*International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682*
*Vol.1, No.7, November 2012 DOI:10.15693/ijaist/2012.v1i7.1-6*

# Human Action Recognition based on motion and appearance

Thanikachalam.V
Assistant Professor/Department of IT
SSN College of Engineering, Chennai, India

Thyagharajan.K.K
Dean (Academic)
RMD Engineering College, Chennai, India

***Abstract*— This paper presents a method to recognize the action being performed by a human in a video. Applications like video surveillance, highlight extraction and video summarization require the recognition of the activities occurring in the video. The analysis of human activities in video is an area with increasingly important consequences from security and surveillance to entertainment and personal archiving. We propose an action recognition scheme based on motion and appearance. Firstly, we define an Accumulated Frame difference (AFD) from which Intensity histograms are built and normalized for extracting features. Then we compute DFT from the Intensity histograms so that features like mean and variance are obtained. Secondly, we try finding out gradient direction and magnitude by taking a key frame from the video. Again, we extract mean and variance from histogram giving out few more feature vectors. Finally with all the extracted features, we train the system using Dynamic Time Warping (DTW) to recognize the various actions.Public dataset is used for Evaluation.**

***Keywords: Action recognition, Accumulated Frame Difference, Histogram of Oriented Gradient, Intensity Energy histogram, DFT, DTW***

## I. INTRODUCTION

Human activity recognition is an important area in the field of computer vision research. The goal of human activity recognition is to automatically analyze ongoing activities from a video (i.e. a sequence of image frames). Recognizing human actions from videos is an important capability in emerging applications such as visual surveillance, robotics and sport video highlighting. This task is not an easy one because of the increasing demand of high-level scene understanding to analyze the behaviors of humans in the scene. In a case where a video is segmented to contain only one execution of a human activity, the objective of the system is to correctly classify the video into its action category. The proposal here is based on a compact 2D spatio -temporal action representation. Spatio-temporal means the use of both features extracted from static images and from image sequences.

In general, activity recognition aims to recognize the actions of one or more agents from a series of observations on the agent's actions and the environmental conditions. Though much research efforts have been dedicated to human action recognition, it still remains a problem, due to the differences in appearance, movement habit of subjects, the view angle variations and illumination changes.For example, the different clothes and gender yield significant differentiation of appearance in conducting similar actions. Thus, it is worth noting that building an efficient and robust action recognition system is a challenging task.

There are two types of human action recognition models: one is template-based models in which single template (i.e., training-free) is used to find the query action in target video sequences and the other is learning-based models where reliable action dataset is essentially needed to build a classifier. The latter is used in our work for which we take existing dataset like Weizmann dataset for implementation. Actions are single person activities that may be composed of multiple gestures organized temporally, such as walking, jumping, running, etc. Gestures are elementary movements of a person's body part, and are the atomic components describing the meaningful motion of a person, for example, 'stretching an arm' and 'raising a leg'. The proposed approach is based on single-layered approaches that represent and recognize human activities directly based on sequence of images. Because of this nature, our method is suitable for the recognition of gestures and actions with sequential characteristics.

The main contributions in our action recognition scheme are summarized as follows: first, the accumulated frame difference image (AFD) is defined by using image differences to represent the spatiotemporal features of occurring actions. It should be noted that only areas containing changes are meaningful for computing AFD instead of the whole silhouette of human body as in previous methods. Next we find out the Intensity histograms from the AFD and normalize them in order to obtain mean and variance. The normalized histograms are used to compute the DFT for which again mean and variance are obtained. A key frame is extracted from the video. Then the gradient direction is being grouped and implemented to build histogram from which again features are extracted. Finally, all the feature vectors are trained using Dynamic Time Warping (DTW) algorithm so as to classify the different actions.

The rest of this paper is organized as follows: the related work is briefly summarized in section 2.The technical details about the method outlined above are explained in section 3.The experimental work carried out with video datasets are shown in section 4 and followed by the conclusion in section 5.

