Human Activity Recognition System: Using Improved Crossbreed Features and Artificial Neural Network

Asmatullah Chaudhry[1,2], Javed Ullah[3], M. Arfan Jaffar[3], Jin Young Kim[1], Tran Anh Tuan[1]

[1]School of Electronics & Computer Engineering, Chonnam National University, Gwangju, South Korea
[2]HRD, PINSTECH, P.O. Nilore, Islamabad, Pakistan
[3]Department of Computer Science, National University of Computer and Emerging Sciences Islamabad, Pakistan
asmat@jnu.ac.kr, javed.fact@gmail.com, arfan.jaffar@nu.edu.pk, beyondi@jnu.ac.kr, tuantran2610@yahoo.com

Abstract: In this paper, we present an intelligent method of human action recognition based on hybrid features. These features are calculated from the space-time cubes (interest points). The Blocks or cubes are derived from the difference of consecutive frames. These features include average number of blocks per frame, velocity of the blocks i.e. average angle and displacement, some of the kinematic features like divergence and vorticity, and hu features derived from motion energy images (MEI). Principal Component Analysis (PCA) has been used to reduce the dimensionality of the feature vector. For classification, we employ artificial neural networks (ANNs), in which each action video is represented by a bag of features.
[Asmatullah Chaudhry, Javed Ullah, M. Arfan Jaffar, Jin Young Kim, Tran Anh Tuan. **Human Activity Recognition System: Using Improved Crossbreed Features and Artificial Neural Network.** *Life Sci J* 2012;9(4):5351-5356] (ISSN:1097-8135). http://www.lifesciencesite.com. 795

Keywords: Human activity recognition; PCA; ANNs

1. Introduction

Human activities recognition in videos is one of the most promising and emerging applications of computer vision. In the current era, the problem of human activity recognition has caught the attention of researchers from academia, industry, security agencies, consumer agencies and the general public as well [1-8]. In simple words the problem of human activity recognition can be defined as, in a sequence of images having one or more persons performing single or multiple activities, can a system be designed that can automatically recognize the activity (activities) being performed? As simple as the problem seems to be, the optimum solution has been that much difficult to find. The human activity recognition system typically follows a hierarchical approach. At the lower levels background foreground subtraction, tracking and object detection is performed. Action-recognition modules serve as mid-level functions. Reasoning engines are at the high-level. They encode the activity semantics based on the lower level action-primitives. Therefore, it is necessary to have an understanding of both these problem domains to enable deployment of these systems in real-life. Some of the application areas that highlight the potential impact of vision-based activity recognition systems are behavioral biometrics, content based video analysis, security and surveillance, interactive applications and environments, and animation and synthesis [9-12].

In general, approaches used for human activity recognition can be categorized on the basis of representation of the features [1]. Some well-known representations include learned geometrical models of human body parts, space-time pattern templates, region features, shape features, interest-point-based representations, and optical flow patterns.

2. Proposed Method

In this paper, we utilize the motion of human body for the activity recognition. The proposed method consists of three steps which are feature extraction, dimensionality reduction and classification. The system diagram of the proposed approach is given in figure 1.

2.1 Interest Points Based Feature Extraction

We use a set of features for human action recognition. All these features depend on the motion of an actor performing the action. Multiple methods are exploited to extract features. All of the features are computed from the interest points. The success of interest points in object detection, their sparsity, and robustness against illumination and clutter have inspired a number of researchers working in the area of motion analysis and activity recognition. In the first step, we compute interest points and then extract features from these computed interest points. Our method of interest points computation is different from [7, 12] in the way they are computed. We describe next our method of interest points computation in detail. First of all each frame of video is localized and a Gaussian filter is applied on it. The Gaussian filter firstly smoothes out the frame and hence the noise is reduced. Secondly the small changes are ignored due to Gaussian smoothing filters. Each frame is divided into blocks (cubes) of size 3×3 or 5×5. Then each corresponding block in

the two consecutive frames is matched if its difference value is greater than a predefined threshold then that block is considered as a spatial temporal cube (interest point) otherwise it is ignored. The advantage of the threshold is that it further reduces the small individual changes. This process is repeated for all the frames of the video.

