Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-02-35

HUMAN ACTIVITY RECOGNITION BASED ON ANN USING HOG FEATURES

Rajvir Kaur[#1], Sonit Singh[#2] & Harkishan Sohanpal[#3]

***Abstract*—In this paper, we present human activity recognition on static images. First, for feature extraction we employ Histograms of Oriented Gradients (HOG). The HOG is invariant to geometric transformations and photometric transformation such as changes in illumination or shadowing effect. The extracted features are then classified using Back-Propagation Neural Network (BPNN) classifier. Experimental results on Images from Weizmann dataset using proposed methodology show the accuracy of 99.2%. The results show that the human activity recognition can effectively be done using HOG features and BPNN as classifier.**

Keywords- Human Activity Recognition(HAR), Histogram of Oriented Gradients (HOG), Artificial Neural Networks (ANN), Back-Propagation Neural Network (BPNN), Multilayer Perceptron (MLP), feature extraction, classification, etc.

I. Introduction

Human Activity Recognition (HAR) has become one of the most active research areas in computer vision. This is due to promising applications in the field of behavioral biometrics, Content based video retrieval, Security and surveillance, interactive applications, environment & synthesis [15], sports video analysis, human computer interfaces, gesture recognition, robot learning and control. Actions can be referred to as simple motion patterns which are executed by a person individually in a general sense. As far as activity is concerned, it refers to more complex sequence of actions. Various examples of human actions can be jump, skip, gallop, walk, walk sideway, one-hand wave, two-hand wave, gallop, etc. Various examples of human activities can be the interaction of two person, players playing in soccer game, abnormal gestures of players in various sports, violation of traffic rules, elderly fall in old-aged homes.

Rajvir Kaur
Discipline of ECE, Lovely Professional University
India

Sonit Singh
Discipline of ECE, Lovely Professional University
India

Harkishan Sohanpal
Discipline of ECE, Lovely Professional University
India

The aim of activity recognition is to recognize the actions of one or more human agents based on the environmental conditions. As large number of surveillance cameras are being deployed in public places to monitor the activities of human. So there is a need for intelligent security systems which can automatically detect, categorize and recognize the activities of human. In [12], the task of visual tracking is to predict and update the target's position, velocity and size based on the video, while the task of visual action recognition is to classify and recognize the human's action.

Major techniques of HAR falls under four categories[1]: Geometric model based (use of geometric primitives like cones and spheres to model head, trunk, limbs and fingers), Appearance based (use of color or texture information to track the body and its parts), Salient points based (uses changes in entropy in space and time that corresponds to peaks in body activity variation), and Spatial-temporal shape based (which treats human body as gestures as shapes in space-time domain).

II. Related Work

Human activity recognition has been studied and many new techniques has been proposed by the researchers to have real-time HAR with good accuracy. The process of HAR begins with the data acquisition from the video capturing cameras and after this some preprocessing techniques are applied in order to improve the quality of the frames. After this feature extraction is done and the extracted features are given to the classifier for classification.

The term feature extraction can be defined as "a process of identifying valid, useful and understandable pattern in the data". The objective of feature extraction is to keep all the useful features of data and discard all the redundant parts of the data. There are two major objectives of feature extraction [2]: (i). To reduce the amount of data to the manageable level (dimensionality reduction), and (ii). To keep the most important features of the data and eliminate all the redundant features of the data (feature selection). Feature representations are used to map the data to another representation space with the intention to make the classification problem easier. In [3] comparison of various feature selection methods and classifiers is done for accelerometer based floor changing activity recognition. Major features of accelerometer based HAR extracted are mean, FFT energy, FFT domain entropy, variance, skewness, kurtosis and eccentricity. Comparative analysis was done on the basis of various classifiers such as DTP, Naïve Bayes classifier, Multi-Layer Perceptron.

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-02-35

In [4], silhouettes are extracted from the video frames. Feature extraction is done using PCA and ICA. Then, using cluster analysis classification is done for classification of human activities. In [5] CLG optic flow and global shape feature are used as feature vector for HAR. In [6], ICA based classification has been done for HAR which comprises of motion features in the form of Exponential Motion History Image (EMHI) for spatio-temporal representation of motion. HAR using dynamic texture based method [7] has been proposed. In this LBP-TOP (Local Binary Pattern Three Orthogonal Planes) are used to represent human movement. Though this method is computationally simple and utilizes image data rather than silhouettes, but it does not provide better results when background is changing dynamically.

