Elysium Journal of Engineering Research & Management

P-ISSN: 2347-4408
E- ISSN: 2347-4734

ANALYSIS OF HUMAN ACTIVITY RECOGNITION USING ACTION RECOGNITION METHODS

R. Anuradha[1], Dr.S.Ravimaran[2]

[1] *PG Scholar, M.A.M College of Engineering, Trichy, India. Email: psganu@gmail.com*
[2]*Principal, M.A.M College of Engineering, Trichy, India.Email:principalmamce@mamce.org*

Abstract— The labelling of image sequences with action labels is termed as human action recognition and the main goal of this process is the series observations on actions. Human activity recognition aims to make better representation using deep and shallow learning techniques. For humans moving in the scene, we use techniques for tracking, body pose estimation, or space-time shape templates and categorize activities based on the video's over- all pattern of appearance and motion using spatio - temporal interest operators and local descriptors to build the representation. It can be support for different fields of studies such as human computer interaction, or sociology, medicine and other applications. The system uses Binary Motion Image (BMI) to perform human activity recognition after preprocessing and used as input for Convolutional neural network (CNN) algorithm to recognize the human activity. The CNN combined with support vector machine (SVM) algorithm can be used as the classifiers to recognize the human action. The algorithm has three kind of inputs test samples, train samples, train labels. This project proposes a joint learning framework to identify the performance of Deep and Shallow Classifiers in training videos.

1. INTRODUCTION

An emerging active area of research in computer vision with wide scale of applications in video surveillance, virtual reality, computer human interfaces (robotic interaction with humans), sports video analysis etc. is Human activity recognition. This method first takes three-dimensional model of a person and then recognizes the motion using representation and recognition theory that decomposed motion-based recognition describing where there is motion (the spatial pattern) and then describing how the motion is moving. Self – Occlusion due to motion overlapping makes the task daunting for motion recognition methodologies address to recognize and understand varieties of human activities. These methods either bypass this problem or solve this problem in complex manner. There are various approaches for activity recognition such as (i) Spatio-temporal (ii) Frequency based (iii) local Descriptors (iv) Shape Based and (v) Appearance based. In this paper, we concentrate on motion self-occlusion problem due to motion overlapping in various complex activities for recognition.

This paper is aimed to improve the performance of dense trajectories in action recognition with different ways. Recently, many research efforts have focused on recovering human poses, which is considered as a necessary step for viewing invariant human action recognition. However, 3D pose reconstruction from a single viewpoint is a well-known difficult problem in itself because of the large number of parameters that need to be estimated and the ambiguity caused by perspective projection.

2. RELATED WORKS

This paper presents methods along with the BMI method, which is very widely employed method for various applications. For example, using the BMI method, Davis et al. has developed a virtual aerobics trainer that watches and responds to a user as he/she performs a workout. An interactive art demonstration can be constructed from the motion templates. An interactive and narrative play space for children, called Kids Room was developed using the BMI method successfully. Yau et al. has developed a method for visual speech recognition employing the BMI method. The video data of the speaker's mouth is represented using grayscale images named as motion history image. Automatically localizing and tracking moving person or vehicle for an automatic visual surveillance system was demonstrated in by employing the BMI method before employing an extended mean shift approach.

In recent years, various approaches have been proposed for human recognition by gait. These approaches can be divided into two categories: model-based approaches and model-free approaches.

Model-based gait recognition approaches focus on recovering structural model of human motion, and the gait patterns are then generated from the model parameters for recognition. Niyogi and Adelson [14] make an initial attempt in a spatiotemporal (XYT) volume. They first find the bounding contours of the walker, and then fit a simplified stick model on them. A characteristic gait pattern in XYT is generated from the model parameters for recognition. Yoo et al. [19] estimate hip and knee angles from the body contour by linear regression analysis. Then trigonometric-polynomial interpolant functions are fitted to the angle sequences, and the parameters so-obtained are used for recognition. In Lee and Grimson's work [11], human silhouette is divided into local regions corresponding to different human body parts, and ellipses are fitted to each region to represent the human structure. Spatial and spectral features are extracted from these local regions for recognition and classification. Bhanu and Han [5] propose a kinematic-based approach to recognize individuals by gait. The 3D human walking parameters are estimated by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. Human gait signatures are generated by selecting features from the estimated parameters. In these model-based approaches, the accuracy of human model reconstruction strongly depends on the quality of the extracted human silhouette. In the presence of noise, the estimated parameters may not be reliable. To obtain more reliable estimates, Tanawongsuwan and Bobick [17] reconstruct the human structure by tracking 3D sensors attached on

P-ISSN: 2347-4408
E- ISSN: 2347-4734

fixed joint positions. However, their approach needs lots of human interaction which is not applicable in most surveillance applications.