In the next step these interest points are mapped to the interest points generated in the previous frame which reduces the complexity of the algorithm as compared to the computation of optical flow. The mapping process becomes feasible due to small number of interest points as compared to the all pixels of all frames. The mapping process evaluates the benefit of the direction followed by most of the interest points. The histogram of moving parts also increases the accuracy of matching interest points in two consecutive frames. The advantage of the block based correlation is also evaluated to increase the accuracy of matching.

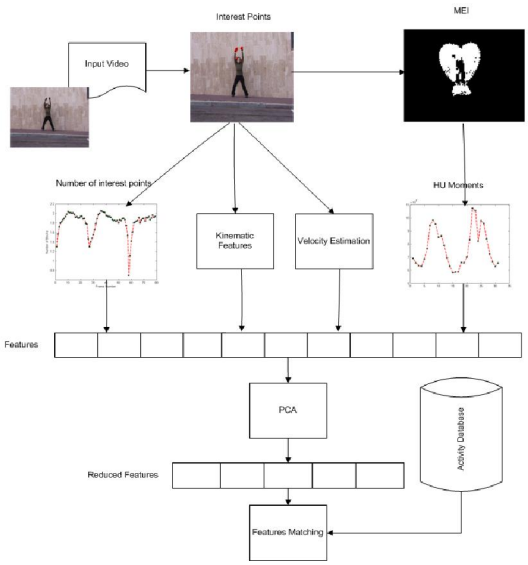

Figure 1. System diagram of human activity recognition

The outliers are removed before the matching of interest points. As shown in figure 1 rest of the features used for activity recognition are computed from these interest points. The first feature is the number of interest points generated for each frame and it is mostly different for various actions performed. Further, each type of activity has a great effect on the variation of interest points produced on each frame. The number of interest points computed for each frame is very close to each other in the running action while during performing the walking action the number of interest points varies regularly. Similarly all other actions have a close relation with the number of interest points generated on each frame of the video as shown in figure 2. Further, number of interest points generated for different actions are plotted against the frame numbers in figure 5. Number of interest points varies considerably for different actions.

2.1.1 *Kinematic Features Extraction and Reduction using PCA*

Kinematic features are those features that rely on the motion of the body parts and not on the forces responsible for motion. These kinematic features are different from the kinematic features used by Shah et al. [10]. We compute these kinematic features from the velocity of the interest points while they [10] have computed it from the optical flow vectors. We consider the interest points while they are using individual pixels of the flow vector. As the number of pixels are more than the number of interest points so our proposed approach reduces the complexity of the algorithm considerably. As a result of the mapping process the horizontal, vertical displacements and direction of these interest points are calculated. These components of velocity are then used for deriving the kinematic features. The kinematic features which we compute are divergence and vorticity. Divergence of a flow field is a scalar quantity which is defined at some point (x, ti) in space and time as the sum of the horizontal and vertical distances i.e. the partial derivatives of the point with respect to horizontal and vertical direction.

Divergence is calculated using equation (1). The importance of the divergence is that it has the ability to capture the amount of local expansion that is why it helps to discriminate between the independent and dependent motions of the different parts of the body. It means that the independent moving parts of the body will have a high value for divergence than the dependent motion of the body parts. The second kinematic feature used for human action recognition is the vorticity. It is a broader concept than rotation; still they are very closely inter-related. Vorticity also called rotation per unit area is the measure of local spin around the axis perpendicular to the plane of the flow field. It can be computed at a point (x, ti) as the difference between the partial derivative of the spatial interest points with respect to x, and y coordinates. It measures the rigidity in motion. It is computed using equation (2). It means that the divergence of bending or hand waving will be greater than that of walking which lack curl or rotation.

$$Dv(x,t_i) = \frac{\partial u(x,t_i)}{\partial x} + \frac{\partial v(x,t_i)}{\partial y} \quad (1)$$

$$Vr(x,t_i) = \frac{\partial u(x,t_i)}{\partial x} - \frac{\partial v(x,t_i)}{\partial y} \quad (2)$$

The velocity itself is also used as a feature for human action recognition because it varies for different actions. For instance, walking action has a less velocity than a running action.