In [13] uses STIP (Space-Time Interest Points) to extract motion features. Shape context descriptors can also be used as a feature vector. The shape context descriptor provides information about the way boundary points are spread out w.r.t one another. These can be classified using SVM (Support Vector Machine) or MLP (Multi-Layer Perceptron) as classifiers.

III. Overview of Human Activity Recognition

Human Activity Recognition generally involves the following set of sequence of steps: preprocessing, object detection, object tracking and behavior classification. In pre-processing, the acquired video is decompressed, removing noise from video or image sequences in order to improve the visual quality of frames of video. Object detection involves background modelling, object localization and distinguishing between various object using classification when multiple objects are present in the image. The complexity of HAR increases with more complex scenes because of cluttered environment, illumination changes, shadowing effect, change in scaling factor, due to camera motion and low background contrast. Broadly, human activity recognition has the following steps:

Step 1: The foremost step for activity recognition is the requirement of the valid dataset. In past years, datasets dedicated to activity recognition have been created and these datasets are used to compare different recognition systems. There are various datasets available for human activity recognition. Some of the commonly used datasets are KTH, Weizmann, UCF101, MSR Action, UT-Interaction, LIRIS human activities, etc.

Step 2: Feature extraction: Feature extraction is an element operation for the recognition of human activities [9]. Feature extraction is used to reduce the dimensionality. When the input data is too large to be processed then the input data will be transformed into a reduced representation set of features and these set of features are also called feature vectors. Transforming the data into feature vectors is called feature extraction. The features set will provide the relevant information from the input data in order to perform the desired task using the reduced representation instead of the full size input.

Step 3: Classification: Classification is the process of identifying to which set of categories the observation belongs to. Classification means to categorize something according to shared characteristics. An algorithm that implements classification is known as a classifier. "Classifier" refers to the mathematical function, implemented by a classification algorithm that maps input to a category.

IV. Proposed Methodology

A. *Histograms of Oriented Gradients (HOG)*

Histograms of Oriented Gradients (HOG) is feature descriptors used in computer vision and image processing for object detection. HOG features are calculated by taking orientation histograms of edge intensity in a local region [10]. An image is divided into N local regions called "blocks". These local regions are divided into small spatial areas called "cell". The HOG features are extracted from local regions with 16×16 pixels. Histograms of oriented gradients with 8 orientations are calculated from each 4×4 local cells. The edge gradients and orientations are obtained by applying Sobel filter.

In [10], each feature is defined by its cell position $C(x_c, y_c, w_c, h_c)$, the parent local region position $B(x_b, y_b, w_b, h_b)$ and the orientation bin number k. Each cell feature f is denoted by $f(C, B, k)$. The gradients at the point (x, y) of image I can be found by convolving gradient operator with the image:

$$G_x(x, y) = [-101] * I(x, y) \quad (1)$$

$$G_y(x, y) = [-101]^T * I(x, y) \quad (2)$$

The strength of the gradient at the point (x, y) is:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad (3)$$

The orientation of the edge at the point (x, y) is:

$$\theta(x, y) = \arctan\left[\frac{|G_y(x, y)|}{|G_x(x, y)|}\right] \quad (4)$$

Divide the orientation range into k bins and denote the value of k_{th} bin to be:

$$\varphi_k(x, y) = \begin{cases} G(x, y), if\ \theta(x, y) \in bin_k \\ 0, otherwise \end{cases} \quad (5)$$

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-02-35

Then the feature value is defined as:

$$f(C,B,k)=\frac{\sum_{(x,y)\in C}\varphi_k(x,y)}{\sum_{(x,y)\in B}G(x,y)} \quad (6)$$

There are two classes of block geometries: one is square or rectangular blocks and the other is circular blocks [11]. These two arrangements are also called as R-HOG and C-HOG. Rectangular blocks are partitioned into grids of squares or rectangular spatial cells and circular blocks in log-polar fashion. R-HOG blocks have many similarities to SIFT descriptors but they are used differently. R-HOG's are optimized for dense robust coding of spatial form. We usually use square R-HOG's, i.e. $\varsigma\times\varsigma$ grids of $\eta\times\eta$ pixel cells each containing β orientation bins, where ς,η,β are parameters. There are two variants of the C-HOG geometry, one with a single circular central cell and the other whose central cell is divided into angular sectors the C-HOG layout has four parameters: the number of angular and radial bins, the radius of the central bin in pixels and the expansion factor for radii.

