Human Action Recognition Using Joint Positions from Depth Videos

Adarsh S, Asha S

Abstract: Human Action Recognition using visual information in a given image or sequence of images, has been an active area of research in computer vision applications. The image captured by conventional camera does not provide the suitable information to perform comprehensive analysis. However, depth sensors have recently made a new type of data available. Most of the existing work focuses on body part detection and pose estimation. A growing research area addresses the recognition of human actions based on depth images. In this paper, the following contributions are made: the proposed method makes an efficient representation of human actions by constructing a feature vector based on the human's 3D joint positions. These locations are extracted fromdepth videos which are taken with the help of Microsoft Kinect sensor. Experiments were performed on a new dataset Kinect Action Dataset (KAD-10). The data set consists of 3D sequences of 10 indoor activities performed by 10 individuals in varied views. Then these feature vectors are given to K-Nearest Neighbour (KNN) classifier to perform the action classification task which results in action labels.

Index Terms*: video surveillance, Depth sensor, Body part labeling, Depth image features, Randomized decision forest, joint position estimation, k-nearest neighbour algorithm.

I. INTRODUCTION

Human action analysis has been widely used in several applications including video surveillance, gaming, human computer interaction, security and health care[1-3]. Behaviour analysis is the foremost stage in intelligence surveillance. As the amount of video data collected each day by surveillance camera has increased, the application of automatic systems to detect and recognize different activities of people and objects has also increased.Besides the complexity of human actions, the large diversity of human body, appearance, posture, motion and illumination changes is a very challenging task in the field of automatically recognizing human actions. Same person will perform the same action differently at different times and even different person will perform the same action in a different manner. Research has been reported on human action recognition using features extracted from RGB images[4]. Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) are the commonly used methods to classify actions[5-8]. However the RGB images captured by conventional camera doesn't provide enough information to perform appropriate analysis on human actions. Moreover, intensity images have many difficulties in robustly performing computer vision tasks such as background subtraction and object segmentation. However depth images overcome the limitations of the intensity images.

Revised Version Manuscript Received on June 27, 2015.

Adarsh S, PG Scholar, Department of Electronics and Communication, SCT College of Engineering, Trivandrum, Kerala, India.

Asha S, Asst. Prof., Department of Electronics and Communication, SCT College of Engineering, Trivandrum, Kerala, India.

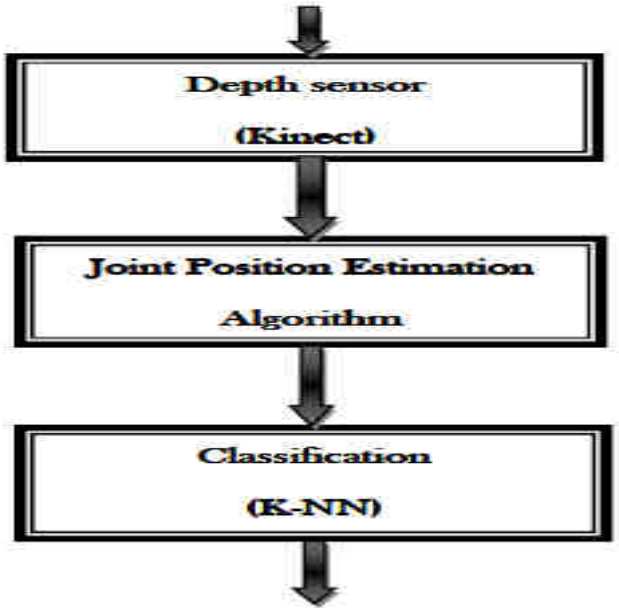

Fig. 1. Overview of the method

In this paper, an efficient representation of human actions based on joint position algorithm is presented. Human performs an action which is captured by the depth sensor which converts it into depth videos. The feature vectors are extracted from the depth videos using joint position estimation algorithm. The results of this algorithm produces the various joint positions of the human body. These joint positions are trained using K-nearest neighbour classifier and finally they are classified into the corresponding action labels.

