Robocontrol 2012
5th WORKSHOP IN APPLIED ROBOTICS AND AUTOMATION
unesp

RWE PATTERNS EXTRACTION FOR ON-LINE HUMAN ACTION RECOGNITION THROUGH WINDOW-BASED ANALYSIS OF INVARIANT MOMENTS

Dennis Romero L.[1 2] , *Boris X. Vintimilla*[2] , *Anselmo Frizera Neto*[1] , *Teodiano Bastos Filho*[1]

[1]Electrical Engineering Department, Federal University of Espirito Santo, Vitoria - Brasil, {dennis;anselmo;tfbastos}@ele.ufes.br
[2]CIDIS - FIEC, Escuela Superior Politecnica del Litoral, Guayaquil - Ecuador, {boris.vintimilla;dromerol}@espol.edu.ec

Abstract: This paper presents a method for on-line human action recognition on video sequences. An analysis based on Mahalanobis distance is performed to identify the "idle" state, which defines the beginning and end of the person movement, for posterior patterns extraction based on Relative Wavelet Energy from sequences of invariant moments.

Keywords: Human action recognition, Relative Wavelet Energy, Window-based temporal analysis.

1. INTRODUCTION

The identification of people movement represents a wide research area, and studies are leading to new approaches in video-based human activity recognition in robotics and human-machine interaction. Moreover, the processing capabilities of new mobile devices are allowing to execute more complex applications, that could provide input data for other devices, especially those developed for human interaction, trend that is increasingly common in several areas, such as motion analysis for providing feedback in patient's physiotherapy, robot control by gesture recognition, bio-mechanical visual analysis, among others. The initial study about the state of the art has identified recent works in the area, as mentioned in [1–4]. Initially, we studied methods for human motion analysis on video sequences, and was referred the problem of extracting features during different activities such as walking, jumping, running, etc. However, the diversity and complexity of human movements led us to seek approaches less restrictive to specific movements, as well as find a better way to extract representative patterns from the found movements. The automatic identification and analysis of people motion involves addressing different sub-problems, such as image segmentation and tracking features. This represents a complex problem due to the parameters variability that conforms the video scene. A single solution would present difficulties adapting the segmentation parameters if the application goals changes. However, there are interesting approaches for tracking [2], in which the target is divided into multiple regions or fragments and after that represented by a Gaussian mixture model (GMM) in a joint feature-spatial space. Modeling of target and background are performed according to Chan-Vese algorithms [5], and the extracted target boundaries are used to learn the dynamic shape of the target over time, keeping the object tracking even in cases of total occlusion. Trivedi et.al. presented in [6] a synergistic approach on the person's body based on the hierarchy of actions, considering static pose, dynamic gestures and body part actions during persons activities, using the standard Baum-Welch algorithm and the Viterbi algorithm [7] in order to encode independent Hidden Markov Models (HMMs). In this way, defined models for the lower body are represented as: Q1 = {"idle", "walk", and "kick"}; Q2 corresponding the set of models for Torso: Q2 = {"idle", "moving left","moving right","moving up","moving down"} and Q3 containing the models for the arms, where Q3={"idle","stretch out", and "withdraw"}, and introducing the concept of *spatio-temporal personal space* to address the different behaviors of people. In this work patterns (9-dimensional) are obtained from three sequences of invariant moments. Our work proposes a method for patterns extraction in time series of features, which could be generated from different tracking methods, like those mentioned above. In this work the (9-dimensional) patterns are obtained from three sequences of invariant moments.

On the other hand, studies based on invariant moments and wavelet transform were performed by [8] for automatic image registration, applied in matching of two different images. Image registration estimates the parameters of the geometric transformation model that maps the sensed images back to its reference image. Feature points from both images are extracted using Mexican-Hat wavelet and control-points is achieved with invariant moments. The properties of the wavelet transform has been recognized and its applications are varied on different research areas, as in [9, 10] where makes use of Relative Wavelet Energy for selecting features in brain electrical activity analysis from EEG (Electroencephalography) signals. In [10] the signal dimensionality is reduced by Linear Discriminant Analysis (LDA) and RWE patterns are obtained for the subsequent classification using Support Vector Machines (SVM).