Son et al. has calculated the MHI and then combined with background model to detect candidate road image. Gait History Image and Gait Energy Image are created for gait analysis based on the concept of the MHI has a threat assessment method for automated visual surveillance with the aid of the MHI. A PDA-based recognition system based on the MHI method is developed by Petras et al. have devised a flexible test-bed for unusual behavior detection and automatic event analysis using the MHI. A recent method by Kellokumpu et al. [20] have proposed a recognition method at the top of the MHI method using texture-based description of the movements by employing local binary pattern (LBP) operator. Later they have used HMM for recognition.

Yuxiao Hu, Liang liangCao,FengjunLv, Shuicheng Yan , Yihong Gong and Thomas S. Huang (2006), "Action Detection in Complex Scenes with Spatial and Temporal Ambiguities", ShugaoMaJianming Zhang NazliIkizler-CinbisStanSclaroff (2013), Action Recognition and Localization by Hierarchical Space-Time Segments", these works employ multi-instance learning (MIL) based Support Vector Machine (SVM) to handle these ambiguities in both spatial and temporal domains. This multi-instance method provides a way to not only recognize the action of interest, but also locate the exact position and time period of the action. The paper called the proposed algorithm as Simulated annealing Multiple Instance Learning (SMILE). The action detection in complex scenes with cluttered backgrounds or partially occluded crowds, it is very difficult to locate human body precisely. In addition, ambiguities may also exist in temporal domain.

Xinxiao Wu Dong XuLixinDuan (2011), Action Recognition using Context and Appearance Distribution Features, this paper first proposes a new spatio-temporal context distribution feature of interest points for human action recognition. Each action video is expressed as a set of relative XYT coordinates between pairwise interest points in a local region. This paper first propose a new visual feature by using multiple GMMs to characterize the spatio-temporal context distributions about the relative coordinates between pairwise interest points over multiple space-time scales. Specifically, for each local region (i.e., sub-volume) in a video, the relative coordinates between a pair of interest points in XYT space is considered as the spatio-temporal context feature. Then each action is represented by a set of context features extracted from all pairs of interest points over all the local regions in a video volume. Gaussian Mixture Model (GMM) is adopted to model the distribution of context features for each video. However, the context features from one video may not contain sufficient information to robustly estimate the parameters of GMM.

3. OVERVIEW OF METHODS

Binary Motion image is built by combining the action sequences into a single image. This is actually a 2D representation of the image retained by View based methodologies where the entire raw image as a single image in HD space. Activity is recognized with the help of this. The outstanding performance of this method compared to others is presented in the section below. This method can be easily scaled to 3D depth maps.

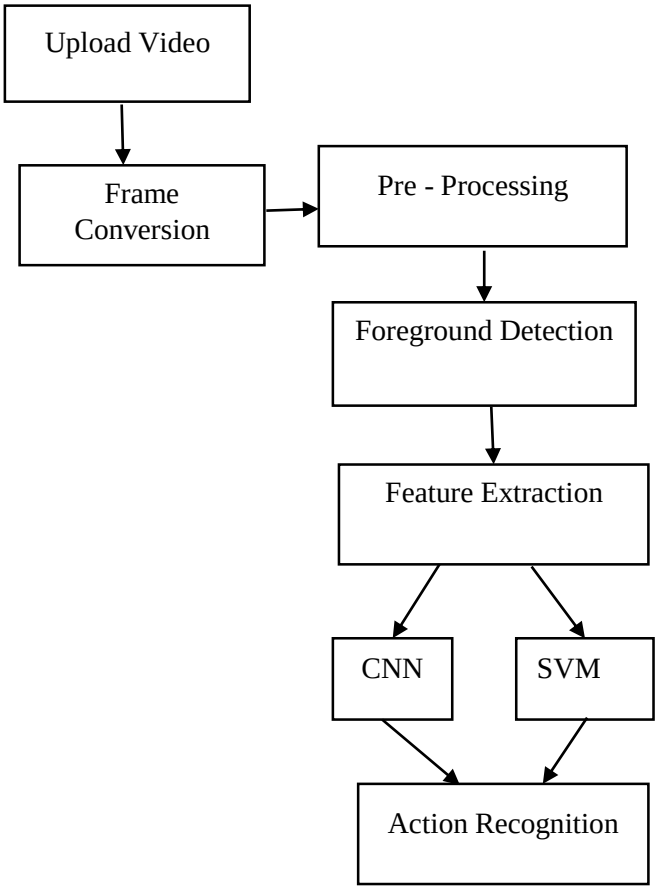

Fig.3.1. Overview of methods for analysis.