II. DEPTH IMAGE

A depth image can be defined as an image that contains information regarding the distance of the surfaces of the scene objects[9]. The Pixels in a depth image indicates the calibrated depth in the scene, rather than a measure of intensity or colour. Depth sensor or depth camera is used to capture a depth image. The latest depth cameras offer numerous advantages over the traditional intensity sensors which includes working in low light levels, giving a measurable scale estimate, being a texture and colour invariant and determining silhouette ambiguities in the pose. Computer vision tasks such as background subtraction and object segmentation can also be simplified to a great extent. One of the depth sensors which has set a remarkable development in the last few years is the Microsoft Kinect. It has reached a consumer price point and has shown atmost reliability on capturing depth images. The merits of the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Microsoft Kinect includes high sample rate and capability of combing visual and depth information. Depth images can be obtained by using three main sensing technologies such as stereo cameras, Time-of-Flight cameras and structured light [10]. The Microsoft Kinectmake uses of structured light technology. The Kinect consist of an RGB camera, IR projector and an IR camera. It provides images at 640 x 480 pixels, 30 frames per second and a practical ranging limit of 0.8 to 4 meters.The RGB image is obtained from the Kinect sensor and its corresponding depth image is shown below:

Fig. 2. RGB image and its depth image.

III. JOINT POSITION ESTIMATION

The various steps in the whole pipeline of the joint position estimation are divided into four. The first step is the body part labellingThe second step deals with the depth image features. Then the phenomenon of randomized decision forest classifier is used. Finally the body joints are hypothesized by local mode-finding approach based on mean shift.

A. Body Part Labelling

The main significance of the work is the intermediate body part representation. Each pixel of the depth image is segmented as a per-pixel classification task. The various body parts that densely cover the body are coded into different colours as shown in Fig 3. The pairs of depth and body part images are used as fully labelled data for learning the classifier. Here 31 body parts : LU/RU/LL/RL head, neck, L/R shoulder, LU/RU/LL/RL arm, L/R elbow, L/R wrist, L/R hand, LU/RU/LL/RL torso, LU/RU/LL/RL leg, L/R knee, L/R ankle, L/R foot. Different parts for the left and right allows the classifier to distinguish the left and right sides of the body.

Fig. 3. Depth image and body parts.

B. Depth Image Features

Simple depth comparison features can be computed for a pixel x as follows:

$$f_\theta(I,x) = d_I\left(x + \frac{u}{d_I(x)}\right) - d_I\left(x + \frac{v}{d_I(x)}\right), \quad (1)$$

where, dI (x) is the depth at pixel x in image I and parameters θ = (u, v) describe offsets u and v. The normalization factor is given by $\frac{1}{dI(X)}$ which ensures that the features are depth invariant. At a given point on the body, if an offset pixel lies on the background or outside the bounds of the image, the depth pixel will have a large positive constant value.

Fig. 4.a Feature locations give large depth difference response.

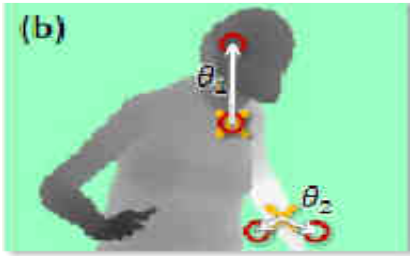

Fig. 4.b Feature locations give small depth difference response.

Fig 4 indicates two features at different pixel locations x. Fig 4.a shows one of the pixel lying outside or in the background resulting in larger depth difference response whereas Fig4.b shows one of the pixel lying in the foreground and hence resulting in smaller depth difference response. Feature $f_{\theta 1}$ point looks upwards. Equation 1 gives a positive value for pixel x near the top of the body and a value close to zero towards the lower part of the body. Feature $f_{\theta 2}$ helps to find the vertical structures such as the arm.

C. Randomized decision forests

This method is considered to be one of the most quick and effective multi-class classifiers for various tasks. This is done with the help of GPU (Graphics Processing Unit).

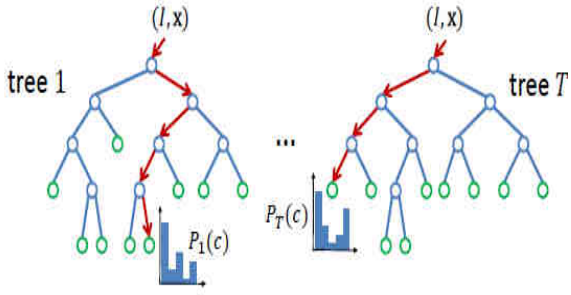

Fig. 5. Randomized Decision Forests. Each tree consists of split nodes (blue) and leaf nodes (green).

