RESEARCH ARTICLE OPEN ACCESS

Temporal Stream based Undisciplined Behaviours Recognition using CNN

Thuzar Tint *, Tin MyintNaing**,Su La PyaePhyoe**
*(Department of Information Science, University Technology (Yatanarpon Cyber City), Myanmar
Email: thuzartint1984@gmail.com)
** (Department of Information Science, University Technology (Yatanarpon Cyber City, and Myanmar
Email: utinmyintnaing08@gmail.com)

----------**********----------

Abstract:

Nowadays, human behaviours detection and recognition has been developed as a significant interesting research in the field of computer vision. This paper is intended to detect and recognize the normal walking and undisciplined behaviours: throwing rubbish and spitting paan in the public area. In the undisciplined behaviours recognition system, the HOG human detector is used to detect the human's position from the incoming CCTV webcam frame. And optical flow (motion features) are calculated for temporal stream. Finally, the GoogLeNet, which is one type of Convolution Neural Network (CNN) is applied to recognize whether the crossed human is the normal walking or performs the undisciplined behaviours. According to experimental results, the proposed system is efficiently able to detect and recognize the human undisciplined behaviours. The undisciplined behaviours recognition system is useful for environmental conservation programs.

Keywords —**Throwing rubbish, Spitting Paan, Walking, Optical flow,GoogLeNet.**

----------**********----------

I. INTRODUCTION

With the coming of new innovations, automatic detection and recognition process for analysing human actions in videos has been developed instead of looking at human eyes. Therefore, vision-based human behaviour recognition is becoming an active research area. Vision-based human behaviours recognition is useful in many application areas that are video surveillance system, sports play analysis, healthcare system, crowd behaviour prediction, human-computer interaction, robotics, environmental conservation and so on. There are many challenging in vision-based human action recognition. Firstly, the actions recognition from video is prohibitive for exhaustive search, without knowing the location, temporal duration and the spatial scale of the action. Secondly, moving cameras or non-stationary backgrounds, cluttered background, variant illumination, occlusion make problem in the action analysis. Thirdly, appearance of human changes because of performing actions keeps on changing depending on the background on which action is performed, clothes and camera angle also play a vital role in appearance of human [1].

The rest of the paper is organized as follows. The related work of human behaviours detection and recognition is reviewed and summarized in Section II.The undisciplined behaviours detection and recognition system is presented in Section III. The experimental evaluation of the proposed system are shown in Section IV, and followed by conclusion in Section V.

II. RELATED WORKS

In vision-based human behaviour recognition researches, there are many different actions such as walking, running, sitting, standing, laying, handshaking, punching, kicking, departing, pointing, drinking, kissing, diving, lifting, riding horse, swing-bench, swing-slide and so on. A variety of human behaviours or activities has been detected and recognized with different approaches in different condition such as single person, multiple person interaction, person-vehicle interactions, person-facility interaction and so on.

Bobick and Davis [2] employed Motion Energy and Motion History Images which is one of the global representations, to automatically perform temporal segmentation. For activity recognition (sitting, arm waving, and crouching), shape moments and temporal template matching approach (Mahalanobis distance) was applied. This approach was tested with video sequences of 18 aerobic exercises. This approach would fail when two people were in the field of view and when one person partially occluded another.

Li et al. [3] proposed the action graph model, which represented activities using several salient postures serving as nodes in action graph. Each posture was characterized as a bag of 3D points from the depth maps. Although computation of all the 3D points was computationally expensive; it was a simple and effective method. Over 90% recognition accuracy achieved by sampling approximately 1% points according to their report.

Vemulapalli et al [4] constructed a kinematic model by using skeletal representation. In this paper, human actions could be modelled as curves in the Lie group with the geometric relationships between various body parts using geometric functions like translation and rotation in 3D space. Dynamic time warping, Fourier temporal pyramid representation and linear SVM were used for classification. This approach was focused only on actions performed by a single person.

Actions of interest are generally application dependent. In the paper, human behaviours which are normal walking and undisciplined behaviour: throwing rubbish and spitting paan, will be recognize based on dense optical flow image with convolution neural network (CNN).