B. *Back-Propagation Neural Network (BPNN)*

The back –propagation learning algorithm [14] is one of the most important algorithm in neural networks. A back-propagation neural network is a multilayer, feed-forward neural network consisting of an input layer, a hidden layer and an output layer. The neurons present in the hidden and output layers have biases, which are the connections from the units whose activation is always 1. The term "bias" also acts as weights. During the back-propagation phase of learning, signals are sent in the reverse direction. The inputs are sent to the BPN and the output obtained could be either binary (0,1) or bipolar(-1, +1). The activation function increases monotonically and is also differentiable.

The training of the BPN is done in three stages: the feed-forward of the input training pattern, calculation and back-propagation of the error, and updation of the weights.

Training Algorithm:

The terminologies used in algorithm are as follows:

x =input training vector $(x_1,....,x_i,....,x_n)$

t =target output vector $(t_1,....,t_k,....,t_m)$

α =learning rate parameter

x_i =input unit i

v_{oj} =bias on jth hidden unit

w_{ok} =bias on kth output unit

z_j =hidden unit j

The net input to z_j is $z_{inj}=v_{oj}+\sum_{i=1}^{n}x_i v_{ij}$ (7)

output is $z_j=f(z_{inj})$ (8)

y_k =output unit k .

The net input to y_k is $y_{ink}=w_{ok}+\sum_{j=1}^{p}z_j w_{jk}$ (9)

and the output is $y_k=f(y_{ink})$

δ_k =error correction weight adjustment for w_{jk} that is due to an error at output unit y_k, which is back-propagated to the hidden units that feed into unit y_k .

δ_j =error correction weight adjustment for v_{ij} that is due to the back-propagation of error to the hidden unit z_j .

Step 0: Initialize weights and learning rate.

Step 1: Perform Step 2-9 when stopping condition is false.

Step 2: Perform Step 3-8 for each training pair.

Phase I (Feed-forward phase)

Step 3: Each input unit receives input signal x_i and sends it to the hidden unit.

Step 4: Each hidden unit z_j *(j=1 to p)* sums its weight input signals to calculate net input:

$$z_{inj}=v_{oj}+\sum_{i=1}^{n}x_i v_{ij} \quad (10)$$

Calculate output of the hidden unit by applying its activation function over z_{inj} :

$$z_j=f(z_{inj}) \quad (11)$$

and send the output signal from hidden unit to the input of output layer units.

Step 5: For each output unit y_k *(k=1 to m)*, calculate the net input:

$$y_{ink}=w_{ok}+\sum_{j=1}^{p}z_j w_{jk} \quad (12)$$

and apply the activation function to compute output signal

$$y_k=f(y_{ink}) \quad (13)$$

Phase-II (Back-propagation of error)

Step 6: Each output unit y_k *(k=1 to m)* receives a target pattern corresponding to the input training pattern and computes the error correction term:

$$\delta_k=(t_k-y_k)\,f'(y_{ink}) \quad (14)$$

Update the change in weights and bias:

$$\Delta w_{jk}=\alpha\delta_k z_j \text{ and } \Delta w_{ok}=\alpha\delta_k . \quad (15)$$

And send δ_k to the hidden layer backwards.

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-02-35

Step 7: Each hidden unit z_j *(j=1 to p)* sums its delta inputs from the output units:

$$\delta_{inj} = \sum_{k=1}^{m} \delta_k w_{jk} \quad (16)$$

Calculate the error term: $\delta_j = \delta_{inj} f'(z_{inj})$ (17)

Update the change in weights and bias:

$$\Delta v_{ij} = \alpha \delta_j x_i \quad \text{and} \quad \Delta v_{oj} = \alpha \delta_j \quad (18)$$

Phase –III (Weight and bias updation)

Step 8: Each output unit y_k *(k=1 to m)* updates the bias and weights:

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$$

$$w_{ok}(new) = w_{ok}(old) + \Delta w_{ok} \quad (19)$$

Each hidden unit z_j *(j=1 to p)* updates its bias and weights:

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij} \quad (20)$$

$$v_{oj}(new) = v_{oj}(old) + \Delta v_{oj} \quad (21)$$

Step 9: Check for stopping condition. The stopping condition may be certain number of epochs reached or when the actual output equals the target output.

V. Results and Discussion

In this research, we recognized human activities based on Weizmann dataset. The dataset consists of nine human activities: bend, run, jump, skip, one –hand wave, two-hand wave, gallop sideway, jumping jack and walk.

Selected frames of each action video have been shown in Fig. 1.1. Experiments are performed with MATLAB R2013a on Intel Core i5-480M processor 2.66 GHz, 4GB RAM running Window 7 home premium (64-bit) operating system. During training 40 frames of each activity are selected to represent an activity sequence. After feature extraction using HOG, Neural Network is trained for these nine classes of activities. The various training parameters for ANN classifier[16] has been listed in the Table-I.