The first section of this paper describes the set of input data, represented by binary images of people in motion. After that, the criterion used for on-line "idle" state identification on invariant moments sequences is detailed, also the approach used for motion's windows capturing and patterns extraction of RWE (Relative Wavelet Energy) to classify movements using Artificial Neural Networks (ANNs). Finally, we present the results, conclusions and future work.

2. METODOLOGY

As mentioned, there are currently robust algorithms for segmentation and people tracking. In our proposal these methods could provide the input data, corresponding to

images of person in motion, which it's subsequently possible to isolate it from the background. For this reason, the initial studies of this proposal were based on the database Muhavi-Mas (Multicamera Human Action Video Data - Manually Annotated Silhouette) proposed by Velastin [11], to evaluate methods of actions recognition. However, in further studies was used a LIDAR sensor (Light Detection And Ranging also LADAR) which provides depth information to contribute in the segmentation process. This made it possible to generate binary silhouettes of people for the desired goals (Figure 1).

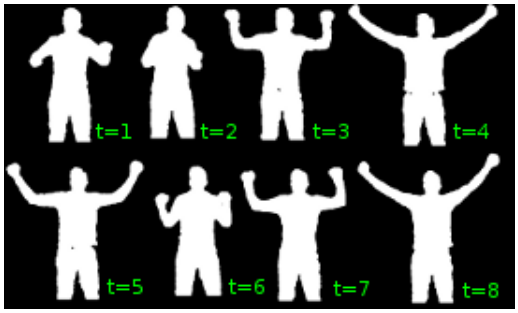

Figure 1: Person silhuette at eight time instants.

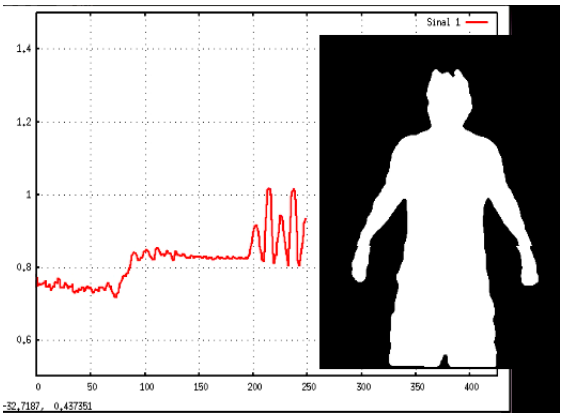

Figure 2: One Hu's moments sequence at different movements.

2.1. By-frame feature extraction

There are several types of invariant moments. Hu's moments [12] are the most commonly used in the literature and the best in terms of their orthogonality, rotation invariance, low sensitivity to image noise, fast computation and the ability to provide a faithful image representation [13]. That's the reason they had been widely applied in image recognition [14, 15]. Hu's moments are region-based invariant features that consider all the image pixels [14]. These seven moments have invariant properties to affine transformations, including changes in scale, translation and rotation, for these reasons we had chosen them for our work. The Figure 2 show one sequence of Hu's moments during at different movements. In order to identify movement patterns, specifically were considered the 2nd, 3rd and 4th moment for each image. The first moment was dismissed for having redundancy with the other moments and for being more sensitive to image noise. The last moment was not considered due to achieving small values in certain movement conditions. Figure 3 shows the moments mentioned during the execution of three different movements ("left arm up-down", "hello signal", "opening-closing arms").