III. METHODOLOGY OF THE PROPOSED SYSTEM

In human undisciplined behaviours recognition system, the HOG human detector is firstly applied to detect and localize the positions of human from the input video frame. And motion features are extracted by applying the optical flow estimation algorithm. Finally, GoogLeNet which is one model of convolutional neural network (CNN) is used to recognize human behaviours which are normal walking or undisciplined behaviours: throwing rubbish and spitting paan. If the system is detected and recognized undisciplined behaviours: throwing litters and spitting paan, the system will produce alarm sound according to the recognition results to the human and recorded these undisciplined behaviours. The overview of the human undisciplined behaviours recognition system is showed in Fig. 1.

A. Human's Position Detection and Localization

In the undisciplined behaviours recognition system, human positions are firstly detected and localized from the incoming video frame by using the HOG human detector. In HOG human detector,

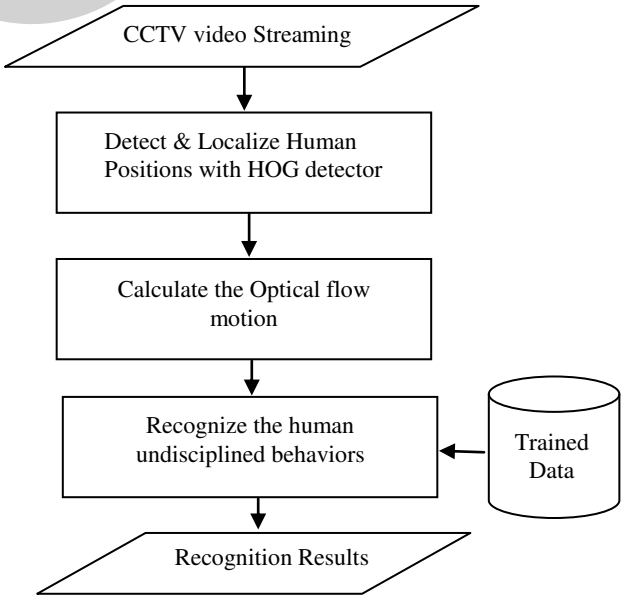


Fig. 1 Overview of the Undisciplined Behaviours Recognition System

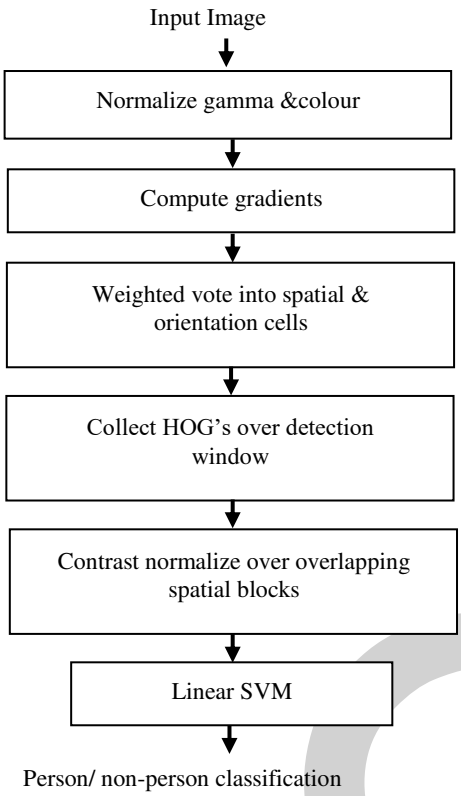


Fig. 2 Overview of the HOG human detection block [5]

the histogram of oriented gradients (HOG) descriptors are extracted as features from the input image and classified these features with SVM classier to detect human. The method is based on evaluating well-normalized local histogram of image gradient orientations in a dense grid. The overview of HOG human detection block is displayed in Fig. 2. In HOG human default detector, the followings are performed.

1) ***Gradient Computation:*** The first step of HOG calculation in many feature detectors is to perform normalized color and gamma values. However, this step can be omitted in HOG default descriptor computation. Therefore, the first step of calculation is to compute the gradient values. In default detector, 1-D centered, point discrete derivative mask is applied in one or both of the horizontal and vertical direction. 1-D centered derivative mask has the following filter kernels:[-1,0,1] and $[-1,0,1]^T$.