TABLE I: Training Parameters for ANN

Number of layers	1 input layer, 1 hidden layer, 1 output layer
Number of neurons in different layers	Input=81 Hidden=20 Output=9
Initial weights	Random
Activation function for hidden and output layers	Log sigmoid
Training parameters for learning	Scaled conjugate gradient back-propagation
Number of iterations	103
Acceptable mean square error	0.005

The input to ANN classifier is 81×360 matrix, representing static data: 360 samples of 81 elements HOG features and target is 9×360 matrix, representing static data: 360 samples of 9 elements. The results of our proposed work has been shown in the form of Confusion matrix listed in the Table II, which shows the actual vs. predicted class of the images. Our proposed method using ANN classifier resulted in an average recognition rate of 99.2% for nine activities.

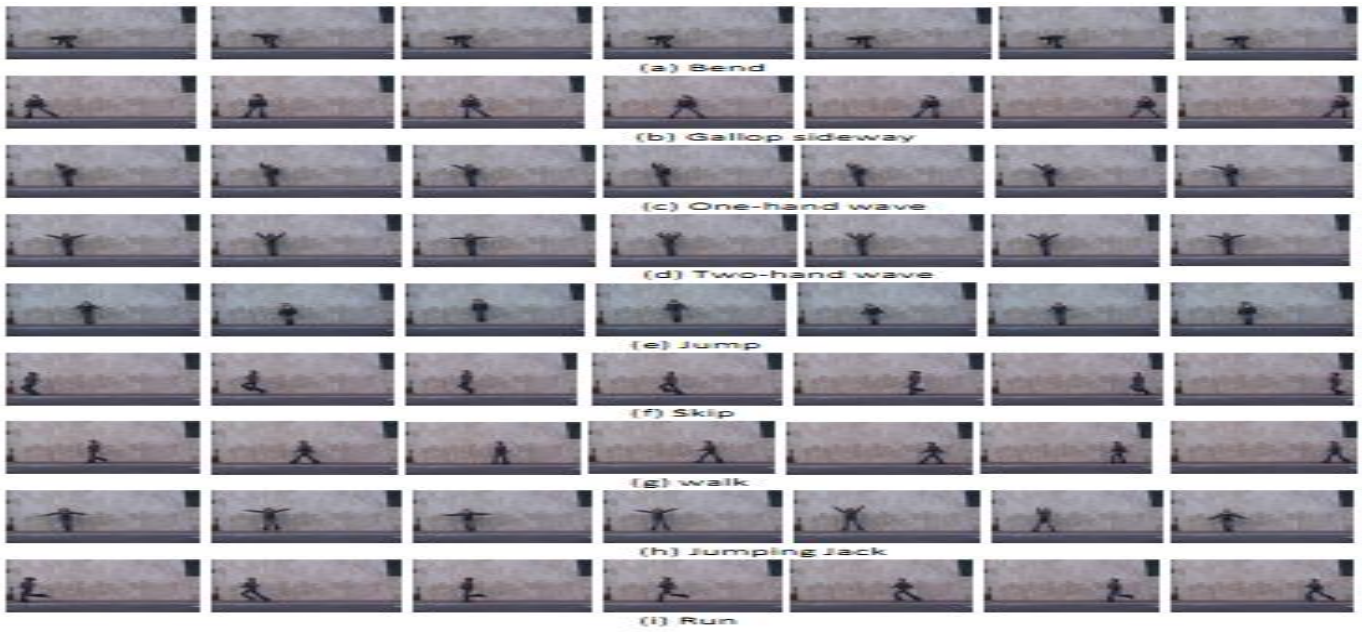

Fig. 1.1 Selected frames of the videos of various human activities from Weizmann Dataset [8], (a). Bend (b). Gallop Sideway, (c). One-hand wave, (d). Two-hand wave, (e). Jump, (f). Skip (g). Walk (h). Jumping Jack & (i). Run [8].

