SEGMENTATION AND CLASSIFICATION OF HUMAN ACTIONS AND ACTOR CHARACTERISTICS WITH 3D MOTION DATA

S. Ali Etemad[1] and Ali Arya[2]

[1]Department of Systems and Computer Engineering, Carleton University
Ottawa, ON K1S5B6, Canada
ali_etemad@carleton.ca
[2]School of Information Technology, Carleton University
Ottawa, ON K1S5B6, Canada
arya@carleton.ca

ABSTRACT

In this paper we have used 3D motion capture data with the aim of detecting and classifying specific human actions. In addition to recognition of basic action classes, actor styles and characteristics such as gender, age, and energy level have also been subject to classification. We have applied and compared three main methods: nearest neighbour search, hidden Markov models, and artificial neural networks. Using these techniques, we have proposed exhaustive algorithms for detection of actions in a motion piece and subsequently classifying the segmented actions and respective characteristics of the actors. We have tested the methods for various sequences and compared the results for a comprehensive evaluation of each of the proposed techniques. Our findings can be largely used for general classification of human motion data for multimedia applications as well as sorting and classifying data sets of human motion data especially those acquired using visual marker-based motion capture systems such as the one employed in this research.

KEYWORDS

Action Recognition; Motion Capture Data; Neural Networks; Hidden Markov Models; Nearest Neighbour Search.

1. INTRODUCTION

Human motion is an essential topic of study and research by physicians, biomedical engineers, and graphics and multimedia experts. Recognition and synthesis of actions are the two main categories of research on human motion. The recognition element involves detecting and identifying primary and secondary motor themes [1]. Primary themes are basic mechanical movements which form basic actions such as walking or jumping, while secondary themes are affective and stylistic variations introduced by the actor with respect to mood, gender age, style, physics, and even genetics. Motion recognition can be used for a variety of applications ranging from human locomotion analysis for biomedical applications to computer vision and surveillance, and more recently for interactive games and simulators [2,3]. On the other hand, interactive virtual environments are rapidly growing for many applications from arts and entertainment to scientific simulation, education, and the service industry [4,5]. Very large human motion data sets, recorded using various means, have therefore been constructed. As a result, it has become largely essential to design and employ systems which automate the interpretation of the contents of different motion files, for both real-time systems (interactive systems) and offline applications (data sets).

DOI : 10.5121/ijaia.2012.3405

This paper aims at classifying basic human actions and their personal variations such as those related to age, energy, and gender. The ability to recognize these actions and variations allow us to not only respond to human movements but also understand and computationally model them for synthesis purposes. Although the focus of this paper is on the classification process, other publications by the authors have demonstrated the use of such models for synthesis and procedural animation of human movements [6].

Recognition and classification of user characteristics can result in more intelligent and efficient responses in many applications. Simulators for highly advanced objectives such as flight or military operations could strongly benefit from gaining knowledge to the personality, mood, gender, age, and other characteristics of the user alongside segmenting, recognizing, and classifying the user actions themselves. Biomedical applications [7] are another quite distinct area where analysis of human motion can be widely applied. The analysis can be conducted with the aim of recognizing human joint, muscle, or bone difficulties for physiology, or even for recognition of specific neurological diseases and motor related disorders. Next generation computer games will be using direct computer vision techniques as the input system and can also benefit from proper recognition of human actions and their personal variations. In most publications (discussed in Section 2), only a specific method for human motion recognition has been employed and in some, highly accurate results have been acquired. What seems to be lacking among the addressed literature, is a comprehensive evaluation and comparison between the different classifiers using the same type of data and successive to similar pre-processing. Also recognition and classification of secondary themes and actor characteristics has not been properly studied. In this research, we have used 3 popular classifiers for segmentation and classification of human actions recorded by a motion capture system and the same systems have then been used for classification of secondary themes. This will provide us with a clear understanding and conclusion regarding the capabilities of different classifiers for segmentation and classification of primary and secondary themes using motion captures data. Also we have studied the effect of the nature of each action on the classification accuracy of different classifiers determining the degree of difficulty for different classifiers to classify the primary and secondary themes. Based on this study, the most suitable classifier can be selected for a specific action class or secondary theme.

In Section 2, a review on the available literature in the field of human motion recognition and analysis is performed. The literatures focusing on human motion classification can be categorized into two groups of studies: studies based on visual (image/video) inputs and studies based on motion capture data. The study presented in this paper is of the second category. The data type used in this research provides us with the benefit of not having to deal with background, light, and clothing issues and the focus can be put on evaluating different classifiers for segmentation and classification primary and secondary themes. Like any other system, pre-processing is carried out for enhancing the data. This is presented in Section 3. The methodologies used for segmentation and classification of primary and secondary themes are described for each technique in Sections 4.1, 4.2, and 4.3. Respectively the methods are nearest neighbor, hidden Markov models, and artificial neural networks. In Section 5, experimental results are presented and discussed, and finally in Section 6, the closing conclusions are provided.

2. RELATED WORK

Many researchers have worked on human motion segmentation and classification. In this section we review some of these works. The focus of this research has been the use of motion-capture data, and so we have focused on researches which have employed such data.

Lv and Nevatia [8] have used the Viterbi algorithm for single view recognition of actions and propose a Pyramid Match Kernel algorithm and compute matching scores between the feature

data sets. They automatically extract key poses for this purpose based on motion energy charts. Their method reaches a classification accuracy of around 80%.

Zhou et al. [9] have focused on segmenting specific actions from motion capture data. They have defined the problem as an extension to practical segmentation techniques. Aligned Cluster Analysis (ACA) has been proposed by the authors for temporal segmentation of actions. This problem is what we, in our research, have referred to as locating actions. Despite the relatively high accuracy achieved in their paper (approximately 97%), a strong drawback is that the exact number of segments must be manually provided to the algorithm to avoid local minima.

Ali et al. [10] model human actions using the theory of chaotic systems. The non-linear dynamic system generating the action is represented by trajectories of specific reference joints: the head, two hands and two feet. An overall classification accuracy of 89.7% is achieved through this technique. Ishiyama et al. [14] use markerless human motion capture samples and describe human motion by a low dimensional linear model. While the proposed model shows to be promising in the sense that it is robust for different body types, there is no indication as to how accurate the classification accuracy is.

Recognition and generation of motion data are carried out using HMMs by Kulic et al. [11]. In their work, different motions are organized in hierarchical tree fashion where nodes closer to the root correspond to more general motion features while further motions represent more detailed motion descriptors.