Our paper has various modules and they are described as follows:

- Frame conversion & Preprocessing
- Foreground detection
- Feature extraction
- Feeding into CNN and SVM
- Action recognition

3.1. FRAME CONVERSION AND PREPROCESSING

Frame conversion is conversion of video to image so that there is a separate value to know about the value of element and the pixel value localization results and compare the image process and

P-ISSN: 2347-4408
E- ISSN: 2347-4734

implement the function. Prior to analyzing main data and extracting of information, Preprocessing is required. This includes operations on images at lowest level of abstraction. Here the input and output are intensity images. This suppresses unwanted distortions and enhances image features that can be given for further processing. The video is uploaded, image values are segmented and ultimately noise is reduced here.

3.2. FOREGROUND DETECTION

The Foreground Detection aims in detecting changes in image sequences. This separates the changes taking place in background. This detects changes when background is set. So the background model should be developed first. In spite of shapes shadows and moving objects, it should be robust to lighting changes, repetitive movements and long term changes. System object compares a color or gray scale video frame to a background model to determine whether individual pixels are part of the background or the foreground. The color is first compared and segmentation is done after element detection. The output of this step is given as input to feature extraction.

3.3. FEATURE EXTRACTION

Feature extraction aims for a transformation of large redundant set of input data into a reduced set of features. These features contain relevant information from input data. This makes an individual to perform the desired action on this feature that is informative and leading to better human interpretations. This is useful for image matching and retrieval. Here we give MEI, MHI, DMHI, BMI and GEI as inputs. The process here is to represent image parts, which is then matched and retrieved as shown in the figure.

3.4. CNN and SVM

A general architecture of CNN is composed of input map such as image, a number of hidden feature maps and output processing layer. To feature map layer is obtained when convolution with a trainable kernel is done. The Gabor like filters obtain edges along different orientations. This is followed by an activation function. This constitutes the first step. The second step is sub-sampling, which involves averaging or max-pooling sub-region and obtaining a spatially down-sampled map. It consists of a single trainable weight and additive bias. This is done to reduce the size of maps and also helps in imparting a small degree of shift and distortion invariance. After a series of convolution and sub sampling layers, a convolution map is developed by randomly selecting a number of trainable weights and obtaining a single matrix. This helps in exploring different features during training. Finally, a linear transformation is applied to obtain an output layer to tell which class is classified. Similar to ANNs, CNNs are trained using Feed Forward and Back Propagation algorithms while keeping in mind the concept of shared weights.

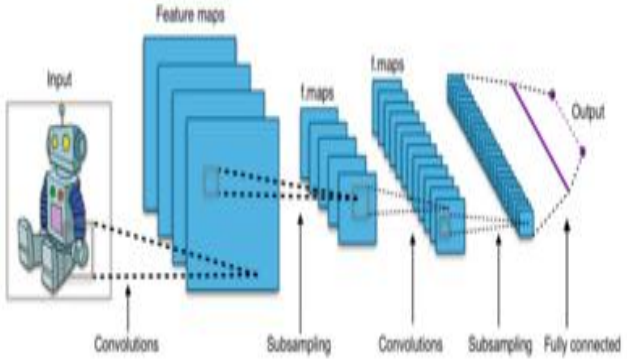

Fig.3.2. CNN Architecture

The above figure shows most common form of a CNN architecture stacing a few layers, following them with POOL layers, and repeating this pattern until the input has been merged spatially to a small size. Convolutional neural networks (CNNs) consist of multiple layers of small neuron collections which process portions of the input image, called receptive fields. Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at some point, it is common to transition to fully connected layers. The last fully connected layer holds the output.

SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible

3.5. ACTION RECOGNITION:

The action recognition will be performed by comparing the outputs from CNN and SVM as mentioned below in the upcoming section.

4. ANALYSIS OF ACTION RECOGNITION METHODS

4.1. VIEW BASED ALGORITHM WITH BMI AND CNN

View based recognitions use visual templates for recognition and do not extract complex features from the image. Instead they retain the entire raw image as a single feature in high dimensional space. These templates are learnt under different poses and illumination conditions for recognition. With this in mind we build an idea of 2-D representation of action sequence by combining the image

P-ISSN: 2347-4408
E- ISSN: 2347-4734

sequences into a single image called Binary Motion Image (BMI) to perform activity recognition. We test our method on Weizmann dataset focusing on actions that look similar like run, walk, skip etc. We also extend our method to 3-D depth maps using MSR Action 3D Dataset by extracting the BMI projections namely front, side and top views.