2) ***Spatial/Orientation Binning:*** After computing the gradient, the second step is to create the cell histograms. For an edge orientation histogram channel, each pixel computes a weighted value based on the orientation of the gradient element centered on it, and the values are added into orientation bins over local spatial regions that is called cells. The cells can either be rectangular or radial in shape, and the histogram channels are evenly spread over 0 to 180 degrees or 0 to 360 degrees, depending on whether the gradient is "unsigned" or "signed". In default detector, linear gradients are computed by voting into 9 orientation bins in 0 to 180 degrees.

3) ***DescriptorBlocks:***To consider for changes in illumination and contrast, the gradient strengths must be locally normalized. For normalization, the cells are grouped together into larger, spatially connected blocks. In HOG descriptor, all of the block regions are linked together to get the concatenated vector of the components of the normalized cell histograms. For the default descriptor, the default parameter are 16x16 pixel blocks of four 8x8 pixel cells.

4) ***Block Normalization:*** Three are four different methods for block normalization. In default descriptor, L2-Hys is used with block spacing stride of 8 pixels. L2-Hys can be calculated by first taking the L2-norm, and clipping the result. It is called renormalizing.

$$L2\text{-}norm, v \rightarrow v/\sqrt{\|v\|_2^2 + \epsilon^2}$$

In above equation, v be the unnormalized descriptor vector, $\|v\|_k$is its k-norm for k=1,2, and ϵbe a small constant [5].

5) ***Classification:*** After extracting the normalized Histogram of Oriented Gradient (HOG) features, the next step is to classify these features with machine learning algorithm. By default, linear SVM with SVM Light (slightly modified to reduce memory usage for problems with large dense descriptor vectors) is used for classification. Results of testing video 1 for human position detection is displayed in Fig. 3.

B. Dense Optical Flow Calculation

After detecting and localizing the humans' positions, the second step is to estimate the temporal stream, which is motion features between consecutive frames by applying dense optical flow algorithm. A dense optical flow algorithm is an algorithm to find a set of displacement vector fields between the two consecutive frames t andt+1.

(a) (b)

Fig. 3Result of Testing Video 1 (a) Original Image

(b) Human Position Detection

Let M_1 and M_2 be the visible mask of a layer at frame $I_1$and I_2. (u_1, v_1)be the flow field from I_1 to I_2 and (u_2, v_2) be the flow field from I_2 to I_1. Estimating layer-wise optical flow [6] consists of the followings:

$$E(u_1, v_1, u_2, v_2) = \sum_{i=1}^{2} E_{data}^{(i)} + \alpha E_{smooth}^{(i)} + BE_{sym}^{(i)}$$

$$E_{data}^{(1)} = \int g * M_1(x,y)|I_1(x+u_1, y+v_1) - I_2(x,y)|$$

where g is a Gaussian filter.

$$E_{smooth}^{(1)} = \int (|\nabla u_1|^2 + |\nabla v_1|^2)^{\eta}$$

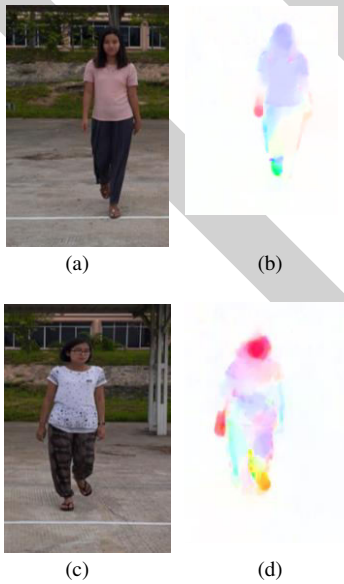
whereη varies between 0.5 and 1.

$$E_{sym}^{(1)} = \int |u_1(x,y) + u_2(x+u_1, y+v_1)| + |v_1(x,y) + v_2(x+u_1, y+v_1)|$$

After that, the optical flow image is getting from the consecutive video frames. Optical flow results from the testing video 1 is showed in Fig. 4.