In [12], Song et al. use the concepts of maximum likelihood to build an unsupervised system applicable to both greyscale images and motion capture data for recognition of human motion. Their system employs an algorithm which uses differential entropy of the variables to locate the optimal structure of the decomposable model which is claimed to perform superior to manually constructed models. Their model however, maintains a trade-off between model complexity and accuracy and has only been tested on walking sequences and it is not clear on how it will perform with multiple action classes. A different type of probabilistic models called Switching Linear Dynamics (SLD) is employed for human action classification by Shimosaka et al. [13]. SLDs are a combination of HMMs and Linear Dynamics (LD) which perform stochastically. They also demonstrate that utilizing their proposed kernel for training support vector machines shows to be very accurate for performing the task where 6 classes of action have been employed for this approach. Nevertheless as the goal of the research was the design of specific kernals, a multi-class classifier is not designed and no results provided. It can also be noted that the 6 chosen classes of action or mostly (except for walk and run), static motions.

Mori et al. [15] have employed motion capture data to train HMMs for segmentation of actions. In this paper the authors have used the evaluation of the starting and ending frames for segmentation and an online SVM-based classifier is used for calculating action probabilities. Then the action probabilities are analyzed by HMMs for determining whether a specific frame is to be segmented or not. As the proposed algorithm shows to be robust for segmentation, it is not clear whether classification of the segmented actions is foreseen by the algorithm and if so, how accurately the task is performed.

Using a rather different approach towards human action recognition in [16], Parameswaran and Chellappa have proposed creating a 3D-invariance space for each action. Each action is characterized in this space by a curve, and test actions are probabilistically analyzed in this space with respect to the defined curves representing each action. The technique is tested for 5 different viewpoints for each action in both 2D and 3D, while the former maintains a mean accuracy of nearly 91% and the latter shows a mean accuracy of approximately 89%.

Brand and Kettnaker [17] have used the concepts of entropy minimization of joint distributions to train HMMs by motion capture data. The classifiers have then been used for classifying actions from normal video sequences. The human figures are first converted to 2D silhouettes, which are then employed by the HMMs. Thus this interesting approach is applicable for ambiguous scenes as well.

Through a rather simpler approach compared to [17], Li and Fukui [18] have trained HMMs by means of factorization of motion capture data for view-invariant classification of human actions. Despite using motion capture data which provides 4D data, they have not used the *Z* information during the experiments yet claiming view invariance for the system. However, they have accomplished a classification accuracy of nearly 95%.

3. DATA ACQUISITION AND PRE-PROCESSING

To record the movement of human motion in the form of digital data is often referred to as motion capture which dates back to 1970's. The recorded motion can be used for on the spot or afterwards playback to be used in computer animation applications. The Vicon MX40 motion capture system, located at the School of Information Technology in Carleton University, provides all the required input data for this research. Vicon MX40 is a marker-based optical motion capture system which applies basic light reflection laws for precise recording of motion. To acquire the required sequences of actions, the standard system setup and calibrations were carried out based on documents on Carleton University website: http://mocap.csit.carleton.ca/.

In this research, 45 physical markers were used for capture sessions which were translated to 17 virtual markers by the system software. The virtual markers are placed on critical joints of the body such as wrists, elbows, shoulders, neck, head, spine, hips, knees, and ankles. The exact motion trajectories of the markers placed on the body are tracked and recorded, and virtual marker data in the form of $\boldsymbol{Data}_{mocap} = \left\langle [\mathbf{D_1} \cdots \mathbf{D_m}]^{\mathbf{T}} [\mathbf{\Theta_1} \cdots \mathbf{\Theta_m}]^{\mathbf{T}} \right\rangle$ are created. In this equation $\mathbf{D}_i$ and $\mathbf{\Theta}_i$ are the position trajectories and rotation trajectories respectively for the *i*th frame. $\mathbf{D_i} = \left[d_i^{(x)}, d_i^{(y)}, d_i^{(z)}\right]$ where $d_i^{(x)}$, $d_i^{(y)}$, and $d_i^{(z)}$ are the three components for the location of the body at posture *i*. $\mathbf{\Theta_i} = \left\langle \left[\theta_i^{(x1)}, \theta_i^{(y1)}, \theta_i^{(z1)}\right] \cdots \left[\theta_i^{(xn)}, \theta_i^{(yn)}, \theta_i^{(zn)}\right] \right\rangle$ where $\theta_i^{(xj)}$, $\theta_i^{(yj)}$, and $\theta_i^{(zj)}$ are the three rotation components for the *j*th marker at the same posture. In this paper, we symbolize the *i*th temporal rotational trajectory for *m* frames by $\boldsymbol{\theta}^{(i)} = \left[\theta_1^{(i)} \cdots \theta_m^{(i)}\right]^{\mathbf{T}}$ where $1 < i < 3n+3$.

The relative scalar values are converted to feature vectors through Eqs. 1 and 2 where $\boldsymbol{v}_{D(i)}$ and $\boldsymbol{v}_{\Theta(i)}$ are the displacement and angular velocity vectors for the *i*th frame, and *fr* is the frame rate. Velocity vectors are employed as opposed to displacement vectors for recognition due to the outcome of our experimental investigations showing their superior in training the classifiers.

$$\mathbf{v}_{D(i)} = fr \cdot (\mathbf{D}_i - \mathbf{D}_{i+1}) \quad (1)$$

$$\mathbf{v}_{\Theta(i)} = fr \cdot (\mathbf{\Theta}_i - \mathbf{\Theta}_{i+1}) \quad (2)$$

Since the frame rates are equal and consistent throughout the research, the *fr* value simply acts as a scaling factor which can be ignored. Figure 1 (left) illustrates the raw 4D (3D vs. time) values of the LeftLegRoll marker (which is one of the markers placed on the left leg). Figure 1 (right) represents the differentiated formats of the motion capture data after being converted to three 2D trajectories (for frames 1 to 43 which is the length of this particular sequence) and applying Eqs. 1 and 2.