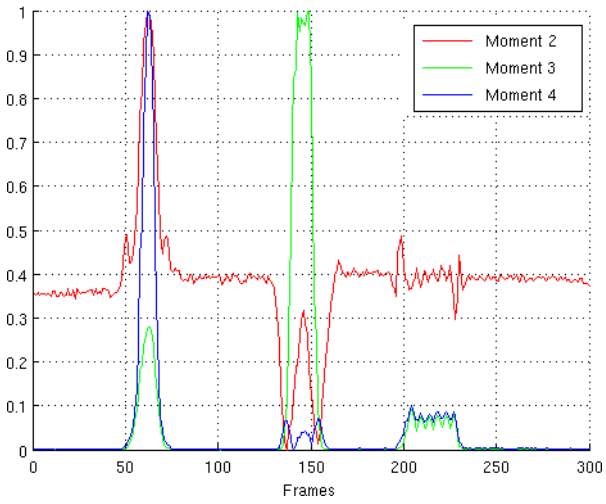

Figure 3: 2nd, 3rd and 4th Hu's moments, in 3 different movements

In previous studies[16], we performed a stochastic analysis of Hu's invariant moments, where it was possible to identify stationary characteristics after applied the 1st derivative [17]. However, in order to preserve and facilitate the identification of detailed moments variations as well as reduce the computational cost, the application of the 1st derivative was not considered. These mentioned variations include a proportional increment or decrement of the Hu's moments after upper and lower body movements, respectively.

2.2. "idle" state identification

The extracted invariant moments for each image allowed us to represent the person's posture in time instants of his movement, on this work, in a three-dimensional space $Pos_{\delta t} = [\phi_1, \phi_2, \phi_3]$, where ϕ represents moments with order 2, 3 and 4. For a time interval τ, we consider each dimension as unidimensional signal, where $S_i = [\phi_{i0}, \phi_{i1}, \phi_{i2}, \phi_{i3}, ..., \phi_{i\tau-1}]$. This unidimensional approach is used to identify the "idle" state, in which just S_4 is considered for verifying initial and ending postures of person movement, because it presents more sensibility on movement presence.

Since invariant moment sequences are updated approximately at the camera's frame rate (25-30 fps), a temporal-windows analysis to identify the beginning and end of the person action was considered. The temporal window's size was defined by optimizing the Mahalanobis distance (1), which can be defined as a dissimilarity measure between two random vectors $\vec{W}_{i-1}$ and $\vec{W}_i$ with the same distribution, and with covariance matrix Σ:

$$d(\vec{W}_{i-1}, \vec{W}_i) = \sqrt{(\vec{W}_{i-1} - \vec{W}_i)^T \Sigma^{-1} (\vec{W}_{i-1} - \vec{W}_i)} \quad (1)$$

The Mahalanobis distance has been used in several applications, such as [18], where it was applied in signature recognition. In the Figure 3 it's possible to observe an homogeneous 4th Hu's moment at movement absence, in this situation the Mahalanobis distance may take value lower than 0.007, under controlled light conditions. However, the tolerance's parameter for "idle" identification may vary depending on light conditions or noise in the capture process. Five frame windows were enough to establish distance measurements that support noise in the capture and binarization process, providing a significant marking of the beginning and end of motion. These windows W are treated on-line as shown in Figure 4.

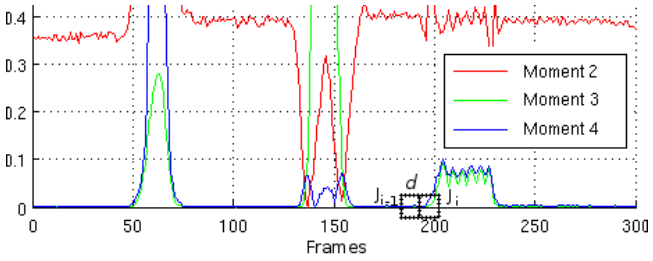

Figure 4: Ilustration, temporal windows of 4th Hu's moment, in a previous instant to the movement.

The Mahalanobis distance method was compared with the auto-correlation criterion between temporal windows, getting more consistent measures with Mahalanobis. The 5-size windows allows to identify a "idle" state in $size_{win}/fps \approx 0.2seg$.