Fig.5illustrates a forest as collection of T decision trees each consisting of a split nodes and leaf nodes. Each split node contains a feature $f_\theta$and a threshold value τ. In an image I, the pixel x is classified starting from the root node and equation 1 is evaluated repeatedly. Here each split node branches towards the left or right depending upon the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

threshold value τ. Once the leaf node is reached, a learned distribution $P(c|I,x)$over body part label c is stored. The final classification of all trees in the forest is given by:

$$P(c|I,x) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,x), \quad (2)$$

D. Mode-finding approach based on mean shift

The global 3D centers of probability mass for each part of the human are accumulated using the known calibrated depth. However, the outgoing pixels degrade the quality of the global estimate. Therefore, a local mode-finding approach based on mean shift with a weighted Gaussian kernel is used. Hence a density estimator per body part may be defined as:

$$f_{c(\hat{x})} \propto \sum_{i=1}^{N} w_{ic} \exp\left(-\left\|\frac{\hat{x}-\hat{x}_i}{b_c}\right\|^2\right), \quad (3)$$

where $\hat{x}$ - a coordinate in 3D world space.
N -number of image pixels.
w_{ic}- pixel weighting.
$\hat{x}_i$ - reprojection of the image pixel x_i
b_c- is a learned per-part bandwidth

The pixel weighting function considers both the inferred body part probability at the pixel as well as the world surface area of the pixel as:

$$w_{ic} = P(c|I,x_i).d_I(x_i)^2, \quad (4)$$

Depending on the labelling of body parts, the posterior$P(c|I,x_i)$, can be pre-accumulated over a small set of parts. Here the thirty-one body parts covering the human body are merged to twenty joint positions. Each joint is represented by its 3D coordinates.

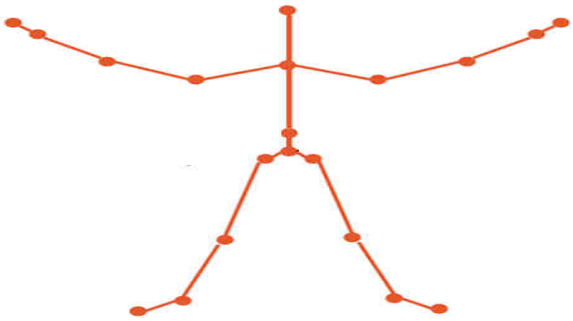

Fig. 6. Twenty joint locations of human body

Twenty joint positions of human body are: Hip center, Spine, Shoulder center, Head, Shoulder left, Elbow left, Wrist left, Hand left, Shoulder right, Elbow right, Wrist right, Hand right, Hip left, Knee left, Ankle left, Foot left, Hip right, Knee right, Ankle right, Foot right.

IV. K-NEAREST NEIGHBOUR CLASSIFIER

It is a type of non-parametric pattern classification method. In the nearest neighbour algorithm, the class is predicted to be the class of the closest training set (ie, when k=1). K-NN is a type of instance-based learning and the functions are only approximated locally. The object is classified according to the majority vote of its neighbours, and it is assigned to the class most common among its k-nearest neighbours.

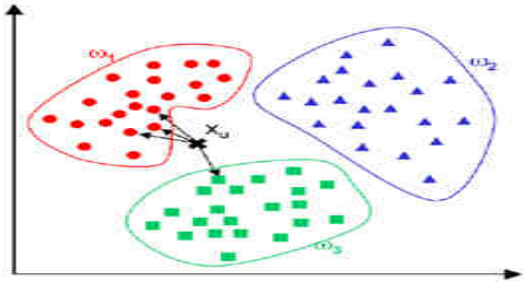

Fig. 7. A unknown template x_u

Fig.7 shows the situation when k is taken 4. The unknown point will be compared with the four nearest neighbours and the result determines as to which set, the point should belong to. The training sets are vectors in a multi-dimensional feature space, each with a class label. The feature vectors and class labels of the training sets are already assigned in the training phase. Usually Euclidean distance is used as the distance metric for classification. Consider a multi-dimensional space having two points x and y where each point an n-dimensional vector, ie, $x = \{x_1, x_2.x_3, \ldots.x_n\}, y = \{y_1, y_2.y_3, \ldots.y_n\}$,
The distance function can be defined as $d_E(x,y)$ between two points by measuring their distance according to Euclidean formula.

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad (5)$$

Our approach uses Euclidean formula to calculate the distance between any given 3-dimensional points. This is based on the minimum distance from the test samples to the training samples to determine the nearest neighbours

V. EXPERIMENTAL SETUP

The experiments were implemented on a 2.5GHz Intel Core i5 PC with 4GB memory, running under Windows7Enterprise. The algorithm is coded using MATLAB R2014a.Experiments were performed on a new dataset Kinect Action Dataset (KAD-10).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Fig. 8: Sample images from videos of the 10 activities in our database. We show RGB image frames as well as the corresponding skeleton joint positions.Action type from left to right, top to bottom: high arm wave,clap hands, hand wave,sit down, dinking, hands up, pick up, walk, phone call, bend.