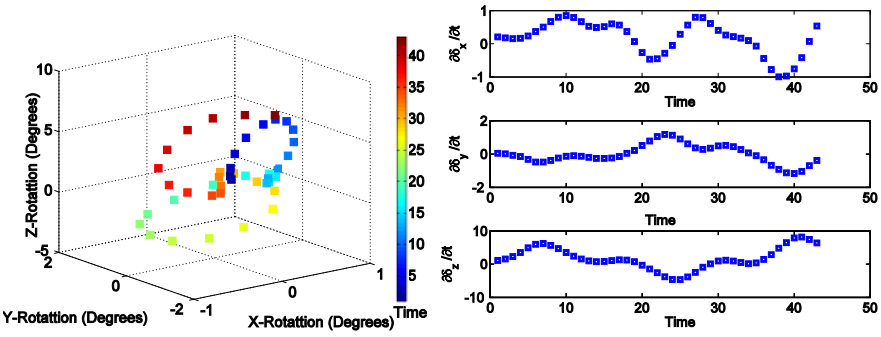

Figure 1. LeftLegRoll data for masculine jump, raw scalar 4D format (left), 2D angular velocity trajectories (right).

The dimensionality of the data is then reduced to decrease the runtime and complexity. Similar to [19] where entire actions have been recognized using the head alone, we have proposed that there is no need to use all the motion data for segmentation and recognition since the task can be carried out using the data of only some feature-rich parts of the body for simplifying the problem. In this research we have carried out quantization and then used the motion data of the lower half of the body (hips, legs, and feet) only. This simplification of the data is also in line with [30] where we have illustrated that approximately 40% of the perceptual information of the human motion lays in the legs and feet section alone.

Like any other data recording process, our data too contain noise. The noise can be originated from a variety of different sources, ranging from imperfect initial calibration to existence of light reflective objects in the room. Moreover, noise in motion capture data is mostly generated when errors occur in the system while tracking a particular marker. The effect of this artefact appears as high-frequency spikes in the trajectories which are reduced using low-pass filtering.

4. METHODOLOGY

Three main popular types of classifiers, namely Nearest Neighbor (NN), Hidden Markov Models (HMM), and Artificial Neural Networks (ANN), have been used for segmentation and classification of primary and secondary themes in human motion data. In this section a brief description of the theory of each classifier is given along with the methodology and algorithm used to employ each classifier. The algorithms are designed to best suit the classifier properties as well as providing the system with the ability to automatically segment and classify both themes. For training the different classifiers, a training set of 90 samples has been employed (approximately 2700 frames). The overall structure of the system is illustrated by Figure 2.

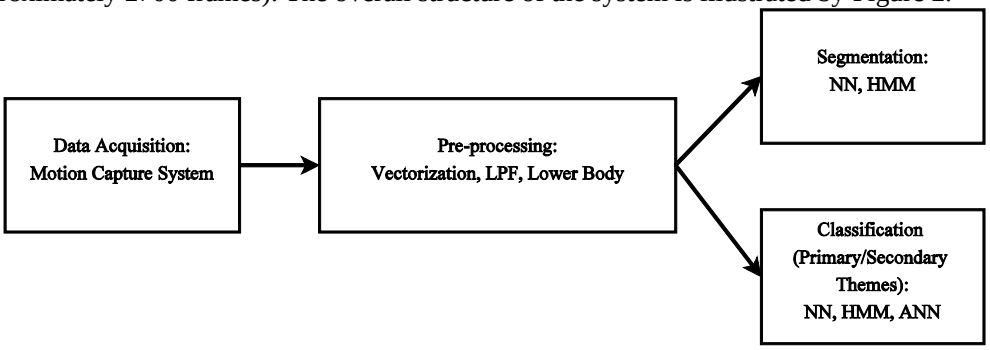

Figure 2. Overall structure of the system.

For recognition of secondary themes, a separate network for each style as well as each primary class was trained. For instance, a separate NN, HMM, and ANN classifier was trained for feminine walk, masculine walk, tired walk, energetic walk, young walk, and for old walk. Similar approach was taken for jump and run primary classes. The test data were fed through all networks and the system resulting in the least error or maximum likelihood determined the primary and secondary themes of the test sequence.

4.1. Nearest Neighbour

By definition, NN search is a supervised classification technique which uses some kind of distance -in most cases Euclidean distance, in the feature space to determine and classify the test samples based on the closest trained samples. In terms of functionality of this classifier, it is a binary search algorithm which is also called similarity search [20]. This classifier has previously been used in some literature for human action recognition and classification [19, 21-23].

We have employed NN search for both segmentation and classification of themes. In order to do so, Algorithm 1 was developed.

Algorithm 1. Segmentation and classification using NN search.

1:	carry out pre-processing of the data
2:	*for* all frames of $\boldsymbol{T}$ *do*
3:	measure e_s (Eq. 3)
4:	return j which satisfies $min(e_s)$
5:	*end for*
6:	for all frames of $\boldsymbol{T}$ (starting from the returned j by line 4) *do*
7:	measure e_c (Eq. 4)
8:	return p which satisfies $min(e_c)$
9:	*end for*

To employ NN for classification, we name the pth training sequences $\mathbf{R}_j^{(i)}(p)$ where i represents the marker and j represents the frame. Based on Eq. 1, three coordinates (x, y, and z) are available for each marker value resulting in $\boldsymbol{R}_j^{(xi)}(p)$, $\boldsymbol{R}_j^{(yi)}(p)$, and $\boldsymbol{R}_j^{(zi)}(p)$. The test sequence is denoted by $\boldsymbol{T}$, using similar notations as $\boldsymbol{R}$. In this algorithm the search initiates from the first frame. The test sample frames are compared to all the starting frames of all the training samples. The test frame that proves to be the nearest neighbour (least distance) is selected as the starting point (for segmentation). Once the initial frames of all the test samples are located, the following frames are used for distance measurements. The distance is calculated by Eq. 7 when the nearest neighbour of a single frame is being calculated. In this formula, n denotes the number of markers. For an entire sequence, Eq. 8 yields the distance where m is total number of frames. For either case, the goal is to satisfy $min(e)$.

$$e_s = \left(\mathbf{T}_j^{(i)}(p) - \mathbf{R}_j^{(i)}(p)\right)\left(\mathbf{T}_j^{(i)}(p) - \mathbf{R}_j^{(i)}(p)\right)^T \quad for\ all\ p \tag{3}$$

$$e_c = \sum_{j=1}^{m}\left(\mathbf{T}_j^{(i)}(p) - \mathbf{R}_j^{(i)}(p)\right)\left(\mathbf{T}_j^{(i)}(p) - \mathbf{R}_j^{(i)}(p)\right)^T \quad for\ all\ p \tag{4}$$

When performing classification, the training sample which then returns the minimum e_c is selected as the sequence containing an action of the same class as $\boldsymbol{R}$ for both primary and secondary themes.