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-02-35

TABLE II : Confusion Matrix for the classification of Nine Human Activities

Predicted / Actual	Bend	Run	Jump	One-Hand Wave	Two-Hand Wave	Walk	Skip	Gallop Side Way	Jumping Jack	Accuracy
Bend	40	0	0	0	0	0	0	0	0	100%
Run	0	38	0	0	0	0	0	0	0	95%
Jump	0	0	40	0	0	0	0	0	0	100%
One-Hand Wave	0	0	0	40	0	0	0	0	0	100%
Two-Hand Wave	0	0	0	0	40	0	0	0	0	100%
Walk	0	1	0	0	0	40	0	0	0	100%
Skip	0	1	0	0	0	0	40	0	0	100%
Gallop Side Way	0	0	0	0	0	0	0	40	1	100%
Jumping Jack	0	0	0	0	0	0	0	0	39	97.5%
Total Accuracy										**99.2%**

VI. Conclusion & Future Work

In this research, human activity recognition has been performed based on ANN classification using HOG features. The method is able to recognize nine different human actions giving 99.2% average recognition accuracy results. As HOG operates on localized cells, it is invariant to geometric transformations, changes in illumination and shadowing. Hence, HOG is having few advantages over other feature extraction methods. This shows that HOG features can be used for human activity recognition with reasonable good results. The present work shows the results of Human Activity Recognition based on static images. In future, the proposed method can be applied for Human Activity Recognition based on real-time streaming videos.

References

[1] Xianping Huang, Lili Zheng, Ronhua Liang, Wanliang Wang, "Human Action Recognition Based on SVM using multiple features", International Conference on Artificial Intelligence and Soft computing, vol.12, 2012.

[2] Sahak I. Kaghyan and Hakob G.Sarukhanyan, "Time Domain Feature Extraction and SVM Processing for Activity Recognition using Smartphone signals", Mathematical Problems of Computer science 40, 44-54, 2013.

[3] Sara Khalifa, Mahbub Hassan and Aruna Seneviratne, "Feature Selection for Floor-changing Activity Recognition in Multi-Floor Pedestrian Navigation", Seventh International Conference on Mobile Computing and Ubiquitous Networking(ICMU), 2014.

[4] Jyotsna E, Akhil P.V, Arun Kumar, "Silhouette based Human Action Recognition using PCA and ISOMAP", International Journal of Advanced Research in Computer and Communication Engineering, vol.2, issue 11, ISSN:2319-5940, november 2013.

[5] Mohiuddin Ahmad, Seong-Whan Lee, "Human Action Recogition using Shape and CLG-motion flow from multi-view image sequences", Pattern Recognition 41(2008) 2237-2252.

[6] Du-Ming Tsai and Wei-Yao Chiu, "A real-time ICA based activity recognition in video sequences", MVA2013 IAPR International Conference on Machine Vision Application, May20-23, 2013, kyoto, Japan.

[7] Vili Kellokumpu, Guoying Zhao and Matti Pietikainen, "Human Activity Recognition using a Dynamic Texture based method", Machine Vision Group, University of Oulu.

[8] www.cs.utexas.edu/~chaoyeh/web_action_data/dataset_list.html.

[9] Haiyong Zhao, Zhijing Liu, "Human Action Recognition based on Non-linear SVM Decision Tree", Journal of Computational Information System 7:7(2011) 2461-2468.

[10] Qing Jun Wang and Ru Bo Zhang, "LLP-HOG:A New Local Image Descriptor for Fast Human Detection", IEEE, pg no 640-643, 2008.

[11] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection", CPVR 2005.

[12] Wei-Lwun Lu, James J.Little, " Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor", Department of Computer Science, Vancouver.

[13] M.Niresh Kumar, Dr.K.Madhavi, "Improved Discriminative Model for View-Invariant Human Action Recognition", International Journal of Computer Science and Engineering Technology(IJCEST), ISSN: 2229-3345, vol.4 no.09 sep 2013.

[14] S.N.Sivanandam, S.N.Deepa, "Principles of Soft Computing",Second edition.

[15] Pavan Turaga, Rama Chellapa, V.S. Subrahmanian and Octavian Udrea, "Machine Recognition of Human Activities: A Survey", IEEE, 2008.

[16] Neural Network Toolbox, www.mathworks.in/product/neural-networks/

About Author (s):

Rajvir Kaur has been working as a Lecturer in the Department of ECE, GES College, Hoshiarpur, Punjab. She has been working towards her M.Tech in the field of digital image processing from Lovely Professional University, Punjab. Her research interests include human activity recognition, video content analysis and content based image retrieval.

Sonit Singh has been working as an Assistant Professor in the Department of ECE, Lovely Professional University, Phagwara, Punjab, India. His research interests include pattern recognition, machine learning, digital image processing, machine vision, soft computing and computational neuroscience.

Harkishan Sohanpal has been working as an Assistant Professor in the Department of ECE, Lovely Professional University, Phagwara, Punjab, India. His research interests include neural networks, fuzzy logic, genetic algorithms and embedded systems.