Bobick et al. [13] use motion energy images (MEI) to recognize many types of aerobics exercises. The binary cumulative motion image of a sequence of frames is known as motion energy images. Let I(x, y, t) be a sequence of images and B(x, y, t) be a binary sequence of images showing the region of motion. Then the motion energy image ET (x, y, t) can be computed as given in equation (3). The MEI can be calculated using frame differencing. In equation (3) and equation (4) represents the temporal extent of an action.

$$E_T(x,y,t) = \bigcup_{i=0}^{T-1} B(x,y,t \quad i) \quad (3)$$

Obviously, if periodic motion exits in the video sequence, we can find similar poses in different frames. In other words, high correlation value can be obtained when comparing the foreground blobs in different frames within a specific period of time. In simple words the change of two consecutive frames is accumulated to form MEI. So in our algorithm we compute the MEI by accumulating the interest points of the frames till the last frame of the image sequence. Mathematically equation (4) summarizes this process. In Equation (4), 'A' represents the number of space time interest points of a particular frame. The MEI can also be achieved by thresholding the motion energy images above zero. In MHI pixel intensity is a function of temporal history of motion at that point. It is a scalar valued image. In MHI [14] more recently moving pixels are brighter than previous moved pixels. The MEI generated in our algorithm is shown in figure 3.

$$MEI(x,y,t) = \bigcup_{i=0}^{T-1} \bigcup_{j=0}^{A} I(x,y,t-i) \quad (4)$$

This MEI is then used to compute the statistical moment based features. We are using Hu moments as a part of our feature set. The use of moments as invariant binary shape representations was first proposed by Hu in 1961 [11]. He successfully used these features to classify handwritten characters. The regular moment of a shape in an M by N binary image is defined in equation (5).

$$u_{pq} = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} i^p j^q f(i,j) \quad (5)$$

In equation (5) i ε {1, . . . , M} and j ε {1, . . . , N} and p = 0, . . . , M – 1 and q = 0, . . . , N – 1. These MN moments ($u_{p,q}$) are sufficient to uniquely determine the discrete image f(i,j) with MN pixels, p+q is the order of moment. As this calculation is a function of the distance between shape pixels and the origin, so the measurements are taken relative to the shapes centroid to make it translational invariant. Therefore the coordinates of the centroid can be determined using equation (6).

$$\bar{i} = \frac{u_{10}}{u_{00}} \quad and \quad \bar{j} = \frac{u_{01}}{u_{00}} \quad (6)$$

Relative moments are then calculated using equation (7) designed for the computation of central moments.

$$u_{pq} = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (i-\bar{i})^p (j-\bar{j})^q f(i,j) \quad (7)$$

These moments individually do not have the descriptive power to uniquely discriminate arbitrary shapes. Further they do have the invariance characteristics. In this experimental study, we derive Hu feature set of 7 invariant features by combining different moments. These features can be calculated using equations (8-14).