C. Human Behaviours Recognition

After getting the segmenteddense optical flowimage, which is motion feature, the next step is to recognize the human behaviours. In the human behaviours recognition step, GoogLeNet which is one of the convolutionalneural network (CNN)



(a) (b) (c) (d)

Fig. 4Result of Testing Video 1 in Figure 3 (a) Segmented Human Region for first one (b) Optical flow Image from a (c) Segmented Human Region for second one (b) Optical flow Image from c

architectures, is used to extract features and recognize the human behaviours which are normal walking or undisciplined behaviours: throwing litters and spitting paan. A Convolutional Neural Network (CNN) is a powerful deep learning algorithm due to the use of multiple feature extraction stages that can automatically learn representations from the image and can classify features in images for computer vision. CNNs are constructed with several kinds of layers. The layers are made up of repeated blocks of convolutional, ReLU (rectified linear units), and pooling layers. The convolutional layers convolve their input with a set of filters. The filters were automatically learned during network training. The ReLU layer adds nonlinearity to the network. Pooling layer performs downsampling operation that minimizes the dimensionality of each map but maintains the important information. The convolution, pooling, and ReLU layers are typically interleaved in a neural network in order to increase the expressive power of the network. Fully connected layer combines the features to classify the images. There are many popular CNN architectures. Among them, GoogLeNet is used to train our model.

1) ***GoogLeNet:*** GoogLeNet consists of 22 layers and it lessens the number of parameters from 60 million (AlexNet) to 4 million. There are numerous inception modules connected together to go deeper. The basic idea of the inception module is that key information in the images is available at different levels of detail. 1×1 convolution is used as a dimension reduction module to reduce the computation before the expensive 3x3 and 5x5 convolutions [7]. In deep learning applications, transfer learning is commonly used to retrain a convolutional neural network toclassify a new set of images.

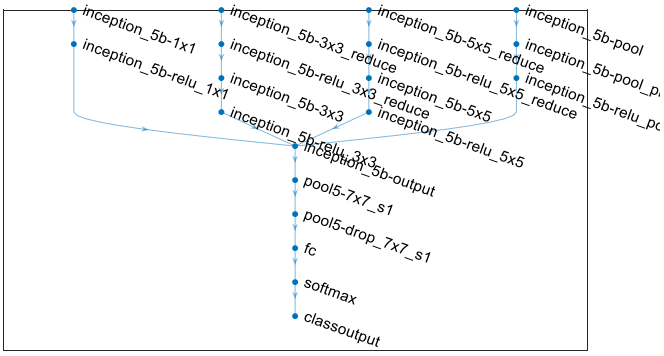


Fig. 5 Layer graph from the trained GoogLeNet Network

Layer graph from the trainedGoogLeNet Network is described in Fig. 5.

Before training, new training images which are three types of human behaviours are firstly resized to the 224x224 in the RGB color space. In training phase, firstly load a pretrainedGoogLeNet network and new images for training. And replace 'loss3-classifier' and 'output' layers with new layers adapted to the new data set to retrain a pretrained network to classify new images. In the system, specify the training options, including 3e-4 initial learn rate, 10 mini-batch size, 3 max epochs and 3 validation frequency. Set the final fully connected layer to have the same size as the number of classes in the new data set. After specifying the training options, retrain the new training images.

In the testing phase, human position is firstly detected and localized in the incoming video. And dense optical flow is segmented from the localized region to find the temporal stream. Then the segmented image is classified with GoogLeNet classier based on the retaining data. Finally, the system will produce alarm sound according to the recognition results to the human and recorded these undisciplined behaviours if the system is recognized undisciplined behaviours: throwing rubbish and spitting paan.

IV. EXPERIMENTAL EVALUATION

In the section, the proposed system is employed and evaluated to show the efficiency of the proposed method. Since there is no standard benchmarking dataset available for throwing rubbish and spitting paan, videos which contains normal walking, standing and human undisciplined behaviours: throwing rubbish and spitting paan, are captured using ip Wi-Fi camera.

First of all, frames (images) are extracted from the captured videos in order to carry out the training process. And then, motion features are calculate and extracted by using optical flow algorithm. After that, these optical flow images are trained in GoogleNet and store the trained result in the system. In the training videos, there are one person walking, one person throwing rubbish and one person spitting paan. In this system, there are 600 normal walking images, 500 throwing rubbish images and 500 spitting paanimagesSome sample optical flow images for training dataset are described in Fig. 6.

