and imperfection. Accordingly, a confusion matrix is conducted, and the derivative criteria are inspected to further analyze the experimental results, in which the analysis shows very promising results and performance. Future work should focus on learning with HMM-related approaches and comparing with real-world testing for more useful applications.

Author Contributions: T.-W.K.: Investigation, Conceptualization, Methodology, Writing—Original draft preparation and Editing; S.-P.T. and C.-W.C.: Data curation, Visualization, Validation; J.-F.W.: Conceptualization, Writing—Reviewing, Supervision, Project administration; C.-A.S.: Software, Data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were partially funded by Sanda University Grant number [2021ZD06] and were partially funded by Changzhou College of Information Technology under the contract [2019KYQD03].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dang, L.M.; Min, K.; Wang, H.; Piran, M.J.; Lee, C.H.; Moon, H. Sensorbased and vision-based human activity recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *108*, 3.
2. Bodor, R.; Jackson, B.; Papanikolopoulos, N. Vision-based human tracking and activity recognition. In Proceedings of the 11th Mediterranean Conference on Control and Automation, Rhodes, Greece, 18–20 June 2003; Volume 1.

3. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A review on human activity recognition using vision-based method. *J. Healthc. Eng.* **2017**, *2017*, 3090343. [CrossRef] [PubMed]
4. Bux, A.; Angelov, P.; Habib, Z. Vision based human activity recognition: A review. In *Advances in Computational Intelligence Systems*; Springer: Cham, Switzerland, 2007; pp. 341–371.
5. Chen, L.; Hoey, J.; Nugent, C.D.; Cook, D.J.; Yu, Z. Sensor-based activity recognition. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2012**, *42*, 790–808. [CrossRef]
6. Liu, Y.; Nie, L.; Liu, L.; Rosenblum, D.S. From action to activity: Sensor-based activity recognition. *Neurocomputing* **2016**, *181*, 108–115. [CrossRef]
7. Kuan, T.W.; Tseng, S.P.; Wang, J.F.; Chen, P.J. A happiness cups system for holding-cup motion recognition and warming-care delivery. In Proceedings of the 2016 International Conference on Orange Technologies (ICOT), Melbourne, Australia, 17–20 December 2016.
8. Yousefzadeh, A.; Orchard, G.; Serrano-Gotarredona, T.; Linares-Barranco, B. Active perception with dynamic vision sensors. minimum saccades with optimum recognition. *IEEE Trans. Biomed. Circuits Syst.* **2018**, *12*, 927–939. [CrossRef] [PubMed]
9. Luan, P.G.; Tan, N.T.; Thinh, N.T. Estimation and Recognition of Motion Segmentation and Pose IMU-Based Human Motion Capture. In *International Conference on Robot Intelligence Technology and Applications*; Springer: Cham, Switzerland, 2017; pp. 383–391.
10. Haescher, M.; Matthies, D.J.; Srinivasan, K.; Bieber, G. Mobile assisted living: Smartwatch-based fall risk assessment for elderly people. In Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction, Berlin, Germany, 20–21 September 2018; pp. 1–10.
11. Chanthaphan, N.; Uchimura, K.; Satonaka, T.; Makioka, T. Facial emotion recognition based on facial motion stream generated by kinect. In Proceedings of the 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Bangkok, Thailand, 23–27 November 2015; pp. 117–124.
12. Bouissou, M.B.; Laffont, J.J.; Vuong, Q.H. Tests of noncausality under Markov assumptions for qualitative panel data. *Econom. J. Econom. Soc.* **1986**, *54*, 395–414. [CrossRef]
13. Li, S.Z. *Markov Random Field Modeling in Computer Vision*; Springer Science & Business Media: Berlin, Germany, 2012.
14. Ramage, D. *Hidden Markov Models Fundamentals*; CS229 Section Notes; Stanford University: Stanford, CA, USA, 2007.