$$M_1 = (u_{20} + u_{02}) \quad (8)$$

$$M_2 = (u_{20} + u_{02})^2 + 4u_{11}^2 \quad (9)$$

$$M_3 = (u_{30} + u_{12})^2 + (3u_{21} + u_{30})^2 \quad (10)$$

$$M_4 = (u_{30} + u_{12})^2 + (u_{21} + u_{03})^2 \quad (11)$$

$$M_5 = (u_{30} + 3u_{12})(u_{30} + u_{12})((u_{30} + u_{12})^2 - 3(u_{21} + u_{03})^2) + (3u_{21} + u_{03})(u_{21} + u_{03})(3(u_{30} + u_{12})^2 - (u_{21} + u_{03})^2 \quad (12)$$

$$M_6 = (u_{20} \mid u_{02})((u_{30} \mid u_{12})^2 \quad (u_{21} \mid u_{03})^2 \mid 4u_{11}(u_{30} \mid 3u_{12})(u_{21} \mid u_{03}) \quad (13)$$

$$M_7 = (3u_{21} - u_{03})(u_{30} + u_{12})((u_{30} + u_{12})^2 - 3(u_{21} + u_{03})^2) - (u_{30} - 3u_{12})(u_{21} + u_{03})(3(u_{30} + u_{12})^2 - (u_{21} + u_{03})^2) \quad (14)$$

These moments provide reasonably discriminative shape based features. The Hu invariant moments include area, centroid, and information about its orientation. They are less sensitive to noise, rotation, and translation.

After extracting all the features, the feature vector has a high dimensionality which may increase the computational complexity of the classification. Considering of this fact, we employ Principle Component Analysis PCA to reduce the curse of

dimensionality. PCA is a technique that reduces data dimensionality by performing a covariance analysis between factors.

2.1.2 Classification using ANN

After extracting features and applying PCA to reduce the feature vector dimensionality, an Artificial Neural Network (ANN) is trained on a set of given action videos to recognize human activity. The whole action video data set is divided into training and test sets. The training data was used to train the network through back propagation algorithm. The testing data was used to check the performance of the trained neural network. The classification of testing videos gives an indication of how well the network generalizes the classification for new unseen data set. A 3-fold cross validation method was used in which the whole data was divided into 3-folds. Each fold is composed of equal number of videos of different actions. Out of the three folds, two were used for training and the rest one is used for testing the accuracy of the neural network. The training and testing process was repeated for maximum possible combinations of the three folds.

The number of input neurons in our experimentation depends on the number of inputs. The input vector is a combination of all features which we derive as mentioned in previous section. For a 25 frames of an action sequence the feature vector has 131 data elements. It contains average distance, angle, vorticity, and divergence of all interest points of each frame. It also includes the number of interest points generated in each frame and 7 Hu moments calculated from the average motion energy images. So without using PCA the input vector has length of 131 elements. So the input neuron will be 131. After extensive experimentations 5 hidden layers having 50 neurons each, 6 neurons in output layer which represents the number of classes of action videos. The experimentation also proves that the tansig works well as activation function. Using PCA 80 input features works considerably well for the recognition of different actions. So after feature vector dimensionality reduction through PCA, number of input neurons is taken 80. Besides, the experiments are conducted on the features of each frame as well. In this case the number of input neurons depends on the number of interest points, velocity i.e. magnitude and direction of each interest point, vorticity and divergence of all interest points, and Hu moments of the Motion Energy Image generated from each frame of action sequence. It increases the computational complexity a little bit.

Table 1. Confusion matrix for our algorithm

Activity	Hand Clapping	Hand Waving	Jogging	Boxing	Running	Walking
Hand Clapping	88.5	2.2	8.13	1	0	0
Hand Waving	2.2	92.1	1.2	2	1.9	1.1
Jogging	3.5	6.4	88.6	2	0	0
Boxing	0	2.1	1	88	5.34	3.2
Running	0	2.61	1.2	3	88.1	4.87
Walking	0	0	1.1	5	7.92	86.31

3 Results & Discussion

In order to verify the performance of our algorithm, we performed experiments on many datasets including KTH dataset and [2, 6]. The KTH dataset consists of six different actions which includes hand clapping, hand waving, boxing, running, walking and jogging. It is a challenging data set due to the zooming in and out, and change in the size of the actor. Confusion matrix obtained through our proposed method for the KTH data set is presented in table 1.