4.2. Hidden Markov Models

HMMs are statistical models often considered as the simplest type of Bayesian networks. In HMMs various states are defined which are hidden and not visible while the outputs which depend on the states are visible, thus the overall model is considered to be visible. The states possess probability distributions over the outputs. In this research, the Baum-Welch [24] algorithm for likelihood estimation through expectation maximization is used to configure the model.

For an HMM network with n hidden states $\{s_1, s_n, \ldots, s_n\}$, the transition probability is $P(s_j(t+1)|s_i(t))$. Transition property is defined as the probability of the HMM being in state s_i at $(t+1)$ if it has been in state s_j at (t). Another definition in HMM is emission probability, where it is the probability of the HMM producing a certain symbol (observation) at a certain state [25]. The goal of the algorithm used is to maximize the expectation of posture emissions using which the model has been trained.

For each class of action and related style variations, a different HMM is created which calculates the total emission probability for generating the sequence of feature vectors corresponding to each action. The test data will be employed by the created networks and the HMM resulting in the most likelihood will determine the class of the test sample.

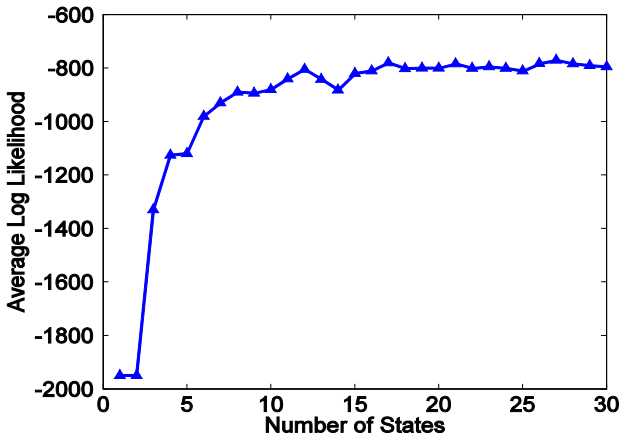

Figure 3. Average Log Likelihood for Different Number of States.

After constructing the HMM networks, the number of hidden states, as well as iterations - in order to reach maximum learning without occurrence of over fitting, must be determined. Figure 3 shows the learning accuracy curve for different number of states in a 50 iteration process. The highest accuracy is obtained for 17, 21, 25, 26, and 27 states. Yet the difference between a 17 state and a 27 state hidden Markov model network in this research is insignificant, and therefore to minimize the computations and runtime, 17 states is assigned to the network. Furthermore, different tests show that the likelihood does not noticeably improve beyond 35 iterations.

Successive to training the HMMs for each action class and secondary theme, an iterative search is carried out to select the location and duration of the sequence returning the highest likelihood by each HMM. The sequence resulting in the highest likelihood will determine the segment and class of action as well as the secondary theme.

4.3. Artificial Neural Networks

ANNs are practical tools in the field of pattern recognition which are inspired by the functionality of biological neural networks for processing of data. In this research the efficient and practical feed-forward Multi Layer Perceptron (MLP) neural network is employed. MLP employs multiple layers and maps inputs and outputs, resulting in non-linear classification where required. The learning algorithm of back-propagation is adopted. Back-propagation is considered a supervised method of learning which utilizes *activation functions*.

Based on the discussion provided in Section 3 the data for the lower half of the body is employed (legs, feet, and hips). As a result, the dimension of the input vectors will be 21 (three dimensions for each of the 7 marker points). There are some different proposed methods for setting the optimal number of hidden nodes and layers [26, 27] with respect to the dimension of the input vectors. Based on these facts and after a number of different trials, 2 hidden layers, each containing 25 neurons were assigned to the network.

Each network is trained to perform as an anticipator. Anticipators are systems which anticipate a specific type of data and any data different from what they have been trained for will result in high error values. Each network is trained for a specific action class containing a specific secondary theme. The data is fed into each network and the Mean Square Error (MSE) is calculated to evaluate the compatibility of the test sequence to the neural network. The network returning the smallest error will determine the themes of the sequence.

For optimization and updating the weights in neural networks, different techniques are proposed and employed. As an alternative to the Levenberg-Marquardt back-propagation training procedure which is quite popular, the very high-speed Resilient back-propagation (RPROP) training technique [28] was used. From [29] we can observe that RPROP is more suitable for 3D image, 3D object, and most likely in our case, 3D motion analysis. The RPROP technique holds several advantages for this research such as fast convergence and the ability to escape local minima having gone through sufficient learning epochs. The main difference between this learning scheme and other more common ones is that the adaptation of RPROP is not affected by the magnitude of the gradient and only takes into account the behaviour of the sign of the gradient. Although this property might be considered a setback for applications where more adaptability is required, in this research, it would be an asset due to uncertainty of human actions even when carried out by the same actor. In RPROP, only the sign of the gradient is considered for weight updates as opposed to other methods where the size of the gradient is a determining factor and this property opens room for more uncertainty in the data.

In each stage of learning, the weight for neuron j to neuron i is updated by the amount of $\Delta w_{ji}(k)$ as shown in Eq. 9 where $A_{ji}(k)$ is the update value and $E(k)$ is the error function [28].

$$\Delta_{ji}(k)=\begin{cases}-A_{ji}(k) & if\ \dfrac{\partial E}{\partial w_{ji}}(k)>0\\ +A_{ji}(k) & if\ \dfrac{\partial E}{\partial w_{ji}}(k)<0\\ 0 & else\end{cases} \tag{5}$$

Therefore we can conclude Eq. 10 where η is the increase factor, μ is the decrease factor, and $0 < \mu < 1 < \eta$ holds true [28].

$$\Delta_{ji}(k)=\begin{cases}\eta A_{ji}(k-1) & if\ \frac{\partial E}{\partial w_{ji}}(k)\times\frac{\partial E}{\partial w_{ji}}(k-1)>0\\ \mu A_{ji}(k) & f\ \frac{\partial E}{\partial w_{ji}}(k)\times\frac{\partial E}{\partial w_{ji}}(k-1)<0\\ 0 & else\end{cases} \quad (6)$$