2.3. *Movement windows extraction*

Once identified the movement beginning (which was determined with the 4th Hu's moment sequence), the new moment's vectors (3-dimensional) are temporarily stored into dynamic windows, between the beginning mark $t - size_{win}$ until the end mark defined by the new "idle" state. This allows that the corresponding motion data could be centered into the window, as shown in Figure 5

2.4. *Movement pattern extraction*

Due to the periodic nature of certain human movements (walking, running, jumping, etc.), in the early stages of this research a frequency analysis was considered, by using the Fast Fourier Transform (FFT) after application of the first derivative to the invariant moments sequence. However, the discrete wavelet transform (DWT) specifically allows the discrimination of non-stationary signals with different frequency characteristics, so the DWT was chosen with Daubechies (db4) family, as a method for posterior features

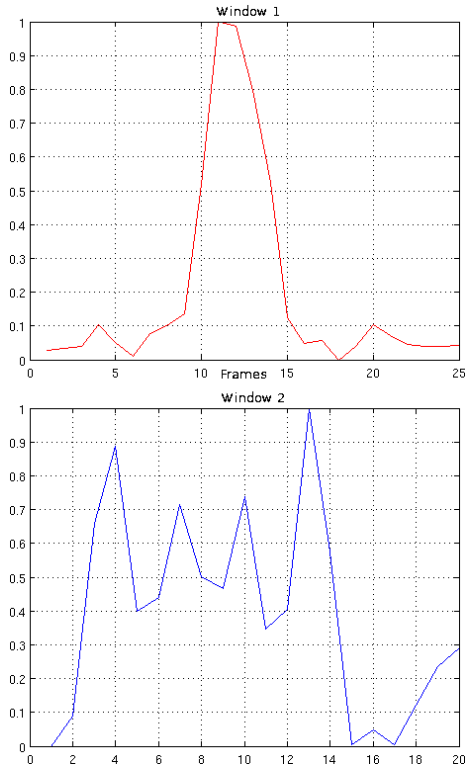

Figure 5: On-line extracted windows during two different movements

extracting of person's movement. The Daubechies wavelet family was chosen by the following properties:

a) Time invariance - If the time series is shifted in time, then its wavelet coefficients are only shifted in time.

b) Fast computation - Daubechies wavelet has self-similar type fractal that directs fast wavelet transform techniques.

c) Filter sharp transition bands - Daubechies wavelet has sharp transition bands which minimize edge effects between the frequency bands.

The DWT is a transformation of the original temporal signal into a wavelet basis space. The time-frequency wavelet representation is performed by repeatedly filtering the signal with a pair of filters that cut the frequency domain in the middle. Specifically, the DWT decomposes a signal into an approximation signal and a detail signal. The approximation signal is subsequently divided into new approximation and detail signals. This process is carried out iteratively producing a set of approximations signals at different detail levels and a final gross approximation of the signal [19].

The detail D_j and the approximation A_j at level j can be obtained by filtering the signal with an L-sample high pass filter g, and an L-sample low pass filter h. Both approximation and detail signals are down-sampled by a factor of two. This can be expressed as follows:

$$A_j[n] = \mathbf{H}\langle A_{j-1}[n]\rangle = \sum_{k=0}^{L-1} h[k]A_{j-1}[2n-k], \quad (2)$$

$$D_j[n] = \mathbf{G}\langle D_{j-1}[n]\rangle = \sum_{k=0}^{L-1} g[k]A_{j-1}[2n-k], \quad (3)$$

where $A_0[n], n = 0, 1, ..., N-1$ is the original temporal sequence, while **H** and **G** represent the convolution/down sampling operators. Sequences $g[n]$ e $h[n]$ are associated with wavelet function $\psi(t)$ and the scaling function $\varphi(t)$ through inner products:

$$g[n] = \langle \psi(t), \sqrt{2}\psi(2t-n)\rangle, \quad (4)$$

$$h[n] = \langle \varphi(t), \sqrt{2}\varphi(2t-n)\rangle. \quad (5)$$

As mentioned above, the selected wavelet $\psi(t)$ is Daubechies 4 (db4) with 3 decomposition levels, and since it is an orthonormal basis for L, the concept of energy is linked with the usual notions derived from Fourier's theory. The wavelet coefficients are given by $C_j(k) = \langle Win_{mov}, \psi_{j,k}\rangle$ which can be interpreted as the local residual errors between successive signal approximations at scales j and $j+1$, and the energy, at each level of decomposition $j = -1, ..., -N$, will be the energy of the detail signal [20],

$$E_j = \|r_j\|^2 = \sum_k |C_j(k)|^2 \quad (6)$$

where $r_j(t)$ is the residual signal at scale j, and the energy at time instant k will be:

$$E(k) = \sum_{j=-N}^{-1} |C_j(k)|^2 \quad (7)$$

Consequently, the total energy can be obtained by,

$$E_{total} = \|Win_{mov}\|^2 = \sum_{j<0}\sum_k |C_j(k)|^2 = \sum_{j<0} E_j \quad (8)$$

Finally, the normalized values p_j are defined. They represents the relative wavelet energy, where $\sum_j p_j = 1$ and the distribution $\{p_j\}$ can be considered as a time-scale density, constituing a suitable tool to detect and characterize specific phenomena in time and frequency planes [9].

$$p_j = \frac{E_j}{E_{total}} \quad (9)$$

The relative energy of three decomposition levels is calculated for each movement's window, it results in a 9-dimensional pattern. The Figure 6 presents the mentioned procedure's diagram:

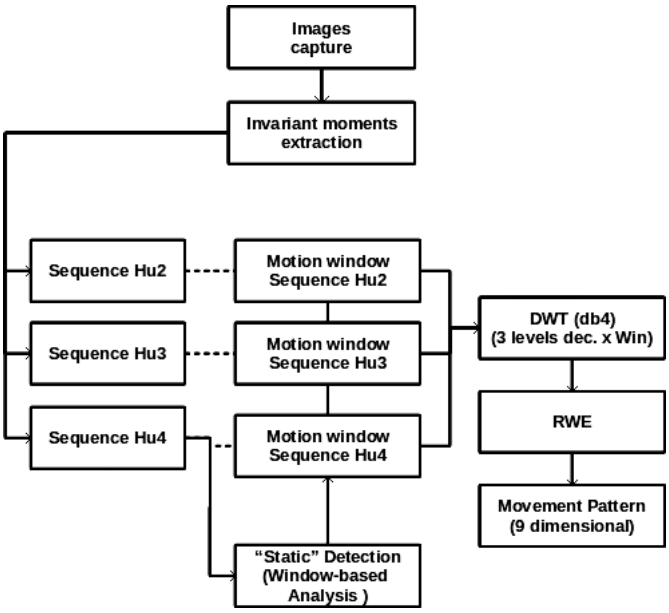

Figure 6: Proposed methodology Diagram

3. RESULTS

The proposed method of temporal windows analysis, allowed the dynamic feature extraction of person's movements, facilitating the obtaining of motion windows that can be directly processed for feature extraction. Figure 7 shows screenshots of the developed application with the explained methods, for on-line identification of different duration movements. We proposed an on-line approach window-based for feature extraction from tracking data of different movements. As input data images captured from a LIDAR Sensor were used to get a better person silhouette, and invariant region-based moments to extract the initial features for tracking. The Figure 7 show movement patterns extracted after completed three different actions. In the training phase, patterns are stored in a relational database, and they are used in a test phase, where each new movement patterns is presented to the neural network for prediction. The Figure 8 presents the confusion matrices, considering 60 patterns aprox. for each movement, in which were divided for training (50%), validation (25%) and testing (25%). In the test confusion matrix it's possible to appreciate some misclassification, mainly between the "up-down arm" and "opening-closing arms" movements (1st and 3rd class at Figure 8) due to presence of signal similarities (ex. one signal burst vs two bursts that could be the first one in two small time instants), this happen with closer kind of movements, which is less likely to occur with more complex movements or longer duration, as such as movement 2, which is the greeting gesture by shacking one hand in the air.