15. Yakowitz, S.J. Nonparametric density estimation, prediction, and regression for Markov sequences. *J. Am. Stat. Assoc.* **1985**, *80*, 215–221. [CrossRef]
16. Benouareth, A.; Ennaji, A.; Sellami, M. HMMs with explicit state duration applied to handwritten Arabic word recognition. In Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006; Volume 2, pp. 897–900.
17. Russell, M.J.; Jackson, P.J.; Wong, M.L. Development of articulatory-based multilevel segmental HMMs for phonetic classification in ASR. In Proceedings of the EC-VIP-MC 2003 4th EURASIP Conference Focused on Video/Image Processing and Multimedia Communications (IEEE Cat. No. 03EX667), Zagreb, Croatia, 2–5 July 2003; Volume 2, pp. 655–660.
18. Aarno, D.; Kragic, D. Layered HMM for motion intention recognition. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 5130–5135.
19. Glodek, M.; Bigalke, L.; Schels, M.; Schwenker, F. Incorporating uncertainty in a layered HMM architecture for human activity recognition. In Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, Scottsdale, AZ, USA, 1 December 2011; pp. 33–34.
20. Broumandnia, A.; Shanbehzadeh, J.; Nourani, M. Handwritten farsi/arabic word recognition. In Proceedings of the 2007 IEEE/ACS International Conference on Computer Systems and Applications, Amman, Jordan, 13–16 May 2007; pp. 767–771.
21. Dehghan, M.; Faez, K.; Ahmadi, M. A hybrid handwritten word recognition using self-organizing feature map, discrete HMM, and evolutionary programming. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 27 July 2000; Volume 5, pp. 515–520.
22. Yasuda, H.; Takahashi, K.; Matsumoto, T. A discrete HMM for online handwriting recognition. *Int. J. Pattern Recognit. Artif. Intell.* **2000**, *14*, 675–688. [CrossRef]
23. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3288–3297.
24. Lo Presti, L.; La Cascia, M. 3D skeleton-based human action classification. *Pattern Recognit.* **2016**, *53*, 130–147. [CrossRef]
25. Guo, H.; Yang, Y.; Cai, H. Exploiting LSTM-RNNs and 3D skeleton features for hand gesture recognition. In Proceedings of the 2019 WRC Symposium on Advanced Robotics and Automation (WRC SARA 2019), Beijing, China, 21–22 August 2019; pp. 322–327.
26. Elaoud, A.; Barhoumi, W.; Zagrouba, E.; Agrebi, B. Skeleton-based comparison of throwing motion for handball players. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 419–431. [CrossRef]
27. Palanimeeraa, J.; Ponmozhib, K. Techniques used to Capture Skeletal Information and their Performance Accuracy: A Literature Review. In Proceedings of the Second International Conference on IoT Social, Mobile, Analytics and Cloud in Computational Vision Bio-Engineering (ISMAC-CVB 2020), Tamil Nadu, India, 29–30 October 2020; pp. 480–486.

28. Raptis, M.; Kirovski, D.; Hoppe, H. Real-time classification of dance gestures from skeleton animation. In Proceedings of the 2011 ACM SIGGRAPH/Euro Graphics Symposium on Computer Animation, New York, NY, USA, 5–7 August 2011; pp. 147–156.
29. Goldberg, M.; Sun, H. Image sequence coding using vector quantization. *IEEE Trans. Commun.* **1986**, *34*, 703–710. [CrossRef]
30. Nasrabadi, N.M.; King, R.A. Image coding using vector quantization: A review. *IEEE Trans. Commun.* **1988**, *36*, 957–971. [CrossRef]
31. Chahid, A.; Khushaba, R.; Al-Jumaily, A.; Laleg-Kirati, T.M. A Position Weight Matrix Feature Extraction Algorithm Improves Hand Gesture Recognition. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montréal, QC, Canada, 20–24 July 2020; pp. 5765–5768.
32. Kästner, M.; Strickert, M.; Villmann, T.; Mittweida, S.G. A sparse kernelized matrix learning vector quantization model for human activity recognition. In Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 24–26 April 2013.