When dealing with human actions, during a walking sequence for instance, the actor may unintentionally take a faster step, resulting in larger rotation values in the feature space. Therefore, this type of algorithm, where the gradient magnitude influence on the weight change is eliminated, is more suitable for action recognition purposes based on the fact that RPROP is suitable for data with an amount of uncertainty. Having used the typical gradient based back-propagation techniques would have resulted in larger weight changes in the network which are inaccurate. Using this method, however, results in the system learning the pattern of the movement and ignoring such uncertainties in human actions.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the experimental results for each of the methods explained in Section 3 are presented. Three different action classes are selected to be tested for each technique. The actions are: walking (W), running (R), and jumping (J). These actions are not randomly chosen, and are in fact selected this way to represent a wide range of different types of actions. Walking and running represent closely related displacement-based and similar actions which can easily confuse the classifier and be misclassified. Jumping, on the other hand, represents a wide range of actions which can be referred to as in-place actions. Other examples of this type of actions are kicking, waving, punching, collapsing, and etc. The 3 mentioned classes of actions are carried out in 6 different styles (secondary themes). These themes are masculine (M) and feminine (F) for gender, tired (T) and energetic (E) for energy, and young (Y) and old (O) for age properties. The variations used in this study are presented in Table 1. Overall, each system is tested for 18 distinct actions.

Table 1. Motion variations used in this study

Primary Theme	**Secondary Theme**	
Walk	*Feminine*	*Masculine*
	Tired	*Energetic*
	Young	*Old*
Run	*Feminine*	*Masculine*
	Tired	*Energetic*
	Young	*Old*
Jump	*Feminine*	*Masculine*
	Tired	*Energetic*
	Young	*Old*

Figure 4 presents the three primary themes along with the secondary themes for a walking sequence (masculine walk, feminine walk, young walk, old walk, tired walk, energetic walk) as well as masculine jump, masculine run. For both segmentation and classification, 5 samples of each class are considered.

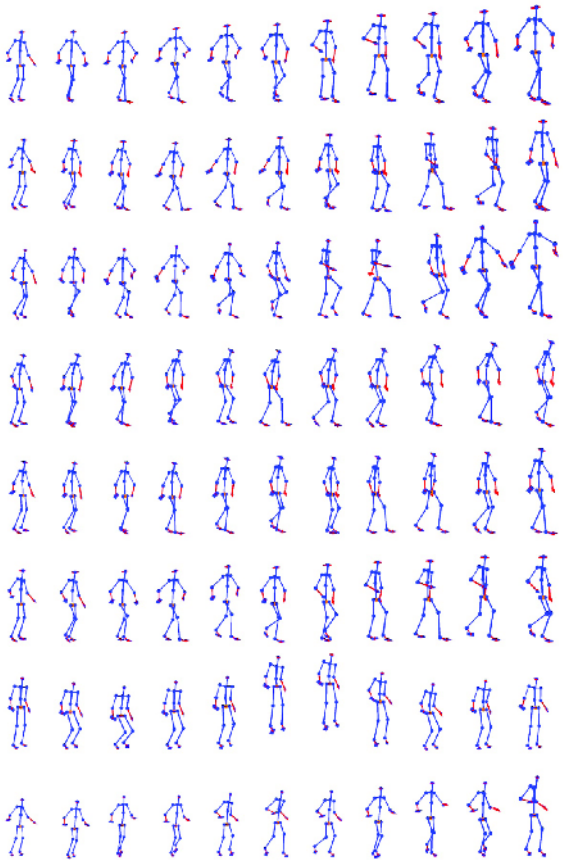

Figure 4. From top to bottom: Masculine Walk, Feminine Walk, Young Walk, Old Walk, Tired Walk, Energetic Walk, Masculine Jump, Masculine Run.

To test the two proposed segmentation techniques, nine of the eighteen sequences contain one class of the mentioned actions and a number of meaningless frames before and after the target segment while another nine set of the sequences contain a meaningful action (which the system is not trained for) before or after the intended action. Successive to training, segmentation using NN is carried out and evaluated with respect to manual segmentation of the actions. The error is calculated by the difference of the true starting frame number and the frame number recognized by the system as the starting frame of an action. Similarly the error for the ending frame is determined and the average error is calculated. This is presented in Table 2. We can also conclude that locating an action sequence from within a sequence containing excessive meaningless data is less complicated compared to the cases where other action sequences have been added to the original action. This is simply because the added action class may contain similar frames to the target action which might confuse the locating algorithm. This trend holds true for most of the test samples except for few where the error has increased by a small amount.

The second proposed method is through employing HMM. The same HMM which is later used for classification of actions is utilized for segmentation of the actions. This is done by selecting an adjustable sliding window which scans through the motion sequence, returning the window with the most likelihood as the segmented action. Table 2 shows that this approach is more accurate compared to NN search.

Table 2. Segmentation using NN and HMM

Sample	Action Class	Action Style	Additional Frames	NN: Mean Frame Difference	HMM: Mean Frame Difference
1-5	Walk	Feminine	Meaningless	6.0	2.6
6-10	Walk	Masculine	Punch	6.8	1.7
11-15	Walk	Energetic	Meaningless	7.6	2.3
16-20	Walk	Tired	Kick	5.6	0.8
21-25	Walk	Young	Meaningless	0.8	0.9
26-30	Walk	Old	Pickup	9.4	1.2
			Mean Error	6.03	1.58
31-35	Run	Feminine	Meaningless	2.0	1.7
36-40	Run	Masculine	Punch	3.2	2.3
41-45	Run	Energetic	Meaningless	9.8	0.8
46-50	Run	Tired	Kick	8.2	2.6
51-55	Run	Young	Meaningless	4.6	2.8
56-60	Run	Old	Pickup	9.0	2.2
			Mean Error	6.13	2.07
61-65	Jump	Feminine	Meaningless	5.8	2.8
66-70	Jump	Masculine	Punch	6.0	2.4
71-75	Jump	Energetic	Meaningless	0.2	2.3
76-80	Jump	Tired	Kick	5.2	0.7
80-85	Jump	Young	Meaningless	8.4	1.3
86-90	Jump	Old	Pickup	10.8	0.9
			Mean Error	6.07	1.73
			SD:	2.06	1.03
			Overall Mean Error	6.08	1.79

The actions which have been previously segmented are not directly fed to the classification subsystems for classification, since any amount of error in locating and segmenting the actions would result in unreal results of the classification system. The goal here is to perform a correct evaluation of the proposed classification techniques, therefore, manually segmented actions are used for a correct evaluation.

Similar to segmentation, 5 samples for each of the 18 action sequences are provided for evaluation. These 90 samples have not been used in the training process of the classification subsystem and are new to the system. All the primary and secondary theme combinations are included in the samples.