## II. PRIOR WORK

Human action recognition has various challenges that have been widely studied for last several decades. Good tracking and segmentation still remains open research questions and are not able to deliver satisfactory performance. Another drawback of these methods is that they are not robust in the presence of occlusion.

Wonjun kim et al.[1] proposed a method for recognizing human actions from a single query action video. They gave an action recognition scheme based on ordinal measure of accumulated motion which does not require any preprocessing task such as learning and segmentation. Francesco and Carlo[2] presented a new motion descriptor based on a sparse optical flow computed by interest point tracking. This motion descriptor is by design invariant to scale, camera motion and is not affected by non-stationary background. A histogram of counts is composed considering the position and the motion of each interest point. This is processed to reduce its dimension by using the LSA(Latent Semantic Analysis). Masato, Masanori et al.[3] proposed a new method for incoherent motion recognition from video sequences in which they used time-series spatio-temporal intensity gradients within a space-time patch(ST-patch).

Zhang,Liu et al.[4] presents a method of human activity recognition based on Radon transform and dynamic time warping after the key frame is extracted from the cycle. For a key binary human silhouette,Radon transform is employed to represent low-level features.Bobick and Davis [5] proposed the temporal templates as models for actions. They construct two vector images, that is, motion energy image (MEI) and motion history image (MHI), which are designed to encode a variety of motion properties. Finally, these view-specific templates are matched against the model of query actions.

In [6],the paper addresses learning and classifying human actions on embedded low-dimensional manifolds. Jia and Yeung proposed a novel manifold embedding method, called Local Spatio-Temporal Discriminant Embedding (LSTDE).Yang et al. [7] consider the problem of human action recognition from a single clip per action. Using a patch based motion descriptor and matching scheme, we can achieve promising results on three different action datasets with a single clip as the template.

Ikizler et al. [8] proposed to use lines and optical flow histograms for human action recognition. In particular, they introduce a new shape descriptor based on the distribution of lines fitted to the silhouette of human body. Schuldt et al. [9] use space-time interest points proposed in [10] to represent the motion patterns and integrate such representations with SVM classification schemes.

Hu et al.[11] use the MHI along with foreground image obtained by background subtraction and the histogram of oriented gradients(HOG)[12] to obtain discriminative features for action recognition. Then they build a multiple-instance learning framework to improve the performance. Zhu et al.[13] integrate the cascade-of-rejectors approach with Histograms of Oriented Gradients (HoG)features to achieve fast and accurate human detection system.

## III. PROPOSED METHOD

The proposed method consists of five stages: AFD computation, determination of Intensity histograms and DFT, grouping of gradient direction to build histogram, feature extraction and training with DTW classifier. The overall procedure is shown in a block diagram Fig.1.

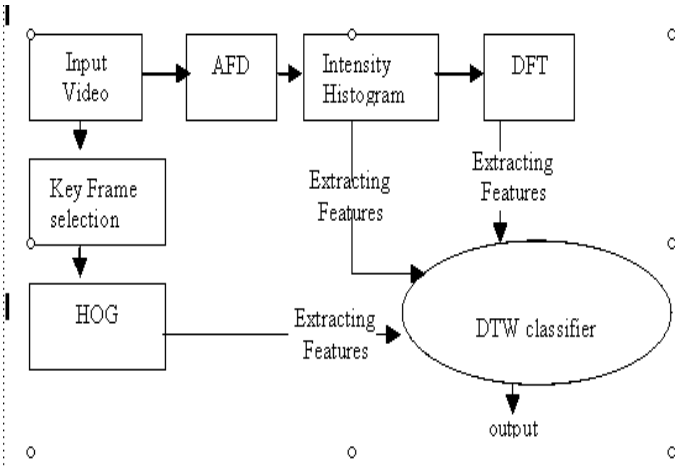


Figure 1 Block diagram of proposed method

### *A. Accumulated Frame Difference Image (AFD)*

A video is divided into number of frames. Each frame is processed using frame differencing to obtain the accumulated motion with variations noticed only in the moving pixels. AFD can be considered as a discriminative feature for recognizing human actions because the accumulated motion is differentiable across different actions. The accumulated frame difference image handles the variations of appearance such as different clothes, gender and also movement of subjects in opposite directions.