A. Experimental Results

In the section, experimental results of the human undisciplined behaviours recognition will be discussed.

Fig. 8, 9 prove that the system can detect and recognize more than one person in real time testing. Fig. 7 shows that one person split paan in the street and it can be well detected and recognize. In Fig.8, there are twopeople walking and one walks in normal and the other one splits paan in the

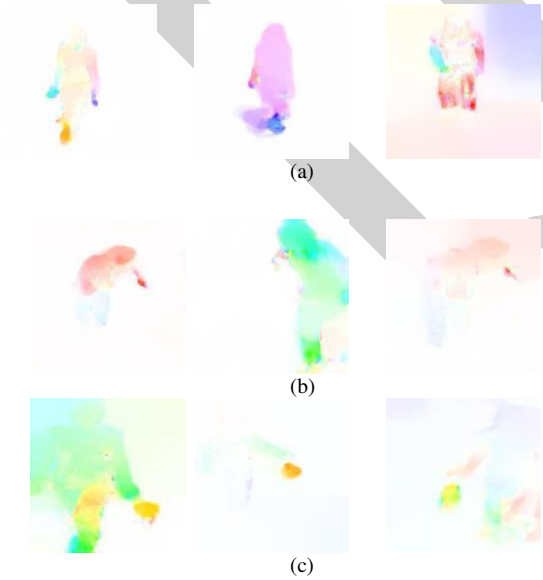


Fig. 6 Sample Optical Flow Image from Training Dataset (a) Normal walking (b) Spitting paan (c) Throwing rubbish

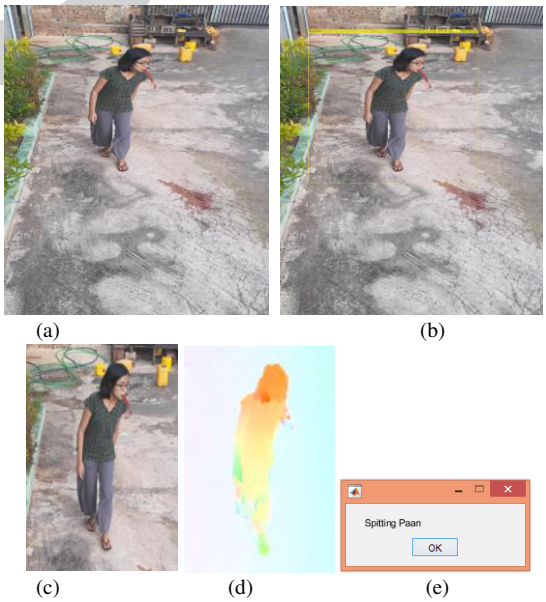


Fig. 7 Results of Testing Video 2 which contains only one person (a) Original Frame (b) Human Position Detection Frame (c) Segmented Human Region (d) Optical flow Image (e) Recognition Result