This BMI representation has a number of characteristics like being local, the features have robustness to viewpoint changes and occlusions; being relatively sparse, they can be stored and manipulated efficiently. Further, by including both dynamic and static components (e.g., optical flow and gradient histograms), they can capture not only what kind of motion occurs, but also what kind of context and actors are present, without requiring reliable tracks on a particular subject. Various developments building on this general framework have yielded impressive results for realistic activities in Hollywood movies or YouTube videos.
BMI combines the image sequence using the following equation:

$$BMI(x,y) = \sum_{t=1}^{n} f(t)\, I_{xy}(t)$$

Where BMI(x,y)is the BMI, $I_{xy}(t)$ is the binary image sequence containing the ROI and f(t) is the weight function which gives higher preference to more recent frames and n is the total number of frames. Here, the quadratic function t^2 is used as a weight function for best looking results. Lastly, a bounding box around the image is used to extract only the region of interest in the image and to discard the black background. The BMI is post processed by applying a normalization operation. The weight function provides a means of depicting the flow of motion in an action or its optic flow. In this way, both the spatial and temporal dimensions of the activity performed are modelled using BMI.

4.2. 2-D WEIZMANN and MSR ACTION 3D DATASET

The Weizmann database is selected for testing purpose. It contains activities performed by 9 individuals from which we selected 5 actions namely Jump,Run, Side and Walk. These are selected so as to judge our method on similar looking actions. Forthis Database BMI is calculated as described aboveand this will serve as input to CNN classifier. Matlab is used for extracting BMI. The below figure showsthe side and skip action and corresponding BMI.

For Human Activity Recognition from 3-D data, this database is used. Consisting of 10 people performing 20 actions with each action performed by an individual, 3 BMI s ate obtained. The first image is the depth map of the forward kick and the second is from which three BMI s are calculated for front, side and Top view respectively.

4.3. ACTION RECOGNITION USING SVM

The system designs a latent support vector machine (SVM) model in which the spatial and temporal extents of the action of interest in positive examples are treated as latent variables. Through iterative learning, spatial and temporal extents of the action in each positive example can be learned simultaneously with the model parameters. During the detection phase, we first apply our split-and-merge algorithm to process trajectories extracted from a query video. Then, action detection is conducted based on the matching quality between the trained latent SVM model and each foreground moving object in the query video.

4.4. OVERCOMING DISADVANTAGES USING BMI

In above system the system uses Binary Motion Image (BMI) and Convolutional Neural Networks to perform human activity recognition. Image re sampling, noise reduction and color changes apply the pixel value noise. Gray image value is to change the original image and image filtering and improve the visualization and brightness of the particular place and also to improve the visual appearance and manual datasets. After this the median filtering to find the value average for the image to analysis. But the action recognized using SVM is not as clear as that of BMI and may be subjected to self-occlusion.

5. CONCLUSION:

This paper presents the performances when inputs are fed into CNN and SVM. This has also shown the robustness of SVM representation. Though the performance of the SVM and the employed feature vectors for recognition are satisfactory, we find that this method cannot perform well or faces difficulties in some cases. When two or more than two persons are in-view, this method cannot recognize properly, especially when all of them are moving in different directions. Even with several templates per individual, if they are from similar conditions, the learned features may be overfit to the training templates.

In this paper, we have chosen a new view-based algorithm for recognizing human activities. Our method stacks all video frames into a single image to form the BMI which demonstrates the flow of motion of the action and is invariant to holes, shadows and partial occlusions. This method is then extended for activity detections using 3-D depth maps. The performance shown by our algorithm on both 2-D and 3-D datasets support our hypothesis. Our method includes a slight level of invariance to translation, rotation and scale changes mainly due to sub-sampling layer in CNN. Due to the use of binary foreground masks, the method is independent of dress style worn by individuals. Also the method is invariant to speed of the action performed. So we conclude that BMI is very efficient when compared to SVM.