This feature AFD is motivated by the gait energy image (GEI) popularly used for the individual recognition and gender classification. When compared to Gait energy image, only areas including changes are used to compute AFD instead of taking the whole silhouette of human body. It is known as accumulated Frame difference image because it shows time-normalized accumulative action energy and higher intensity pixel values represent motions occur frequently at those positions.

To this end, the gray-level AFD is defined by using image differences as follows in (1)

$$\mathrm{AFD} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \left|\mathrm{D}(\mathrm{x},\mathrm{y},\mathrm{t})\right| \qquad (1)$$

where $D(x, y, t) = I(x, y, t) - I(x, y, t-1)$ and $T$ denotes the length of the query action video (i.e., total number of frames). The accumulated frame difference image found will look as in Fig. 2.

### *B. Intensity Energy Histograms*

This proposal takes up in finding Intensity energy histograms horizontally and vertically for the purpose of extracting features. From the AFD computed, Intensity energy histograms have to be found in both the horizontal and vertical directions. First, the horizontal projection is performed to accumulate all the AFD values in each row of the image. The vertical projection is conducted by accumulating all the values in each column of the image. These Intensity energy histograms have to be normalized in order to bring the values to a particular range.

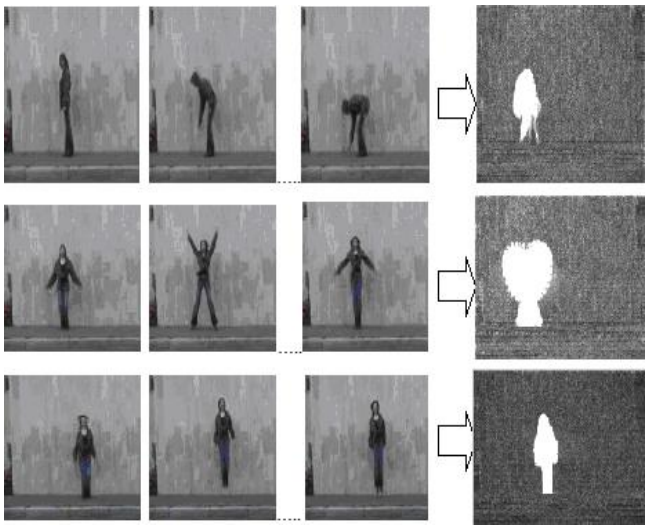

Figure 2. Examples of AFD for various actions from Weizmann dataset: bend , jack, parallel jump respectively

In Image Processing, Normalization is the process that changes the range of pixel intensity values. Accumulated values of each bin are normalized by dividing each value in the bin with the maximum value along both the horizontal and vertical bins. Normalized Intensity Energy histogram for each direction are defined as follows in equations (2) and (3),

$$IEH_h(i) = \sum_{j=0}^{W-1} \frac{AFD(i,j)}{max\_AFD(i)}, i = 0,1,..H-1 \qquad (2)$$

$$IEH_v(j) = \sum_{i=0}^{H-1} \frac{AFD(i,j)}{max\_AFD(j)}, j = 0,1,...W-1 \qquad (3)$$

where *H* and *W* denote the height and width of the accumulated motion. max _AFD (·) denotes the maximum value among AFD values belonging to the $i^{th}$ or $j^{th}$ bin in each energy histogram. Normalization is carried out to minimize redundancy.

### *C. Discrete Fourier Transform(DFT)*

The discrete Fourier transform (DFT) is a specific kind of discrete transform, used in Fourier analysis. It transforms one function into another, which is called the frequency domain representation, or simply the DFT, of the original function. The Fast Fourier Transform(FFT) is an efficient algorithm for computing the DFT. The input to the DFT is a finite sequence of real or complex numbers, making the DFT ideal for processing information stored in computers.