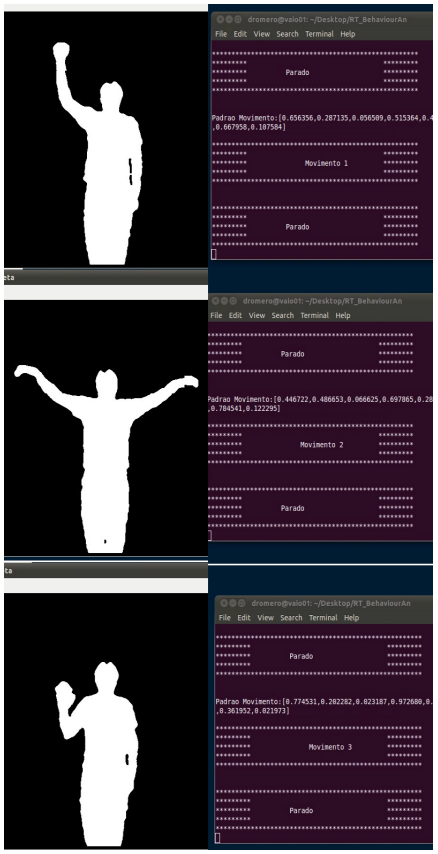

Figure 7: Screenshot during an action recognition test

The wavelet transform and the relative energy computed from the decomposition levels have allowed a rapid features extraction for movement classification. The patterns dimensionality (3 levels for each movement window) joined to the wavelet transform wavelet properties, give as result valuable patterns for movement identification. The Figure 8 show a confusion matrix from classification of three different movement using a Multilayer perceptron (MLP) neural network [21].

4. Conclusions and Future Work

This paper presented a methodology for on-line human actions recognition on video sequences. The on-line action recognition is a step for further studies on vision-based human tracking and activity recognition in robotics and human-machine interaction. The method used for actions recognition could be applied to different multi-dimensional tracking data, in this case was applied on 3 sequences of Hu's moments, obtaining 9-dimensional patterns for classification, in this case, using artificial neural networks (ANN's).

As a future work, an probabilistic approach to track different actions will be study, in order to reach a higher level of human-machine interaction and contribute to studies on automatic human behavior analysis.

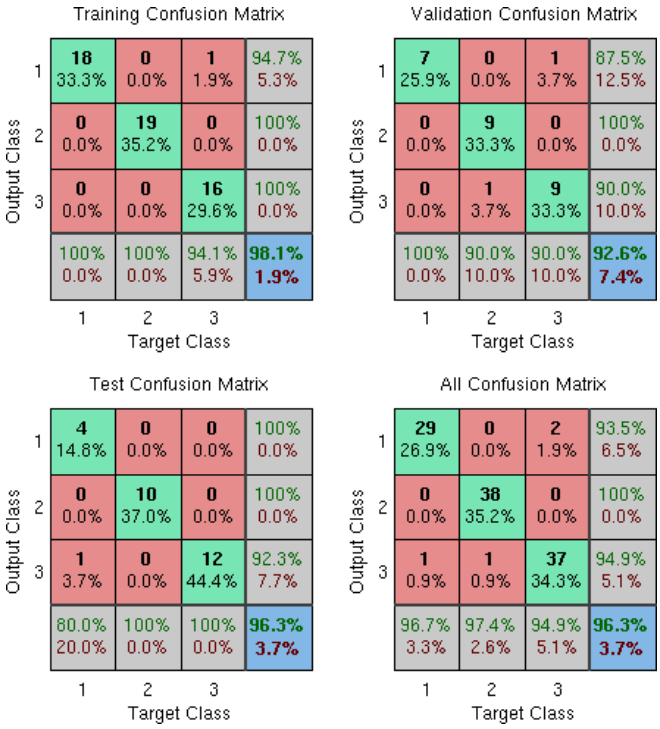

Figure 8: Confusion matrix from classification of three different movements

References

[1] Hee-Deok Yang, A-Yeon Park, and Seong-Whan Lee. Gesture spotting and recognition for human ndash;robot interaction. *Robotics, IEEE Transactions on*, 23(2):256 –270, april 2007.