NN is the first technique which we used for classification of actions. This is done by calculating the numerical distance between the test action sequence and the training action set. The training sample which returns the least distance, determines both the primary and secondary class of the action. The drawback of the proposed method is the fact that the length of the test sample must be exactly equal to all the training samples for acquiring the best results, which is an impossible requirement. Thus, the accuracy rate in some cases drops to as low as 20% which is considered a significant decrease. This can be remedied using dynamic time warping (DTW). The goal, however, is to compare the performance of each technique without significant pre-processing.

Table 3 presents the classification results. It is observed that the accuracy of the action classification is considerably higher than that of style classification. This is not unexpected since the secondary theme features are quite insignificant and numerically small compared to the primary theme features.

Next, HMMs are employed for classification. Successive to training, the test samples are fed to each of the 18 HMM networks and classified based on maximum log likelihood measures.

Table 3. Classification using NN, HMM, and ANN

Sample	Action Class	Action Style	NN: Primary Theme	NN: Secondary Theme	HMM: Primary Theme	HMM: Secondary Theme	ANN: Primary Theme	ANN: Secondary Theme
1-5	Walk	Feminine	80%	80%	100%	80%	80%	60%
6-10	Walk	Masculine	80%	20%	80%	60%	80%	80%
11-15	Walk	Energetic	80%	60%	80%	80%	100%	80%
16-20	Walk	Tired	60%	60%	60%	60%	80%	80%
21-25	Walk	Young	60%	40%	80%	60%	80%	80%
26-30	Walk	Old	80%	40%	80%	60%	80%	60%
	Mean Accuracy:		73.33%	50.00%	80.00%	66.67%	83.33%	73.33%
31-35	Run	Feminine	80%	60%	60%	40%	80%	80%
36-40	Run	Masculine	60%	40%	60%	60%	80%	80%
41-45	Run	Energetic	60%	40%	60%	60%	80%	60%
46-50	Run	Tired	80%	20%	80%	40%	80%	80%
51-55	Run	Young	60%	60%	80%	60%	100%	100%
56-60	Run	Old	60%	60%	40%	40%	60%	60%
	Mean Accuracy:		66.67%	46.67%	63.33%	50.00%	80.00%	76.67%
61-65	Jump	Feminine	100%	60%	100%	60%	100%	80%
66-70	Jump	Masculine	100%	60%	100%	80%	100%	100%
71-75	Jump	Energetic	100%	80%	100%	80%	100%	100%
76-80	Jump	Tired	100%	20%	100%	60%	100%	60%
81-85	Jump	Young	100%	60%	100%	60%	100%	80%
86-90	Jump	Old	100%	60%	100%	60%	100%	80%
	Mean Accuracy:		100.00%	63.33%	100.00%	66.67%	100.00%	83.33%
	SD:		16.80%	18.43%	18.75%	12.78%	12.15%	13.53%
Overall Mean Accuracy:			80.00%	53.33%	81.11%	61.11%	87.78%	77.78%

Table 3 illustrates that he overall action (primary theme) classification accuracy using HMM shows a slight improvement compared to the nearest neighbour search. Style (secondary theme) classification, is also more accurate than the nearest neighbour technique. Similar to NN, style classification is carried out with lower accuracy compared to the primary theme classification.
With regards to the accuracy of different primary and secondary themes being classified, the performance of the system shows to be superior for Jump when the primary theme is to be classified, while Walk and Jump show similar performances and higher than that of Run when being classified for secondary themes.

The third and last proposed method for classification of actions is ANN. A network of 21 ANNs, 3 networks for primary action classes and 18 for secondary classes, in the form of MLP for each action and different styles are constructed and trained using the Resilient Back-Propagation (RPROP) technique. Two hidden layers with 45 to 90 neurons in each hidden layer were provided for each network. Various numbers of experiments proved the final orientation of 45, 70, 50, 45 neurons for layers one to four to be most effective. For different ANN blocks, slight changes were made on the number of hidden neurons based on the number of frames of actions used to train the respective ANN. Training was then carried out in all the neural networks using the RPROP training technique as well as Batch training with weight & bias learning rules, Powell-Beale conjugate gradient back-propagation, Gradient descent back-propagation, Scaled conjugate

gradient back-propagation, and Levenberg-Marquardt back-propagation, where none proved to be more precise for this application compared to RPROP. The RPROP, however, displayed the most number of local error maxima during training, which could result in erroneous results if the training stops within one of the maxima. Providing the networks with large number of training epochs significantly reduces the probability of being trapped in local error maxima. Due to the large size of ANN inputs, outputs, and the ANNs themselves, up to 25000 training epochs were completed. Also different training functions were tested for different layers where Log-Sigmoid, Tan-Sigmoid, Tan-Sigmoid, and Pure-Linear were adopted for layers one to four, respectively. Test samples are then fed to each network and the networks showing the least MSE determine the classification result. Table 3 presents the results for classification of 90 samples (similar to previous sections). In this table, jumping has been classified with greater accuracy for both primary and secondary themes. Also generally, classification of the primary themes has been more accurate with respect to that of secondary themes.

Figure 5 illustrates the segmentation results for different action lasses using NN and HMM. Evidently, the overall segmentation performance of the HMM is more accurate than that of the NN. While the HMM shows a more robust routine with respect to different primary and secondary themes, the variations of NN is significant in this regard.

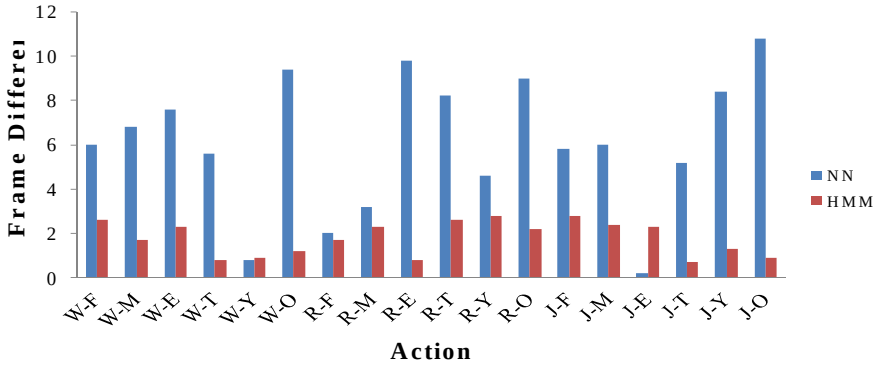

Figure 5. Segmentation error rate for different actions using NN and HMM.