The sequence of N spatial complex coefficients $x_0, x_1,\ldots\ldots,x_{N-1}$ is transformed into the sequence of *N* frequency complex coefficients X(0), X(1),…….., X(N-1) by the DFT according to the formula as given in equation (4)

$$X[K] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\Pi nk}{N}}, K = 0 \text{ to } N-1 \qquad (4)$$

where n=0, 1… N-1 and k=0, 1… N-1.

But the DFT requires an input function that is discrete and whose non-zero values have a limited (finite) duration. Such inputs are often created by sampling a continuous function. Unlike the discrete-time Fourier transform (DTFT), it only evaluates enough frequency components to reconstruct the finite segment that was analyzed. Using the above equation (4), we compute DFT for the normalized energy histogram vectors in both directions (i.e., horizontal and vertical).

### *D. Histogram of oriented Gradient*

An image gradient is a directional change in the intensity or color in an image. Image gradient may be used to extract information from images. Gradient is used for a gradual blend of color which can be considered as an even gradation from low to high values. In order to analyze the direction of movement of the pixels, we try to find the gradient of motion of the object.

The algorithm we use to find the gradient of motion in our proposal is as follows,

```
for each pixel (x,y) in an image I
 {
 find the gradient of the pixel by absolute differencing
      dx = I(x,y)-I(x+1,y)
      dy = I(x,y)-I(x,y+1)
}
find gradient direction or orientation = arctan(dx,dy)
find gradient magnitude = sqrt(dx*dx+dy*dy)

if(gradient magnitude >threshold)
   {
    find the group of gradient direction
                and implement the frequency
   }
```

In the above statements, arctan means arctangent which is used to calculate the angles of a right triangle. With the gradient direction and magnitude obtained as per the above algorithm, we group the directional values based on a condition involving magnitude values. Then the grouped values are used in building a histogram.

### *E. Feature Extraction*

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. If the features extracted are carefully chosen, it is expected that the feature set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Best results are achieved when an expert constructs a set of application-dependent features.

a) ***Mean:*** For a data set, the mean is the sum of the values divided by the number of values. The mean describes the central location of the data. Mean is defined as follows in equation (5)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (5)$$

Where 'x bar' is the arithmetic mean. The mean has to be computed for normalized energy histograms and DFT in two directions as mentioned earlier.

b) ***Variance:*** Variance is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean. The variance has to be computed for normalized energy histograms and DFT in both horizontal and vertical directions, which is defined as the average of the squared differences from the mean which is given by the formula (6)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2 \qquad (6)$$

Where 'x bar' is the arithmetic mean.

### *F. Dynamic Time Warping (DTW)*

DTW is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected even if in one video the person was walking slowly and if in another he or she were walking more quickly. This method allows a computer to find an optimal match between two vectors with certain restrictions. The sequence is warped non linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. Its advantage is that the optimal result can be obtained in low computational complexity otherwise it is dynamic time warping in matching time sequences.

## IV. IMPLEMENTATION DETAILS

### *A. Dataset*

For evaluating the proposed algorithm, this paper uses the Weizmann dataset, which is relatively larger in terms of the number of subjects and actions. It includes 81 low-resolution videos (180-by-144, 25fps) from 9 different people, each performing 10 natural actions(periodic and non-periodic actions, and stationary and non-stationary motions along both horizontal and vertical directions), i.e., bending (bend), jumping jack (jack),jumping-forward-on-two-legs (jump), running (run),jumping-in-place-on-two-legs (pjump), walking (walk),galloping-sideways (side),skipping (skip),waving-one-hand (wave1), and waving-two-hands (wave2).Many actions are similar in the senses that the limbs have similar motion paths, and this high degree of similarity among actions makes discrimination more challenging. Also, each action in this dataset is performed by different people with different physical characteristics and motion styles, thus providing more realistic data for the test of the versatility of the proposed method.

### *B. Extraction of mean and variance*

With the computed accumulated frame difference image, we compute intensity energy histograms in horizontal and vertical directions. After normalizing, feature vectors like mean and variance are calculated. The values are shown for various actions in table 1.