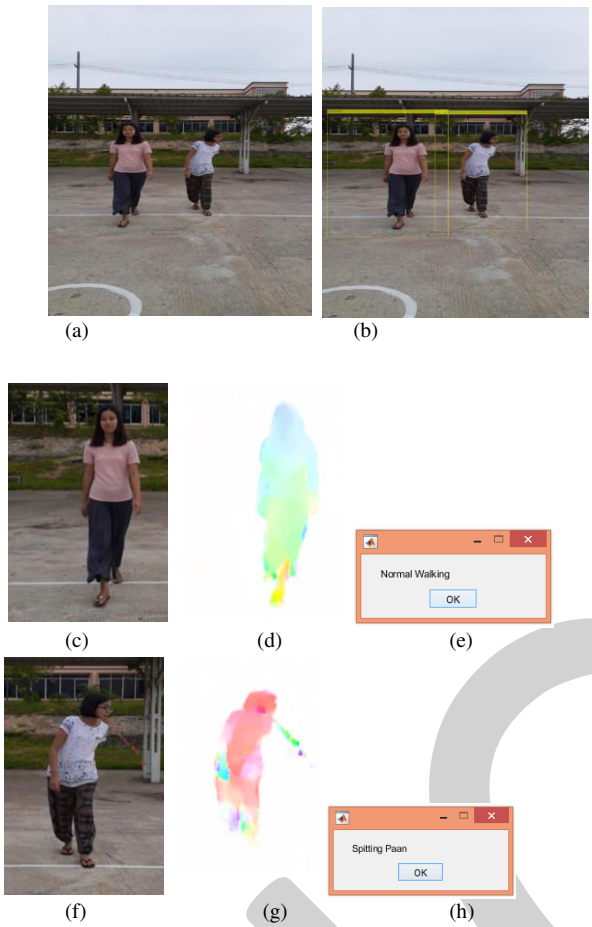


Fig. 8 Results of Testing Video 3 which contains two people (a) Original Frame (b) Human Position Detection Frame (c) (f)Segmented Human Regions (d) (g)Optical flow Images (e)(h)Recognition Results

street. The system can detect and recognize both two kind of action in this case. In Fig. 9, there are three people and the two walks in normal and the third one throw rubbish in the street. In this case, the system can perform correctly for walking and throwing rubbish done by that three people. The system can detect and recognize throwing and spitting behaviours in any human position. However, there are misclassification results in some frames in the testing. Thereason is that the system may perform in some completed background and if the colour of rubbish or paan is similar to background. Moreover, HOG human detector cannot detect when human are far from camera.

Fig. 9Results of Testing Video 3 which contains three people (a) Original Frame (b) Human Position Detection Frame (c) (f) (i) Segmented Human Regions (d) (g) (j) Optical flow Images (e) (h) (k) Recognition Results

B. *Performance Evaluation*

The performance evaluation for the proposed system is computed according to the following formulas:

$$Precision(C) = Correct\ of\ C / Total\ as\ C$$

$$Recall(C) = Correct\ of\ C / Total\ of\ C$$

$$Accuracy = Correct\ Perdiction / Data\ Size$$

TABLE 1

CONFUSION MATRIX OF RECOGNITION RESULTS

	Total Frames	Walking	Throwing	Spitting
Walking	385	360	13	12
Throwing	150	20	115	15
Spitting	105	15	11	79

TABLE 2

PERFORMANCE MEASURE OF RECOGNITION RESULTS

	Precision	Recall
Walking	0.91%	0.93%
Throwing	0.82%	0.76%
Spitting	0.74%	0.75%

The performance measure is calculated based on 30 videos which contain normal walking, throwing rubbish and spitting paan behaviours frames. Confusion matrix of recognition results is shown in Table 1 and Performance measure of each behaviour is described in Table 2. According to the experimental results, the proposed system achieves 0.86 % overall accuracy, 0.823 % overall precision and 0.813% overall recall.

V. CONCLUSIONS

In the paper, the human undisciplined behaviours recognition system is proposed to capture undisciplined people. HOG human detector is employed to find the human position and optical flow algorithm is applied to compute the temporal motion data. And convolutional neural network (GoogLeNet) is used. According to the testing results in order to recognize human behaviours correctly. The overall accuracy result is acceptable but it is not the best for human activity recognition system. The system can detect and recognize more than one person.

REFERENCES

[1] C.J. Dhamsannia and T.V. Ratanpara. "A Survey on Human Action Recognition from Videos," in Proc. Online International Conference on Green Engineering and Technologies (IC-GET), 2016.

[2] A.F. Bobick and J.W. Davis. "The Recognition of Human Movement Using Temporal Templates," in Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23,No.3, March 2001,pp.257-267.

[3] Li, W., Zhang, Z. & Liu, Z. "Action recognition based on a bag of 3D points," IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, pp. 9-14.

[4] R. Vemulapalli, F. Arrate and R. Chellappa "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group" in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[5] N. Dalal, B. Triggs. "Histograms of Oriented Gradients for Human Detection," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.

[6] C. Liu, W. T. Freeman, E. H. Adelson, "Human-Assisted Motion Annotation," in Proc. IEEE Conference on Computer Vision and Pattern Recognitionn, 2008.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, "Going Deeper with Convolution ," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.