P-ISSN: 2347-4408
E- ISSN: 2347-4734

REFERENCES

[1]Analysis of Motion Self-Occlusion Problem Due to Motion Overwriting for Human Activity Recognition Md. Atiqur Rahman Ahad, JooKooi Tan, HyoungSeop Kim and Seiji Ishikawa Faculty of Engineering, Kyushu Institute of Technology, Fukuoka, Japan, JOURNAL OF MULTIMEDIA, VOL. 5, NO. 1, FEBRUARY 2010
[2]Application of SAD Algorithm in Image Processing for Motion Detection and SIMULINK Blocksets for Object Tracking, Menakshi Bhat1, Pragati Kapoor2, B.L.Raina3 1Assistant Professor, School of Electronics & Communication Engg.
[3]Individual Recognition Using Gait Energy Image , Ju Han and BirBhanu Center for Research in Intelligent Systems University of California, Riverside, California 92521, USA
[4]Human Activity Recognition using Binary Motion Image and Deep Learning, Tushar Dobhal, Vivswan Shitole, Gabriel Thomas, Girisha Navada.
[5]Arandjelovic R. and Zisserman A. (2012), "Three things everyoneshouldknow to improve object retrieval", in IEEE Conference, pp. 2911–2918.
[6]Brendel W. and Todorovic S. (2011), "Learning spatiotemporal graphs ofhuman activities," in IEEE International. Conference, pp. 778–785.
[7]Brox T. and Malik J. (2011), "Large displacement optical flow: Descriptormatching in variational motion estimation," IEEE vol. 33, pp. 500–513.
[8]Efros A.A., Berg A.C, Mori G., and Malik J. (2003), "Recognizingactionata distance,"inIEEEconference vol. 2, pp. 726–733.
[9]Hu Y., Cao L., Yan S., Gong Y., and Huang T. S. (2009), "Action detection in complex scenes with spatial and temporal ambiguities," IEEE pp.128–135.
[10]Jhuang H., Gall J., Zuffi S., Schmid C., and M.J. Black (2013), "Towardsunderstanding action recognition," IEEE International Conference, pp. 3192
[11]A. Bobick and J. Davis, "The recognition of human movementusing temporal templates", IEEE Trans. On PAMI, vol.23, no.3, pp. 257-267, March 2001.
[12]Md. Atiqur Rahman Ahad, J.K. Tan, H.S. Kim, and S. Ishikawa,"Human activity recognition: various paradigms",International Conference onControl, Automation and Systems, pp. 1896-1901, Oct. 2008.
[13]D.Gavrilla, "The visual analysis of human movement:asurvey", Computer Vision and Image Understanding, vol.73, pp. 82-98, 1999.
[14]M. Pantic, A. Pentland, A. Nijholt, and T.S. Hunag, "Humancomputing and machine understanding of humanbehavior:asurvey", Int. Conf. on Multimodal Interfaces, pp. 239-248, 2006.
[15]R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image understanding, vol.108, no.1-2, pp. 4-18, Oct. 2007.
[16]Md. Atiqur Rahman Ahad, T. Ogata, J.K. Tan, H.S. Kim ,andS.Ishikawa, "Motion recognition approach tosolveoverwritingcomplex actions", 8th Int. Conference on Automatic Face and Gesture Recognition, Amsterdam, 6 pages, Sept. 2008.
[17]J. Liu and N. Zhang, "Gait history image: a novel temporaltemplate for gait recognition", IEEE Int. Conf .on Multimediaand Expo, pp. 663-666, 2007.
[18]H. Meng, N. Pears, and C. Bailey, "A Human ActionRecognition System for Embedded Computer VisionApplication", 3nd Workshop on Embedded ComputerVision(with CVPR), pp. 1-6, June 2007.
[19]R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image Understanding, vol. 108, no. 1-2, pp. 4-18, Oct. 2007.
[20]J. Liu and N. Zhang, "Gait history image: a novel temporal template for gait recognition", IEEE Int. Conf. on Multimedia and Expo, pp. 663-666, 2007.
[21]J. Han and B. Bhanu, "Individual recognition using gait energy image", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pp. 316-322, 2006.
[22]T. Jan, "Neural network based threat assessment for automated visual surveillance", IEEE Int. Joint Conf. on Neural Networks, vol. 2, pp. 1309-1312, July 2004.
[23]K. Leman, G. Ankit, and T. Tan, "PDA-based human motion recognition system", Int. J. Software Engineering and Knowledge, vol. 2, issue 15, pp. 199-205, Apr. 2005.
[24]V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Texture based description of movements for activity analysis" Third Int. Conf. on Computer Vision Theory and Applications (VISAPP 2008), Madeira, Portugal, vol. 1, pp. 206-213.
[25]H. Meng, N. Pears, and C. Bailey, "Motion Information Combination for Fast Human Action Recognition", 2nd International Conference on Computer Vision Theory and Applications, Spain, Mar. 2007.
[26]H. Meng, N. Pears, and C. Bailey, "Recognizing Human Actions Based on Motion Information and SVM", 2nd IEE International Conference on Intelligent Environments, pp. 239-245, 2006.
[27]Zhou, Q.; Aggarwal, J. K.; "Tracking and classifying moving jects from video", Proc of 2nd IEEE Intl Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2001), Kauai, Hawaii, USA (December 2001).