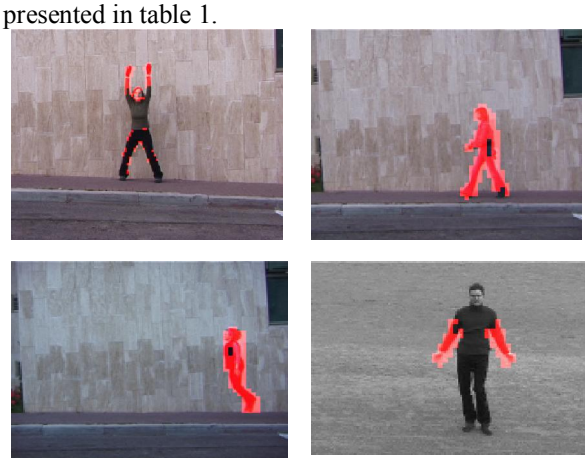

Figure 2. Interest points generated for different activity sequences

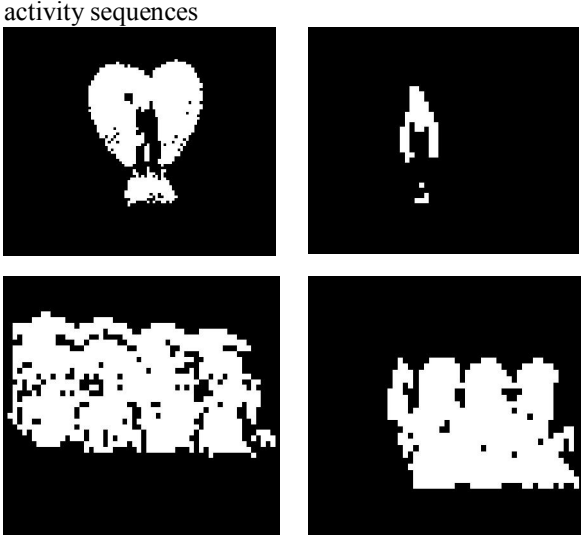

Figure 3. Motion energy images for different activity sequences

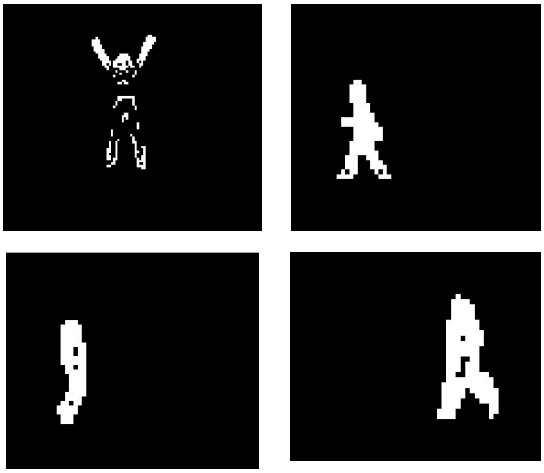

Figure 4. Interest points generated for a single frame of different activity sequences

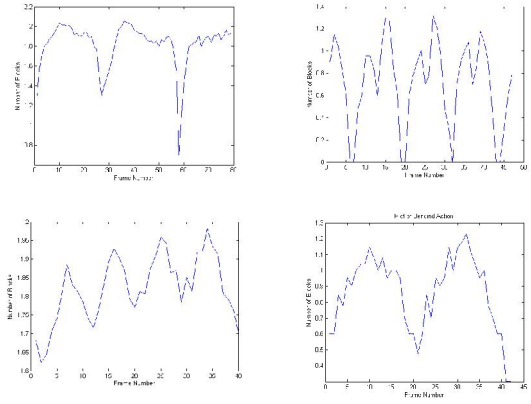

Figure 5. Number of interest points generated for 25 frames for different activity sequences