From the normalized intensity energy histograms, we try to compute DFT in both horizontal and vertical directions. With which we obtain the 1st and 2nd order moments (mean and variance) being shown for some actions in Table 2.

With the gradient direction and gradient magnitude obtained, histogram is built for the grouped directional values based on the condition related to gradient magnitude values. The histogram coefficients are used to extract mean and variance features which is shown in the Table 3.

TABLE 1 MEAN AND VARIANCE FOR INTENSITY ENERGY HISTOGRAMS

| | **Horizontal Direction** | | **Vertical Direction** | |
|---|---|---|---|---|
| **Actions** | **Mean** | **Variance** | **Mean** | **Variance** |
| Bend | 0.8305 | 0.0106 | 0.6580 | 0.0127 |
| Jack | 0.8142 | 0.0178 | 0.6710 | 0.0236 |
| pJump | 0.7876 | 0.0160 | 0.5056 | 0.0200 |
| Side | 0.5824 | 0.1023 | 0.8552 | 0.0262 |
| Walk | 0.6763 | 0.0699 | 0.9147 | 0.0146 |
| Wave1 | 0.7529 | 0.0086 | 0.7306 | 0.0079 |

TABLE 2 MEAN AND VARIANCE FOR DISCRETE FOURIER TRANSFORM

| | Horizontal Direction | | Vertical Direction | |
|---|---|---|---|---|
| **Actions** | **Mean** | **Variance** | **Mean** | **Variance** |
| Bend | 1.0000 | 89.3285 | 0.6943 | 78.9528 |
| Jack | 0.7075 | 69.1052 | 0.4604 | 49.9793 |
| PJump | 0.9830 | 58.8715 | 0.5955 | 134.8497 |
| Side | 0.4246 | 57.7676 | 0.4800 | 112.8522 |
| Walk | 0.5203 | 58.8715 | 0.5955 | 134.8497 |
| Wave1 | 0.9818 | 85.6275 | 0.7588 | 92.4666 |

When comparing the mean and variance values, they are differentiable across each different actions carried out on a single person. The Fig.3 shows how the action is recognized by our method. With all the feature vectors obtained, we train the system using Dynamic Time Warping technique which correctly classifies the action category by measuring similarity. Our system tracks the action fastly and efficiently

TABLE 3 MEAN AND VARIANCE FOR GROUPED GRADIENT DIRECTIONS

| **Actions** | **Mean** | **Variance** |
|---|---|---|
| Bend | 46.1834 | 1024 |
| Jack | 53.7834 | 921 |
| pJump | 62.1067 | 925 |
| Side | 62.2475 | 825 |
| Walk | 52.7822 | 1028 |
| Wave1 | 47.5403 | 1252 |