The data set consists of 3D sequences of ten indoor activities performed by ten individuals. The sequences are taken using a single stationary Kinect. The RGB images and depth maps were captured at 30 frames per second. The 10 actions include: *High Arm Wave, Sit Down, Hand Wave, Walk, Hands Up, Clap Hands, Bend, Drink, Pick, Phone call.* The length of each action ranges from 0 to 100 frames.Fig. 8 shows the RGB images of ten human actions and their corresponding skeletal joint position images which has been derived from their respective depth videos. The experimental results of the various actions are shown below:

ACTIONS	ACCURACY RATE
High arm wave	90%
Sit down	87%
Hand wave	90%
Walk	82%
Bend	80%
Clap hands	87%
Hands up	88%
Drink	83%
Phone call	84%
Pick	84%

Table.1

Table shows the efficiencies of the corresponding 10 human actions. The highest recognition rate is 90% (High arm wave and Hand wave), while the worst is 80% (Bend). Finally the human actions are classified and are labelled accordingly.

VI. CONCLUSION

In this paper, a low dimensional representation of human actions by developing a feature vector based on the human joint positions which are extracted from depth images is proposed here. Then, passing these feature vectors to K-nearest Neighbour (K-NN) to perform the classification task. The experimental results demonstrate the superior performance of the proposed approach to the state-of-the-art methods. In the future, more action categories can be included as more diverse and complicated movements. Hence more complex activities can be explored to exploit the effectiveness of this technique. Furthermore, this method can be implemented in real-world applications also.

ACKNOWLEGMENT

For the successful completion of this paper, there are people behind the screen who contribute a great deal. I take this opportunity to express my sincere and profound gratitude to my thesis guide AshaS(Assistant Professor, Department of Electronics and Communication) for her expert advice, suggestions and encouragement. I would also like to thank my P G coordinator MrLibish T M (Assistant Professor, Department of Electronics and Communication) for his valuable support and guidance. I would also like to extend my gratitude to my family for their cooperation in all measure. Above all, I would like to thank the Almighty for his immeasurable blessings showered upon me to make this

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

venture a success

REFERENCES

[1] Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta, "A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-Assisted Living" In the International Journal of Expert Systems with Applications, Volume 39, Issue 12, September 2012.
[2] Ronald Poppe, "A Survey on Vision-Based Human Action Recognition," In Image and Vision Computing Journal,Volume 28, Issue 6, pp 976-990, June 2010.
[3] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. In *ACM Computing Surveys*, 2011.
[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. CVPR*, Anchorage, AK, USA, pp. 1–8, 2008.
[5] N. Dalal and B. Triggs, "Histograms of oriented gradientsfor human detection," in *Proc. IEEE Computer Soc. Conf. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.
[6] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines,"*Mach. Learn.*, vol. 46, no. 1, pp. 131–159, 2002.
[7] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2–3, pp. 107–123, Sept. 2005.
[8] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas, "Multi-view Human Action Recognition: A Survey," In Proceedings of the 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '13), Beijing, China, pp. 522-525,October 2013.
[9] Vennila Megavannan, Bhuvnesh Agarwal, and R.Venkatesh Babu, "Human Action Recognition using Depth Maps," In proceedings of the International Conference on Signal Processing and Communications (SPCOM),Bangalore, India, pp. 1-5, July 2012.
[10] Lulu Chen, Hong Wei, James Ferryman, "A Survey of Human Motion Analysis Using Depth Imagery," In Pattern Recognition Letters, Volume 34, Issue 15, pp. 1995–2006,November 2013.

Adarsh S is currently pursuing the M.Tech. degree in Signal Processing with the Department of Electronics and Communication Engineering, SCT College of Engineering,Thiruvananthapuram, Kerala. She received the B. Tech degree from the University of Kerala, Thiruvananthapuram, in 2011 in Electronics and Communication Engineering. His research interests include signal processing, image processing, Pattern Recognition and video processing.

Asha S is the Assistant Professor, Dept. of Electronics and Communication Engineering., SCT College of Engineering, Trivandrum. Kerala. She received the M.Tech degree in Signal Processing from College Of Engineering Trivandrum, in 2013. She received B.Techdegree in Electronics and Communication Engineering from University of Kerala, India in 2002.She has more than 13 years teaching experience. She has technical paper publication in an International Journal, Conferences and National conferences.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.