[2] P. Chockalingam, N. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1530 –1537, 29 2009-oct. 2 2009.

[3] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram. A novel approach for recognition of human actions with semi-global features. *Machine Vision and Applications*, 19:27–34, 2008. 10.1007/s00138-007-0074-2.

[4] Sangho Park and Mohan M. Trivedi. Understanding human interactions with track and body synergies (tbs) captured from multiple views. *Comput. Vis. Image Underst.*, 111(1):2–20, July 2008.

[5] T.F. Chan and L.A. Vese. Active contours without edges. *Image Processing, IEEE Transactions on*, 10(2):266 –277, feb 2001.

[6] Sangho Park and Mohan Trivedi. Multi-person interaction and activity analysis: a synergistic track- and body-level analysis framework. *Machine Vision and Applications*, 18:151–166, 2007. 10.1007/s00138-006-0055-x.

[7] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 –286, feb 1989.

[8] Jignesh N. Sarvaiya. Automatic image registration using mexican hat wavelet, invariant moment, and radon transform. *IJACSA - International Journal of Advanced Computer Science and Applications*, (Special Issue):75–84, 2011.

[9] O.A Rosso, M.T Martin, and A Plastino. Brain electrical activity analysis using wavelet-based informational tools (ii): Tsallis non-extensivity and complexity measures. *Physica A: Statistical Mechanics and its Applications*, 320(0):497 – 511, 2003.

[10] Zhao Haibin, Wang Xu, and Wang Hong. Feature selection using relative wavelet energy for brain-computer interface design. In *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, pages 1434 –1437, may 2008.

[11] S. Singh, S.A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 48 –55, 29 2010-sept. 1 2010.

[12] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179 –187, february 1962.

[13] Zhihu Huang and Jinsong Leng. Analysis of hu's moment invariants on image scaling and rotation. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, volume 7, pages V7–476 –V7–480, april 2010.

[14] Rafael C. Gonzalez. *Digital Image Processing, 2nd Edition*. McGraw-Hill, 2010.

[15] Qing Chen, Emil Petriu, and Xiaoli Yang. A comparative study of fourier descriptors and hu's seven moment invariants for image recognition. In *Electrical and Computer Engineering, 2004. Canadian Conference on*, volume 1, pages 103 – 106 Vol.1, may 2004.

[16] Dennis Romero, Teodiano Bastos, and Anselmo Frizera. Movement analysis in learning by repetitive recall. an approach for automatic assistance in physiotherapy. In *ISSNIP Biosignals and Biorobotics Conference 2012*, page 4, jan 2012.

[17] Andreas Antoniou. *Digital Signal Processing*. McGraw-Hill, 2005.

[18] Yu Qiao, Xingxing Wang, and Chunjing Xu. Learning mahalanobis distance for dtw based online signature verification. In *Information and Automation (ICIA), 2011 IEEE International Conference on*, pages 333 –338, june 2011.

[19] Ling Guo, Daniel Rivero, Jose A. Seoane, and Alejandro Pazos. Classification of eeg signals using relative wavelet energy and artificial neural networks. In *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, GEC '09, pages 177–184, New York, NY, USA, 2009. ACM.

[20] Osvaldo A. Rosso, Susana Blanco, Juliana Yordanova, Vasil Kolev, Alejandra Figliola, Martin Schurmann, and Erol Basar. Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods*, 105(1):65 – 75, 2001.

[21] B. Varone, J.M.A. Tanskanen, and S.J. Ovaska. Response analysis of feed-forward neural network predictors. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 4, pages 3309 –3312 vol.4, apr 1997.