The figure 4 shows the foreground background subtractions and is achieved through the proposed method from space time interests' points. It also plays a vital role in recognizing different human actions. These images can be also used to form Motion Energy Histograms for extracting further discriminating features as well. The further effects of the number of space time interests points is displayed in figure 5. Our method of space time interest points form images that are very similar to the low resolution images. The main advantage of the proposed algorithm is that it is computationally less expensive. We have implemented the proposed technique in Matlab 7. Firstly, it processes 13 frames per seconds, which indicates that its processing speed is more than state of the art approaches. Secondly, our proposed approach is totally dependent on motion features which are computed from space time interest points generated through the proposed technique. The accuracy of our algorithm is 88.61%, which is quite good on the basis of that much compromise on computational complexity. This accuracy is a little bit good from state of the art approaches of human activity recognition. Its main advantage is the least computational complexity as compared to those methods.

4. Conclusion

Neural network classifier based human activity recognition in videos has been proposed. Providing the ability to see and understand as humans do has fascinated scientists, engineers and even the common man. In this paper we have explored the use of kinematic features, derived from motion information of the interest points [10] for the task of human activity recognition in videos. As motions in human activity videos become more complex the optic flow is more difficult to discern, while our features depends on the interest points instead of optical flow. The numbers of interest points are also playing an important role in the recognition of different human actions and we use appearance based technique. We can find out the continuity and repetition of an action using these interest points and this makes it computationally more feasible. The dominant features are selected by employing PCA on each feature set. The selected features are then used to train and query a neural network, and the recognition rate of the human activities is used as an evaluation measure of the system. In future work, we intend to explore new features sets which may further improve accuracy of human activity recognition system. Our method of interest points generation can be used for background foreground subtraction with some amendments. It has the ability to handle the background noise efficiently.

Acknowledgements:

This research work is supported by BK21, South Korea under Postdoc fellowship.

Corresponding Author:
Dr. Arfan Jaffar
Department of Computer Science,
National University of Computer and Emerging Sciences, Islamabad, Pakistan
E-mail: arfan.jaffar@nu.edu.pk

References

1. D. M. Gavrila, The visual analysis of human movement: A survey, Computer Vision and Image Understanding, Vol. 73, no. 1, January 1999, pp. 82–98.
2. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, Actions as Space-Time Shapes, IEEE ICCV, Vol.2, 2005, pp. 1395–1402.
3. Y. Ke, R. Sukthankar, and M. Hebert, Efficient Visual Event Detection using Volumetric Features, IEEE ICCV, Vol. 1, 2005, pp. 166–173.
4. J. Liu, S. Ali, and M. Shah, Recognizing Human Actions Using Multiple Features, IEEE CVPR, 2008, pp. 1–8.
5. P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
6. C. Schuldt, I. Laptev and B. Caputo, Recognizing Human Actions: A Local SVM Approach, IEEE ICPR, Vol. 3, 2004, pp. 32–36.
7. A. Efros, A. Berg, G. Mori and J. Malik, Recognizing Action at a Distance, IEEE ICCV, Vol. 3, 2003, pp. 726–733.
8. Y. Yacoob and M. Black, Parameterized Modeling and Recognition of Activities, Computer Vision and Image Understanding, Vol. 73, 1999, pp. 232–247.
9. J. Little and J. Boyd, Describing Motion For Recognition, Proceedings of International Symposium on Computer Vision, 1995, 235–240.
10. Ali. S, Shah. M, Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 31, Issue 2, Feb. 2010, pp. 288–303.
11. M. K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Information Theory, Vol. 8, pp. 179–187, 1962.
12. D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision (IJCV), Vol. 60, Issue. 2, 2004, pp. 91–110.
13. A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. IEEE Trans. on PAMI, Vol. 23, Issue. 3, 2001, pp. 257–267.
14. D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision (IJCV), Vol. 60, Issue. 2, 2004, pp. 91–110.

12/2/2012