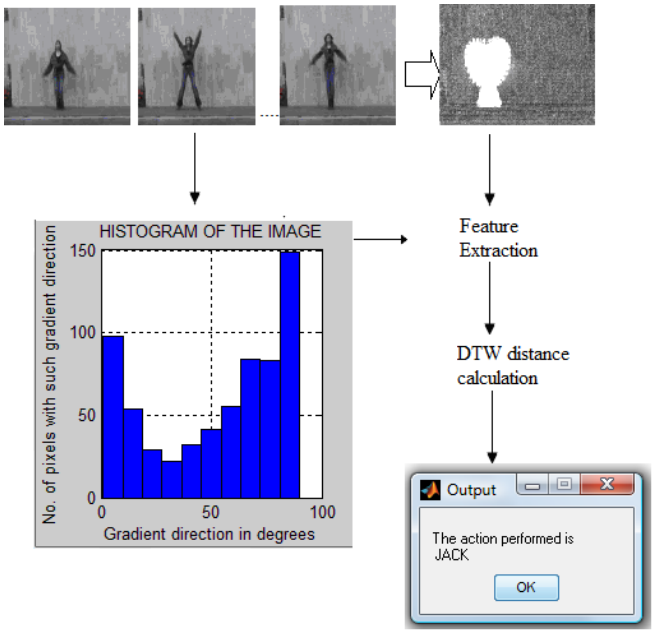


Figure 3. Output

## V. CONCLUSION

A novel method for human action recognition is proposed in this paper. Compared to previous methods, our proposed algorithm is performed very fast based on the motion and appearance. To this end, AFD computation is carried out which is differentiable across various actions and regarded as a discriminative feature. This helps in handling the variations of appearance and clearly shows the moving pixels. Histogram of gradients allows to detect object and its movement of direction with improved accuracy. Feature vectors like mean and variance have been obtained from accumulated frame difference and gradient, with which the system is trained using DTW. The trained system classifies the actions appropriately with fastness and efficiency. Our future work could handle more than one object in a video and occlusions

## REFERENCES

[1] Wonjun Kim, Jaeho Lee,Minjin Kim, Daeyoung Oh and Changick Kim,"Human action recognition using ordinal measure of accumulated motion",in EURASIP Journal on Advances in Signal Processing, 2010.

[2] Francesco Monti and Carlo S.Regazzone,"Human action recognition using the motion of interest points",IEEE 17th International Conference in Image Processing,2010.

[3] Masato kazui,Masanori Mujoshi,Shoji Muramatsu,Hironobu Fujiyoshi,"Incoherent Motion Detection using a time-series gram matrix feature",in 2008.

[4] Hao Zhang,Zhijing Liu,Haiyong Zhao and Guojian Cheng,"Recognizing human activities by key frame in vidoe sequences",inAugust2010.

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

[6] Kui Jia and Dit-Yan Yeung,"Human Action Recognition using local spatio-temporal discriminant embedding",in 2008.

[7] Weilong Yang, Yang Wang, and Greg Mori, "Human action recognition from a single clip per action",in IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops,2009.

[8] N. Ikizler, R. G. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms", in Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08), pp. 1–4, Tampa, Fla, USA, December 2008.

[9] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04), vol. 3, pp. 32–36, Cambridge, UK, August 2004.

[10] I. Laptev and T. Lindeberg, "Space-time interest points",in Proceedings of the 9th IEEE International Conference on

Computer Vision, vol. 1, pp. 432–439, Nice, France, October 2003.

[11] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang,"Action detection in complex scenes with spatial and temporal ambiguities", in Proceedings of International Conference on Computer Vision (ICCV'09), October 2009.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection",in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893, San Diego, Calif, USA, June 2005.

**[13]** Qiang Zhu, Shai Avidan, Mei-Chen Yeh and Kwang-Ting Cheng,''Fast Human Detection using a cascade of Histograms of Oriented Gradients'',in IEEE Computer Society Conference on Computer Vision and Pattern Recognition,June 2006.

### Authors Profile

Mr.V.Thanikachalam, Assistant Professor in the Department of Information Technology in SSN College of Engineering. He has 14 years of teaching experience. He received his B.E (Computer Science and Engineering) from Bharathidasan University, M.E(CSE) from Anna University Chennai. He is currently pursuing Ph.D. (part time) at Anna University chennai. He has published 3 papers in International conference. He has involved in many UG projects and PG Projects in the area of Image Processing.

Dr. K.K. Thyagharajan obtained his B.E., degree in Electrical and Electronics Engineering from PSG College of Technology (Madras University) and received his M.E., degree in Applied Electronics from Coimbatore Institute of Technology in 1988. He also possesses a Post Graduate Diploma in Computer Applications from Bharathiar University. He obtained his Ph.D. degree in Information and Communication Engineering (Computer Science) from College of Engineering Guindy , Anna University in 2007.He is the Dean (Academic) of R.M.D. Engineering College. His current interests are Multimedia Networks, Content Based Information Retrieval, Mobile Computing, Web services, Data Mining, e-learning, Image Processing, Microprocessors and Microcontrollers. He is reviewer for many International Journals and Conferences.