Figure 6 illustrates the classification accuracy for each of the 18 classes carried out by means of the 3 classifiers and for different themes. This figure enables us to investigate each of the 18 actions and determine the actions for which the classifiers perform with higher or lower accuracy.

Based on Figure 6, it can be determined that classification of the primary theme *jump* takes place with absolute robustness for all classifiers. This is due to its distinct features with respect to walk and run. Classification of the secondary themes for all three classes of action is carried out with a higher error rate with respect to their respective primary themes. This was anticipated since these primary theme features are minute and more difficult for the classifiers to recognize.

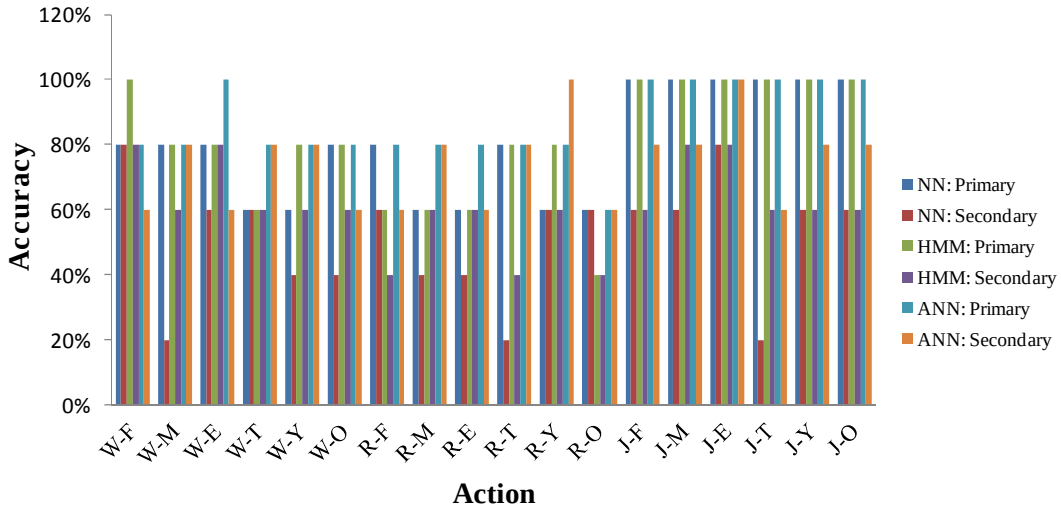

Figure 6. Classification accuracy for different primary and secondary themes using NN, HMM, and ANN.

Figures 7 and 8 are derived from Figure 5 and Figure 6, which illustrate the classification performance of each method for segmentation and classification of primary and secondary themes (regardless of the action type), as well as the standard deviation obtained by the different methods. From Figure 7 it can be concluded that segmentation of actions is carried out with higher accuracy and a smaller deviation when HMM is employed. The NN performs with a significantly higher error rate and with a much larger standard deviation when compared to the HMM technique. This can also be verified from Figure 5 where the higher error and variations in the NN performance are noticeable.

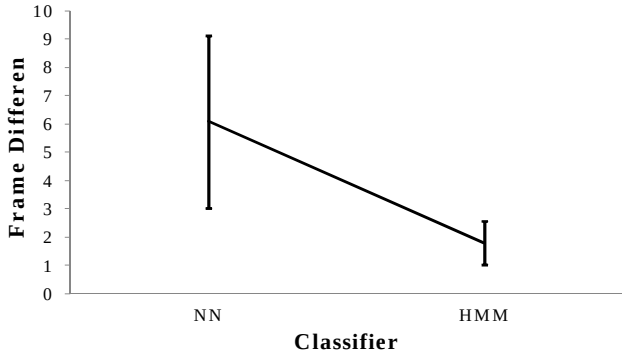

Figure 7. Segmentation error rate and standard deviation for NN and HMM.

Figure 8 presents the accuracy and standard deviations of classification of primary and secondary themes using the three discussed methods. We conclude that overall the accuracy of primary theme classification is significantly higher than that of secondary themes. Also each technique has a higher correct classification rate for primary themes with respect to secondary themes. The ANN is the most accurate means for classification with the least deviation for primary themes, the HMM however, shows a lower deviation for secondary themes. The improvement shown for classification of secondary themes shows a more significant improvement when carried out by means of HMM instead of NN. This however, is not the case for the primary theme where NN and HMM perform almost alike.

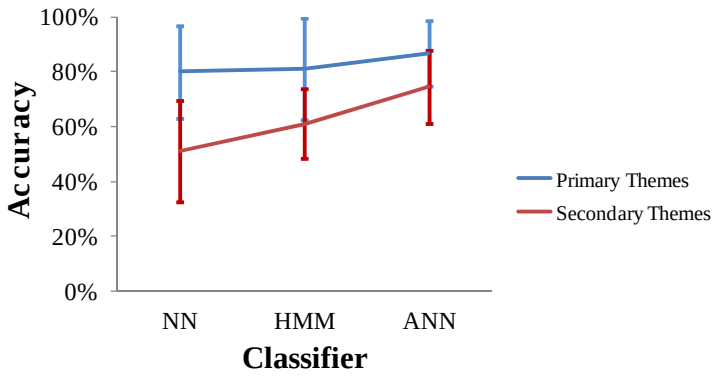

Figure 8. Classification accuracy and standard deviation for NN, HMM, and ANN for primary and secondary themes.

6. CONCLUSION

The goal of this paper has been to conduct a conclusive and comprehensive study of different techniques used for segmentation and classification of actions using motion capture data. The primary themes in the motion data as well as the secondary themes resulting from actor characteristics have both been subject to classification. The findings of this study can largely be utilized for interactive (as well as non-interactive) multimedia and data base applications. The data employed for this research have been captured by means of a Vicon motion capture system. Different actors have been asked to perform the needed actions. Pre-processing is carried out on the data for noise elimination, dimensionality reduction, quantization, and vectorization of the data.

The focus of the first section of this research has been on locating and segmenting the actions by means of NN and HMM, where HMM proved to be more robust for different primary and secondary themes and performed with a higher overall accuracy as well. The second task of this research has been classification of individual actions for both primary and secondary themes. Three different classifiers were employed: NN, HMM, and ANN. The NN classifier performs with higher error rate and with a larger deviation for both themes. Therefore the use of NN is ruled out for classification of motion capture data. The HMM, for classification of primary themes, performs with a similar accuracy to that of the NN. For secondary themes, however, the accuracy increases noticeably. The deviation of the HMM for both themes is also smaller than that of the NN which is a desired trend. The last technique used for classification is ANN. RPROP training is used for the ANN which proved to be a suitable method for this purpose. This technique proved more accurate for classification of both themes. The deviation is also decreased with respect to HMM for primary theme classification. The deviation of secondary theme classification using ANN however, is increased which is undesired. Overall for segmentation of motion capture data, HMM is suggested over NN. For classification purposes, ANN shows a higher accuracy for both themes; However if a lower deviation is desired, for primary themes, ANN and for secondary themes, HMM is the suitable choice of classifier.

REFERENCES

[1] A. Hutchinson, (1996), Labanotation, Dance Books.

[2] T. B. Moeslund & E. Granum, (2001) "A Survey of Computer Vision-Based Human Motion Capture", Computer Vision and Image Understanding, Vol. 81, No. 3, pp. 231-268.

[3] D. M. Gavrila, (1999) "The Visual Analysis of Human Movement: A Survey", Computer Vision and Image Understanding, Vol. 73, No. 1, pp. 82-98.

[4] A. Arya, N. Nowlan, & N. Sauriol, (2010) "Data-Driven Framework for an Online 3D Immersive Environment for Educational Applications", Proceedings of the International Conference on Education and New Learning Technologies.

[5] M. N. Boulos, L. Hetherington, & S. Wheeler, (2007) "Second Life: an overview of the potential of 3-D virtual worlds in medical and health education", Health Information & Libraries Journal, Vol. 24, pp. 233-245.

[6] S. A. Etemad & A. Arya, (2009) "3D Human Action Recognition and Style Transformations Using Resilient Back-propagation Neural Networks", Proceedings of 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 296-301.

[7] B. Carelsen, R. Jonges, S. D. Strackee, M. Maas, P. van Kemenade, C. A. Grimbergen, M. van Herk, & H. J. Streekstra, (2009) "Detection of In Vivo Dynamic 3-D Motion Patterns in the Wrist Joint", IEEE Transactions on Biomedical Engineering, Vol. 56, No. 4, pp. 1236-1244.

[8] F. Lv & R. Nevatia, (2007) "Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8.

[9] F. Zhou, F. De la Torre, & J. K. Hodgins, (2008) "Aligned Cluster Analysis for Temporal Segmentation of Human Motion", 8th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1-7.

[10] S. Ali, A. Basharat & M. Shah, (2007) "Chaotic Invariants for Human Action Recognition", 11th IEEE International Conference on Computer Vision, pp. 1-8.

[11] D. Kulic, W. Takano, & Y. Nakamura, (2007) "Incremental on-line hierarchical clustering of whole body motion patterns", 16th IEEE International Symposium on Robot and Human interactive Communication, pp. 1016-1021.

[12] Y. Song, L. Goncalves, & P. Perona, (2001) "Learning Probabilistic Structure for Human Motion Detection", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,Vol. 2, pp. 771-777.

[13] M. Shimosaka, T. Mori, T. Harada, & T. Sato, (2005) "Marginalized Bags of Vectors Kernels on Switching Linear Dynamics for Online Action Recognition", Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 3072-3077.

[14] R. Ishiyama, H. Ikeda, & S. Sakamoto, (2006) "A Compact Model of Human Postures Extracting Common Motion from Individual Samples", Proceedings of the 18th International Conference on Pattern Recognition, Vol. 1, pp. 187-190.

[15] T. Mori, Y. Nejigane, M. Shimosaka, Y. Segawa, T. Harada, & T. Sato, (2005) "Online Recognition and Segmentation for Time-Series Motion with HMM and Conceptual Relation of Actions", 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3864-3870.

[16] V. Parameswaran & R. Chellappa, (2003) "View Invariants for Human Action Recognition", Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 613-619.

[17] M. Brand & V. Kettnaker, (2000) "Discovery and segmentation of activities in video", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 844-851.

[18] X. Li & K. Fukui, (2008) "View Invariant Human Action Recognition Based on Factorization and HMMs", IEICE Transactions on Information and Systems, Vol. E91-D, No. 7, pp. 1848-1854.

[19] A. Madabhushi & J. K. Aggarwal, (2000) "Using head movement to recognize activity", Proceedings of the 15th International Conference on Pattern Recognition, Vol. 4, pp. 698-701.

[20] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, & A. Y. Wu, (1998) "An optimal algorithm for approximate nearest neighbor searching fixed dimensions", Journal of ACM, Vol. 45, No. 6, pp. 891-923.

[21] S. Ali & M. Shah, (2009) "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 2, pp. 288-303.

[22] A. A. Efros, A. C. Berg, G. Mori, & J. Malik, (2003) "Recognizing Action at a Distance", Proceedings of the 9th IEEE International Conference on Computer Vision, Vol. 2, pp. 726-733.
[23] L. Wang, (2006) "Abnormal Walking Gait Analysis Using Silhouette-Masked Flow Histograms", 18th International Conference on Pattern Recognition, Vol. 3, pp. 473-476.
[24] L. R. Rabiner, (1989) "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of IEEE, Vol. 77, No. 2, pp. 257-286.
[25] G. D. Jr. Forney, (1973) "The viterbi algorithm", Proceedings of IEEE, Vol. 61, No. 3, pp. 268-278.
[26] D. Chester, (1990) "Why two hidden layers are better than one", Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 265-268.
[27] N. Wanas, G. Auda, M. S. Kamel, & F. Karray, (1998) "On the optimal number of hidden nodes in a neural network", IEEE Canadian Conference on Electrical and Computer Engineering, Vol. 2, pp. 918-921.
[28] M. Riedmiller & H. Braun, (1993) "A direct adaptive method for faster backpropagation learning: the RPROP algorithm", 1993 IEEE International Conference on Neural Networks, Vol. 1, pp. 586-591.
[29] E. Besdok, (2007) "Neurovision with Resilient Neural Networks", Advances in Visual Information Systems, Springer Berlin/Heidelberg, Vol. 4781/2007, pp. 438-444.
[30] S. A. Etemad, A. Arya, & A. Parush, (2011) "Spatial Perceptual Weights of Energy-related Features in Animation of Human Motion", Proceedings of Computer Graphics International, S15.